EMPIRICAL STUDIES ON PRICE DETERMINANTS OF ONLINE AUCTIONS WITH MACHINE LEARNING APPLICATIONS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF SOCIAL SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY
BY

EMRAH ÖZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF ECONOMICS

FEBRUARY 2019

Approval of the Graduate School of Social Sciences

____

Prof. Dr. Tülin Gençöz
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy.

____

Prof. Dr. Meltem Dayıoğlu Tayfur
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy.

____

Assoc. Prof. Dr. Esma Gaygısız
Supervisor

Examining Committee Members

Prof. Dr. Erdal Özmen	(METU, ECON)	____________________
Assoc. Prof. Dr. Esma Gaygısız	(METU, ECON)	____________________
Assist. Prof. Dr. Ayşe Özgür Pehlivan	(Bilkent Uni., ECON)	____________________
Assoc. Prof. Dr. Serhan Duran	(METU, IE)	____________________
Assoc. Prof. Dr. Özge Sezgin Alp	(Başkent Üni., MFY)	____________________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Emrah ÖZ

Signature:

ABSTRACT

EMPIRICAL STUDIES ON PRICE DETERMINANTS OF ONLINE AUCTIONS WITH MACHINE LEARNING APPLICATIONS

Öz, Emrah
Ph.D., Department of Economics
Supervisor: Assoc. Prof. Dr. Esma Gaygısız

January 2019, 208 pages

Current technological developments have changed our trading habits and the importance of e-commerce in our lives has grown rapidly in the past decade. This new economic and technological environment generates massive, cheap, easily accessible and invaluable data. One of the important topics in electronic trade is the price estimation. Electronic trade takes place usually through two sales methods. The first is auctioning and the second is Buy-it-Now (BIN) sales. This dissertation concentrates on the determinants of online auction end prices in the smartphone markets. In this context, 444 auction and 676 BIN sales realized between March-July 2018 were analyzed with current Machine Learning (ML) algorithms. As a new contribution to the literature, vendors' descriptions are analyzed with Natural Language Processing (NLP), prices of similar products are taken into account and the effect of information from the bids in the initial stage are investigated with image processing algorithms. The analyses show that vendor's descriptions, prices of similar products, information from the bids in the initial stage, auction length, the day auction starts, product accessories, the number of visits to the vendor profile have positive effects on auction prices. On the other hand, sellers' reputation, especially negative reviews adversely affect auction prices.

Keywords: E-Commerce, Online Auctions, Price Determinants, Sentiment Analysis, Machine Learning Applications

ÖZ

MAKİNE ÖĞRENMESİ UYGULAMALARI İLE ONLİNE İHALE FİYATLARINI ETKİLEYEN FAKTÖRLER ÜZERİNE DENEYSEL ÇALIŞMALAR

Öz, Emrah
Doktora, İktisat Bölümü
Tez Yöneticisi: Doç. Dr. Esma Gaygısız

Şubat 2019, 208 sayfa

Bilgi çağı olarak adlandırdığımız günümüzde teknolojik gelişmeler ticaret alışkanlığımızı da değiştirmiş ve hayatımızda elektronik ticaretin yeri hızla artmıştır. Bu yeni ekonomik ve teknolojik durum ile birlikte müthiş büyüklükte ucuz, kolay ulaşılabilir ve çok değerli veri oluşmaktadır. Elektronik ticaret alanında en çok dikkat çekici konulardan biri de fiyat tahminidir. Online alışveriş genellikle iki satış metodu ile gerçekleşmektedir. Birincisi ihale, ikincisi ise şimdi-satın-al (ŞSA) satış yöntemleridir. Bu tez akıllı telefon piyasasında yapılan online ihalelerde son fiyat belirleyicilerini analitik olarak incelemektedir. Bu kapsamda Mart-Temmuz 2018 arasında gerçekleşen 444 adet ihale ve 676 adet ŞSA satış verileri güncel makine öğrenme algoritmaları ile analiz edilmiştir. Literature katkı olarak, ürünler için satıcılar tarafından girilmiş açıklamalar doğal dil işleme yöntemleri ile analiz edilmiş, benzer ürünlerin daha önceki satış fiyatları dikkate alınmış, ihalenin ilk döneminde verilen tekliflerden elde edilen bilgiler de görüntü işleme yöntemleri ile incelenmiştir. Yapılan analizlerle birlikte benzer ürünlerin fiyatları, ilana girilmiş açıklamalar, ihalenin başladığı gün, ihalenin uzunluğu, aksesuarlar, satıcı profilinin ziyaret edilme sayısı ve başlangıç döneminde verilen tekliflerle ilgili bilgilerinin ihale bitiş fiyatını pozitif etkilediği bulunmuştur. Diğer bir taraftan da satıcının ünü, özellikle aldığı negatif yorumların ise satış fiyatını negatif etkilediği tespit edilmiştir.

Anahtar Kelimeler: E-ticaret, Online İhale, Fiyat Belirleyicileri, Duygu Analizi, Makine Öğrenmesi Uygulamaları

to the One I Love...

ACKNOWLEDGMENTS

I wish to express my deepest gratitude to my supervisor Assoc. Prof. Dr. Esma Gaygısız for her guidance, advice, criticism, courage and insight throughout the research.

I wish to express my thanks to Assist. Prof. Dr. Ayşe Özgür Pehlivan for her invaluable support, insightful comments and criticisms from the beginning of the dissertation.

I wish to express my gratitude to Prof. Dr. Erdal Özmen for his insightful and inspiring comments and criticisms and feedback.

I wish to express my thanks to Assoc. Prof. Dr. Serhan Duran for his insightful comments, criticisms and feedback.

I wish to express my thanks to Assoc. Prof. Dr. Özge Sezgin Alp for her insightful comments and criticisms and feedback.

I would also thank to the love of my life, Esra ÖZ, for her invaluable help, support and motivation.

I would also express my deepest gratitude to my family for their invaluable support, guidance and patience throughout my education life.

I would also thank to the TUBİTAK BİDEB scholarship program for supporting me with graduate student funding throughout this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AC	Autocorrelation
AI	Artificial Intelligence
ANN	Artificial Neural Network
AP	Auction Price
APL	Auction Price Lags
APM	Auction Path Models
BIN	Buy it Now
BLUE	Best Linear Unbiased Estimator
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CV	Common Value
DF	Dynamic Features
DT	Decision Tree
eBay	Ebay Inc. is a Multi National E-Commerce Corporation
EF	Economic Features
HC	Heteroscedasticity
IA	Information Aggregation
IPV	Independent Private Value
LR	Logistic Regression
MAP	Maximum A Posteriori
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MNB	Multi nominal Naive Bayes
NB	Naive Bayes
NLP	Natural Language Processing
OLS	Ordinary Least Square

PCA	Principal Component Analysis
PFA	Principal Feature Analysis
PPE	Price Prediction Error
RET	Revenue Equivalence Theorem
RFE	Recursive Feature Elimination
SF	Static Features
SML	Supervised Machine Learning
SVC	Support Vector Classification
TF-IDF	Terms Frequency-Inverse Document Frequency
UML	Unsupervised Machine Learning
URL	Uniform Resource Locator
USA	United States of America
WCSS	Within Clusters Sum of Square
WTP	Willingness to Pay

CHAPTER 1

INTRODUCTION

Electronic commerce is growing at an unprecedented rate all over the globe and online shopping is rapidly replacing conventional commerce. Widespread use of the internet, the development of electronic payment systems, the impact of the new generation production network systems are indicated as the main factors behind this phenomenon. There are a number of national and international companies that market their products online instead of the conventional shop-windows. Those vendors have various means to easily publicize their products online and publish product and price information for their customers. On the other side, buyers can easily search a product, compare alternatives and prices easily and choose the one that fits their preferences online. For this reason, investment in e-commerce companies are now favorable and the market values of those companies are leapfrogging. One of the international electronic commerce companies has already reached a market value of $1.000 billion. This metric is even higher than the gross domestic product of many countries. The figure below shows the e-commerce sales volume between 2010 and 2017 in the USA, which was $170 billion in 2010, increased 2.5 times in 7 years and exceeded $450 billion in 2017. The share of e-commerce in total retail sales approximately doubled in this period. It is estimated to exceed $700 billion by 2022 implying that the importance of electronic commerce will grow with a steeper curve. In the USA, e-commerce appears to be dominated by big companies like Amazon and eBay. Amazon's 2017 sales volume was $178 billion. Similar to Amazon, eBay is another leading company in electronic commerce. As of the second quarter of 2018, eBay reached 175 million active users and more than 25 million vendor profiles worldwide. Its annual gross merchandise volume in global marketplaces has increased in the last five years and reached 88.4

(36.3 in the USA) billion dollars at the end of 2017. Ebay mobile app has been downloaded by more than 370 million users and there are more than 1 billion products which can be screened in the eBay system at a time. Such a large-scale trade through internet and the resultant massive amount of data generated by traders have drawn the attention of both economists and data scientists.

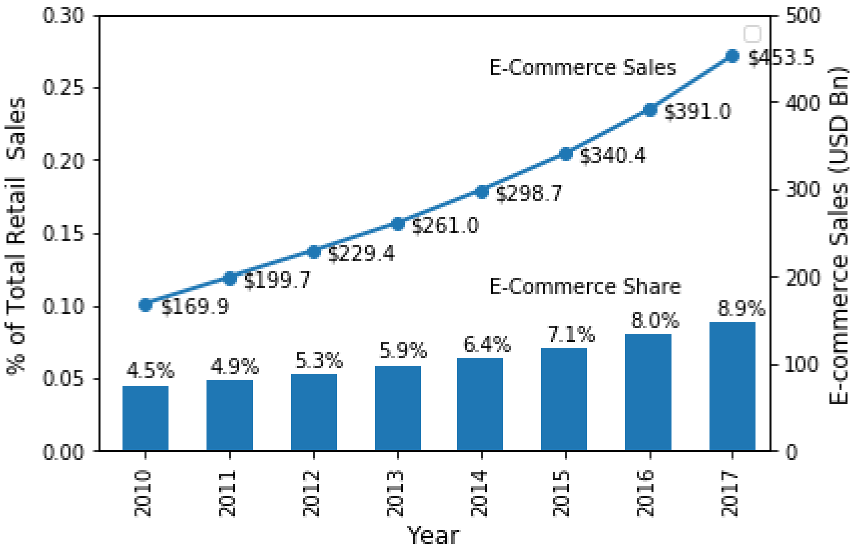


Figure 1.1: E-Commerce Sales Volume in the USA

Source: U.S. Census Bureau, emarketer.com

Similarly, the value of e-commerce for both developed and developing countries, as well as for Turkey is increasing. The Figure 1.2 below shows the ratio of electronic commerce in total retail. In the last 5 years, the ratio of electronic commerce in retail market has almost doubled for all countries. Indeed, the e-commerce is expected to increase further in the near future. The Figure 1.3 below shows the current place of the electronic commerce and its expected level in the future. Internet trading, which was 2.3 trillion dollars in 2017, is expected to double by 2021 and reach 4.8 trillion dollars globally. Similarly, while the ratio of online trade to total trade in the world is 10.2% in 2017, it is estimated to be 17.5% in 2021. The graphs show that there will be a new economy in the future and recommend deep researches in this field.

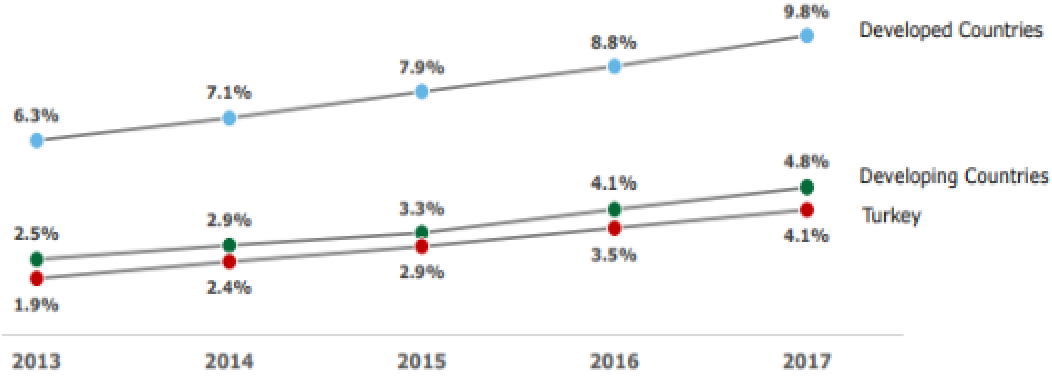


Figure 1.2: Percent of Online Retail in Total Retail

Source: Deloitte 2018

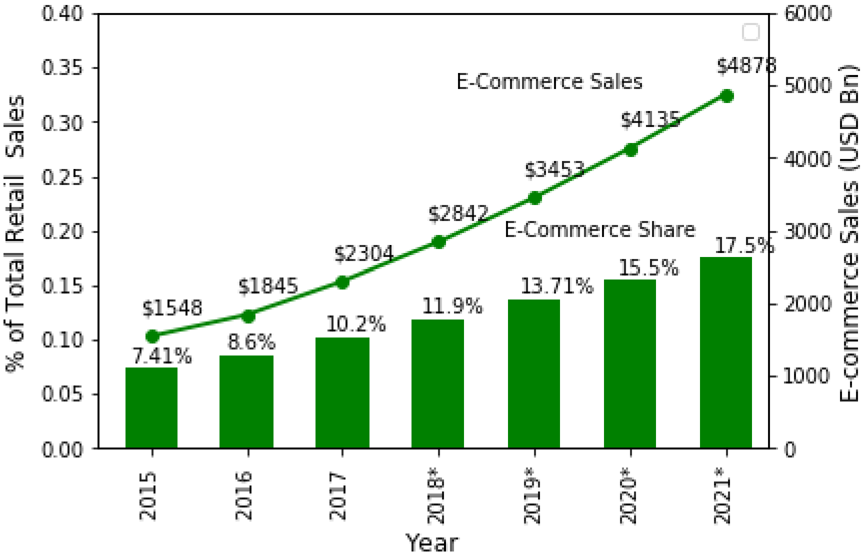


Figure 1.3: Global E-Commerce Sales Volume and the Future

Source: Statista.com 2018

There are two pricing strategies commonly pursued by online traders. The first one is Auction Pricing and the other is Buy it Now (BIN) pricing. Under BIN pricing, sellers publicize their products with a certain price tag and wait for the customers to arrive. However, the auction end price is determined by the competition of the buyers provided that it is above a certain reserve price. The seller must sell the product once a quotation surpasses the reserve price in the auction. Yet, the marketer does not have to sell the product until the tag price is offered in BIN pricing strategy. Sellers can harness both strategies and largely similar goods or services can be transacted at different prices within the same day despite having almost same features. For this reason, the influencing factors on the prices have become crucial and research on predicting prices before the auction ends is now one of the trending topics. Today, information and data analysis technology have made a significant progress and the processing power of the computers, ubiquitous internet, efficient open source machine learning algorithms are now far beyond the past. As most of the events we encountered in our daily lives, big data has been generated for electronic commerce and it has led to the emergence of issues which can only be resolved by Artificial Intelligence (AI). In this dissertation, it is aimed to shed a light on the factors affecting the online auction prices in smart phone market by using Machine Learning (ML) algorithms and developing predictive models for auction end prices.

1.1 The Motivation for This Research

Research on e-commerce is a trending topic of contemporary economic research literature and it is also expected to develop in the close future. Both in conventional and electronic trade; price, quality, delivery terms, warranty and trust are the most important factors. The “price” may sometimes appear to be the most important one among them. Determining the factors that affect the price for any kind of trade and predicting the equilibrium price within sufficient time-frame are essential for both the sellers and the buyers. When sellers identify the factors that affect the price, they can increase the probability of sales execution, shorten the sales cycle and optimize the profit by setting optimal sales strategies. When buyers know the factors that affect the price, they can refrain from giving an excessive price for the requested product. So buyers can maximize their utility with minimum cost. If there is an arbitrage opportunity between the real value of the product and the sales price stated in the product page, it is possible to make a reasonable price estimate and counsel the traders to balance market and remove the market friction. In addition to these, it is also possible to ensure the sales execution in a shorter cycle for both parties, sellers and buyers. All of these and more motivates me to make this research.

1.2 Research Objectives and Aims

Main objectives of the study are as follows:

- What are the factors which affect eBay type online auctions that repeat over time for same or similar products end prices?
- Can we estimate the end price of an auction with a high accuracy rate at the beginning of auction?
- What does Machine Learning bring to the predictive modeling of auction prices?

Sub-objectives of the study are as follows:

- What are the static features which are known at the beginning of the auction and do not change within the auction time? What is the relationship between static features and the end price of the auction? Can the auction end price be estimated with a reliable accuracy by the help of these features?
- What are the dynamic features that are clearly identified during the auction time, but not obvious at the beginning of the auction? What is the relationship between those features and end price of the auction? Can the auction end price be estimated with a reliable accuracy by the help of these features?
- Can an automated price attribute be created by performing sentiment analysis on the product description information written for the product in auction?
- Can information aggregation that occurs when a certain amount of time has passed for the auction, give a reliable idea about the result of the auction?

- Can the approximate value of a product in auction, such as market value, the benchmark value or opportunity cost be determined? What is the impact of these metrics on the end price of the auction?

- What is the relationship between economic features outside the auction and the auction end price? Can the auction end price be estimated with the help of these features?

- Which Machine Learning algorithms can be useful for the auction end price models?

1.3 Research Contributions

- In addition to the present studies in the field of online auctions, the prices of similar products were also investigated in predictive models. The models have been established through brand-new machine learning applications.

- Auction paths for online auctions have been modeled and estimated by curve fitting applications and stages of auctions were determined.

- The relationship between the average price of the auctions, BIN and past auction prices has been analyzed.

- The product descriptions written at the auction pages have been included in the auction price models using text classification algorithms as a variable for the first time in literature. That is, the relationship between the product description and auction end price has been analyzed.

- Supervised models have been trained by the textual product descriptions submitted in BIN sales. The descriptions of the products sold with auctions have been classified by the trained supervised model and class attributes have been included in the predictive end price model. Such a study is also the first one in literature as far as my literature review.

- The impact of the auction start day has been analyzed in detail.

- Whatever the length of the auction, the number of bids given at the first and the last stages of the auction are higher. There are fewer bids between them. The bids given at the initial stage of the auction (up to 20% of the auction time) have been incorporated as information aggregation variable for the first time.

- The bidder reputation in the auction have been included in the price model for the first time. That is, the impact of bidder ratings have been investigated.

- Extra items which are accessories given with the products such as case, screen protector, charger, cables etc. have been also included in auction price prediction models.

- Auction path graphs have been plotted with the data generated in the information aggregation stage. These figures have been clustered by image classification algorithms and a price prediction method has been proposed. This is the first price prediction methodology with image classification for the online auction literature as far as my review.

1.4 Organization of the Thesis

In the second chapter of the study, similarities, differences and the nexus between auction and buy it now prices will be highlighted. Relevant information about these two pricing strategies will be summarized. After basic information, theoretical and practical studies on the auction prices will be explained in the third chapter, and the milestones that the literature has progressed from past to present will be presented. The gap in the literature will also be pinpointed and the contribution of this research to the literature will be specified. In Chapter 4, the data used in the analysis and the characteristics of the data will be explained. In Chapter 5, the factors affecting the prices of online auctions will be investigated i.e. auction end price will be modeled and estimated with the help of machine learning algorithms. This chapter will consist of 3 sections. First, the product description written on the product auction page will be clustered with unsupervised learning algorithms and relevant description attribute will be generated. Then, the second clustering process will be done by some features affecting the auction price and a multiple linear regression model will be specified. In the next section, the textual product description written on the product page of BIN sales will be used and supervised clustering model will be trained to generate a new product description attribute for auctions. Second clustering with features and multivariate regression process described above will be conducted again to specify new auction price models. In the last section, auction path figures which are formed at the first stage of the auctions will be used for image classification by the help of the machine learning algorithms. Then, the end price of the auctions will be predicted without using any other data. With this analysis, a part of the auction chart can be interpreted and used for the end price prediction, similar to the case an artwork expert can guess who the artist is when he sees a piece of an artwork. Model results obtained in previous sections will be compared in Chapter 6. The conclusions and the future research opportunities will be discussed in Chapter 7. References and appendices will be provided at the end.

CHAPTER 2

BASIC ONLINE AUCTIONS AND BUY IT NOW CONCEPTS

There are mainly two types of pricing strategies in e-commerce namely auction and buy-it-now (BIN) prices for consumer goods and there are several views why this segmentation exists. For instance, Ockenfels et al. (2006) pointed out in general that sellers and buyers prefer auction sales when they do not have a definite idea about the true price of a marketable product. Bajari and Hortacsu (2004) explained why sellers and buyers prefer auction sales. First, online auctions provide an easy and inexpensive environment for special products that do not have a massive trade volume. Second, it is an alternative to traditional markets for the vendors selling idiosyncratic and special products. Third, online auctions are regarded as a type of game or entertainment. Bajari and Hortacsu (2003) also mentioned that in order to reach a judgment about the product value before entering an auction, first, it is necessary to do research. This brings both time and effort costs to the stakeholders. They defined this as the cost of auction entry and gave an example to make it numerical. For a product having a book value of $1000 and with a minimum bid of $600, the computed entry cost is $66. That is, there is a hassle cost of 6% on the average which makes BIN sales more favorable for the customers looking for convenience. However, we can claim that this cost is decreased nowadays with the utilization of fast internet technology, search algorithms and bid snipping applications used on behalf of customers. Similarly, Milgrom and Tadelis (2018) emphasized that auctions have become a common mechanism to reap significant gains from trade when price discovery is essential. Another contemporary study examining the auction and BIN sales in detail is Einav et al. (2018). This paper covers eBay posted price (another name for BIN sales) and auction sales data for the period 2003-2015. Sales of a wide range of products have been examined on a large set of data and the researchers have found compelling results. Earlier, the auction sales were dominant in eBay, but BIN sales have been domineering lately. Those who preferred the auction sales stated that they preferred price discovery and those who preferred the posted prices were more impatient and preferred convenience. They mentioned that auctions are better for special and idiosyncratic products as well as used items. BIN sales are more favorable for standard products.

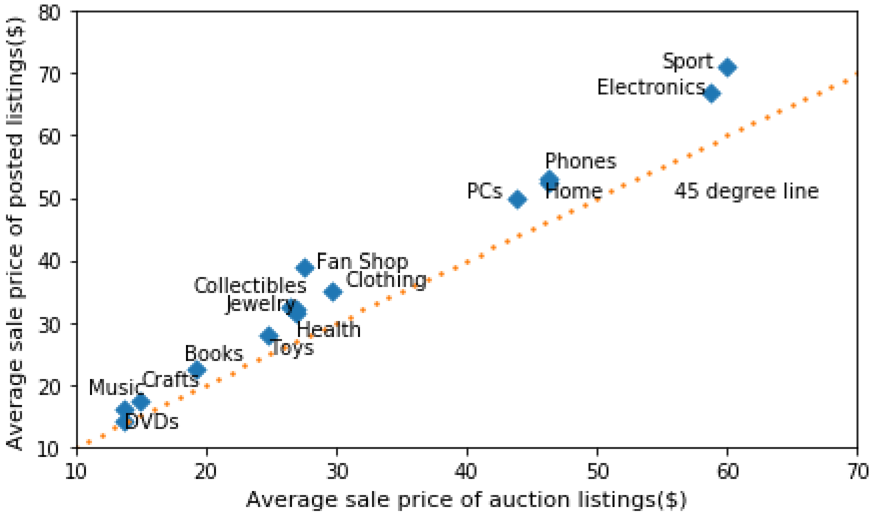


Figure 2.1: Auction and BIN Sales Price

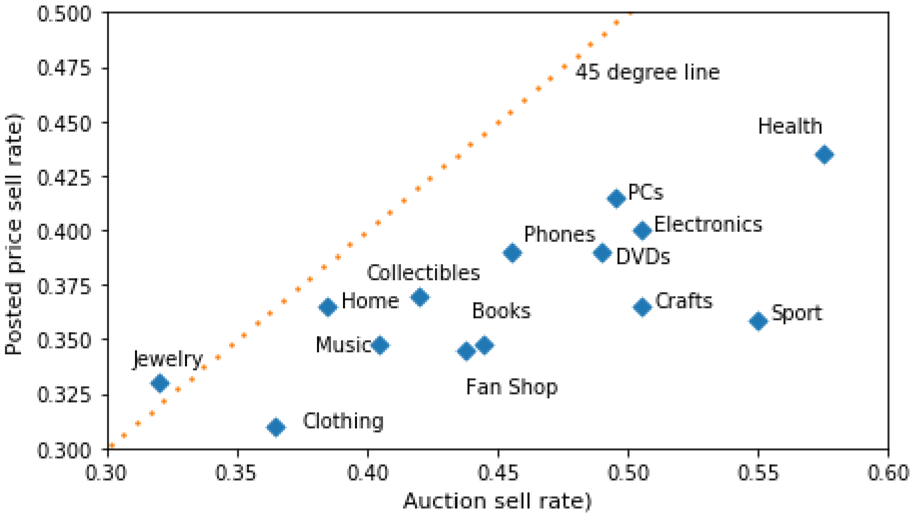


Figure 2.2: Auction and BIN Sale Rates

Source: Einav et al.(2018)

The figures above show the sales prices and rates by auction and BIN sales according to Einav et al. (2018) paper. The figures prove remarkable results; selling price in the auctions are lower than the posted price, but the sales probability is much higher. In other words, Figure 2.1 shows that, for specific kind of products, the BIN price is bigger than the auction price. On the other hand Figure 2.2 illustrates that the sales rate for BIN sales are lower than the auction sales rate. In addition to these, it has been shown that the frequent buyers, i.e. the experienced buyers, prefer the auctions with the pursuit of more profit, and the people who have more sales experience prefer BIN sales. For this reason, the authors claimed that there is also a market segmentation providing two sales methods continue in e-commerce. As can be understood from the above statements, the auction and the BIN sales can be substitute markets and prices have a large correlation with each other. Both sellers and buyers can observe the prices of sales in both markets, which gives them information on the approximate market price of the product. We can deduce that reference price information will have an important role in the product price modeling. Since the focus of the dissertation is the auction price, we will leave BIN sales market aside and give some brief information about eBay auctions below:

To begin with, auctions are theoretically divided according to bidders valuations into two groups namely, private value auctions and common value auctions. Ockenfels et al. (2006) described this distinction in simple terms. The auctions in which each bidder has different valuation (his maximum willingness to pay) for the auctioning product, are called private value auctions. On the other hand, in the common value auctions, the value of the product is the same to all bidders, but bidders have different information about it. In case there is an auction for a mining license in a special area, the economic value for the license is the same for each of the participants. However, each auction participant is likely to forecast a different price. Bidders may get different “signals” about whether the area has high mine reserve or not. In such situations, bidders typically tend to change their valuations when they receive the competitors’ signals. When we think that the resale prices of the electronic products are the same for all bidders, and the uncertainty in valuation caused by the use of the products, we can evaluate eBay auctions as common value auctions. Theoretically, Bajari and Hortacsu (2003) showed that the equilibrium in the eBay auction model is a second-price sealed-bid common value auction.

After a brief theoretical basis for eBay auctions, we will provide some technical details for the auctions running on eBay below. These auction characteristics will be also essential to identify price determinants. An eBay type online auction is held for a product on a fixed schedule. That is the characteristics are determined at the auction start and cannot change over time.

- Characteristics which can be determined by vendors are the starting time of the auction, length of the auction, the title of the announcement, item location, shipping cost, product description, product photos, starting price, price increments, and reserve prices.
- There are also auction policies determined by eBay.
- On eBay, the auction length can be either 1,3,5,7 or 10 days. When this predetermined duration is over, the highest price bidder is the winner. Duration is fixed and cannot change as they do Amazon auctions.
- The winner pays the price of the second highest price plus bid increment. That is, eBay auction is a kind of second price Vickrey (1961) auction.
- Customers can monitor their and competitors' bids in real time on the auction website.
- Actual bidders' names are hidden but the bidders can see the bids, bidder reputations and bid time.
- If any customer gives the highest bid, the eBay system displays it on the auction website a bid price slightly above the price given by the 2nd highest.
- The owner of the highest bid (leader) is automatically updated as soon as the system receives a new and higher offer, up to the leader's maximum bid.

- The maximum bid price given by the leader is not displayed on the system. The other players in the auction cannot screen the leader’s maximum price without making more bids than the leader.

- If any bidder bids more than the leader’s maximum price, they are identified as the new leader. The above steps will reiterate for the new leader.

- The auction is completed when the designated auction time is over.

- The price that appears as a winning bid is actually the price that the buyer should pay. That is, winning bid is reflected slightly above the second highest price. The leader’s maximum bid or leader’s valuation is not shown anywhere.

In this research, the auction price of iPhone 7 Plus product, which is sold in eBay system, has been examined and the market structure can be summarized as follows: The item in the analysis is a widely used product with many alternatives and mass-produced in the market. Buyers and sellers in this market are mostly ordinary people. Vendors and buyers only come together online and do not see each other face to face. The products are delivered to the buyers with the cargo system and payments are made online. Buyers rate sellers by the online system after shopping. In the eBay system, the products are offered either in auction or BIN form. If the seller and buyer profiles are evaluated, experienced and professional sales people sell their products in the form of BIN sales. On the other hand, experienced buyers prefer more auction sales. In the light of these explanations, it can be emphasized that the market analyzed has a specific structure and the price model developed is specific to this market. Although it is difficult to generalize the price model for other products, the proposed process can be used to create price models for any kind of products. In other words, product condition ratings created by UML and SML models can be included in the price models as a factor affecting the product. Other static, dynamic and economic variables affecting the product can be included in the model to create price models and the performance of price models can be increased by clustering algorithms.

CHAPTER 3

LITERATURE REVIEW AND THEORETICAL BACKGROUND

3.1 Literature Review

Evolution of trade and big data have attracted the attention of the academic researchers as well as the big data analysts in business. In this prolific topic, researchers have published a lot of articles. In the following paragraphs, the importance of the research carried out in the auction economy, the players' strategic behaviors in the auctions, structural auction models and empirical studies which are more favorable due to their ease in application also the role of the Machine Learning (ML) algorithms in these studies will be explained. Then, the gaps in literature and contribution of this study will be emphasized.

First of all, it should be stated that online auction-themed studies are based on two main reasons as identified by Bajari and Hortacsu (2004). The first is the abundance of high quality and available data, and the second one is the natural setting where the theory of auctions can be tested. Auction economics is a broad and deep scientific field and it is also called "auctionomics". However, it was not possible to reach sufficient amount of data for every kind of auction until online auctions have emerged. By the new auction era we have reached by the virtue of internet, it has also been observed by the studies that the behaviors and actual equilibrium in the real auctions may be different from the theory. For instance, the most basic concept in the auction science is the Revenue Equivalence Theorem (RET), that is, under certain conditions, the seller's income is expected to be same regardless of the type of auction. However, as Bajari and Hortacsu (2004) implied, the use of ascending auctions by a large number of firms in the online auction house sector shows that the theory may not be validated in practice. For this reason, it is a worthy area of study to determine the factors that affect online auction mechanisms, the players' strategies and the factors affecting the price.

Ockenfels et al. (2006) clearly identified why sellers prefer auctions, stating that when sellers are less certain about the market value of their products, they prefer online auctions for price discovery. The expected returns of price discoveries have increased by virtue of the possibilities offered by technology i.e. the decrease in the transaction costs, search costs and hassle costs. In addition to these, especially the companies such as eBay, Amazon, which makes sales through auctions, and the companies such as Google, as Milgrom and Tadelis (2018) lately mentioned, which determine the internet search advertisement prices through real-time auctions financed and supported this kind of research studies. For this reason, the auction studies are still among the trending topics in both business and academy.

We can discuss the studies in this field in two fundamental categories. The first one is the examination of online auctions in a game theoretical framework. In online auctions, both sellers and buyers are engaged in strategic behaviors. Sellers determine their strategies by determining product and auction characteristics as revealed by the eBay system. Buyers determine their bidding decisions based on both the vendor and their competitors' strategies. For this reason, some studies focus on a game theoretical framework of online auctions, focusing on the equilibrium of games and players' strategies. Buyers' strategies can be analyzed through the bids. It is noteworthy that most of the bidders place a bid just before the end of the auction, many players place a bid only once, and some of them place a bid as soon as the bidding starts, they enter the game with small bids and incrementally increase their bid price over time. The former is called bid snipers and the latter is called incremental bidders in literature.

Ariely et al. (2005) identified that the players choose late bidding strategy that is they place a bid in the last seconds only in hard-close auctions. In other words, in the case of Amazon-type auctions, where auction closing time can be extended, such a behavior does not appear as an equilibrium. Furthermore, the authors stated that the tendency of last second bidding increase with the bidder experience. Ockenfels and Roth (2006) elaborated this issue and stated in their paper that a lot of bids were placed in the last seconds and most winners placed a bid in the last 3% of the auction time. They also argued that bid snipping is the best response to incremental bidding strategy.

The auctions on eBay are common value auctions by and large. That is, the product in auction has the same value for everyone, but this value is unknown by the bidders. In such auctions, Bajari and Hortacsu (2004) mentioned that bidders cut down on their bids strategically in order to avoid the "Winner's Curse" and bid in the last seconds to avoid a potential "bidding war". They also found that it was a weakly dominant strategy for the bidders to bid his reservation value for eBay type auctions. Similarly, in the paper written by Bose and Daripa (2017), late bidding strategies appear as an equilibrium in eBay auctions. The sellers' strategic decisions are concentrated on auction characteristics such as auction minimum bid, reservation price, auction length, auction duration. These studies give us important information on how dealers act in various auction mechanisms. However, in the case of an unspecified number of bidders during the auction period, identification of these strategies does not suffice to develop a predictive price model.

The second group studies in auction literature are the auction price models namely, structural and empirical models. Bajari and Hortacsu (2003) and Bajari and Hortacsu (2005) specified structural auction models by the distribution of bidder valuations and the number of bidders. As Bajari and Hortacsu mentioned, structural auction models are based on 3 strong assumptions. First, the bidders try to maximize their utility. Second, the bidders can determine the probability of winning the auction by the bid they place. Third, the bidders given their beliefs accurately maximize their utilities. However, it is noteworthy that the bidders do not always satisfy the above assumptions in actual auctions. Therefore, researchers critical thinking about the applicability of these strong assumptions were emphasized in the paper. In addition, Rezende (2008) criticized that structural models are usually established for specific auction types and the estimations are very complicated. Bajari and Hortacsu (2003) were agreed the difficulty in the estimation process of common value auctions stating that the structural model is technically very challenging. Therefore, many of the applied researchers refrained from using structural models because most assumptions are overly strong and they come with many technical difficulties. After brief information about those, it can be said that game theoretical framework and structural auction model issues are out of scope in this dissertation due to their difficulty to produce an easy and speedy solution for price estimation process. To be clear, the aim of the study is to produce simple and fast solutions and determine empirical factors affecting the auction closing price. In order to achieve this goal, we can use today's bountiful data supplies and algorithmic capabilities. We will mention a few more papers below that inspired us to conduct an empirical study.

First of all, Athey and Imbens (2015) expressed that non-parametric estimation methods offer much more successful results with big data in the predictive model business. It is also a fact that in these models, prediction performance can be improved by trial and error method and cross-validation methods. In other words, models developed by Machine Learning (ML) algorithms are not only easy to implement but also reliable. For this reason, ML models are frequently used in sectors i.e. technology, finance, statistics, genetic science and neuroscience which pay attention to forecasting performance. Frongillo (2015) also stated that for today's decentralized market economy, distributed nature of recent ML optimization algorithms is a much better match. Athey (2017), also supported this claim saying that ML methods are frequently used in many fields satisfying data abundance, computational techniques and resources. She also explained that the simplicity of these methods comes from the fact that they depend on only a few assumptions. However, she also mentioned the risk of disregarding the causality element in these methods. Some of the empirical studies conducted in the field of auction literature have been tested experimentally by researchers. Therefore, following the literature, there might not be a causality problem in the methods we have employed. Milgrom and Tadelis (2018) explained that marketplace and retailers are able to determine customer demand, target market segmentation and supply accurately by means of high forecasting performance obtained by AI and ML methods. They also emphasized that these methods will play a big role in future market design with prompt and attentive responses to emerging problems. For instance, in internet search ads, the ranking should realize within milliseconds. Even the models that solve such a dynamic problem in just a few seconds are slow and can be solved merely by ML methods. Therefore, it is clear that there is a need for algorithms for parallel and decentralized processing. Moreover, Natural Language Processing (NLP) algorithms, which are widely used today, have brought a new breath to economic models. For example, Milgrom and Tadelis (2018) pointed out that customer satisfaction can be determined by examining the messages exchanged between the buyer and the seller by electronic commerce firms. This can also reduce the friction in the market and increase the efficiency. All these and more show the applicability and unparalleled success of the new generation ML methods. We have chosen to use ML methods in this dissertation because of their high performance, high accuracy, speed, and ability to extract meaningful information from textual data.

In this context, we have gone through a rich literature. The following is a table of empirical studies with conventional and ML methods which provides a brief summary in this context.

Table 3.1: List of Empirical Studies on Auction Prices and Findings

Reference	Items Sold	Type of Study, Model Specification	Covariates	Results
Melnik and Alm (2002)	Gold Coin	Tobit Maximum Likelihood Dep Var: Price	Seller Ratings, Shipping Cost, Insurance, Photo scan, Auction Closing Time, Auction Length, Credit Card, Gold Price	Seller's Reputation And Some Variables Has A Consistent, Significant And Positive Impact On The Price
Lawrence (2003)	Computer	Naive Bayes Classification Dep Var: Win/Loss	Industry Name, Incumbency Number of Employees, Profit Margin, Part Quantity, Financing Opportunity High-Profile Account, Internal Advocate At Buyer Part Revenue Opportunity, RFQ Revenue Opportunity Identity Of Competitors, Services Opportunity	Approximately 98% Of The Asymptotic Accuracy Is Captured By The First Seven Features
Bajari and Hortacsu (2003)	Coin	Structural Econometric Model Tobit Regression Dep Var: Winning Bid / Book Value	Number of Bidders, Minimum Bid/Book Value, Secret Reserve, Blemish, In (Negative Feedback), In (Overall Reputation), Average Bidder Experience	Blemishes And Negative Reputation Decrease Price But Overall Reputation Increases The Sales Revenues. Bidder Experience Do Not Affect The Price
Ghani and Simmons (2004)	PDA	Linear, Polynomial Regression, CART Multi Class Classification Dep Var: Price	Seller Features, Item Features, Auction Features, Temporal Features	Model Performances Are Linear Regression (MSE 5.9) Cart (MSE 5.4) Multi Binary Classification Is Better (96% Accuracy)
Wang et al. (2004)	Palm M515 PDA	Functional Data Analysis Polynomial Spline Penalized Smoothing Spline Dep Var: Price	Current Price, Opening Price, Interaction Between The Opening Price And Time, Price-Velocity, Price-Acceleration And Price-Jerk , Current Average Bidder Rating, And Current Total Number Of Bids.	The Forecasted Value Of The Final Price Is 224.37 Dollars, While The Mean Of The Actual Final Price Is 228.34 Dollars.
O'Regan (2005)	Ipod Mini Ipod	Regression Dep Var: Price	Bids, Length, Start, Extend, Day, Xmas, Constant	Extend, Day, Xmas variables are Significant Extended Auction Brings Prices Up
Ockenfels and Roth (2006)	Computers	Regression, Probit Estimation Dep Var: Late Bidding	Number Of Bidders, Ebay Feedback Number, Binary Numbers For Auction House And Item Category	Late Biding Is More Common and Significant On Ebay Than On Amazon

Table 3.1 Continued

Houser and Wooders (2006)	Intel Pentium III Processors	2 Step GLS Dep Var: Price	Seller Reputation, Buyer Reputation, Auction And Product Characteristic Variables, Market Price	Seller Reputation (but Not Bidder) Is A Statistically And Economically Significant
Hou (2007)	LCD Monitors	Nonlinear Regression Dep Var: Price	Starting Bid, Reserve Price, Picture Length Weekend Bidder Expertise, Seller Reputation, Control Variables (Shipping, Bidders, New)	Starting Bid And Seller Reputation Similar Effect On Price, Product Pictures Expertise, Seller Credibility, And Weekends are important
Lucking-Reiley et al. (2007)	Coin	Regression Dep Var: Price	Bookvalue, Minbid, Reserve, Pos, Neg, Numdays, Weekend, Dummy Variables For Duration And Number Of Bidders	Seller's Ratings, Specially, Negative Ones, Have An Effect On Price. Minimum Bids And Reserve Prices Have Positive Effects And Auction Duration Increases The Auction Price.
Raykhel and Ventura (2009)	Laptops	Feature Weighted KNN Dep Var: Price	Brand, Family, Series, CPU Type, Multi-Core Configuration, CPU Speed, Ram Size, Disk Size, Lcd Size, Operating System, Optical Drive Type, Condition, Seller Feedback, Seller Powerseller Level And Duration, Title, Record Some Auction- And Seller-Related Data	Average Prediction Error Is 16%. 562% Increase In Trading Efficiency
Hortacsu et al. (2009)	30 Main Categories	Regression Dep Var: Number Of Transactions	The Description Of The Item, Location, The Shipping And Handling Fee, The Seller’s Feedback Rating, The Insurance And Payment Methods, Listing Time	Distance Is An Important Deterrent To Trade Between Geographically Separated Buyers And Sellers, The Authors Found Strong Home Bias
Jank et al. (2010)	Digital Camera	Beta Model Penalized Smoothing Splines Exponential, Logistic Growth. Dep Var: Price	Opening Price, The Auction Duration Or The Price Velocity	The Beta Model Can Provide Fast And High Accuracy In Price Predictions.
Cabral and Hortacsu (2010)	Coins, IBM Thinkpad, Notebooks , T. Beanie Babies	Cross Section Regression Dep Var: Price	Reputation Measure, Other Demand Factors	Negative Feedback Causes A Drop From A Positive 7% To A Negative 7%; Subsequent Negative Feedback Ratings Causes 25% More Rapidly Than The First One

Table 3.1 Continued

Kaur et al (2012a)	Palm M515 PDA	K-Means Clustering Regression Tree Dep Var: Price	Opening Bid, Closing Price, Number Of Bids, Average Bid Amount, And Average Bid Rate	Clustering Improved the End Price Prediction Performance By 7.46% when the AvgBR is very low. 32.52% when it is medium.5.55% when the AvgBR is high.59.72% when the AvgBR is very high.
Storch (2013)	6000 Auctions	Two-Level Hierarchical Approach Clustering KD-Tree, KNN Linear Regression Poisson Regression Dep Var: Price	Per-Auction Feature Vectors, Per-Auction Feature Vectors Such As The Buy-Now Price Of The Good Being Auctioned And The Retail Category Of The Good. Bid, Bid Time, User	Knn-Lin Has The Largest Rss And Knn-Poi Is The Favored Method.
Grossman (2013)	Sports Autograph Category	Logistic Regression Classification And Regression Trees KNN Dep Var: Sales And Price	Price, Starting Bid, Bidcount, Title, Quantitysold, Sellerrating, Selleraboutmepage, Startdate, Enddate, Positivefeedbackpercent, Haspicture, Membersince, Hasstore, Sellercountry, Buyitnowprice, Highbidderfeedbackrating, Returnsaccepted, Hasfreeshipping	Logistic Regression Provided 86% Success Rate for Predicting the Sale CART Prediction Using KNN for Eliminating Obvious Outliers Does a Good Job Predicting Final Sale Price
Lackes et al. (2013)	Levi's 501 (New)	Neural Network, Decision Tree Dep Var: Price	Presentation-Specific Factors, Seller-Specific Factors, Auction Procedure-Specific Factors, Number Of Pictures, Number Of Ratings, Ratio Of Positive Ratings, Duration Of The Auction, Ending Day, Ending Time, Shipping Costs, Starting Price	Distinguished Between A Below-Average Or Above-Average Price. The authors found The Experience And The Reputation Of A Seller Is The Most Important Factor
Nicholson and Paranjpe (2013)	Music - Records	Multinomial Logistic Regression, Naive Bayes (NB), And Uniform Prior Naive Bayes (UPNB) Dep Var: Sale And Price	Classifier Based Solely On The Item Title Text Item Number, Item Title, Start Time, End Time, Text Description, Detailed Item Specifications1, Seller Feedback Percentage And Total Score, Shipping Costs, Return Policy, And Image Presence.	For Sales Prediction, NB Classifier shows best performance, For Final Price Prediction, UPNB Performs Best Start Price Is Not Always A Good Indicator Of The End Price

Table 3.1 Continued

Lin et al. (2013)	Digital Camera	Neuro Fuzzy Approach Markov Chain Model Expert System, Fuzzy Logic, And Neural Network Dep Var: Price	110 Digital Camera Auctions Starting Price, External Reference Price, And Final Price Of An Auction	Neuro Fuzzy Performs the Best Among These Five Methods External Reference Price Plays an Important Role in Determining the Final Price
Gupta and Pathak (2014)	Various	Dynamic Pricing, Regression K-Means Clustering, Logistic Regression Dep Var: Price	Visit Attribute, Demographic Profile, Context, Purchase History, And Purchase Intentions Derived Variables: Purchase By Offer (POR), Purchase By Category (PCT), Purchase By Quantity (PQT), Purchase By Company (PCY), Purchase By Brand (PBD), And Purchase Channel (PCN).	The Error Rate Decreased For Much Better Price Ranges, Which Is Optimum For Both Customer And Organization
Kaur et al (2014)	Palm M515 PDA	Clustering, Multiple Regression, Fuzzy Reasoning Dep Var: Price	For Classification: Opening Bid, Closing Price, Number Of Bids, Average Bid Amount And Average Bid Rate	Price Prediction Performance Increases By Clustering. The Fuzzy Bidding Strategies Are Proficient In Terms Of Success Rate And Expected Utility. The Model Preserves A Balance Between Bidders Attitude And Auction Competition . Outperforms The Classic Model Of Final Price Prediction.
Mary et al. (2014)	Not Defined	Multinomial Logistic Regression, Naive Bayes (NB), And Uniform Prior Naive Bayes, Linear, Polynomial Regression, CART Dep Var: Sale And Price	Item Title, Price Velocity, Price Accelaration, Static Predictor Variables, Time-Varying Predictor Variables, Price Dynamics, Price Lags. (auction Variables)	NB classifier predicts whether the sales will occur best For Price Prediction, UPNB outperforms others

Table 3.1 Continued

Einav et al. 2015)	Various	Fixed Effect Regression Dep Var: Sale And Price	BIN Price, Starting Price, Duration, Shipping	Price Dispersion Across Equivalent Auctions and The Relationship Between Auction Prices And Equivalent Posted Prices are Estimated, The Shape Of Auction Demand, The Effect Of BIN Prices, And Internalization Of Shipping Fees Were Analyzed. Authors Found Sellers Do Not Converge In Their Listing Behavior For A Given Item
Kaur et al (2017)	Exp.	Regression Negotiation Decision Function Dep Var: Price	Auction Characteristics, Bidder Characteristics	Fuzzy Reasoning With A Regression Approach Outperforms Other Models
Chow (2017)	Plate Prices	A Deep Recurrent Neural Network (RNN) Natural Language Processing Ensembling Dep Var: Price	Characters On The Plate	A Deep Recurrent Neural Network Provides Best Performance RMSE: Combined: .5527, Combined + Extra: .5298 R2: Combined: .8177, Combined+Extra: .8325
Kumar and Rishi (2017)	Various	Regression, Clustering, Hybrid Dynamic Functional Forecasting Model. Regression Trees, Multi-Class Classification, Logistic Regression Dep Var: Price	Opening Bid, Closing Price, Number Of Bids, Average Bid Amount, Average Bid Rate, Auction Ending Behavior And The Seller's Reputation, Social Media	By The HDAM Model, An Online Auction To Participate In Is Determined, Clustering And Regression Process Is Applied For Prediction In View Of The Evaluated Starting Price, State Of Mind Of The Bidders And Competitors' Behaviors
Li (2017)	Canon EOS 6d	Multiple Regression Analysis Dep Var: Auction Price	Seller Reputation, Last-Minute Bidding, Reserve Price, Length Of Auction,	Seller Reputation Has A Significantly Positive Effect On Auction Price; But, This Effect Can Be Moderated By The Last-Minute Bidding Behavior
Chan and Liu (2017)	Xbox	Semiparametric Regression Model B-Splines Dep Var: Price	Condition, Reserve Price, Early Bidding, Jump Bidding, Opening Bid, Winning Bid, Number of Bids, Seller Rating, Bidder Rating	By The Model Price Forecast Can Be Done Dynamically And The Prediction Can Be Updated By Newly Arriving Information

Table 3.1 Continued

Tseng et al. (2017)	Various	Sentiment Analysis AR Model And Moving Average Model, SVM , Regression Analysis (Linear, Non Linear, Logistic) Dep Var: Price	Historical Price Data Attributes Of The Products Keywords About Related Electronic Products In News	Proposed Method Prediction Performance Is Better Than The Traditional Methods
Einav et al. (2018)	Various	Linear Probability Model With Fixed Effects Dep Var: Sale And Price	Auction Indicator Variable, Auction Start Price	Shift From Auctions To Price Posting Decline In The Relative Demand For Auctions Decline In Auctions Was Driven By Changing Seller Incentives.

There are lots of papers on the auction price literature listed above and many more it couldn't be specified here. However, there are still gaps in the literature. For example, Melnik and Alm (2002) stated that if a vendor could not express their product quality properly, there might be a market failure. Benkard and Bajari (2005) elaborated this, if some of the product features cannot be seen, the models can have a bias, for instance, if the CPU benchmark information is missing in the personal computer price model, it will create a large bias. Considering these two studies, it should be stated that all possible features of the products need to be included in the price models to prevent the market or model failure.

Ockenfels et al. (2006) also stated that it is quite difficult to bid on a common value auction. First of all, it is necessary to estimate the expected value of the product. In order to avoid “Winners Curse” the estimation should be done correctly and the right bid should be determined which is not easy. Ebay provides a proxy bidding system and recommends their customers to use this system. Ebay says that the client can enter the maximum Willingness to Pay (WTP) values for the item and that the proxy bidding system will automatically increase their bids on behalf of them and this will prevent them from being unfairly outbid in the last seconds. Surprisingly, it is found that customers do not tend to use this system. This is mostly due to the fact that it is difficult for them to accurately estimate the value of the product, in other words, it is very difficult for them to determine WTP. Customers observe their competitor's bids as signals and update their valuations in the bidding time upon their newly-formed interpretations. For this reason, we can say that there is still a need for a system that makes it easier for customers to predict the value of the product.

Houser and Wooders (2006) also expressed that they tried to include all kinds of information about the product in their models, however, they were only able to add the textual information in the product title to their models as for product conditions. In this case, the product description including the broader information, accessories, warranties and shipping cost could not be included in their models causing the risk of an inefficiency problem. In fact, the lack of these variables in the models will create a bias problem in addition to inefficiency. Similarly, Lucking-Reiley et al. (2007) mentioned that some dealers have entered more accurate product descriptions, which is likely to have an impact on customers yet the authors did not include this information in their models. Therefore they were in doubt about a bias problem in models. To summarize, the descriptions written in the sales advertisement provide an important piece of information about the product. This importance was emphasized in the literature, but it was not easy to incorporate it into the models by the technological means of that period. However, with NLP algorithms developed by today's technology, these explanations can be turned into numerical variables and added to price models easily. We aim to fill this gap in the literature with this study. That is, with this dissertation, we have developed an easy, high-speed and reliable price forecasting model using all available variables including product descriptions.

When we review the literature, reference prices and approximate market values are missing in price estimation models developed with machine learning algorithms. These studies merely cover static or dynamic features of the auction but do not include any reference price values. We all know that the products have market values and we need to use some benchmark values in the price models. Customers collect data about the products by conducting a market research and they form a valuation about the product. Thus, some economic features are necessary to predictive auction price models. The studies conducted without economic features can be considered as biased because they do not contain adequate variables correlated with the auction prices.

To sum up, in the literature, the product description written by the sellers have not been analyzed up until now. The effect of accessories such as boxing, warranties, battery chargers, and connector and cables specified in second-hand electronic product sales have not been investigated either. The bidder's rating for the product has rarely been used. Although the auction end time has been studied frequently, auction start time has been disregarded in the literature so far. In addition to the reference or market value of the products, the features, which are essential for the price formation, will be identified by exploratory analysis. Moreover, as mentioned over the coming chapters, there is valuable information at the initial stage of the auction, which indicates the first 20 percent of the bidding time. Can the information aggregation that occurs at this stage of the auction be useful in auction price models? This question, which has not been addressed in the literature up to now, is also examined and the bids, bidders, bidder ratings available at the beginning of the auction will be profoundly analyzed to shed a light on auction price determinants.

3.2 Theoretical Background

In this research, it is aimed to determine the factors that affect the auction end price and to develop a model that can predict the price. Current auction models can be divided into two namely structural and empirical models. Theoretical models are mostly called as structural models which are based on the valuation distributions of the bidders. In the context of the auction theory, the structural models mentioned by Levin (2004) will be briefly explained below. In empirical models, all variables that can affect the bid price are included in the models linearly or non linearly. Although empirical models sometimes are criticized in the literature, they are used for price estimation because of ease of application. In addition, article proposed by Rezende (2008) that can fill the gap between theoretical and empirical models has been also published and will be mentioned below to form a theoretical background for this study.

Below, structural auction models described by Levin (2004) will be summarized. In general, auctions can be divided into two groups according to the information and valuation sets. The first one is the Independent Private Value (IPV) auction models and the other is the Common Value (CV) auction models. On the other side, there are generally two types of auctions according to the end prices. The first of these is the "first price auction" and the other is the "second price auction" which is also called as Vickrey (1961) auctions.

The Independent Private Value (IPV) Model

One object is on sale, there are *n* bidders and bidder *i* observes a signal S_i~F() with realization of $s_i \in [s, \bar{s}]$. Assuming F is continuous, bidders' signals ($s_1, s_2, s_3, ... s_n$) and valuations are independent then bidder *i*'s valuation is $v_i(s_i) = s_i$. The information of bidder *i* i.e the signal is private in the sense that it determines only the bidder's valuation and does not affect any other one's valuation.

Sealed Bid (First-Price) Auction

In this type of auction bidders submit their bids (b_1,..., b_n) as sealed. The highest bidder wins the auction and pays the bid she submitted. Nevertheless, the bidder may not bid her valuation since then she ensures a zero profit. Determining a bid below her valuation may bring some profit and there are two approaches to solve symmetric equilibrium bidding strategies and they are explained below.

A. The "First Order Conditions" Approach

Consider bidder i's problem. Bidder *i* aims to maximize expected payoff as function of bid b_i and signal s_j. $U(b_i, s_i) = (s_i - b_i) * \Pr\left[(b_j = b(s_j) \leq b_i, \ \forall j \neq i\right]$ **(3.1)**

And bidder *i* determines bid b to solve: $\max\limits_{b_i} \ (s_i - b_i) F^{n-1}(b^{-1}(b_i))$ **(3.2)**

Thus the first order condition is at symmetric equilibrium

$$F.O.C \ : \ (s_i - b_i)(n-1)F^{n-2}(b^{-1}(b_i)) f(b^{-1}(b_i)) \frac{1}{b'(b^{-1}(b_i))} - F^{n-1}(b^{-1}(b_i)) = 0 \qquad \textbf{(3.3)}$$

where $b_i=b(s_i)$, FOC becomes a differential equation and dropping subscript *i*,

$$b'(s) = (s - b(s))(n-1)\frac{f(s)}{F(s)} \tag{3.4}$$

by using the boundary condition *b($\underline{s}$)=$\underline{s}$* the above equation can be solved as

$$b(s) = s - \frac{\int_{\underline{s}}^{s_i} F^{(n-1)}(\tilde{s})d\tilde{s}}{F^{(n-1)}(s)} \tag{3.5}$$

B. The “Envelope Theorem” Approach

Envelope theorem is another useful approach to identify the necessary conditions for a symmetric equilibrium. The equilibrium payoff for bidder *i* given signal s_i is can be represented as below:

$$U(s_i) = (s_i - b(s_i))F^{n-1}(s_i) \tag{3.6}$$

Since the player *i* will play best response strategy in equilibrium the above equation can be written as follows:

$$U(s_i) = \underset{b_i}{max}(s_i - b_i)F^{n-1}(b^{-1}(b_i)) \tag{3.7}$$

By the envelope theorem stated by Milgrom and Segal (2002),

$$\frac{d}{ds}U(s)\Big|_{s=s_i} = F^{n-1}(b^{-1}(b(s_i)) = F^{n-1}(s_i) \tag{3.8}$$

and also

$$U(s_i) = U(\underline{s}) + \int_{\underline{s}}^{s_i} F^{(n-1)}(\tilde{s})d\tilde{s} \ \ where\ U(\underline{s}) = 0 \tag{3.9}$$

Combining (3.6) and (3.9) for equilibrium strategy and dropping subscript *i,* equilibrium bids function can be represented as follows which is same as the one FOC condition approach.

$$b(s) = s - \frac{\int_{\underline{s}}^{s_i} F^{(n-1)}(\tilde{s})d\tilde{s}}{F^{(n-1)}(s)} \tag{3.10}$$

We can see that, two approaches give the same bid equations by (3.5) and (3.10)

Common Value Auctions

In this model, bidder *i* can re-evaluate her valuation after learning the bidder *j*'s information which means information of bidder *i* is not independent of the bidder *j*.

There are *n* bidders (*i=1,...,n*) and signals *(S_1,..., S_n)* with joint density *f(.)* where signals are exchangeable and affiliated and bidder *i*'s valuation is $v(s_i, s_{-i})$ where;

$v(s_i, s_{-i}) = s_i$ Value to bidder *i* is s_i and it depends on both bidder's and other bidders' signals

Pure common value is special case in which all bidders have same value, a random variable *V*, the signals $S_1,\ldots, S_n$ are correlated with *V* but independent conditional on it (ε_i is are independent), Then

$S_i = V + \varepsilon_i$

So valuation of bidder *i* is dependent on all of the bidders' signals.

$$v(s_i, s_{-i}) = E(v|s_1, s_2, \ldots, s_n)$$

Vickrey (Second-Price) Auction

In this type of auction bidders submit their sealed bids and the bidders with highest bid wins the auction but pays the amount of second highest bid and weakly dominant strategy for the bidders to bid their valuations $b_i(s_i) = s_i$. (Levin, 2004)

Second Price Auctions

Suppose s_i is the highest signal of bidders *j≠i* and bidder *i* can win if she bids $b_i \geq b(s^i)$ and pays $b(s^i)$.

For a symmetric equilibrium, The bidder *i* problem can be represented as follows:

$$\max_{b_i} \int_{\underline{s}}^{\bar{s}} \left[E_{s_{-i}} \left[V(s_i, S_{-i}) \middle| s_i, S^i = s^i \right] - b(s^i) \right] 1_{\{b(s_i) \leq b_i\}} f(s^i | s_i) ds^i \qquad \textbf{(3.11)}$$

and the first order condition is

$$0 = \left[-\frac{1}{b'(b^{-1}(b_i))} \left[E_{s_{-i}} (V(s_i, S_{-i}) | s_i, S^i = b^{-1}(b_i) \right] - b(b^{-1}(b_i) \right] * f(b^{-1}(b_i) | s_i) \qquad \textbf{(3.12)}$$

simplifying the above equation

$$b_i = b(s_i) = E_{s_{-i}}\left[V(s_i, S_{-i}) \middle| s_i, \max_{j \neq i} s_j = s_i \right] \qquad \textbf{(3.13)}$$

in equilibrium, bidder *i* will submit a bid conditional both on her own signal and all competitors' signals having a signal less than hers.

Considering Econometrics of Auctions by Least Squares:

The models described above are called structural models, and methods described in literature are favorable for equilibrium bid estimation. There are many assumptions such as valuation distribution for structural models. Furthermore, the estimation of these models is quite difficult. Although an equilibrium is satisfied with structural models, price forecast is not so easy. Therefore, researchers in general use linear regression models rather than structural models. Rezende (2008) proposed a method that is theoretically logical and easy to implement against criticisms of linear regression models.

According to the author, using the price information, *p* as dependent variable where *X* stands for explanatory variables, in (3.14) prevents the empirical models to be a structural model.

$$p = X\beta + \varepsilon \qquad \textbf{(3.14)}$$

The author therefore emphasizes that it would be more accurate to have valuation, $V_{i,}$ on the left side of the model like (3.15) instead of price. Customer valuation might also be seen a sense of willingness to pay.

$$V_i = X\beta + \varepsilon_i \qquad \textbf{(3.15)}$$

The author concentrates on estimating the factors that determine the location and scale of the valuation of bidders. Below the assumptions and details of the model proposed by the author are provided.

Assumption 1: Valuation of bidder *i* at auction *l,* $V_{il,}$ can vary around a mean, $\mu_{l,}$ by standard deviation, $\sigma_{l,}$ where ε_{il} are independently and identically distributed with F distribution. So, following formulation can exist: $V_{il} = \mu_l + \sigma_l \varepsilon_{il}$

The aim is evaluating the covariates affecting μ_l and σ_l. Assume X_l and Z_l are the exogenous, deterministic, publicly known regressors affecting μ_l and σ_l.

Assumption 2: The relation is linear and $\mu_l = X_l\beta \text{ and } \sigma_l = Z_l\alpha$

Assumption 3: The total number of bidders n_l is exogenous and known before the auction by all participants.

Assumption 4: The rule in auction such that the good is awarded to the highest bidder and the lowest bidder do not pay anything.

Theorem 1: Under the Assumptions 1, 3, 4 the expected price for the auction l is $E[V_{(2:nl)l}]$ by the all information publicly available at the time of the auction below equality can be established.

$$E\left[p_l \middle| X_l, Z_l, n_l\right] = E\left[V_{(2:n_l)} \middle| X_l, Z_l, n_l\right] \qquad \textbf{(3.16)}$$

The author proposes two estimation procedures according to information on F distribution.

Estimation When F is Known

Under the above mentioned assumptions and available data for auction prices, covariates and number of bidders, expected auction end price relation can be established as follows:

$$\begin{aligned} E\left[p_l \middle| X_l, Z_l, n_l\right] &= E\left[V_{(2:n_l)} \middle| X_l, Z_l, n_l\right] \\ &= E\left[\mu_l + \sigma_l \varepsilon(2:n_l) \middle| X_l, Z_l, n_l\right] \\ &= \mu_l + \sigma_l E\left[\varepsilon(_{2:n_l})\right] \end{aligned} \qquad \textbf{(3.17)}$$

Defining $$a(n) \equiv E\left[\varepsilon_{(2:n)}\right] = n(n-1)\int tF(t)^{n-1}(1-F(t))dF(t) \qquad \textbf{(3.18)}$$

So, expected auction price can be simplified to

$$E\left[p_l \middle| X_l, Z_l, n_l\right] = X_l\beta + \alpha(n_l)Z_l\alpha \qquad \textbf{(3.19)}$$

Conditional expectation of auction price is linear in *β* and *α* which supports unbiased and consistent estimation method for the coefficients. When F is known the author propose simple procedure to estimate *β* and *α.*

Method 1: By using the standardized value distribution F, *a(n$_l$)* for all values of *n$_l$* in the sample can be computed and auction price can be regressed over all variables.

Estimation When F is Not Known

The above mentioned method has a significant disadvantage. The F distribution for the model must be known in advance, but this will not always be possible. In this case, the *p$_l$* model can be used instead of *a(n)* function. So a dummy variable, *d$_{kl,}$* can be created for each bidder number.

$$\mu_l = X_l\beta = \beta_0 + x_l\beta_1 \text{ } then \qquad \textbf{(3.20)}$$

$$\begin{aligned} E\left[p_l \middle| X_l, \{dk_l\}\right] &= x_l\beta_1 + \sum_k d_{kl}\left[\beta_0 + \sigma\alpha(k)\right] \\ &= x_l\beta_1 + \sum_k d_{kl}\delta_k \text{ } where \text{ } \delta_k = \sigma\alpha(n) \end{aligned} \qquad \textbf{(3.21)}$$

$$X\beta = Y\beta_1 + x\beta_2 \text{ } and \text{ } Z\alpha = Y\alpha_1 + z\alpha_2 \text{ } then$$

$$\begin{aligned} E\left[p_l \middle| X_l, Z_l, \{dk_l\}\right] &= x_l\beta_2 + \sum_k d_{kl}Y_l\left[\beta_1 + \alpha(k)\alpha_l + \sum_k d_{kl}z_l\left[\alpha(k)\alpha_2\right]\right] \\ &= x_l\beta_2 + \sum_k d_{kl}Y_l\delta_{k1} + \sum_k d_{kl}z_l\delta_{k2} \end{aligned} \qquad \textbf{(3.22)}$$

$$\delta_{k1} = \beta_1 + \alpha(k)\alpha_1 \text{ } and \text{ } \delta_{k2} = \alpha(k)\alpha_2$$

Method 2: For every number of bidders *k* observed in an auction in the sample, construct dummy variables *d$_{kl}$* for the event *n$_l$=k.*

Based on the theoretical background and the empirical literature we can make a simple model proposition.

With the paper, Rezende (2008) has shown that auction prices can be modeled with linear models. However, according to the article, it was stated that the coefficients were statistically insignificant when the factors affecting both the auction average price and standard deviation were included in the model. The author stated that this might be due to the multicollinearity problem. In addition, there might be homoscedastic variance structure in OLS models, so it is not always necessary to add a variable to explain the model variance. Furthermore, the number of bidders in eBay type auctions is an endogenous variable. According to the information generated within the auction, bidders decide to enter in or exit from the auction. Therefore, it is not possible to know in advance total number of bidders to be formed within the whole period of the auction time. Thus, it will not be correct to include this information in the model. On the other hand, we can include the number of bidders in the initial period of the auction. This information is known to all bidders and can be considered exogenous. However, because there are many bidders at this stage, using dummy variable for each number of bidder will result in loss of degrees of freedom in the model and there might not be enough data in many clusters for model estimation. For this reason, in this study, we will add the level of number of bidders which is formed only in the initial stages.

CHAPTER 4

DATA DESCRIPTION AND FEATURE SELECTION

In previous studies, researchers appear to have used auction data from *modelingonlineauctions.com*. These datasets cover the period of 2003-2004 and they can be called "outdated" to analyze since the technology and preferences of the mobile tech users changed drastically in the past decade. In this dissertation, it is used eBay sales data for a single type of product, which is the Apple iPhone 7 Plus covering a time-frame of 120 days. The product and data description is provided in the table below. Relevant data is gathered with the help of *"import.io"* web crawler.

Table 4.1: Product and Data Description

Title	Description
Product Name	Apple iPhone 7 Plus
Product Condition	Used Item
Product Characteristics	128 GB, any color, any model
Number of Auctions Sales	444 Auctions
Number of Buy it Now Sales	676 BIN
Sales Condition	Sold Items
Delivery Option	Free Shipping
Network	Unlocked
Data Time	March 2, 2018- July 2,2018
Data Source	www.ebay.com

The prices of electronic products change radically as technology develops and companies introduce new models for all kind of products into the market on a regular basis. Thus, second-hand product prices can be affected over time. For this reason, the Apple iPhone 7Plus, which is the closest model to the latest model by the time this study has been carried out, has been selected. A second-hand model is used for analysis to represent some uncertainty in product conditions. Since the dynamics affecting the product price will change as the time elapses, the shortest possible period of time is selected, which can provide adequate amount of data to analyze. Thus, by choosing this time period, it is aimed to alleviate the time impact and focus primarily on the variables affecting the auction price.

4.1 Auction and Buy it Now Data

The description of the auction and BIN prices used in this study are given in the table below and distribution of the price data by time is given in the figure below. For the predetermined time interval, the sales data of 1120 instances including 444 auction and 676 BIN sales has been obtained through the search filter mentioned above. Although the prices of these two sales types are slightly different, their distribution over time is similar. Nevertheless, contrast with the literature, the average price in auction sales are slightly higher than BIN sales price. In order to confirm statistically, whether the average of the two price series is equal, two-tailed t-test will be used where the null hypothesis proposes equal means.

H_0: $\mu 1 = \mu 2$

H_A: $\mu 1 \neq \mu 2$

The result of the test is as follows: t statistic is 3.45 and p-value is 0.00058. Thus H_0 is rejected and we can see that the average of these two price series is not equal even in the 99% confidence interval. Although the average prices are different, it cannot be drawn any general conclusion for auction and BIN sales by merely looking at the test of these two series. Yet, when we look at the price difference between these sales it is only 13 dollars or 2.6%. Then, it can be concluded that the prices of these two types of sales are very close to each other. Although the items have been chosen by filtering with the same features, selected products are not exactly the same. First, the product may not be the same since it may indicate the different levels of deformation, there may be accessories offered with the product, and the characteristics of the sales and the sellers might be different, which can significantly affect the sales price. This may seem a drawback in analyzing second-hand electronic items. On the other hand, the presence of uncertainty in prices will give us an advantage in modeling the prices. Otherwise, it won't be easy to find out the factors affecting price since there will be no price change.

Table 4.2: Auction and BIN Price Description

	Auction Price	BIN Price
count	444.00	676.00
mean	495.44	482.77
std	60.38	59.66
min	276.00	249.99
25.00%	460.00	450.00
50.00%	500.00	484.99
75.00%	530.00	519.95
max	730.00	685.00

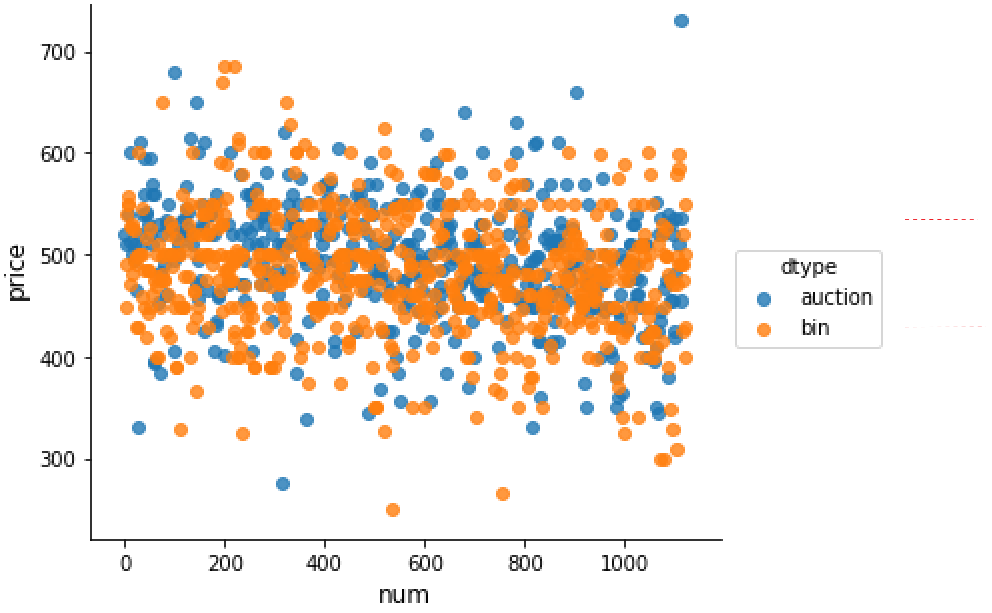


Figure 4.1: Auction Price, BIN Price

When the price data distributions given in the Figure 4.1 to Figure 4.3 are examined, the prices appear to have a distribution around a mean of 490$ and 70% of the prices are distributed between 450$ and 550$ for both type of sales. The remaining 30% of the prices are either less than 450$ or above 550$.

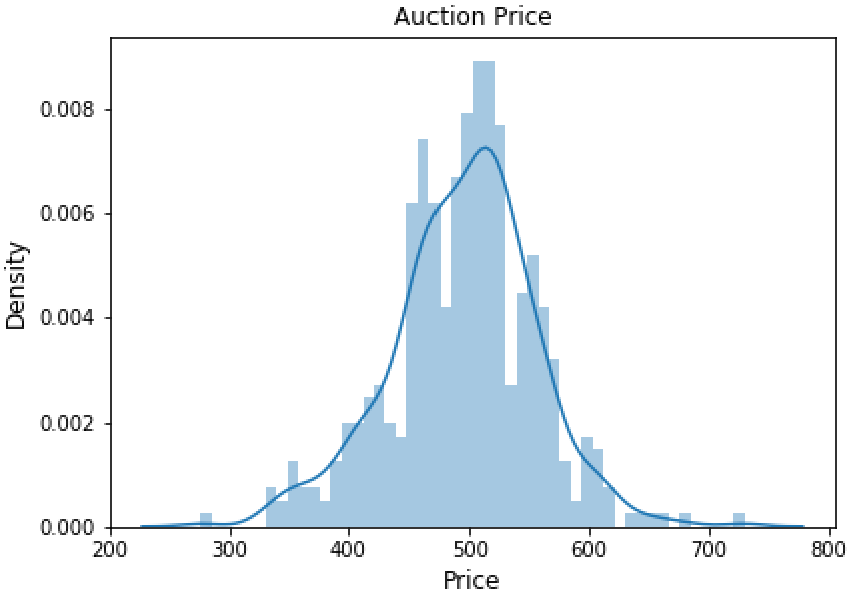


Figure 4.2: Auction Price Data Distribution

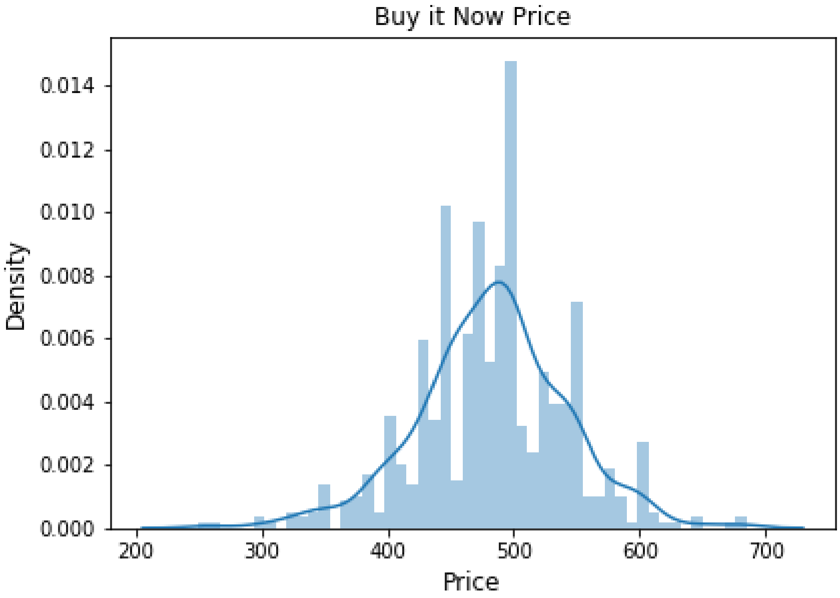


Figure 4.3: Buy it Now Price Data Distribution

According to normality tests based on D'Agostino and Pearson (1973) normality assumption is rejected for price series studies. However, it should be noted that the price data is adequate to assume normal distribution. To support this claim, the algorithm developed in python to predict the best suitable distribution to data is used. With this algorithm, normal distribution suits the data best among different distributions i.e. normal, exponweib, weibull_max, weibull_min, pareto, genextreme distributions. For these reasons, the price data is assumed to be distributed normal. In BIN sales, the sellers determine a fixed price tag and wait for the customer but the price is expected to be discovered by the buyers during the auction. Nevertheless, these two distinct sales strategies have similar price distributions. This implies that experienced customers have a strong belief in the real value of the product in auction by referring to the price information that earlier occurred in auction or BIN sales. This prevents the auction market separating from the BIN market. Therefore a price balance between these two markets can occur on eBay.

4.1.1 Static Features of Sales

These are the features of vendor or presentation-specific characteristics determined at the beginning of the auction and do not change during bidding time frame. These features are the same for Buy it Now sales. The list of static features is provided in the Table 4.3 below. Missing features in either pricing strategies are marked with ✗ a cross sign.

4.1.2 Dynamic Features of Auction Sales

These are the features clearly determined during the auction time, which are not obvious at the beginning of the auction. They are often based on the competitive behavior of the bidders in the auction. Buy it Now sales do not have these features. The list of dynamic features is given Table 4.4 below.

Table 4.3: List of Static Features of Sales

Static Features of Sales	**Auction**	**BIN**
Title of Sale Announcement	✓	✓
Date time of Sale	✓	✓
Day of Sale	✓	✓
Item Location	✓	✓
Seller Name	✓	✓
Seller Rate Total	✓	✓
Seller Rate Positive	✓	✓
Seller Rate Negative	✓	✓
Seller Rate Neutral	✓	✓
Seller Rate Percentage	✓	✓
Seller Top-rated	✓	✓
Seller Store	✓	✓
Seller Note for Product Description	✓	✓
Seller Followers	✓	✓
Seller Collections	✓	✓
Seller Guides	✓	✓
Seller Profile Views	✓	✓
Seller Collections	✓	✓
Seller Membership Duration	✓	✓
Product Release Date	✓	✓
Auction Sale Start Date time	✓	✗
Auction Start Day	✓	✗
Auction Start Price	✓	✗
Auction Duration	✓	✗
Extras	✓	✓
Shipping Cost	✓	✓

Table 4.4: List of Dynamic Features of Sales

Dynamic Features of Sales	**Auction**	**BIN**
Bids	✓	✗
Number of Bids	✓	✗
Number of Bidders	✓	✗
Number of Bids for Each Bidder	✓	✗
Bidder Rate	✓	✗
Bid time	✓	✗
Time Difference Between Bids	✓	✗
Difference Between Bids	✓	✗
Number of Bids for Winner	✓	✗

4.1.3 Economic Features or Reference Prices During Sales

Experienced buyers are conscious of their preferences and make detailed price research in advance. Buyers can view both the auction and buy it now prices of the previously sold products via eBay system. In this way, they can see the reference price for a product they would like to buy from the market. Then, they can determine the price valuation for the product by the earlier auction records. Economic features can be also considered as opportunity costs, best alternatives of similar products. In addition to this, the new product price available on Apple official website can be ceiling sales price for the auctions. Economic features used in this thesis are listed in the table below.

Table 4.5: List of Economic Features During Sales

Economic Conditions	**Auction**	**BIN**
New Apple iPhone 7 Plus Online Official Price	✗	✓
Apple iPhone 7 Plus Historical BIN Price	✗	✓
Apple iPhone 7 Plus Historical Auction Price	✓	✗

4.2 Data Preprocessing, Exploratory Analysis, Outlier Detection

4.2.1 Data Preprocessing

Ebay retains historical auction and BIN price data for traders up to past 4 months. We exploited the *import.io* web crawler to retrieve data from eBay database. This web crawler provides automatic macros for capturing data from websites. There is no need for extra coding for this, it has a kind of artificial intelligence and users need to indicate just by clicking which kind of data they need to crawl. When users create macros called *"extractors"* and train with relevant data, the system automatically detects remaining data in the tables and learns what to download. The crawler visits the URL list provided by the user and retrieves predetermined data. We made a product research under certain options mentioned above and listed products to get their url addresses. We created another web crawlers for static features, dynamic features, and seller information and obtained relevant data both for auction and BIN prices. Researchers can also use eBay API system to get data whereas we used import.io web crawler since it provides relatively an easier way. We created 5 different extractors at import.io. The first one is to get URL lists for static features, the second one is to get URL list for bid history, the third one is to retrieve static feature extraction, the fourth one is to obtain bid history data extraction and the fifth one is to get seller rating data from seller web pages on eBay. All data obtained from extractors can be saved as csv or excel files. The data obtained from web pages cannot be used directly in analysis most of the time and we need to preprocess the data before making any estimation. Otherwise, we can not use data as it is in Python or any other program. To prepare data for analysis, we have removed the URL links in the data table, converted the numbers which are entered as text type, removed the US and $ signs from the price table and changed date information to a form that the python system could interpret. Afterwards, we extracted the start and end day of the auction from date data. Then, we combined the title of the announcement, short seller note and long product description into a single text variable. We retrieved the seller and buyer rating information entered in parentheses next to the seller or buyer name as a new variable such as the seller rate, negative and positive rating and bidder ratings. These ratings are in fact reputation measures. We cleared some unnecessary gaps and signs in the data table. We carried out all these necessary preprocessing with the help of LibreOffice Calc and Python software and saved the all data table in csv format so that it can be used by Python.

4.2.2 Exploratory Analysis

The distribution of important variables is shown in the following histograms in order to provide an overview of the data used.

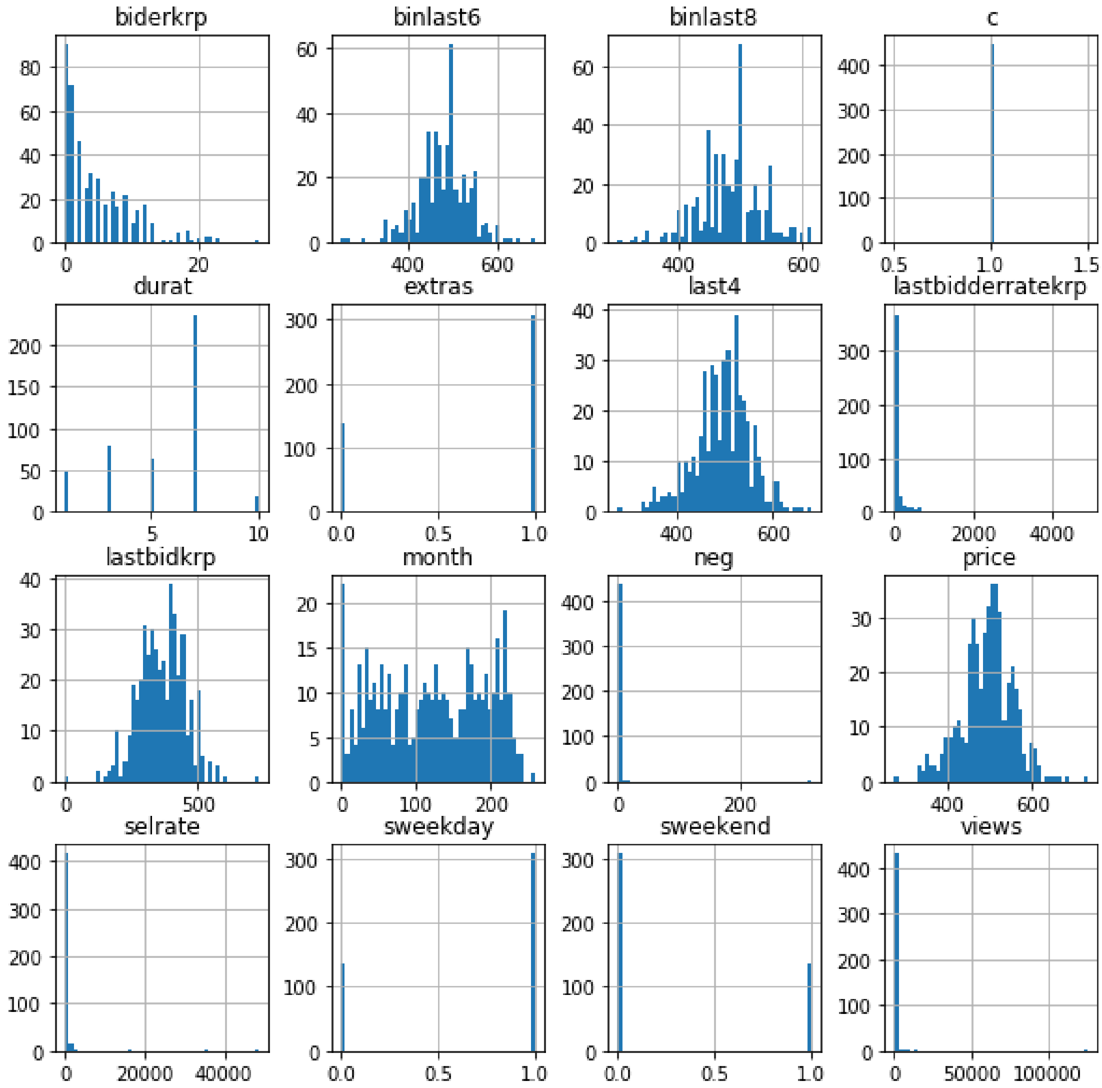


Figure 4.4: Histogram of Some Features

The number of bids and timing of bids for all of 444 auctions are represented in the figures below. The illustrations imply that bidders submit a number of bids at the beginning and at end of the auctions and this matter will be elaborated below.

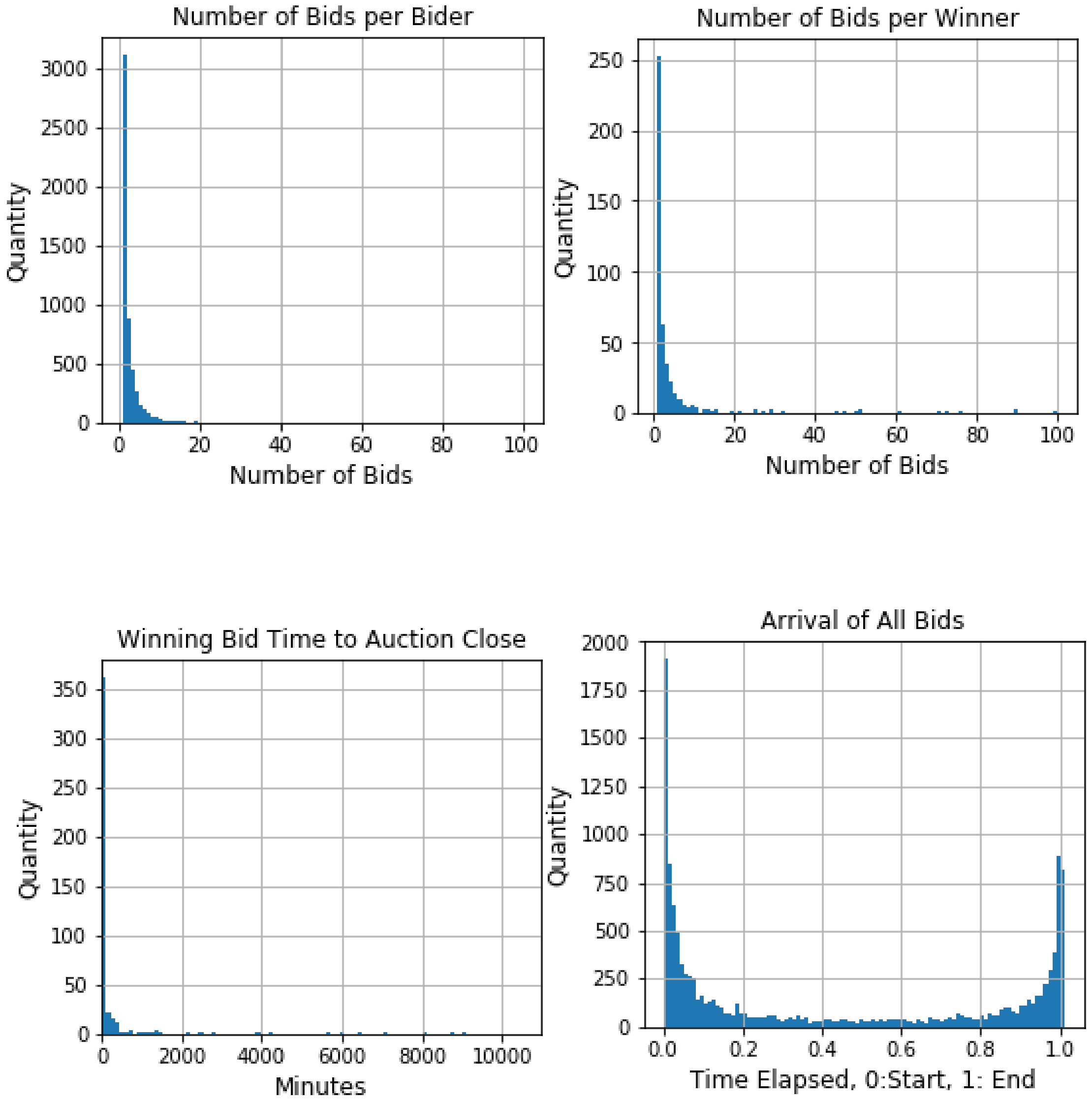


Figure 4.5: Density of Bids

Ebay allows vendors to organize fixed time auctions for many products. In this system, sellers decide when the auction will start, how long it will last and when it will sharply end. These time-frames are fixed and binding and cannot be intervened during the auction. The fact that the buyers can see all bids given in the system earlier and all participants know the deadline that the auction will end in a certain interval lead the customers to demonstrate certain strategic behaviors. The above-given graphs represent the density of the bids and timing of bids for all 444 auctions. These graphs provide crucial information in terms of understanding the behaviors of bidders in eBay type auctions. For example, in the chart at the top left, the majority of customers who bid on the product appear to only bid once during the auctions. Some customers, on the other hand, have tried to increase the chances of winning the auctions by bidding multiple times. The top right chart shows how many bids are placed by the winners. Similar to the previous graph, a large number of people who won the auction seems to have won the auction by only one bid. However, some of the winners won much more than only one bid. In the graph shown at the bottom left, time of winning bids as seconds to the auction close are represented. As can be seen clearly from this graph, most of the auction winning bids were given at a time very close to the deadline: Winning bids and a lot of competitor bids were given in the last second or before a few seconds to end. There are also winners who placed their bid long before the auction ends but these winners are not substantial. From these 3 charts, the inference is that there are two types of customers following the auctions. The first of these contains the people who follow the auction from the beginning to end and bid more than once during the auction time. Some of the customers bid 2-3 times and some of them offer as much as 20-30 bids. When multiple bidding customers are investigated, they are relatively less experienced buyers who demonstrate low customer ratings. In some articles such as Kaur (2014), they are called desperate buyers. In other words, the bidders sometimes regards this auction process as a kind of game and they always follow up the product and increase their bid price with the fear of losing this "auction game". In such auctions, competition is intense and the end prices may be higher than others as expected. However, there are also cases with high-price auctions which result with the cancellation of the purchase or the non-payment of the purchased goods. For this reason, some experienced sellers leave notes in the write-up in advance saying the product will not be sent to the buyers with zero ratings, and they insist that those with a zero rating should not participate in their auctions. Other types of customers are the ones who place a bid only once, merely very close to the auction end. These customers are generally experienced ones and do not bid before the last portion of the auction time-frame. They bid in the last seconds in order not to reveal their valuation and not be exposed to any kind of competition. This process is called last-minute bidding and customers who pursue this strategy are called bid-snipers. Bid snipping can be conducted manually from the eBay system but is usually done via third-party websites. Customers register to a website and receive a bidding service to be realize in the last seconds of auctions. The bid snipping process prevents unnecessary competition and unmotivated price increase, but there is also a drawback of this behavior since it jeopardizes the probability of winning the auction. For example, technical reasons that prevent bids reaching to the eBay system may crop up. There is a probability of failure to update the valuation and increase the bid price promptly after a competitor's next action is visible. In short, when all of the bids for 444 auctions are examined, by and large, experienced customers have won the auctions with their one bid in the last seconds. No matter how long the auction interval is and no matter how many people follow the auction, the experienced bidder does not reveal herself, they will observe current or past auction data with buy it now sales information, evaluate the product and make a proposal at the last second.

To sum up, the above figures imply that the bidders use two types of strategic moves to get the item. The first one is, following the auction from the beginning to the end and bidding continuously to prevent new competitors come in and enforce existing competitors to exit the auction. The second strategic move is bidding in the last seconds in order not to reveal their product valuation and avoid the competition.

In addition, the last figure at the lower right corner provides a crucial message. This graph shows the total number of bids given during the auction time frame. In the chart, 0 shows the time when the auction starts and 1 when it ends. All bids submitted for 444 auctions are summarized in this graph. This figure actually supports the other three graphs in the sense that an auction actually consists of 3 stages. As soon as the auction begins, many bids are submitted and similarly many bids are submitted close to the auction end. In the interim period, very few bids are submitted and this does not change with the length of the auction. Therefore, we can say that the first 20% of the auction time is the initial stage of the auction, the last

20% is the final stage which has vibrant and intense participation where the "auction fever" is high. The interim period is the middle stage with low participation. This kind of segmentation in an auction period is independent of the duration of the auctions.

The bids submitted in the last stage are the vital decisions that determine the final auction prices. However, it is not possible to determine them and use in the price models since they are also kind of endogenous variables. Similarly, the bids submitted in the initial stage will also provide essential information about product valuations and they are not directly linked to the last stage variables which makes them as kind of exogenous variables. The initial stage bids may provide some insights on what customers think about the real value of the product. Obtaining certain features from a part of the auction can be called as feature extraction. From this stage some features namely, number of bids, number of bidders, the last bid, and the rating of the last bidder etc. can be extracted and can give useful insights. For example, if there are a high number of bidders in the first stage of the auction and the last bid price is high, this may signal a high product value. Similarly, if the rating of the bidders in the initial stage is high, it will give another idea of how the experienced people perceive that specific product. Therefore, the initial stage of the auction can be treated as an information aggregation stage. The customers can gather some useful information about other customers' beliefs on product valuation and make some strategic actions during the auction period. These helpful signals might be used in price models knowing the fact that they are also exogenous variables since they differ from the bids in the last stage.

4.2.3 Outlier Detection

A few instances which do not fit into the general distribution of the data, but included in the dataset due to some reasons are called outliers. The inclusion of outliers in the analysis distorts the model parameters and test statistics. Hence, it is usually necessary to remove these instances. When we read some comments written by the sellers on the sales of eBay auctions, the following cases are likely to be encountered: The customers with very low ratings have submitted very high bids to guarantee to win the auction. When an inexperienced buyer gives another high bid in a similar way, then the system will automatically result in a much higher than the real value of the product. In this case, when customers have to pay the high price, they either cancel the purchase or do not pay. Therefore, some sellers notify the buyers with a caveat such as low rating customers should not bid on their auctions, otherwise, they will not sell the product in such cases. Although this is not the case in BIN sales, these kind of situations might appear in auction sales. Yet, it is not easy to identify such instances one by one. Similarly, a product might be broken and almost inoperable, the seller might be needy for an extremely urgent sale or do not implement a reserve price, then the auction may finalize with a very low price. Although these two events do not appear very often, they can still be found in the dataset, even though they do not fit into the overall data distribution. Hence, overly high prices or overly low prices might be regarded as outliers. In addition, there are sellers who have very high seller ratings in the BIN market by means of their serial sales for a long time. These sellers do not frequently sell their items in auction markets. The number of customer visits to vendor profiles are also the same and disrupt the data structure of the auction sales. There are a number of common techniques for outlier detection and a general approach has been employed in this study. Data points which deviate by 3 standard deviations from the average were basically excluded from the dataset as outliers for three features namely, price, seller rate and views. The excluded data constitute approximately 4% of all the data used in the analysis. This process only handled in the training data in order to measure prediction performance well. Outlier removing algorithm is in fact, a type of filtering process and given below.

```
X: all available features dataset
for feature in ["price","views","selrate"]
    μ=np.mean(X[feature])
    σ=np.std(X[feature])
    X=X[(X[feature]>μ-3*σ) & (X[feature]<μ+3*σ]
```

4.3 Data Normalization, Data Scaling

Data scaling or data normalization is a part of data preprocessing. Data needs to be scaled to decrease the magnitude according to a fixed ratio. Although it is not a mandatory process, it is practically helpful. Data scaling helps to reduce computational time and it helps to improve the model performance. Data scaling process is particularly necessary for clustering algorithms. If the scaling is not done before the clustering process, the algorithm outweighs large data chunks and the algorithm assigns a lot of data to specific clusters, in which case the clusters form a skewed structure in terms of data distribution. Scaling will be used before clustering in order to prevent this drawback. There is no such need for regression models. Two most common scaling formulas are given below. The first one is the min-max scaling and the second one is the normalization formulas. In Min-Max scaling process, the minimum value of the data set is subtracted from each data and divided by the difference of the maximum and minimum values of the data. In the normalization process, the data is divided by the standard deviation after subtracting the mean of the data set. Let *X* is any vector, μ is average and σ is standard deviation for that vector:

Min-Max Scaling:

$$X_{sc} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad \textbf{(4.1)}$$

Normalization:

$$z = \frac{X - \mu}{\sigma} \quad \textbf{(4.2)}$$

4.4 Auction Graphs, Auction Paths and Auction Path Models

To begin with, the bids in auctions need to be understood well. The figures below show the time and bidder distribution of the consecutive bids in the 11th auction.

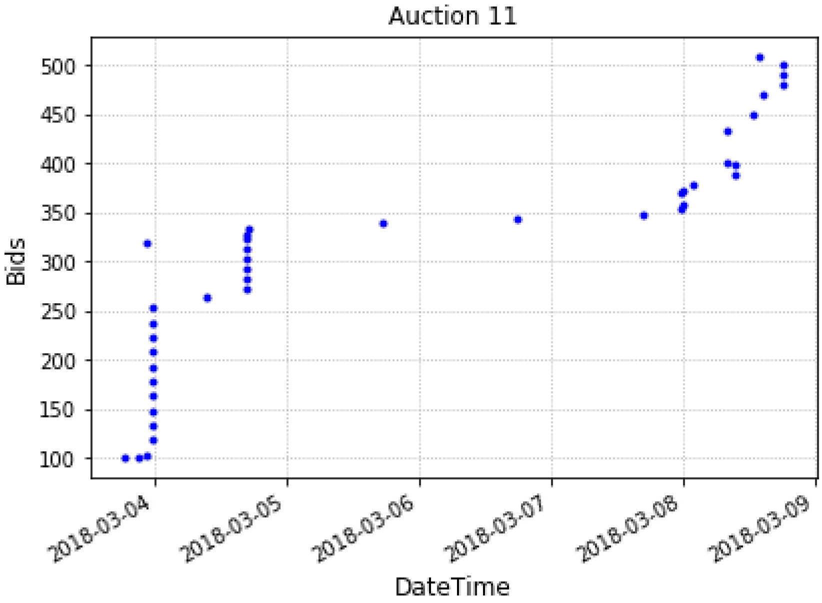


Figure 4.6: Bids During Auction 11 by Time

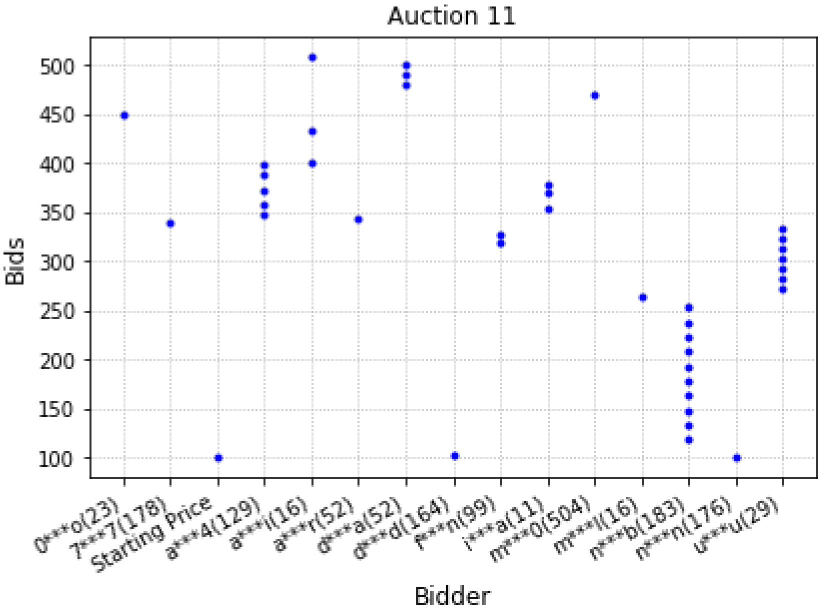
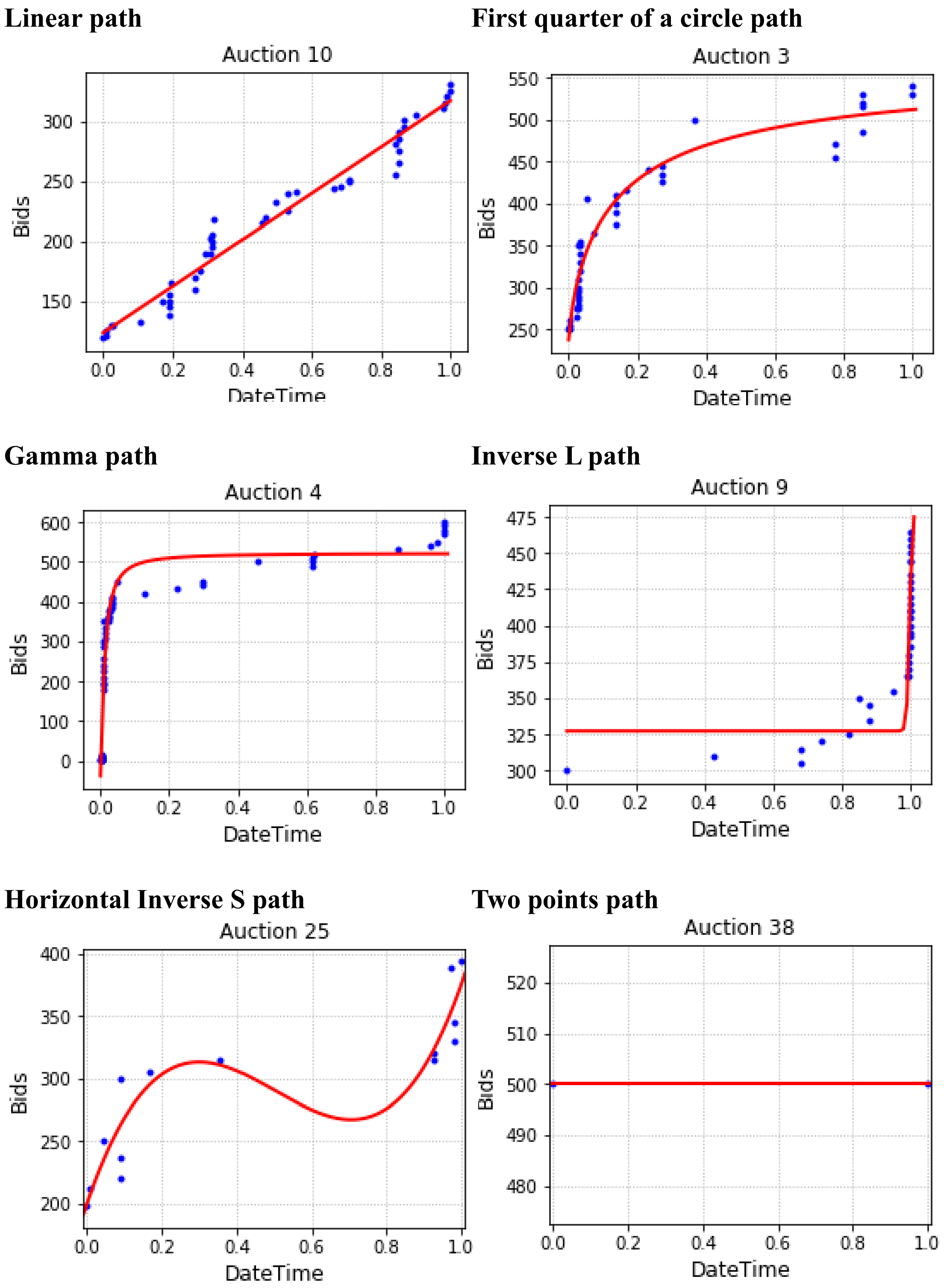


Figure 4.7: Bids During Auction 11 by Bidder

The 11th auction started on March 3, 2018, and as a 5-day auction it ended on March 8, 2018. The auction ended with a price of $509 and awarded to a person having a username of a***i(16). This buyer submitted 3 different bids and made his first offer 10 hours before the end of the auction. His last and winning bid was submitted 5 hours before the auction end. This person seems to have followed the first strategy as mentioned earlier. In other words, by following the auction, he increased his price several times in order to win the item and won it by giving a high bid for a long time ago. This player seems to have a fear of squandering the chance of buying the item and does not hesitate to disclose his private valuation for the product. On the other hand, the number 16 that appears next to the username indicates the rating of this user, which accounts for the players' experience and this player has relatively less experience. From this point onward, this player might be regarded as slightly hasty in the auction and offered a competitive price before the auction time ends. Players who demonstrate this strategy are relatively scarce. Most players prefer to win the game with their bids in the last minutes. In this game, a total of 40 bids were submitted by 14 different bidders. Since the starting price of the auction was a price of $100, a lot of players could be easily attracted to the auction. Some players have bid only once, although some have bid several times, for example, the player n***b(183) has bid 10 times in the first section of the auction. But then s/he couldn't stand the competition and s/he exited. In addition to this, this auction also follows the general segmentation story we've mentioned before. In other words, while there were many bids in the first and last stage of the auction, there were a few bids in the interim period. In the last stage, the competition or auction fever has developed significantly and the price chart has formed a convex shape at the end.

The bids submitted by all of the bidders across the auction time frame carves a. bidding path according to eBay bidding policy. Throughout the dissertation, this will be called as "auction paths". In the figures below, frequent auction paths encountered in all of the auction datasets are illustrated. Although not all auctions can be classified only by these paths here, they are the most frequent ones and they will provide useful information about the auctioning process.



Linear path

First quarter of a circle path

Gamma path

Inverse L path

Horizontal Inverse S path

Two points path

Figure 4.8: General Auction Paths

The first of these is the “linear path”. In auctions where there is intense competition from the beginning to the end, the bids are continuously updated. This builds a linear auction path. As the velocity of price is constant along with the continuous bids, the auction path constitutes a linear chart. The second path is similar to the “first quarter of a circle”. While there is a high competition in the initial stage, it decreases over time and the price development slows down by time. For the third auction path, a high price development with very rapid bidding occurs at the initial stage. In the interim and last stages, bids come with small increments. In short, with initial intense competition, a large part of the end price is formed at the beginning and there is a slight price development in the last stage. Since the figure of this kind of a path is similar to the big gamma symbol, this might be called a “gamma path”. Other conclusions that might be drawn from the last two paths are as follows: Auctions 3 and 4, which appear highly competitive in the initial stage, finalized with higher prices than the others. This will create a signal that the competition in the first stage will result in a high-end price. The fourth path is similar to the “inverse L shape”. In the first and middle stage, the price development is very slow due to fewer bidding activities. Yet, when the auction is in the final period, the bids are fast-moving and frequent under a cut-throat competition. In such auctions, the buyers are usually experienced ones and they monitor the auctions up to the end time without giving any signals about their valuation, hence preventing competition and consequently a potential price increase. The fifth path is a way that resembles an “inverse and horizontal S letter”. In this type of auctions, infrequent or frequent offers are given at the initial and final stage and few bids are submitted in the interim stage. Change of rate in price is different in different stages. In the first stage, the rate of change in price decreases, but it increases in the last stage. In other words, the S path is a combination of a concave function in the first stage and a convex function in the final stage. The final auction path is “two fixed points”. Such auctions start with a high starting price and do not attract a lot of buyers, nor do they bid during the auction period. In this case, if the bidders are eager to buy this product, they offer the starting price in the last minutes and win the product. The auction ends with the initial price and there is no price development.

In previous sections, how the auctions followed a path in the bidding period and what these paths look like were investigated. The segmentation of the auction time into 3 stages namely initial, final, interim stages and different strategic behaviors in each stage give an idea of the tendency of the auction paths to be non-linear. Moreover, by examining all of the auctions figures, we can see that the auctions have non-linear paths by and large, although a few of them follow linear paths. This chapter will focus on how these paths can be modeled. The proposed mathematical models are non-linear auction paths are shown in the table below.

Table 4.6: List of Curve Fitting Models for Auction Paths

Model Name	***Curve Fitting Model***		
Linear Curve	$y=ax+b$		**(4.3)**
Polynomial 2nd Degree Model	$y=ax^2+bx+c$		**(4.4)**
Polynomial 3rd Degree Model	$y=ax^3+bx^2+cx+d$		**(4.5)**
PieceWise Linear 2nd Degree Model	$y=ax+b$ $y=cx+d$	$x < T$ $x >= T$	**(4.6)**
PieceWise Linear 3rd Degree Model	$y=ax+b$ $y=cx+d$ $y=ex+f$	$x < T_1$ $T_1 =< x < T_2$ $x >= T_2$	**(4.7)**
4PL Logistic Model Sigmoid	$y=\frac{(a-d)}{(1+(\frac{x}{t})^b+d)}$		**(4.8)**
Isotonic Regression	*Minimize:* *Such that:*	$\sum_i w_i(y_i-\hat{y}_i)^2$ $\hat{y}_{\min}=\hat{y}_1 \leq \hat{y}_2 \ldots \leq \hat{y}_n=\hat{y}_{\max}$	**(4.9)**

The linear and two-point paths can be easily modeled with a simple linear model. Circle type auction paths can be modeled by 2nd degree polynomials, the gamma and inverse L paths can be modeled with a 4-variable logistic (sigmoid) model.

The inverse S-shaped auction paths can also be modeled with a 3rd degree polynomials. Since the auctions usually consist of 3 different periods, the paths can be modeled by a 2nd degree or 3rd degree piecewise linear functions as well. In addition, a model called the isotonic regression can be used in python. Isotonic regression is a model that is a combination of many linear models. The number of linear models and their junction points are automatically determined by the algorithm. Non-decreasing bid structure during the auction period allow the use of isotonic regression. Although the isotonic regression model appears to be consistent with high performance, it uses a very different number of linear models for each auction. This prevents us from generalizing this model for auctions. Therefore, the results of the isotonic regression models will not be a part of this thesis. Following figures for the auction 11 are given as a sample for the auction path models. First, this auction has a type of an inverse S-shape auction. In other words, this sample auction is in 3-stage structure. Therefore, those non-linear models will be compatible with the example. The Price Prediction Error (PPE) statistics showing the price estimation performance and R^2, the compliance of the model with the data are given in the table below. In this example, the best performance is obtained by an isotonic regression model. However, since it is very difficult to generalize this model, it would be more suitable to evaluate the performance of other models. When we look at both PPE and R^2, the 3rd degree piecewise function and 3rd degree polynomial function indicate the best performance. These models have 0.9 R^2 and forecast the end price at a performance of 98%. To sum up, the following inferences can be drawn from this example and its structure: The auctions satisfying the with the 3-stage structure can also be estimated by the third-order functions. In the models, only the bidding time is used as an exogenous variable and tried to predict the path of the auction. In this specific case, the use of the whole time frame of the auction is needed. Therefore, it is not possible to estimate the end price before the deadline of the auction by these models and hence more proactive price models are necessary which will be introduced in the following sections.

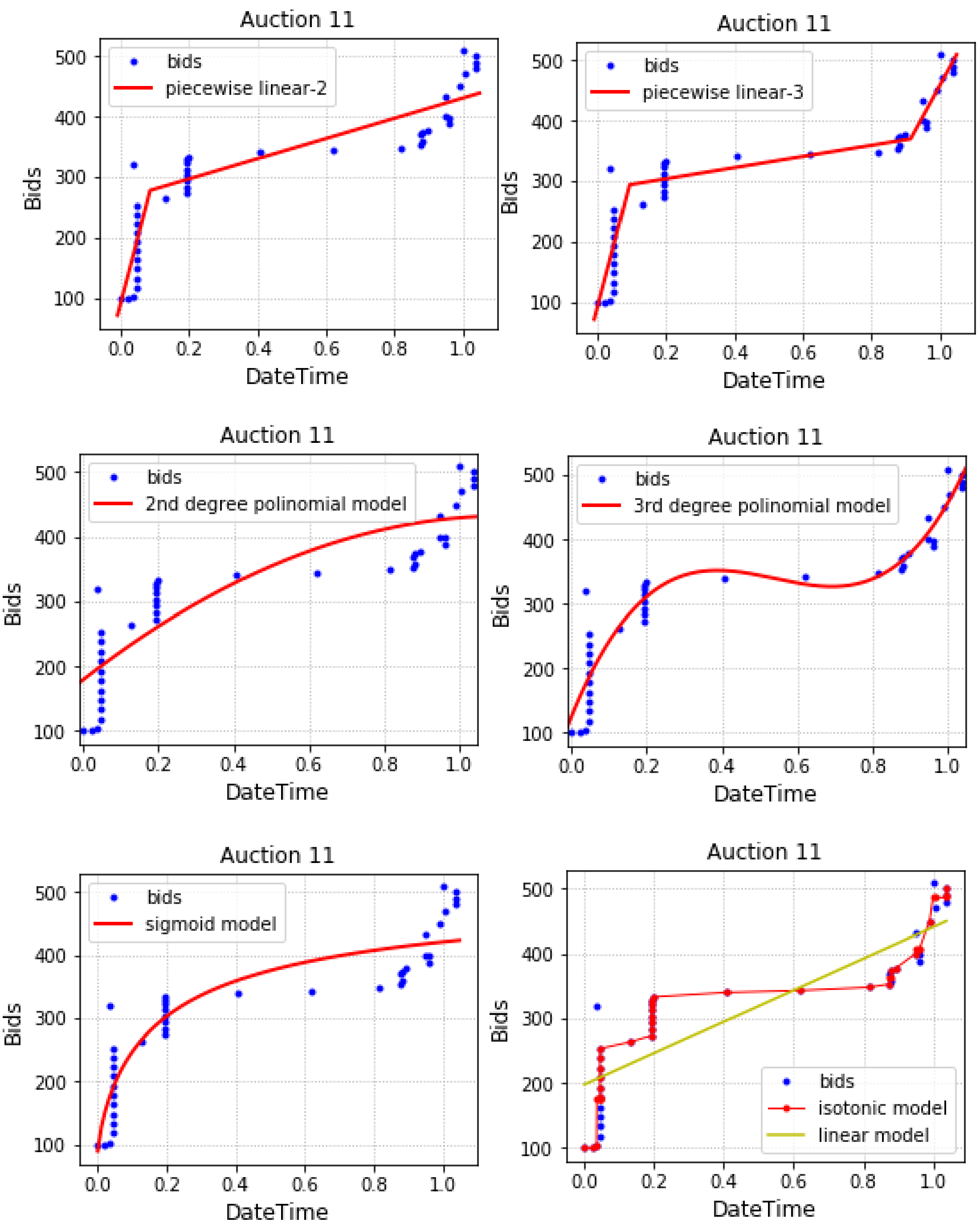


Figure 4.9: Non-Linear Curve Fitting Models for Paths of Auction11

Table 4.7: Performance of Curve Fitting Models for Auction 11

Model Types	Auction End-Price	End Price Prediction	PPE	R^2
Linear	509.00	450.00	11.59%	0.76
PieceWise Linear 2nd Degree	509.00	436.80	14.18%	0.85
PieceWise Linear 3rd Degree	509.00	498.70	2.02%	0.90
Polynomial 2nd Degree	509.00	430.80	15.36%	0.78
Polynomial 3rd Degree	509.00	498.20	2.12%	0.90
Logistic 4PL (Sigmoid)	509.00	422.70	16.95%	0.83
Isotonic	509.00	500.00	1.77%	0.95

The following two tables display which models provide the best cases for all auctions by the R^2 and PPE statistics. So, the most accurate model seems to be the 3rd degree piecewise function model. For this model, the average error of the last price estimates is 2.48% and the model complies with the auction data with a high performance of R^2 which is 0.93. Analyzing these tables, non-linear curve fitting models, especially models with different structures for 3 different periods, can be said as a useful tool for modeling the auction paths.

Table 4.8: Comparison of Curve Fitting Models for all Available Auctions

	PieceWise Lin. 3rd Degree	PieceWise Lin. 2nd Degree	Sigmoid Model	Polynomial 3rd Degree	Polynomial 2nd Degree	Linear	Total
1st Best							
PPE	182	105	12	87	12	46	444
R^2	394	26	7	3	0	14	444
2nd Best							
PPE	120	136	27	108	13	40	444
R^2	34	261	50	99	0	0	444
3rd Best							
PPE	63	73	70	135	43	60	444
R^2	1	133	124	172	14	0	444

Table 4.9: Performance of Curve Fitting Models for all Available Auctions

	PieceWise Linear 3rd Degree	PieceWise Linear 2nd Degree	Sigmoid Model	Polynomial 3rd Degree	Polynomial 2nd Degree	Linear
PPE						
Min	0.00	0.00	0.00	0.00	0.00	0.00
Max	15.22	23.85	30.75	42.18	50.24	90.04
Average	2.48	3.81	7.61	4.20	7.47	8.18
R^2						
Min	0.50	0.35	0.18	0.25	0.15	0.01
Max	1.00	1.00	1.00	1.00	1.00	1.00
Average	0.93	0.89	0.84	0.86	0.78	0.68

4.5 Feature Engineering, Feature Correlations, Feature Selection

Big data technological opportunities can create many features for each instance. For example, in this thesis, 100 features were created for each auction and BIN instances and the size of the dataset can be seen in a large matrix structure of 1120×100. In such a large data set, the curse of dimensionality crops up. In other words, the use of a large number of features in the models increases the variance of estimation, on the other hand, using less than the necessary variables develop a bias problem. These two problems lead to the fact that the obtained statistics may not be meaningful and the interpretations would be inaccurate. It may also affect the performance of the forecasts. In the literature, this is called the over fitting, under fitting or bias-variance problem. A model with high in-sample performance and including many variables can create serious problems when it comes to the out of sample prediction. The bias problem that occurs in case of omitted variables indicates that the coefficient estimates can be significantly different from the real means. In addition, the average of the estimation error can be different from zero. Thus, one of the crucial steps in the predictive model development is the feature selection. For this stage, it is essential to determine the optimum number of features and feature subset. In this study, due to the different structures and theoretical background of regression and classification models, separate feature selection methods will be used.

4.5.1 The Feature Selection for Regression Models

Although the variables such as number of bids, number of bidders, the last bid, the last bidder rate, max bidder rate, min bidderrate, mean bidderrate, etc. are all directly related to the auction price, they will create an endogeneity problem since they are also among the parts of auction end price decisions. In addition, these variables will not work well in a predictive model since the data related to these features cannot occur before the auction ends. The initial stage of the auction can actually be regarded as the information aggregation stage for the product in auction. At this stage, the auction participants submit strategic offers based on their instinctive beliefs about the product value. Some relevant features and useful information can be extracted in this stage. Therefore, the initial stage features will be used as instrumental variables instead of the aforementioned features.

As stated earlier, although the static and dynamic features of an auction affect the end price, it cannot be much higher than the market or the economic value of the product. Very high auction end price may be theoretically likely but this is not true in today's economic conditions where similar products are offered through a lot of sales. Even after the products are sold, the cancellation of the purchase process under certain conditions prevents the product from being sold at a price that is too high or too low. This provides a level of balancing in the auction markets. As shown in the table below, the average values of the auction price are highly and positively correlated with past auction and buy it now prices and the correlation increases as time increases. Detailed correlation Heatmap for average auction and BIN prices is provided in the Appendix A. Below simple OLS models also show that the average auction price can be modeled with past auction and BIN prices.

Table 4.10: Correlation Table for Auction and BIN Prices for Day Average

	7day	6day	5day	4day	3day	2day	1day
Auction & BIN	0.693	0.658	0.611	0.568	0.502	0.364	0.243
Auction & APL	0.626	0.560	0.472	0.390	0.326	0.285	0.176

The relationship between the auction and BIN prices can be specified by OLS model. The endogenous variable of the model is the average auction price. Exogenous variables are the average of Auction Price Lag (APL) and past BIN prices. For instance, A7 stands for average auction prices realized in the last 7 days. Length of the exogenous variables is the same with the endogenous variable. Estimated coefficients and the p-values are represented below.

OLS models for average auction prices can be represented as follows:

$$y = X\beta + Z\theta + u \qquad \textbf{(4.10)}$$

$$ap = \beta_0 + \beta_1 * binp + \beta_2 * apl + u \qquad \textbf{(4.11)}$$

Table 4.11: OLS Results for Auction and BIN Prices for Day Average

AucPrice	C	Past BIN	pValue	Past APL	pValue	RSquare
A7	29.33	0.5602	0.000	0.3917	0.000	0.53
A6	47.44	0.5630	0.000	0.3526	0.000	0.47
A5	88.31	0.5454	0.000	0.2874	0.000	0.38
A4	133.85	0.5216	0.000	0.2191	0.007	0.31
A3	165.49	0.4839	0.000	0.1920	0.027	0.24
A2	220.56	0.3514	0.003	0.2095	0.025	0.14
A1	306.04	0.2534	0.029	0.1274	0.168	0.063

As this analysis suggests longer the time, higher the importance of the historical prices likewise, as the time shortens, the correlation decreases and it can be assumed that the properties of the product and the characteristics of auction have more impact on the price. In short, it can be concluded that the prices of past auction and buy it now prices should be included in the auction end price models to account for the market value of the product.

When there are too many features for each instance, the probability of high correlation between exogenous features also increases. In case of high correlation between the variables, any change in an exogenous variable also affects other exogenous variables and this makes it difficult to interpret the coefficients of the estimates. This is called multicollinearity problem which leads to unstable and unreliable estimates of regression coefficients. However, it is not easy to get rid of multicollinearity and there is no standard solution to this problem. Rule of thumb applications such as removing variables according to a correlation threshold or variance inflation factor can be used to eliminate these problems.

On the other hand, as stated by Neter et al. (1996), when the aim is the predictive performance of the models, or if Multicollinearity is not among the variables that should be carefully monitored, then this problem can also be ignored. The correlations among features are represented below in the Heatmap. In this study, a simple algorithm given below is used and the variables which have correlation higher than 0.8 are removed from the feature set by rule of thumb. The algorithm for removing highly correlated features can be summarized as follows.

```
Set of all available features: $S_1=\{f_1,f_2,f_3,,,,f_n\}$
for i in range (0,n):
        for j in range(0,n):
                if j==i:
                        pass
                else:
                if corr(fi, fj)>0.8
                drop fj from $S_1$ and create $S_2$
create new set of features: $S_2=\{f_1,f_2,f_3,,,,f_m\}$
```

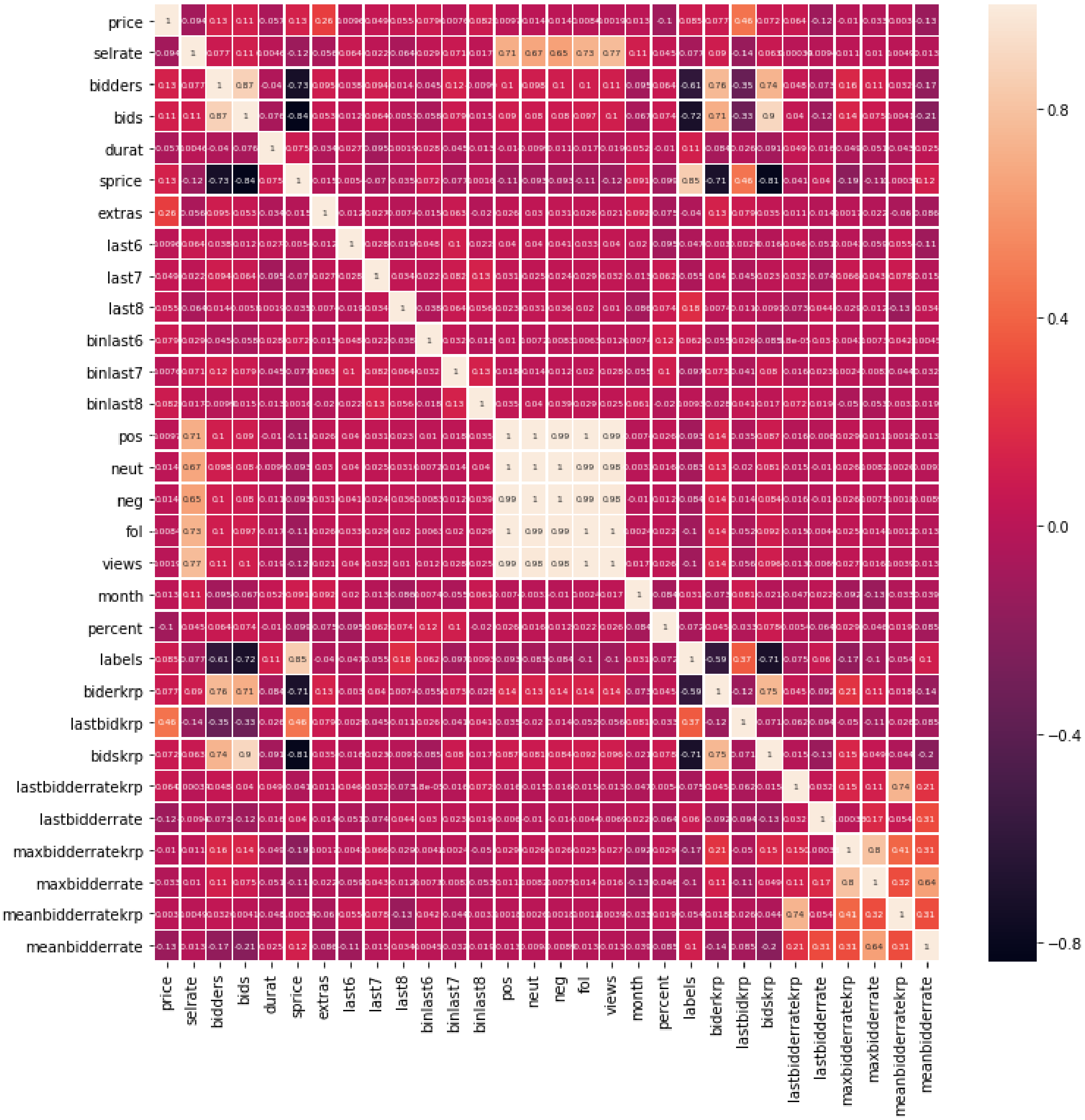


Figure 4.10: Heatmap for Auction Prices and Some Features

In regression models, the omission of relevant variables creates bias and inclusion of irrelevant variable into models creates variance problem. This is called bias-variance trade-off. Omitted variable bias is more important problem than variance problem comparatively and the bias problem can be prevented by top-down feature selection methodology. Top down feature selection starts with the widest set of features and by the help of p-value analysis, insignificant features are dropped from the model one by one. When all features are statistically significant, that is, each feature has a pvalue less than a certain threshold value (0.1 in this case) the feature elimination process is completed. The following is a simple feature elimination algorithm used for the top-down approach.

```
initial $p_{max}$ :1
initial $X=S_2$                #Set of all available features
Endogenous variable : price
Exogenous variables : X

while $p_{max}$ >0.1:
        fit OLS model with (Price, X)
        compute $p_{max}$
        remove feature with $p_{max}$from X
        update X
create new set of features $S_3=X=\{f_{1,}f_2,f_3,,,,,f_m\}$
```

4.5.2 Feature Selection Methods for Classification and Clustering

In clustering algorithms, the aim is to group the instances with a certain degree of similarity and classification algorithms aims to label each instance with predetermined classes by some similarity indices. Thus, we have to know which features to use to create the groups. There are several common feature selection algorithms yet there is no clear clue for which one is the best. In this dissertation, 4 different feature selection methods have been used and the best one was selected by predictive performance comparisons. Feature selection is basically done with 3 processes namely, filtering, wrapping, and embedding. Filtering methods use the properties of data, wrapping methods use a classifier and embedding methods use more complex algorithms to determine the best features to use in clustering approach.

For analysis, sklearn application of scikit-learn developed by Pedregosa et al. (2011) was utilized. As stated by the authors, scikit-learn is python module combining lots of ML algorithms for medium size supervised or unsupervised problems. For feature selection purpose, SelectKbest, Recursive Feature Elimination (RFE), specially developed Principal Feature Analysis (PFA) based on Principal Component Analysis (PCA) and grid search methods were exploited. Although we will not deep dive into technical details here, these methods will be clarified one by one briefly below.

4.5.2.1 SelectKbest

SelecKBest method is a kind of wrapper of sklearn application and it uses a classification function and computes scores for all features. The user defines the number of features, k, to select and the algorithm removes all the features with lowest scores except for k highest scoring features.

4.5.2.2 Recursive Feature Elimination (RFE)

Recursive feature elimination is also kind of wrapper of sklearn application and eliminates weakest features one by one as its name suggests. In this method, firstly a model is estimated with an estimator and feature weights are defined. For example, this may be the coefficients in the linear models. In this method, the desired number of features are also determined by the user and the most insignificant features are eliminated one by one until they reach the desired number. This methodology is actually akin to the p-value analysis we performed with the top down methodology in the regression models.

4.5.2.3 Principal Feature Analysis (PFA)

The dimensionality reduction provides a significant processing and predictive advantage in the high dimensional data. When dimensionality reduction is mentioned, Principal Component Analysis (PCA) comes first. The principal component analysis is meant to be creating uncorrelated feature space with enough variance. But in this case, the original features are lost and it becomes difficult to see which one is effective and more important. In this thesis, the dimensionality reduction methodology which was proposed by Lu, Y. et al (2007) is used. With this method, a subset of the original features that contains the most essential information, using the same criteria as PCA is determined. This method is called Principal Feature Analysis (PFA) by the authors.

The number of features and the clusters cannot be determined automatically for any of SelectKbest, RFE, PFA algorithms. For this reason, clustering has been done both for a certain number of clusters and for all possible features. Then, the best feature subset has been determined which provides the best in-sample and out of sample performance. Here, the number of clusters is chosen between 3 and 9. If the number of clusters is less than 3, the expected benefit from clustering is less and if it is more than 9, there may not be enough data to perform OLS estimation. The available number of features that can be used for clustering is 50. This is restricted by feature availability and Multicollinearity problem. The following algorithm was used to achieve the above-mentioned objectives.

```
for  i in range (0,50):                 # n is number of all feature
        for j in range(3,9):            # try for all number of clusters between 3 and 9
                select features separately by SelectKbest, RFE or PFA method
                fit clustering models with selected features
                fit OLS models for all clusters
                compute in-sample and out of sample performance for all models
select the feature subset and number of clusters which has maximum performance
```

4.5.2.4 Grid Search

This method is based on an assumption that classification decisions can be made based on one or a few essential criteria. This is in line with human decision strategies such that when there are lots of criteria, most of them are naturally ignored and some need to be focused. With this assumption, only one feature was selected from each feature type namely, static, dynamic, economic features and a feature subset was created with a total of 3 features. Clustering models were run for all possible subsets and the best feature subset was determined according to the performance results. The number of clusters is chosen among 3 and 4 since predictive performance is higher. Below specially developed algorithm which is utilized for grid search process is provided.

```
Dynamic Feature Set   :   $S_1=\{f_1,f_2,f_3,,,,f_k\}$
Static Feature Set    :   $S_2=\{f_1,f_2,f_3,,,,f_l\}$
Economic Feature Set  :   $S_3=\{f_1,f_2,f_3,,,,f_m\}$

for i in range (0,k)        :          # choose 1 dynamic feature
        for j in range (0,l)      :       # choose 1 static feature
                for j in range (0,m) :        # choose 1 economic feature
                        features subset $s_{ijk}=\{f_i, f_j, f_k\}$
                        fit clustering models for number of clusters 3,4 with $s_{ijk}$
                        fit OLS models for each cluster
                        compute in sample and out of sample performance
select the feature subset and number of clusters which has maximum performance
```

CHAPTER 5

DETERMINANTS OF ONLINE AUCTIONS AND END PRICE PREDICTION WITH MACHINE LEARNING APPLICATIONS

In this chapter, the factors affecting the end price of fixed-term online auctions will be determined and the price models will be developed to predict the final price before a certain period of auction closing time. To accomplish this aim; static, dynamic and economic features will be analyzed and the degree of effect of the variables on the price will be investigated.

The product descriptions written by the vendors under the product auction pages contain perfect information about the condition of the product. However, this information has not been included yet in the price models to the best of my literature review. Product descriptions will be included in price models for the first time with this dissertation. Vendors write a short product description below the title of the product and write a longer description at the bottom of the advertisement. A sample description written for an auction is given in the Table 5.1 below.

First, the name of the product and all other text descriptions have been incorporated as a new description variable. Within these explanations, sometimes the sellers score their products on a scale they have prepared before. For example, sellers post a score table from 1 to 10 and make a rating of 8 out of 10 for the product they sell. However, the vast majority of sellers publish their comments about the product as a text. In this case, it is necessary to make product information as a numerical variable to use in the price models. Thus, text classification methods developed within the scope of machine learning applications need to be used.

Table 5.1: A Sample Product Description Written by a Vendor

End Price	US $551.00
Bids	68
Bidders	32
Duration	7 days
Date	06/28/18 09:44 PM
Seller	4wdmall (1564)
Item Title	Apple iPhone 7 Plus - 128GB - Gold Unlocked T-Mobile Sprint
Short Description	Amazing condition used with case and screen protector
Long Description	This auction is for used iPhone 7 Plus 128GB. It was purchased from Sprint Originally and when we moved to T-Mobile Sprint Unlocked it for us. So at this point its unlocked and we guarantee it works on Sprint AND T-Mobile. Does not come with SIM CARD since we had to use that on our new phone. This phone was always used with the case and also had screen protector on it. We removed screen protector for pictures and are including a BRAND NEW Screen protector with the listing now. - Comes with Box - Comes with Charger and Head phones never opened - Comes with Case - Comes with Screen Protector …

In addition to this, it has been stated in the papers Kaur et al. (2011), Kaur et al. (2012b) and Kaur et al. (2014) that the estimating separate price models after clustering similar products increase the performance of the predictive models. In this chapter, features that will be used for clustering will be determined by certain feature selection algorithms and several clustering algorithms will be applied to group auctioned products by using these variables. Then, the performance of the price models for all sample and each group will be measured.

In this chapter, 3 different research studies will be carried out in order to realize the above-mentioned purposes. First two of them are about text classification and multiple linear regression models after the clustering process of similar products and the remaining is about an image classification model for the auction graphs. But first of all, unsupervised and supervised models will be defined in general.

To begin with, definitions of unsupervised and supervised machine learning algorithms are summarized below.

Machine learning algorithms usually fall into two categories according to whether the data they use is “labeled” or “unlabeled”. The label is called the information of the expected result of the instance in the dataset. For instance, if the goal is to predict gender from person's height, weight and age information, the true gender information of the person is called the label. In the case that the true gender information is among the features in the dataset, then the data is called labeled data. Machine learning models developed by using labeled data are called Supervised Learning Models (SML) where the supervisors are the labels. The models are developed by utilizing the true results i.e. labels during the training period and they have the ability to estimate the outputs for the sets they have never seen. The algorithms developed with datasets without any labels, i.e. no expected true output information, are called Unsupervised Machine Learning (UML) models. For example, if the correct gender information of individuals is not in the dataset, the process of grouping the most similar height, weight and age information is called unsupervised learning. Since there is no gender information, similarities in only raw data are used to create the groups. But in this case, according to the similarity index of the dataset, the number of groups becomes a variable, i.e. more than 2 groups can be suggested for the gender output by this methodology. Hence, the modeler has to restrict the number of clusters from the very beginning or run this model several times according to certain evaluation criteria and determine the best number of clusters. In the first section of this chapter, an unsupervised text clustering model will be used among machine learning applications. In this method, the texts that have been written for all of the auction and buy it now sales that are most similar to each other without any labels will be clustered. The group numbers of the clustered texts will be included in the feature set as a product condition variable. After this stage, auctions will be clustered second time according to the relevant features and the proposed price model will be estimated for all sample and each cluster separately. In the second section of the chapter, algorithms of SML methods will be used for the purpose of text classifications. In this model, the product descriptions and the prices of the products gathered from buy it now sales will be used for training purpose. Certain clusters will be created according to definite intervals of the product prices which can be used to label the product descriptions. A text classification model will be developed from buy it now sales and then the classes for product descriptions entered for auction prices will be estimated. Later, the performance of the price models will be calculated by repeating the process described above. In the third section, image classification models will be proposed which can estimate the auction end price from the graphs obtained from the information that is gathered only from the initial stage of the auctions. SML algorithms will be used for these models. As the supervisor, the groups' number created from the end price segmentation will be used. In addition to this, auction paths defined in the previous sections will be used as cluster labels and price model performance will be measured in this section. The following table shows data partitioning. Approximately 86% of the data was used for the model training and 10% of the data was used for out of sample performance evaluation. Remaining 4% of data was removed from the data set outlier.

Table 5.2: Data Partition for Analysis

Data Partition	Number of Observation
Train	380
Test	43
Removed as Outlier	21
TOTAL	444

5.1 Unsupervised Machine Learning Algorithm For Auction Prices

In this section, the process of categorizing and rating the product descriptions written by vendors without using any labels and including the ratings into the models as a factor affecting the price will be explained. The process map of the works carried out within this scope is given below. Briefly, vendor descriptions for all auctions and BIN sales written will be clustered and class numbers will be included in the feature set as a "product condition rating" attribute. The specific features will be determined by feature selection algorithms that will work to make another clustering. The multiple linear price model will be proposed by determining the relevant variables with the top-down selection method mentioned before. Lastly, price models will be run for the whole datasets and each cluster to measure in-sample and out-of-sample performance to get the best auction price model.

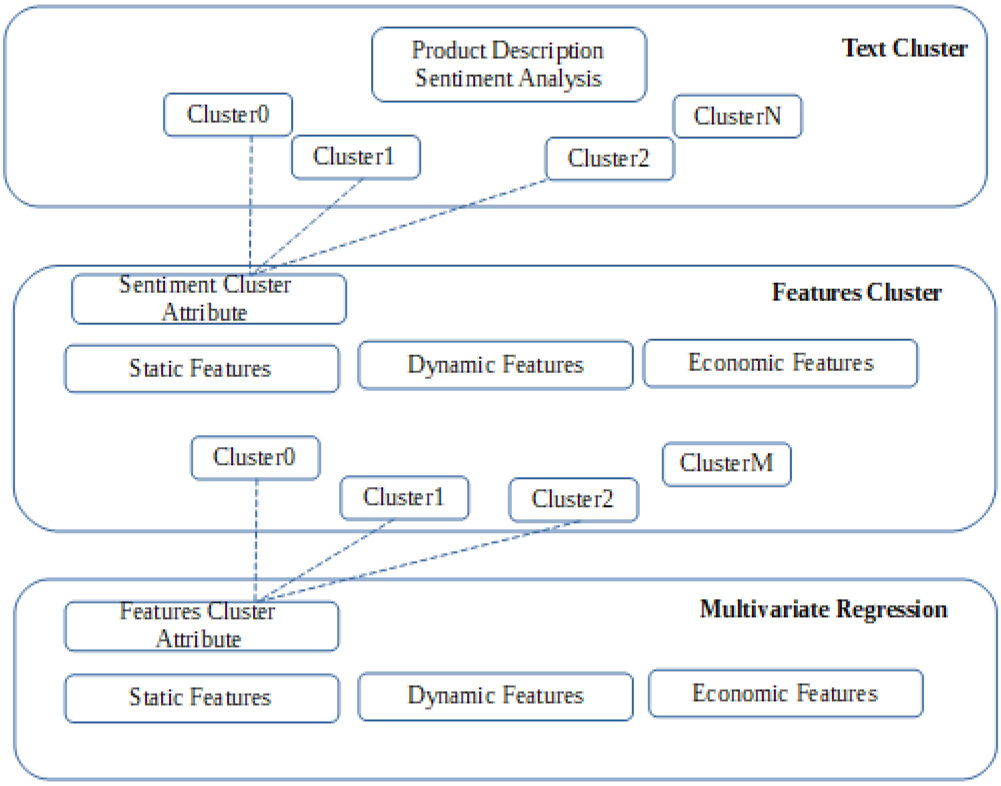


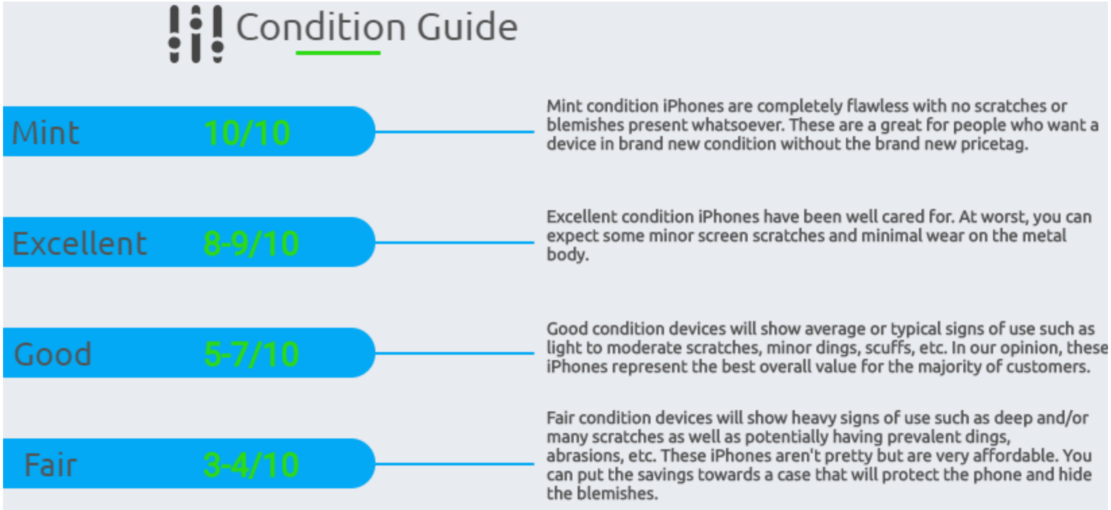
Figure 5.1: Unsupervised Machine Learning Process Map for Auction Prices

5.1.1 Natural Language Process, Sentiment Analysis and K-Means Text Clustering for Product Condition Rating

Natural Language Processing (NLP) is the process of computer algorithms analyzing the language used by humans and making a clever sense of it. This is also called information engineering technology. With this technology, summaries of texts, translation from one language to another, recognition of speech and sentiment analysis are carried out. People sometimes do not express everything in their speeches or writings, or they do not say everything exactly. Therefore, it is a very difficult task to make sense of speech or writings. NLP is not just a word processor. In this method, in addition to the words are examined, the phrases formed by a few words, the sentences formed by the phrases and the meanings that these sentences imply are analyzed as well. It would be very useful to have the correct result labels in the dataset to make a sense. One of the ways of extracting the meaning from the text is "Sentiment Analysis" which aims to determine whether a written text is positive or negative. A lot of postings and reviews are written on social media every day. Sentiment analysis is assumed to be a binary classification method such as good-bad, but it can be done for many classes, for example, positive, neutral or positive. Then we can create a score for reviews or comments by classification models and prepare a rating that can be used. Similarly, we can also classify the product descriptions written by the sellers for our price model to obtain a "product condition rating". To do this, K-Means text classification algorithm can be used as an unsupervised machine learning model in which no labels are used for texts. The product descriptions are classified based on similarities of the texts. As a first step, product descriptions obtained from both the auction and BIN sales were assigned to a variable by combining the product title, short description for the product and all other long explanations. As a second step, data cleaning process for unnecessary information from the texts was handled. For example, numbers, punctuation marks, stopwords were removed from the text, and upper-case and lower-case sensitivity is eliminated. Then, the numerical feature extraction is made, since the K-Means algorithm can only use numerical variables. For this, the algorithm developed in sklearn named TfidfVectorizer (TF-IDF) application is used by python. With this application, word frequency and inverse document frequency are determined. Each word in the text is determined as a term and calculates the frequency of this term together as a vector such as <(t1, f1), (t2, f2), (t3, f3),…, (tn, fn)>. In this case, the terms are words, and f shows the frequency of the word in the text. These vectors will be used as input for the K-Means model.

K-Means algorithm is one of the easiest and most widely used unsupervised learning algorithm for clustering by which the most similar instances are grouped without using any labels. Several distance metrics such as euclidean are used to measure similarity. For this process, first of all, it is necessary to determine the number of centroids, i.e. the number of clusters, indicating how many groups the data will be divided into. Although there is no definite method to determine the number of clusters, we can take the opinions of field experts into account or use the statistics such as the silhouette coefficient and the elbow method. In this study, the product condition tables determined by the sellers (experts) will be taken into consideration for the classification of the texts and the condition of the products.

Some of the sellers on eBay provide condition guides on their websites about the product they sell. That is, they rate their own products according to some predetermined scales. A sample condition guide is given in the figure below.



Condition Guide

Mint 10/10

Mint condition iPhones are completely flawless with no scratches or blemishes present whatsoever. These are a great for people who want a device in brand new condition without the brand new pricetag.

Excellent 8-9/10

Excellent condition iPhones have been well cared for. At worst, you can expect some minor screen scratches and minimal wear on the metal body.

Good 5-7/10

Good condition devices will show average or typical signs of use such as light to moderate scratches, minor dings, scuffs, etc. In our opinion, these iPhones represent the best overall value for the majority of customers.

Fair 3-4/10

Fair condition devices will show heavy signs of use such as deep and/or many scratches as well as potentially having prevalent dings, abrasions, etc. These iPhones aren't pretty but are very affordable. You can put the savings towards a case that will protect the phone and hide the blemishes.

Source: https://www.ebay.com/itm/Factory-Unlocked-iPhone-6S-Plus-Gold-Silver-Space-Gray-ATT-Verizon-16-64GB-128GB/282100249169?epid=237849564&hash=item41ae7c3251:m:ml-60zJA6YJIllTit5yN5Bg

Figure 5.2: Product Condition Guide Provided by a Seller at eBay

This vendor has made a rating from 1 to 10 for the product he or she sells. The rating is divided into 4 groups namely, “mint”, “excellent”, “good” and “fair”. A product that can be a perfect match for those who want a product such as brand new, no blemishes, no scratches, has been identified as 10/10 with a class of “mint condition”. On the other hand, a product which has a lot of scratches and shows heavy signs of use has been identified as 3-4 / 10 points with a class of “fair”. Other classes of are products have a condition in between these two. Taking this condition guide and similar ones into account, we determined K=4 classes for the product descriptions.

The following steps of K-Means clustering algorithms are applied one by one and the product descriptions written for each auction have been converted into a product condition rating. K-Means algorithm represented below minimizes the below euclidean distance cost function and this algorithm has mainly three steps, namely initialization, assignment and moving the centroids.

Objective Function:

$$\sum_{i=0}^{n} \min_{\mu_i \in C}(\| x_j - \mu_i \|^2) \qquad \textbf{(5.1)}$$

- *Initialization step: Randomly choose k samples as initial centroids*
- *Cluster assignment step: Assign each data point to the cluster with Euclidean distance to cluster centroids.*
- *Moving centroid step: Compute new centroids by the mean of all data assigned to clusters,*
- *Repeat above steps, until the difference between old centroids and new centroids is less than a certain threshold value.*

A total of 994 product descriptions were used to train K-Means text classification model. Out of this pool, 395 were written for auction sales and 599 were for BIN sales. The rest will be used for evaluating the out of sample performance stage, which is not used in the model development stage.

The figure below shows how much observation is assigned to each of the product condition ratings by the K-Means model. Most of the products were labeled as Cluster1, while a few were labeled as Cluster2. Although the number of labels is in a skewed structure, the data segmentation provides a wide variety to use in the price models.

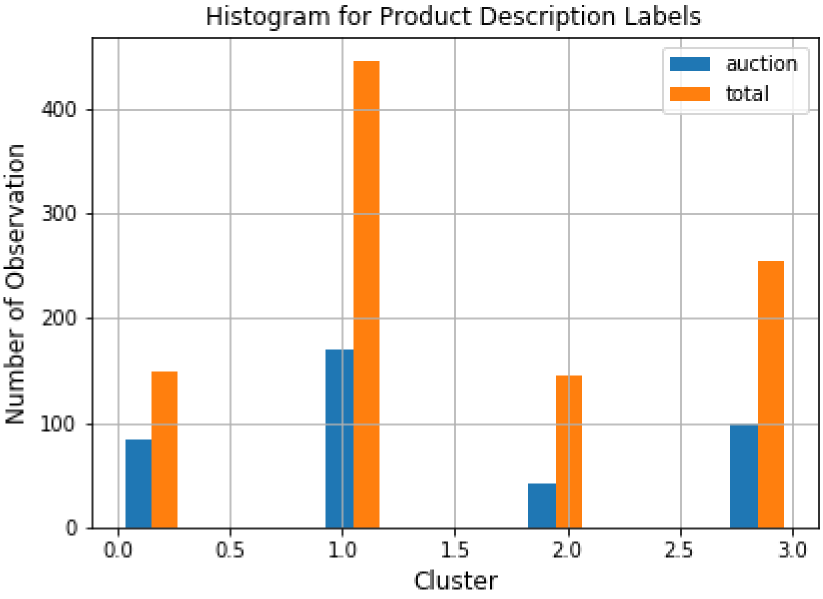


Figure 5.3: Histogram for Estimated Labels for Product Descriptions

The following table shows the first 15 words determined by the K-Means model that are most frequently mentioned in the product descriptions for the clusters. This table provides us with essential information for the product condition ratings. The words which are indicated by bold and gray fills imply a significant difference for the condition of the product. The order of words in the same column shows how often they appear within the group. For example, the words shown in the first row in the same column appear to be the most common within this group.

Table 5.3: Top 15 Terms for Each Cluster

	Cluster0	Cluster1	Cluster2	Cluster3
	protector	phone	phone	iphone
	case	**scratches**	mobile	**original**
	screen	condition	gsm	box
	used	unlocked	pictures	plus
	phone	screen	unlocked	unlocked
	condition	**good**	iphone	condition
	comes	iphone	work	apple
	great	used	shipping	128gb
	glass	device	condition	**excellent**
	new	item	ship	**new**
	iphone	**wear**	plus	black
	box	**great**	apple	charger
	scratches	charger	free	comes
	excellent	clean	factory	used
	kept	included	carriers	phone
Sentiment	Neutral(+)	Negative(-)	Neutral(0)	Positive(+)
Price Effect	Medium	Low	Medium	High

First of all, we see the words “original“, “excellent”, and “new” as the most important ones for the Cluster3. These words are helpful to draw a positive meaning about the product condition. There is also no term for a negative perception of Cluster3. Therefore, we can have a positive sentiment for the condition of the products labeled as Cluster3. Looking for Cluster1, the term “scratches” comes first among the words that can determine the product condition. So it can be figured out that there may be too many scratches on the products in this group. There is also a similar word “wear” which shows us that the product has been used and worn for a long time. While the words “good” and “great” appear to be positive in this cluster, the negative evaluation becomes more dominant since the words that have a negative meaning are shown earlier in the same column. Thus, a negative sentiment can be expected for the condition of products labeled as Cluster1. When we look at the most frequent words for Cluster0, we see the words “great”, “new” and “scratches” and "excellent" in the middle rows of the same column. From this sequence, we can reach both positive and negative evaluation for Cluster0. Yet, it cannot be concluded that one of these is dominant because of the fact that no words appear in the first rows. On the other hand, the positive words outnumber, an overall positive evaluation has a little more chance than the negative one. When we look at the words obtained from the explanations for Cluster2, we do not see any word about the condition of the product. In this case, a neutral opinion might be resultant for the condition of the products in this group.

Based on the above information, if we put the product condition rating as a variable on a linear price model, it might be claimed that the label for Cluster3 can have the highest effect on the price and the label for Cluster1 can have the lowest one. Similarly, we can foresee that the labels for Cluster0 and Cluster2 will have an effect between the other two but that the effect of Cluster0 may be slightly more than Cluster2. All these inferences will be evaluated in the results of the price models that we propose in the following sections.

5.1.2 Grouping Similar Products by Variable Clustering

The clustering of the products and the estimation of the proposed price model parameters for each group were reported to increase the performance of price estimation models in the literature as mentioned before. Such an approach adds a non-linearity to the linear models which can also capture market segmentation. In this section, grouping similar products with various clustering methods namely, K-Means and Hierarchical K-Means methods and determining the features to be used for this process will be explained. When the data are used without any scaling, the large numbers in the data distort the clustering due to the cost function we aim to minimize. Therefore, the data is scaled with min-max scaling algorithm before the clustering operation. By this way, the data became more uniform to prevent outweighing some particular instances.

5.1.3 Hierarchical Variable Clustering

In this study, two types of unsupervised clustering methods will be used. The first one is the K-Means clustering method as described in the previous section, and the second one is the hierarchical clustering method. The hierarchical clustering method has two different versions: Divisive clustering from whole-to-piece and agglomerative clustering from piece-to-whole. In the piece-to-whole method, each item is defined as a cluster for an initial step, according to some distance metrics, the similarities of each item to the other clusters are calculated and closest clusters are combined. The divisive method is the opposite, and all data is defined as a single set, and new clusters are created by dividing large clusters according to certain distance metrics. This continues until the system becomes stable. The working principle of the agglomerative clustering algorithm is shown below.

The Agglomerative Hierarchical Clustering algorithm is based on Johnson's (1967) article and consists of 4 steps. For N items, it is summarized below:

- *Assign each item to its own cluster and compute a NxN distance matrix*
- *Combine two closest items*
- *Create (N-1) x (N-1) distance matrix*
- *Repeat steps 2 and 3 until a single cluster remains*

This algorithm produces step-by-step dendrograms and creates clusters. The figure below is a dendrogram as an example. Firstly, each auction data is defined as a cluster and then the similarity is calculated according to distance metrics (e.g. single, complete, average, ward, weighted etc), then A1 and A5 which are the closest in the auction set are merged, this set is followed by A3 after that A6 is combined with A2. This process continued until all the auctioned items are turned into a single cluster.

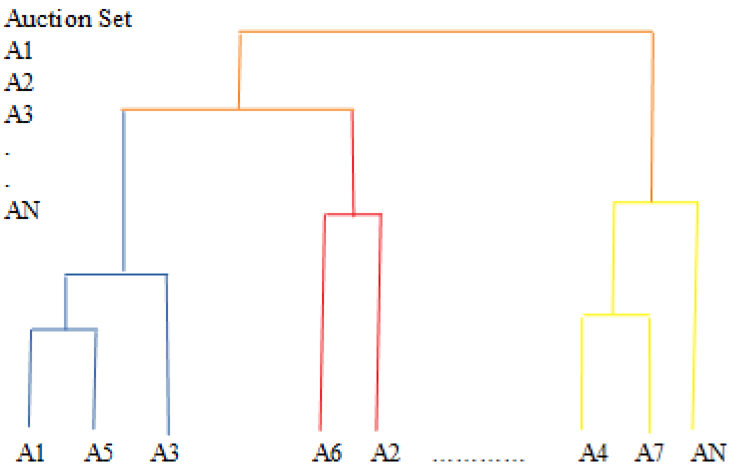


Figure 5.4: Sample Dendrogram for Hierarchical Agglomerative Clustering

In order to group similar products, it is necessary to decide on some issues. In particular, it is necessary to determine in advance which clustering model will be used, which features will be used and how many groups will be created. To do so, we will employ SelectKbest, RFE, PFA and Grid Search feature selection algorithms, and K-Means and Hierarchical Agglomerative Clustering algorithms. The number of features that can be used for clustering is 50. With these features, the cluster number should be between 3 and 9 to assign adequate amount of data to each cluster for the linear price model. There is no clear evidence in the literature on which method will give the best performance. Therefore, the best practice cannot be determined without trying all of the combinations of models and features mentioned here. For this reason, a multivariate OLS price model is necessary where the variables are determined by the top-down method which will be mentioned thoroughly in the following sections. This model has been run about 10,000 times for each of the aforementioned combinations of the features and clustering algorithms then model parameters were estimated to calculate the performance. Unsupervised model performance of the feature selection methods for K-Means and Hierarchical clustering methods are provided in Appendix B and Appendix C respectively. The table showing the performances of the best cases is provided below. So, the Grid

Search feature selection algorithm can be claimed to provide the best performance. If the auction dataset is clustered with the features of the duration of the auction (d3), the average bidder rates of auction initial stage (meanbidderratekrp) and BIN sales (binlast5) for 4 groups, then the K-Means clustering application gives the best in-sample and out-of-sample performance, that can be achieved. The relevant results of the best cases will be presented in more detail in the following sections.

Table 5.4: Performance of Feature Selection Methods for Unsupervised Model

Feature Selection Method	Clustering Type	Number of Clusters	Number of Features	In Sample MAPE	Weighted In Sample MAPE	Out of Sample MAPE	Weighted Out of Sample MAPE
SelectKBest	K-Means	3.00	6.00	7.14	7.05	7.35	7.17
RFE	K-Means	3.00	36.00	6.98	7.14	8.46	8.36
PFA	K-Means	4.00	18.00	6.32	6.93	8.15	8.35
SelectKBest	Hierarchical	8.00	20.00	6.12	6.10	7.46	7.98
RFE	Hierarchical	3.00	13.00	6.88	7.13	7.74	8.39
PFA	Hierarchical	3.00	42.00	6.88	7.15	8.01	7.91
Grid Search	**K-Means**	**4.00**	**3.00**	**6.00**	**6.94**	**4.65**	**6.91**
Grid Search	Hierarchical	4.00	3.00	6.80	6.81	9.82	8.26

To sum up, we can continue with K-Means clustering method with the optimum number of K is 4 and selected features are d3, meanbidderratekrp, binlast5. After selecting the method and the features to use in the clustering process, the next step is to figure out whether the recommended number of K matches the structure of the data. But, we do not need to change the number of clusters that give the best performance. We will only reinforce the argument by providing more evidences to support the decision. There are 2 most common methods to determine the number of clusters to be used for the K-Means model. The first is the elbow method. For this, the sum of the square of the distance of each data point, x, from the centroid of the cluster is calculated. This is also called Within Clusters Sum of Square (WCSS) and the mathematical formulation is given below. This statistic is kept within the K-Means algorithm as “kmeans.inertia_” in scikit-learn and it is called distortion.

$$WCSS(k) = \sum_{j=1}^{k} \sum_{x_i \in clusterj} \left\| x_i - \overline{x_j} \right\|^2 \quad where\ \overline{x_j}\ is\ sample\ mean\ in\ cluster\ j \quad \textbf{(5.2)}$$

In order to visualize it, WCSS is compiled for the number of clusters that can be created (between 1 and 10) and an elbow curve is provided below. These graphs are called elbow curve due to their shape. In the literature, where the bending of elbow occurs is suggested as the optimal number of clusters. When we examine the figure there are 2 important breakpoints. The distortion drops rapidly until 2, and after that, the speed of the fall is slightly slower, but the distortion drops slightly at 4 clusters again. After that point, the reduction rate in the distortion is very slow. So we can conclude that clustering up to 4 groups can ensure a significant advantage and there is no benefit for creating more than 4 clusters. As explained in the previous sections, we cannot reach the variety of data we expect from the clustering when the cluster number is 2, so the number of 4 clusters suggested above seems to be a good choice. In other words, we can support the decision of the previously determined number of clusters with this elbow method.

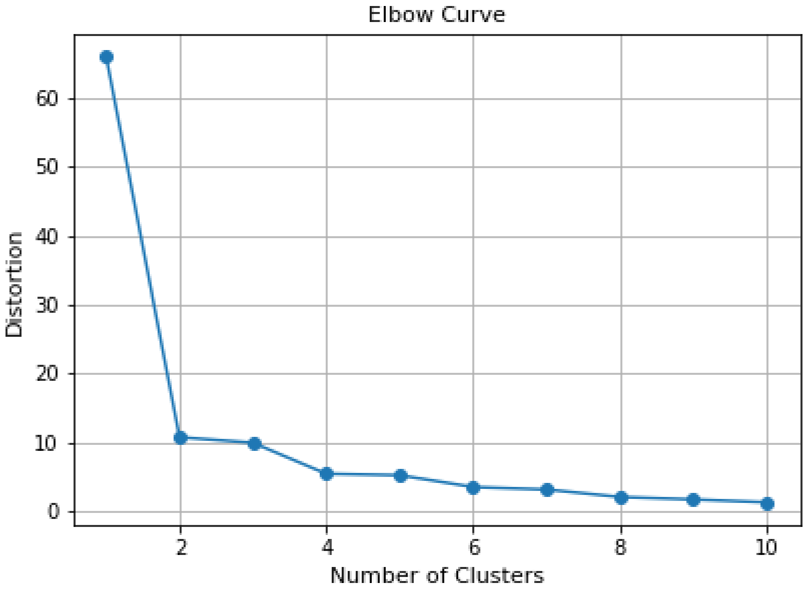


Figure 5.5: Elbow Curve Plot for Clusters

The Elbow curve analysis is, in fact, a descriptive analysis and it can sometimes cause confusion in determining the number of clusters. Therefore, the silhouette statistic proposed by Kaufman and Rousseeuw (2009) is often used to determine the number of clusters in the literature. This statistics shows the quality of the clustering process. That is, it shows how well the data fits in each cluster. This statistic is calculated from -1 to +1. The fact that statistics are close to +1 indicates that the data in this cluster is quite remote to neighboring clusters. Conversely, the statistic -1 implies that data actually more suitable for neighboring clusters rather than the cluster it is in. Having a statistic of 0 points out that the data set is very close to the decision boundaries. Therefore, statistics close to 1 are preferable but not the others. The silhouette graph illustrated for the K-Means clustering process is displayed for 4 clusters. For each cluster, the calculated silhouette statistic is closer to +1 and the average silhouette coefficient is calculated as 0.52. Thus, according to this statistic, the number of chosen clusters seems logical.

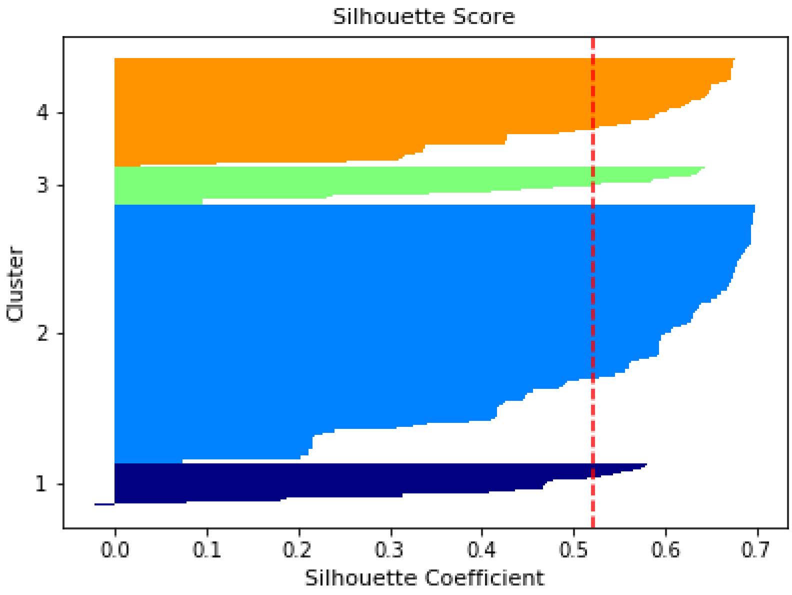


Figure 5.6: Silhouette Coefficients Plot for Clusters

The following histogram shows how much auction data are assigned to each cluster by clustering operations. As can be seen from this chart, a large part of the auction data is assigned to cluster 1. In other words, the distribution of data in clusters is not uniform. However, since a specific set of features have been used when estimating clusters, we did not expect a uniform distribution. We do not need to worry about data distribution since there are also adequate amount of data to estimate the price model in each cluster.

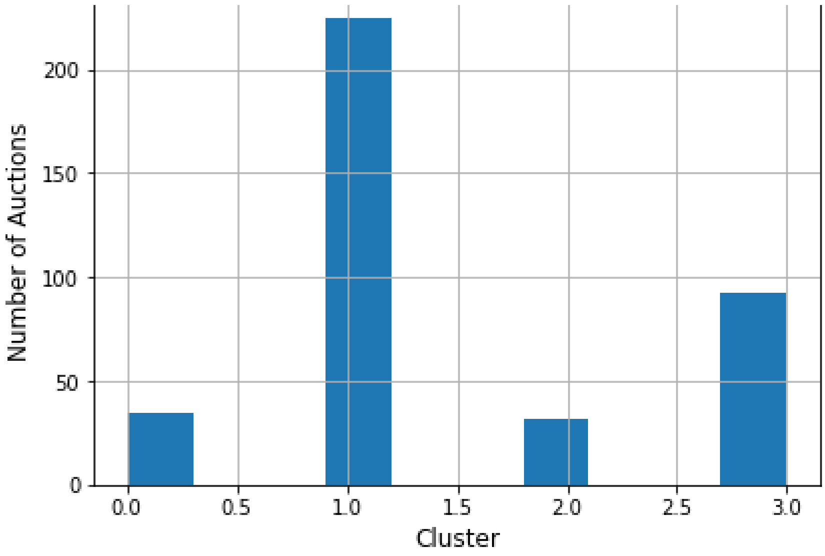


Figure 5.7: Number of Auctions for Each Cluster

5.1.4 Past Auction and BIN Price Clustering

Up to now, we included economic variables to represent the market value of the product in the price model. To be more clear, the historical price of products sold by the auction or BIN sales is also included in the feature set. These economic variables can be added to the price models in two ways. First, the last prices can be determined according to the time of sales only, so economic variables can be used in the clustering set as well. We used this approach in previous sections. Another method is to aggregate similar products and determine the last prices within the same clusters and add them to the price models. In this case, economic variables cannot be used in feature selection algorithms of clustering algorithms. In other words, static and dynamic features can only be used there. The second method is a long and time-consuming process. In addition, there might not be enough data to use for the price model according to the feature selection algorithm and the clustering methods. Therefore, in order to measure the performance of this approach, clustering was performed with K-Means algorithm for the number of 3 to 9 clusters using only the features selected from the 20 available features with the Recursive Feature Selection method. With 3 clusters and 4 features, the best case in-sample performance is calculated as 6.87% and out of sample performance is 7.34%. This calculated performance is not better than the performance calculated in the previous sections. Determining past prices after clustering process seemed interesting, but we’ve found, rather surprisingly that there seems to be little gain from doing this. Therefore, we will continue with the approach for determining the last prices as proposed earlier.

5.1.5 Multivariate Linear Regression and Regularization

Following the literature, linear price models will be used in this dissertation. Although the linear models are simple, they are very successful in price models. In the case of too many variables in the variable set, the researchers allow the model to determine the variables to be used. Specifically, all variables in the variable set are included in the model, but variables to be used are penalized with certain functions. Such models are called regularized regressions. The mathematical formulations of cost functions (Equation 5.3, 5.5) and estimators (Equation 5.4, 5.6) for the Lasso and Ridge Regression models are shown below.

Lasso Regression

$$L(\beta,\lambda)=\sum_{i=1}^{n}(Y_i-X_i\beta)^2+\lambda\sum_{j=1}^{p}|\beta_j| \qquad \textbf{(5.3)}$$

$$\beta^{lasso}=\underset{\beta\in R^p}{\arg\min}\,||Y-X\beta||_2^2+\lambda\,||\beta||_1 \qquad \textbf{(5.4)}$$

Ridge Regression

$$L(\beta,\lambda)=\sum_{i=1}^{n}(Y_i-X_i\beta)^2+\lambda_1\sum_{j=1}^{p}(\beta_j)^2 \tag{5.5}$$

$$\beta^{ridge}=\underset{\beta\in R^p}{\arg\min}\,||\,Y-X\beta\,||_2^2+\lambda\,||\,\beta\,||_2^2 \tag{5.6}$$

In Lasso regression, the absolute values of the coefficients of the variables are included in the cost function. In Ridge regression, the square of the coefficients is added. Thus, while the model is estimated, the model can automatically determine the most suitable set of variables because the coefficient of each variable will add a marginal cost. These models can be used as a quick solution when there are many variables and only one model should be estimated. However, we know that when one of the necessary variables is missing the model will create a bias problem. Therefore, even though it costs more time and effort, we will use the following OLS model (Equation 5.7, 5.8) to predict the model more accurately. By the model results, the most insignificant variable will be removed. This step will be repeated several times by removing only one variable at a time until there is no statistically insignificant variable in the model. This is expected to end up with the most accurate model. This is also called the top-down variable selection method. The dependent variable is auction price, *Y*, and exogenous variables are static features, dynamic features, and economic features represented as *X*. Multiple linear regression model is represented below:

$$Y_{nx1}=X_{nxm}\beta_{mx1}+u_{nx1} \tag{5.7}$$

$$ap=sf*\beta+df*\theta+ef*\delta+u \tag{5.8}$$

$$L(\beta)=\sum_{i=1}^{n}(Y_i-X_i\beta)^2 \tag{5.9}$$

$$\beta^{ols}=\underset{\beta\in R^p}{\arg\min}\left\|Y-X\beta\right\| \tag{5.10}$$

The variables used to estimate linear multivariate auction end price model, their coefficients, types, and statistical significance levels and are shown in the table below:

Table 5.5: Results of Multivariate Regression for All Sample

Dependent Variables	**Auction Price**		
Number of Observations	380		
Independent Variables	**Coefficient**	**pValue**	**Feature Type**
auclast4	0.076	0.078	**economic**
binlast6	0.088	0.060	**economic**
binlast8	0.100	0.054	**economic**
extras	19.985	0.001	static
selrate	-0.016	0.086	static
neg	-6.605	0.020	static
views	0.022	0.000	static
k0	93.943	0.000	sent.cluster atribute
k1	86.175	0.000	sent.cluster atribute
k2	93.521	0.000	sent.cluster atribute
k3	101.598	0.000	sent.cluster atribute
d1	74.986	0.000	static
d3	79.705	0.000	static
d5	87.469	0.000	static
d7	79.363	0.000	static
d10	53.713	0.004	static
smonday	57.884	0.000	static
stuesday	54.928	0.000	static
swednesday	51.098	0.000	static
sthursday	47.935	0.000	static
sfriday	57.709	0.000	static
ssaturday	46.297	0.000	static
ssunday	59.385	0.000	static
biddercrp	1.424	0.008	**dynamic**
Lbidcrp	0.334	0.000	**dynamic**
Lbidderratecrp	0.012	0.091	**dynamic**
Adj. R-square	0.307		

The top-down feature selection method was used to determine the effective variables in order to avoid a substantial bias problem in the regression model. Within the scope of this methodology, a regression model was estimated by including all possible variables. The most statistically insignificant variable is extracted from the feature set and the model is re-estimated. This process was applied until there are no insignificant variables. For the estimation of the model described above, a dataset containing 380 separate auction data was used after the outliers were removed. The adjusted R^2 calculated for the model is 0.307. The endogenous variable for the model we propose is the auction closing price and exogenous variables are a set of 26 distinct features. We will interpret each exogenous variable used for the model below.

In order to add an average market value in the price model, we used economic features such as past auction and BIN prices. The last fourth price of auction sales and the last sixth and eighth product prices sold by BIN were found to be significant in the model. The coefficients of these variables are positive. This result is also economically meaningful. In other words, if price increases in the market for similar goods in the recent period, the effect of this on the auction will be positive. However, if prices in the market have been declining in the recent period, this will have a negative effect on the ending price of the auctions which is consistent with the economic theory. It is also logical that the price of the last items is statistically insignificant and the price of the products previously sold is significant. Customers cannot easily track all auctions and BIN sales in the last minutes. However, they can know some previous sales and evaluate the market value from them.

The variable called "extras" is a binary variable that takes 1 when the phone sale is included with the box, charger, cable, warranty and so on and 0 when there are no any other accessories. The coefficient of this variable is statistically significant and is around \$20. Thus, if an extra item is included with the phone, the marginal effect on the auction price is \$20.

When we look at the total ratings of the sellers, “sellrate”, the significant coefficient is negative but close to zero. In other words, it has not a high impact on the price. On the other hand, the negative rating of the sellers has a coefficient of -$6.6. This is an important result and compatible with the literature. A negative rating significantly damages the seller reliability, which creates a negative impact on the price of the products they sell. For example, if the seller receives an additional negative rating, the average price of the products they sell is reduced by more than $6. We can say that sellers should avoid negative rating as much as possible and give importance to customer satisfaction. The “view” variable, which indicates how many times the seller web profile has been visited by the customers, has a positive but near-zero effect. It is logical that the effect is positive. The fact that buyers visit the vendor profile on a number of occasions shows a reputation for this vendor, meaning buyers are searching for the vendors to follow the product they sell.

The product condition rating variables that we create with clustering methods are named as “k0”, “k1”, “k2” and “k3”. These variables are set to binary numbers according to the cluster numbers the products in. For example, the k0 variable holds the value of 1 for the product condition rating labeled with Cluster0 and 0 for the other clusters. First of all, all of the four variables are statistically significant even in 1% level. From this point of view, there has to be a condition rating variable in the price models of the used goods indicating how well the products were used. Lack of such variables will create a significant bias problem in the models. We will elaborate on this issue below.

In the previous sections, we have predicted that the seller notes labeled as Cluster3 have a positive (+) sentiment and the price effect would be high. The model result table indicates that the coefficient of the k3 variable is $101.6 which is significantly higher than the coefficient of the other 3 variables. For example, in a case of product descriptions including one or more of the words “original” “excellent” or “new” implies the product has a perfect condition and this affects the product price by $102. This effect is above $10 on the average when it is compared to other product condition attributes. Similarly, in the previous sections, we have mentioned that the seller notes written for the Cluster1 products create a negative(-) sentiment which means price will move below the average. The coefficient of the variable defined as k1 is $86.2, which is the smallest value compared to the table. In other words, when one or more of the words such as “scratches”, “good”, “wear” and “great” entered in the product description, a negative perception occurs on the buyers and this causes a drop in the price. In addition, the seller notes for Cluster0 and Cluster2 were predicted to create a neutral perception and that the price effect would be somewhere between the other two effects. For Cluster0, we defined the perception as neutral(+), meaning that the effect in the Cluster0 will be slightly more than that of Cluster2. When we look at the table, the coefficient of k0 is $93.9 and the coefficient of k2 is $93.5 which are in between $86 and $102. In this case, the previous inferences seem correct.

It has already been shown that the descriptions written for the product provide important information to the buyers and have an important role in the price models. Another remarkable point is determining which product descriptions are more effective in the customer segments. Therefore, the relationship between the product condition rating and the bidder ratings is also worth examining.

The following table summarizes the product condition variables estimated by the UML model, the estimated coefficients, the average prices of the products, the average bidder ratings during the entire auction time and the bidder ratings in the initial stage of the auction. The coefficient of the product group, which is classified as k1 and contains a lot of scratches in the product descriptions, is $86.2, as previously emphasized. This value is the lowest value compared to others. The average auction end price of the products of this class is $ 488.05 which is also the lowest one. When the bidder ratings in this class are examined, the average bidder rating in the initial stage of the auction is 65.77 with median of 21.20. These values are much lower than the values of other classes. To put it more clearly, the existence of words that can affect the customers negatively in the product descriptions prevents the experienced customer from bidding on the product. On the other side, the buyers who have less experience do not refrain from bidding in the initial stage for that kind of products. In other words, less experienced buyers cannot fully internalize product descriptions and may miss out on the importance of product descriptions. This situation changes when the bidder ratings are analyzed for the entire duration of the auction, and the relationship between product descriptions and buyer experiences becomes more complex. This may be due to buyers' ability to update their valuation by other customers' bids or competition in the auction.

Table 5.6: Bidder Rate and Product Condition Rating in UML Model

UnsuperVised		Price ($)		meanbidderrate		meanbidderratekrp	
Product Cond.	Coefficient	Mean	Median	Mean	Median	Mean	Median
k0	35.05	507.52	505.00	155.79	99.24	99.02	34.00
k1	27.28	488.05	495.00	147.71	92.83	65.77	21.20
k2	34.63	493.77	494.00	120.08	72.50	207.53	32.00
k3	42.70	508.21	510.00	139.11	90.19	85.60	31.16

Another variable sets in the table is the duration variables that indicate the length of the auction. These variables are listed as d1, d3, d5, d7, and d10. Here d1 is the binary variable created for auctions with a 1-day auction period. 5 of these 5 variables are statistically significant. After reviewing the coefficients clearly, the bidding period has a nonlinear effect on the price. If the auction period is between 1 day to 5 days, duration is positively correlated with price. After 5 days the correlation turns into negative. The highest effect is visible when the duration is 5 days and the lowest effect registers when the duration is 10 days. With this result, it is possible to conclude that the relation between duration and price can be represented by a concave curve and optimum length for the bidding period is 5 days for this kind of product on eBay. This analysis shows us that a period of 5 days is as short as for customers to discover the product and react to competitors' bids and hence reach an optimum time to finalize the auction without waiting too long. Therefore, the highest coefficient pertains to the d5 binary variable. When the auction period is 10 days, the price coefficient goes extremely down. This is due to the fact that the duration of the auction is fixed and the customers are not likely to reach the product before the deadline. In addition, the longer the auction period, the more dealing costs in that customers will spend more time to keep track of the product, i.e. hassle cost increases. For this reason, customers do not prefer such auctions causing demand and competition to be low. Thus, the price ends up lower compared to others. On the other hand, in the 1-day auctions, the duration coefficient is lower than the others. The reason behind this may be that, in such a short period of time, customers are less likely to detect, evaluate, decide and bid on the product. In this case, competition and demand are low, so the price also falls slightly. The effect on the price of 3-day and 7-day auctions is similar but slightly lower than in the 5-day auctions. With these analyzes, it is proven that the duration of the auctions is an important issue for traders who want to use a strategic decision tool. The graph below shows the effect of the duration on auction price. It can easily be determined that the duration of the auction has a nonlinear effect on the price and it is the highest when the auction period is 5 days.

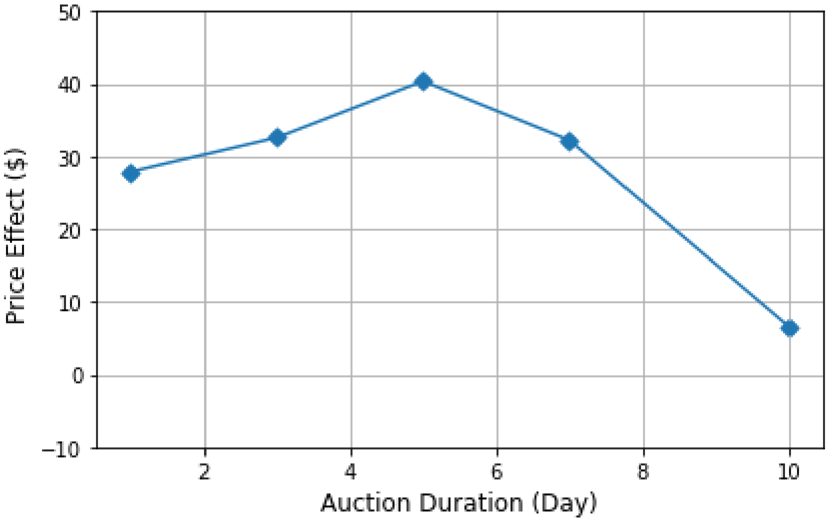


Figure 5.8: The Effect of Auction Duration on Price in UnSupervised Model

In this study, the duration of the auction was considered as categorical variable and dummy variables were added to the model. If the duration of auction are added to the model as its own level, squared or root as a single variable, the coefficient of this variable is not statistically significant. The non-linear relationship between the duration and price of auctions can be seen from the figure above.

This relationship can be formulated in the following form: $y^2 + (x-5)^2 = 5^2$ **(5.11)**

This can be summarized as follows: $y = \sqrt{10x - x^2}$ **(5.12)**

If the duration variable is added to the model according to this formula, the variable is statistically significant, but it does not contribute to the performance of the model. The results of the model are given in Appendix D. In the study, the duration of the auction was added to the model as a categorical variable in order to see the effect of each day.

The next group of variables in the table is the group of binary variables created for the days when the auction begins. In the literature, the effect of the day on which the auction ended has been investigated frequently. Since the buyers of the product used in this thesis are usually ordinary individuals, the ending day of the auction is an important variable. Customers usually work on weekdays, or engaged in daily routines such as education, health etc. on the other hand, at the weekend, they rest, entertain and do the shopping. Assuming that, the bidding on auction sales are a type of game and entertainment, it is most likely that the auctions ended at the weekends attracts more bidders which stimulates competition and ultimately a price increase. On the other hand, the effect of auction start days has not been analyzed yet in the price models. The start and end day of the bidding period are linearly dependent variables by the duration of the auction. For this reason, we only included the auction start days in the model. In previous sections, it was mentioned that a high number of bids are submitted in the initial stages of the auctions. This shows that many customers wait for the new auctions and follow the announcements. As we can see from the table, all of the day variables are statistically significant. The coefficient of the day variable is the highest in the auctions that started on Sunday. This finding is compatible with the expected results. People search for auctions quite a lot on Sundays when leisure time generally hits the peak. There is also a high probability that lots of new auctions can start on Sundays. The second most effective days appears to be Monday and Friday for end prices and this may be because of the first and last working day effect. Since Monday is the first business day, people might be spending more time on product search before they take on their responsibilities. On Friday, people might be slowing down momentum of the week and be searching for new products from the web in the happy Friday time interval and consequently choose the product auctions to bid. In addition, the price coefficients of the auctions start on Wednesday, Thursday and Saturday are considerably lower at about $10. This result is quite understandable. Customers may not find much time for product search during weekdays. Besides, people might be reserving Saturdays to relax or have fun to throw off the fatigue of the week, which can mean that people may miss the Saturday auctions causing relatively less competition in the auctions and lower end prices.

The last group of variables shown in the table is the dynamic features that can be evaluated as the in-auction variables obtained from the initial stages. The first of these is "biddercrp" which stands for the number of people who bid at least once in the initial stage. This variable is statistically significant and its coefficient is 1.4. In other words, every new person participating in the auction in the initial stage increases the end price by $1.4. This variable is a competition indicator for the auction. It is also theoretically reasonable to have a statistically significant and positive coefficient. The "Lbidcrp" variable stands for the last bid given in the initial stage. As mentioned before, auctions enter a rest period in terms of bidding after the initial stage and there are not too many bids in the interim period. In order for this period to start, bidders might need to see a signal. In the analysis, the first 20% of the auction period is used as a signal for this and we determined the initial period according to this criterion. However, the final offer given in this period actually indicates the level at which the price has reached as soon as the auction starts, which can spark valuation signal about the product for the bidders. The statistically significant and positive result of this variable in the model supports the previous interpretation. The most recent variable shown in the table is the rating of the person who made the final bid in the initial period. This number shows the experience of the bidder. This variable has a statistically significant and positive coefficient in the model. Other bidders might see this rating as another positive signal for product valuation. If any experienced person is bidding on the product, the end price of the product will increase. Thus, most of the time, experienced customers do not reveal their own valuations and avoid the price increase by holding their bids back until the last seconds of auctions.

In this study, it is also taken into consideration that the number of alternative products might affect auction prices as soon as the product auction starts. For this reason, number of alternative products was calculated separately at the beginning of auctions and at the end of initial stages and added to the model. According to the results given in the Appendix E, the number of alternative products was not statistically significant.

The following figures show the model in-sample performance with 380 data. In the graph that appears on the top left, the actual end price for each auction, the estimated end price and the difference between the two are illustrated. In this graph, small price difference we can be observed but the variance of the estimated values is close to the real price variance. In other words, the estimated prices follow the actual prices and the model seems to capture the price variations quite well. Moreover, it can be said that the nominal value of errors is not very high and it is concentrated around a mean of zero value.

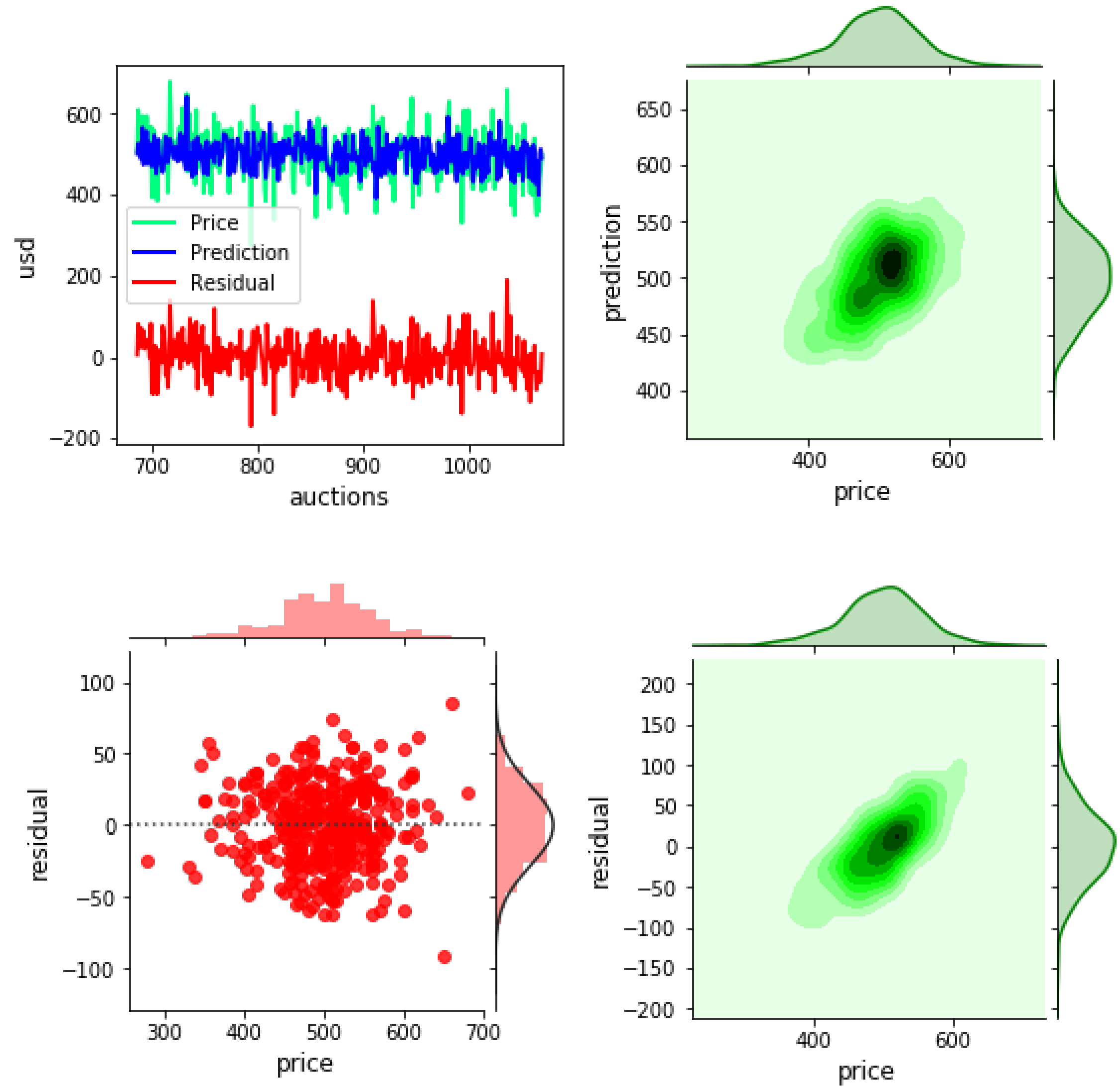


Figure 5.9: In Sample Prediction Performance for All Sample

At the top right, true prices and predictions are illustrated. From dark green colors of this chart, one can infer that the actual prices and projections are concentrated at around a value of $500, which is similar to the contour lines in the form of joint plot chart. At the bottom left, it can be seen that the estimation errors are concentrated around zero and the error values are nominally distributed between maximum values of (-50$, +$50). In the last joint plot shown in the graph below dark green coloration error around the zero implies also the good fit of the model. As can be inferred from the figures, the in-sample performance of the proposed actual price model is quite successful.

Below the figures showing the out of sample performance of the developed model are provided. For out of sample performance, 43 auction data which are not used in the training period of the model were used. From top left figure, one can see that the actual price and the estimated prices are close enough, and it can be concluded that the variations in actual price could be grasped by the model. In the graph, the residuals that are indicated by a red line appear to be concentrated around a mean of zero value. This shows that the model's out of sample performance is quite successful. The joint plot graph on the top right tells us by the dark green colored part that the price and estimates are concentrated around $500. According to the actual price and estimation error graph at the bottom left, it is seen that the errors are close to a normal distribution at around zero value. As to the nominal values of the errors, we see that this residual distribution is between the maximum values of (-60$ and + 60$) the majority of which are close to the zero. This distribution also shows that the estimation errors are not proportionally too high with respect to actual prices. Similarly, by the joint plot chart at the bottom right, the dark green color is concentrated, that is, the area with the most data occurs at zero error zone with a true price of $500.

The analysis of the nominal values of the in-sample and out-of-sample performance of the developed model will be explained in the following sections.

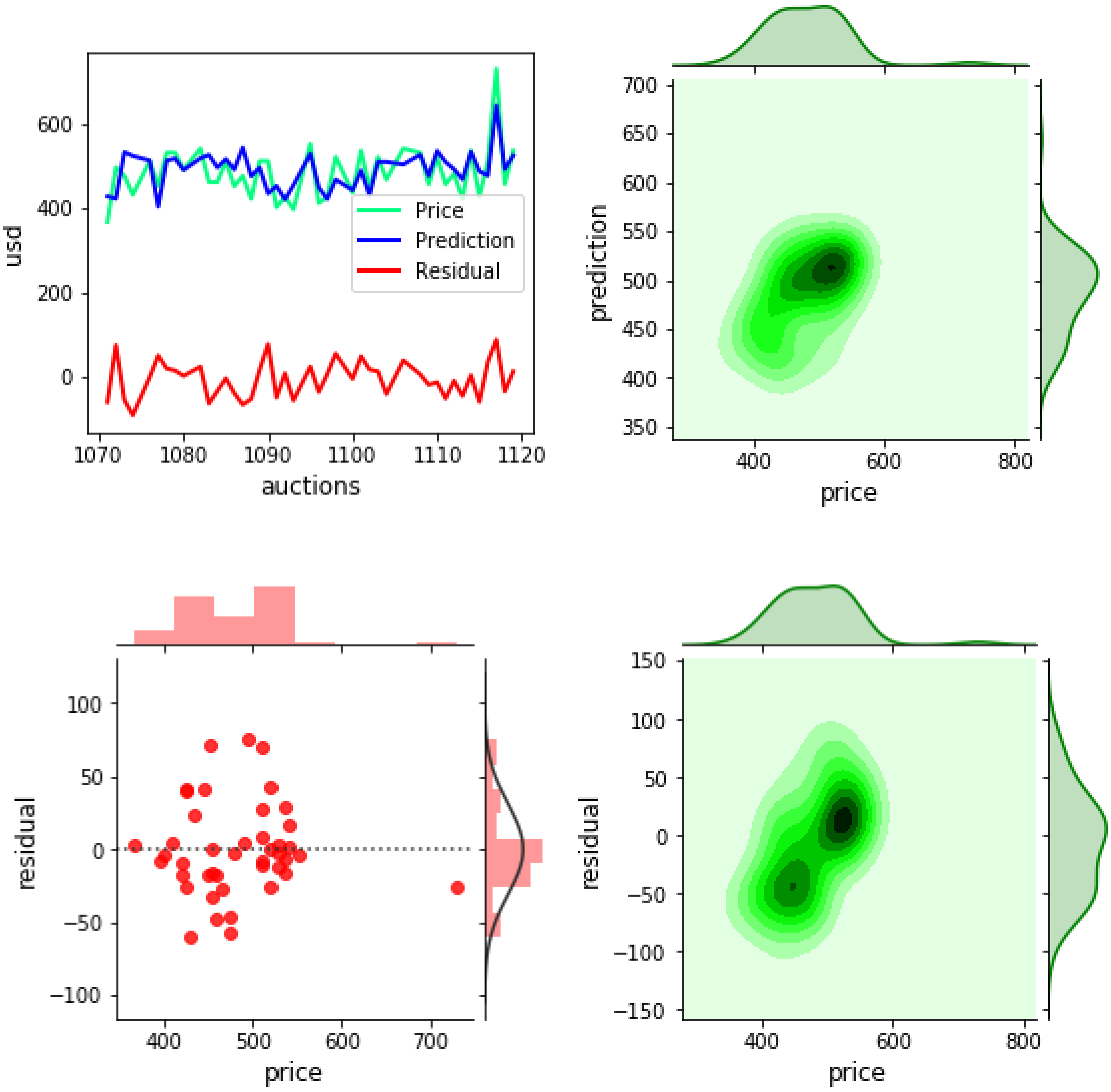


Figure 5.10: Out of Sample Prediction Performance for All Sample

5.1.6 Diagnostic Tests For the Multivariate Regression Model

Let the multivariate linear regression as follows: $Y = \alpha + \beta X + u_{.}$ Y is dependent variable, X is explanatory variable and u is iid error term. Some important and necessary assumptions which are made before estimating linear models are listed below. A brief description of these assumptions and how they are checked will be given in this section.

Linearity: The relationship between endogenous and exogenous variables is linear.

Homoscedasticity: Error variance is constant and finite $var(u_t)=\sigma^2$.

Independence of Error terms: Error terms are statistically independent of each other, i.e. there is no serial correlation or autocorrelation in error terms. $E(u_t u_s)=0.$

Independence of Exogenous variables: Exogenous variables do not have a high level of multicollinearity.

Normality: Error terms are normally distributed.

No Endogeneity: No correlation between error terms with exogenous variables.

Unbiasedness: The average of error terms is zero, $E(u)=0.$

In order to satisfy these assumptions, the error terms is expected to be $u \sim N(0, \sigma^2 I).$

According to the Gauss-Markov theorem, the above assumptions will ensure that the OLS model estimates give us the Best (minimum variance), Linear, Unbiased Estimators. These estimators are also called BLUE by the first letters of the words.

Violation of the above assumptions results in serious problems such as bias, inefficiency, and inconsistency in the OLS model.

If the assumptions are examined in order, modeling a non-linear relationship with a linear model creates functional misspecification problem and disrupts all the analysis from the beginning. Therefore, the structure of the model must be compatible with the relationship between the variables.

Heteroscedasticity is opposite of homoscedasticity and it means uneven distribution of variance among data. Error terms differ across values of some of independent variables. In the case of violation of homoscedasticity, the estimators are expected to be inefficient. Even if the model parameters are consistent and unbiased the estimators will not have the smallest variance. In other words, due to the change in standard errors, the reliability of the t-statistics will be low and hence the statistical inference might be wrong. In this case, the estimators are not “best” ones.

Another important assumption is the independence of the error terms. To test whether this assumption holds or not, autocorrelation tests can be conducted on error terms. If the mathematical form of the model is not specified correctly, if all the relevant variables for the model are not included, or if there is a measurement error in the data, autocorrelation problem might be detected in the error terms. This causes both bias and efficiency problems in the model. Even if the model is specified correctly and all relevant variables are included in the model, i.e. the bias and consistency problems are eliminated, the autocorrelation problem will cause a serious efficiency problem. In other words, the estimated parameters will not have the smallest variance. In this case, as in the heteroscedasticity problem again, the reliability of the inferences will be low.

The next assumption is the no multicollinearity between exogenous variables. In order to assure this assumption, as we have explained that in the previous chapters, we removed the variables having high correlation from the variable set. Violation of this assumption first damages the full rank condition and complicates the inverse of the exogenous variable matrix. Besides this technical problem, a high correlation between the variables makes it difficult to interpret the coefficients in the model results. For example, when there is a high correlation between x1 and x2 exogenous variables, one cannot infer the effect of 1 unit change in x1 data by looking at the coefficient of x1 alone, since a change in x1 will also change x2 and this will have an effect on the model results. It is a bit difficult to get rid of multicollinearity in the models where many variables are used. In these cases, although OLS estimators are BLUE, the standard errors of the coefficients of some variables may be high and the coefficients of the variables can be sensitive even to very small changes in the data. However, in the literature, if the model is to be used for predictive purposes or if the high correlation is not among the variables which need to be interpreted and followed carefully, multicollinearity can be ignored.

Another assumption is that errors are normally distributed. If this assumption holds, statistical inferences can be made more easily and reliable estimation intervals can be created. However, if the errors are independent, have an appropriate variance covariance structure and there are a high number of observations, this assumption can be optional and ignored. In the previous chapters, we have mentioned that all variables except for the binary ones used for modeling appear to have normal distribution. In fact, the normality assumption may be disregarded because the dataset includes a satisfactory number of observations.

Another important assumption for the correct model is that there is no endogeneity problem. High correlations between the estimation errors and the exogenous variables might result in endogeneity problem. This makes the coefficients of the variables inconsistent. In other words, the model parameters do not converge to true population means. This problem is one of the most important problems that needs to be solved, otherwise, the coefficients of the model are inconsistent, biased and the model estimates are wrong. In order to avoid this problem, variables such as the number of the bidders, the total number of bids, the last bid, the rating of the final bidders which are directly linked to the auction end price decisions are not included in the model.

The next assumption is the predicted values are close to the actual values, that is, the proposed model can explain the variance in the endogenous variable and the assumption that the average of the errors is a zero. If this assumption does not hold, the estimates are again biased. In the model, it’s seen that the average of the error terms is a small number close to zero. One can understand whether the model has been specified correctly and the assumptions described above do hold or not by applying several tests. Autocorrelation, heteroscedasticity, normality and linearity tests which are commonly used in the literature, are mentioned below. Here are the null hypotheses, alternative hypotheses, test statistics and the distribution of these statistics:

Table 5.7: Diagnostic Tests

Breush-Godfrey (1978) Autocorrelation Test (AC)

$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \ldots + \rho_q u_{t-q} + \varepsilon_{t,}$

$\varepsilon t \sim$ iid $N(0, \sigma^2)$, $t = 1, 2, ..., n$

$H_{0:}\ \rho_1 = \rho_2 = \ldots = \rho_q = 0$: There is no AC in Residuals

H_A: at least one is non-negative : There is AC in Residuals

Auxiliary regression : $\hat{u}_t = \alpha_0 + \rho_1 \hat{u}_{t-1} + \rho_2 \hat{u}_{t-2} + ... + \rho_q \hat{u}_{t-q} + \gamma X + e$

Test statistic is $LM = nR^2 \approx \chi^2_{(q)}$

Ljung and Box (1978) Autocorrelation Test (AC)

$H_{0:}\ \rho_1 = \rho_2 = \ldots = \rho_q = 0$: There is no AC in Residuals

H_A: at least one is non-negative : There is AC in Residuals

Test statistic : $Q = n(n+2)\sum_{k=1}^{q} \frac{\hat{\rho}_k}{n-k}$

Test statistic follows $\chi^2_{(q)}$ distribution, where n is sample size ρ autocorrelation at lag j and q is the number of lags to be tested

Breusch and Pagan (1979) Heteroscedasticity Tests (HC)

Auxiliary regression : $\hat{u}_t{}^2 = \delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k + e$

$H_0 : \delta_1 = \delta_2 = ... = \delta_k$: There is no HC in Residuals

$H_A : \delta_1 \neq \delta_2$ or $... \neq \delta_k$: There is HC in Residuals

Test statistic is LM and follows $\chi^2_{(k)}$

White (1980) Heteroscedasticity Tests

Auxiliary regression : $\hat{u}_t{}^2 = \delta_1 + \delta_2 \hat{y} + \hat{y}^2 + e$

This is a special type of White Test

$H_0 : \delta 1 = \delta 2 = 0$: There is no HC in Residuals

$H_A : \delta 1 \neq \delta 2 \neq 0$: There is HC in Residuals

Test statistic is $LM = nR^2$ and follows $\chi^2_{(k)}$

Table 5.7 Continued

Anderson and Darling (1954) Normality Test

H₀ : Normal distribution

H_A : Not Normal distribution

Test statistic is A

where $A^2 = - N - S$ and $:S = \sum_{i=1}^{N} \frac{(2i-1)}{N} [\ln(F(Y_i) + \ln(1 - F(Y_{N+1-i}))]$

N is the number of samples

This test has special tabulated critical values for specific distributions.

Shapiro and Wilk (1965) Normality Test

H_0 : Normal distribution

H_A : Not Normal distribution

Test statistic $: \; W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$

$\mathbf{x}_i$ stands for ascending samples,

where x_1 is the smallest, a_i computed from the means, variances and covariances of the order statistics, where sample size is n. This test has special tabulated critical values for specific distributions.

Kolmogorov (1933) and Smirnov (1948) Normality Test

H_0 : Data follows a Normal distribution

H_A : Data do not follow a Normal distribution

Y_i are ordered data series from smallest to largest, N number of samples and F is the cumulative normal distribution

This test has special tabulated critical values for specific distributions.

Test Statistic $: \; D = \max_{1 \le i \le N} \left[F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i) \right]$

Table 5.7 Continued

Utts (1982) Rainbow Nonlinearity Test

Auxiliary regression : $y_i = \beta_0 + \beta_1 x_i + \sum_{h=1}^{q} \theta_h w_{hi} + \varepsilon_i$

H_0 : θ = 0 , Model is correctly modeled as linear

H_A : θ ≠ 0 , Model is not linear

Test statistic is based on F statistics : $F = \frac{(SSE_F - SSE_c)/(n-m)}{SSE_c/(m-2)}$

SSE_F is sum of squares of error for n (all) samples

SSE_F is sum of squares of error for m (central) samples

Lagrange Multiplier Non-Linearity Test*

H_0 : Model is correctly modeled as linear

H_A : Model is not linear

Auxiliary regression : $\hat{u} = X\beta_1 + X^2\beta_2 + \varepsilon$

fit auxiliary regression and compute R^2 and compute LM=nR^2

** This test is developed in statsmodels module of python to match Gretl's linearity test*

The results of the diagnostic tests for the model we propose are given in the table below. According to Breush-Godfrey and Ljung-Box autocorrelation tests, p-values are greater than 0.1. Therefore, the null hypothesis that the errors do not contain AC cannot be rejected. Thus, one cannot say that there is an autocorrelation problem in the error terms. Similarly, by looking at the Breusch-Pagan and White Heteroscedasticity tests, the p-values calculated for the two tests are again greater than 0.1 and the null hypothesis of homoscedasticity cannot be rejected. In short, there is not a heteroscedasticity in the error terms of the model. The table also shows the results of 3 normality tests. These tests are Anderson-Darling, Shapiro-Wilk and Kolmogorov-Smirnov tests respectively. In all of these tests, null hypothesis is the normal distribution of data. In the first two, p-values larger than 0.1 resulting that the normality assumption cannot be rejected. According to the third test, normality claims were rejected. By taking the majority, we can argue that error terms are normally distributed. Last but not least, the proposed model has been tested by using Rainbow and Lagrange Multiplier tests for linearity assumption. The null hypotheses of these tests are linearity of the model. By the test statistics and p-values, the null hypotheses cannot be rejected for these tests. In other words, the assumption that the recommend model is linear.

Consequently, by the above analyzes and tests, the model has successfully passed the diagnostic tests and the estimators seem unbiased, efficient and consistent. This shows that the proposed model is specified quite accurate and estimators are BLUE.

Table 5.8: Results of Diagnostic Tests

OLS Diagnostic Test	**Test Statistic**	**pValue**	**Results**
Autocorrelation Tests			
Breush-Godfrey	15.29	0.50	Ho is not rejected. No AC
Ljung-Box	0.04	0.83	Ho is not rejected. No AC
Heteroscedasticity Tests			
Breusch-Pagan	21.72	0.65	Ho is not rejected. No HC
White	233.87	0.85	Ho is not rejected. No HC
Normality Tests			
Anderson-Darling	0.37	0.43	Ho is not rejected. Residuals are normal
Shapiro-Wilk	0.99	0.13	Ho is not rejected. Residuals are normal
Kolmogorov-Smirnov	0.48	0.00	Ho is rejected. Residuals are not normal
Non-Linearity Tests			
Utts Rainbow	1.13	0.22	Ho is not rejected. Model is Linear
Lagrange Multiplier Test	17.84	0.85	Ho is not rejected. Model is Linear

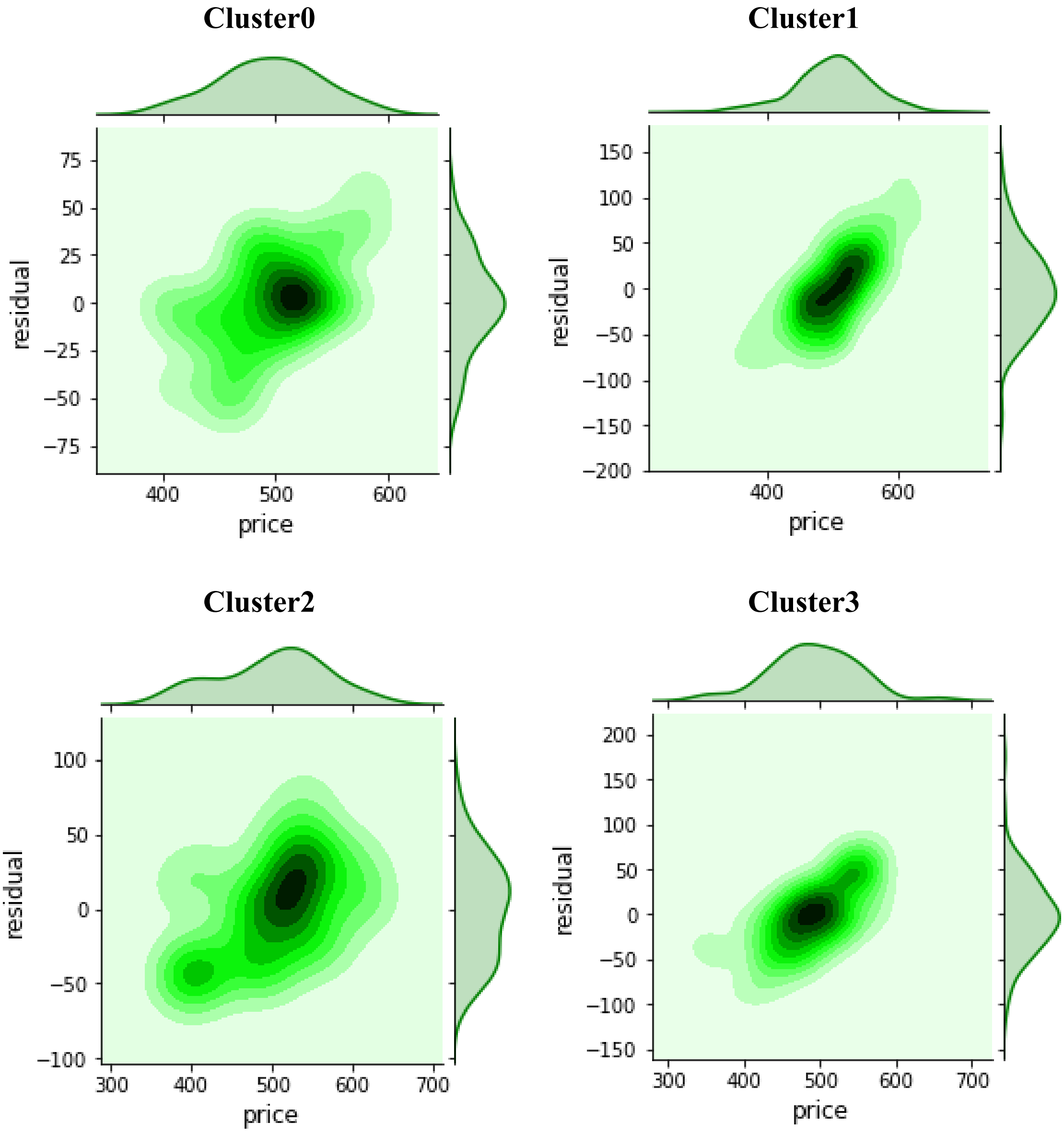
5.1.7 Multivariate Linear Regression for Each Cluster

One of the main objectives of this dissertation is to develop price models that can predict auction closing prices. The developed price model specification was explained in the previous sections. In the literature, it is expressed that the predictive performance of the models can be increased as by clustering process based on certain features. Following this perspective, the price model was re-estimated for each clusters applying some clustering algorithms with the recommended number of clusters and the feature selection algorithms that we mentioned earlier. Regression results of the unsupervised models for each cluster is provided in Appendix F and the p-values showing the statistical significance level of the variables are summarized in the table below. In the new models, many variables do not seem statistically significant and the significant ones are marked bold. This result can actually be expected since similar auctions were grouped by clustering process. According to the silhouette statistics of the auction groups, it was proved that the groups are well separated from each other. Then, the price dynamics in each cluster might be different and the effect of the variables on the price may change. If each cluster models is examined individually; in Cluster0, k1, duration of auctions and the ssunday variables are highlighted. In Cluster1, the variables appear to have an impact on the end price are namely, product condition ratings, auction duration, auction starting days and initial stage variables. In Cluster2 and Cluster3, similarly, economic variables, auction duration, auction start days and initial stage variables have affected the auction closing price.

Table 5.9: Results of Multivariate Regression for Each Cluster

	Cluster0	**Cluster1**	**Cluster2**	**Cluster3**
Number of Observations	35	223	32	90
Independent Variables	**pValue**	**pValue**	**pValue**	**pValue**
auclast4	0.795	0.147	0.627	0.699
binlast6	0.164	0.323	0.845	**0.046**
binlast8	0.809	0.251	0.477	0.136
extras	0.213	0.179	0.588	0.398
selrate	0.783	0.107	0.750	0.221
neg	0.412	0.246	0.274	0.582
views	0.396	0.701	0.735	0.329
k0	0.142	**0.000**	0.515	0.220
k1	**0.071**	**0.001**	0.800	0.274
k2	0.231	**0.000**	0.530	0.177
k3	**0.062**	**0.000**	0.330	0.149
d1	**0.056**	**0.000**	0.551	**0.067**
d3	**0.054**	**0.000**	0.501	0.198
d5	**0.046**	**0.000**	0.477	0.176
d7	0.484	**0.000**	0.542	0.150
d10	0.136	**0.015**	0.542	0.582
smonday	0.360	**0.000**	**0.046**	0.304
stuesday	0.120	**0.000**	0.737	0.221
swednesday	0.257	**0.003**	0.363	0.241
sthursday	0.295	**0.002**	0.563	0.924
sfriday	0.415	**0.000**	0.906	**0.051**
ssaturday	0.885	**0.018**	0.337	0.220
ssunday	**0.100**	**0.000**	0.861	0.253
biddercrp	0.554	**0.013**	0.308	**0.095**
Lbidcrp	0.247	**0.000**	**0.001**	**0.001**
Lbidderratecrp	0.205	**0.046**	0.524	0.922
Adj. R-square	0.238	0.306	0.33	0.226

In order to illustrate the in-sample performance of the models for each cluster, the actual price and the residuals are represented with the joint plot. The fact that the actual prices are around $500 and the residuals shown with a dark green around $0 means that the data is concentrated there giving good tips about the high performance of the model.


Cluster0

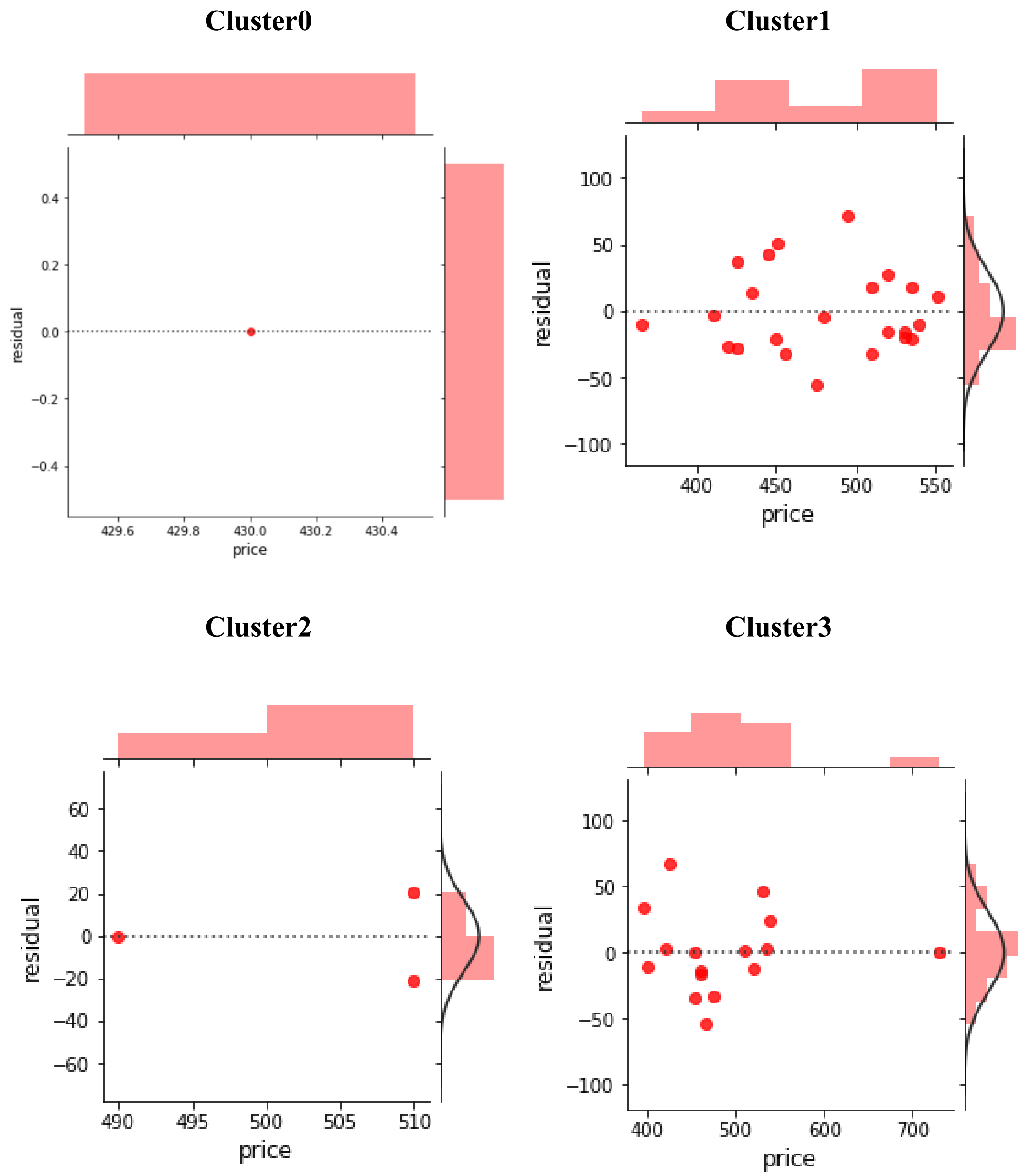
Cluster1

Cluster2

Cluster3

Figure 5.11: In Sample Prediction Performance for Clusters

And the following graphs show the out of sample performance of the models in which a total of 43 auction data were used to measure. The data are shown as dot plots to make it easier to understand with the advantage of not having too many data.


Cluster0

Cluster1

Cluster2

Cluster3

Figure 5.12: Out of Sample Prediction Performance for Clusters

The distribution of data between clusters is not uniform but resembles the distribution of in-sample data. The estimation error of the model is distributed around the average of $0 in the range of (-50$, +$50). For two clusters, Cluster0 and Cluster2 the forecast errors seem much better than these values.

5.1.8 Comparison of Regression Results

The following tables show the statistical information of the estimation error in order to evaluate the in-sample and out-of-sample performances of the models in a profound way. The data in the table are given in terms of Mean Absolute Percentage Error (MAPE) values. First of all, according to the in-sample performance table, the estimation made with the all-sample model has a MAPE of 7.64%. This shows that the variance of real price value is largely grasped by the price model. On the other hand, by the forecast performance of the clusters, one can see a better result where the average estimation error of the cluster models is 6.00% and weighted forecast errors by the number of observations in the clusters are 6.94%. The prediction error could be reduced by more than 1.5% by simply clustering the forecast model without changing any variables in the model at all. In addition, there are two additional important issues. The standard deviation in the average of the error rates of the cluster model is smaller than the standard deviation of the all-sample model. Also, the maximum value of the estimation error is lower in cluster models when compared to all sample model. In short, the clustering process ensures that smaller error terms with less standard deviation and indicates the increase in the predictive performance of the auction end price model.

Table 5.10: OLS In Sample Performance for All Sample and Each Cluster

In Sample	AllSample	Cluster0	Cluster1	Cluster2	Cluster3	Average	Weighted Av
count	380.00	35.00	223.00	32.00	90.00	95.00	
mean	**7.64**	**3.86**	**7.67**	**5.66**	**6.79**	**6.00**	**6.94**
std	6.84	3.48	7.05	3.91	5.95	5.10	6.20
min	0.00	0.00	0.01	0.24	0.03	0.07	0.03
25.00%	2.93	1.06	2.74	2.34	2.42	2.14	2.48
50.00%	5.91	3.40	6.05	5.69	5.52	5.16	5.65
75.00%	10.56	6.00	10.40	8.74	9.40	8.63	9.62
max	61.49	12.27	54.26	14.12	31.56	28.05	41.64

The following table shows the out-of-sample performance of the proposed price model from MAPE perspective. This table also contains similar contents to the previous one. The all-sample model which is created without any clustering could estimate the end price of an auction which is never seen before with an average of 7.59% error. This performance is even higher in cluster models where the average error rate is 4.65%, and the weighted error rate is 6.91%. Similarly, the standard deviation and maximum of errors are also significantly reduced by the clustering process. Based on these values, it can be concluded that the price forecast success of the proposed price model and clustering approach is quite high.

Table 5.11: OLS Out of Sample Performance for All Sample and Each Cluster

Out of Sample	AllSample	Cluster0	Cluster1	Cluster2	Cluster3	Average	Weighted Av
count	43.00	1.00	23.00	3.00	16.00	10.75	
mean	**7.59**	**0.01**	**6.71**	**3.64**	**8.23**	**4.65**	**6.91**
std	5.55	nan	4.90	0.96	4.62	3.49	4.51
min	0.23	0.01	0.42	2.76	0.71	0.97	0.68
25.00%	2.43	0.01	2.22	3.13	4.75	2.53	3.17
50.00%	7.58	0.01	7.28	3.49	7.74	4.63	7.02
75.00%	11.93	0.01	9.14	4.07	11.49	6.18	9.45
max	21.46	0.01	16.30	4.65	15.31	9.07	14.74

5.2 Supervised Machine Learning Algorithm For Auction Prices

The process is called supervised learning if the true result information about the instance to be estimated is also included in the training dataset. The true result information is called the "label" for the data and such datasets are called labeled data. For instance, suppose the aim is classifying each incoming email by spam or not by the help of natural language processing algorithms. If the emails are actually marked as spam or not, this dataset is labeled data and the label is the marks. The labels might be included in the nature of the dataset or can be determined manually by the experts. Briefly, the supervised learning process is modeling the relation between the input *x*, and the output *y* with a function $y = f(x)$ and estimating the parameters of the model in order to predict the results of the cases which were never seen before. The supervisors are the labels in the training dataset. According to label prediction, the cost function value is calculated in an iterative way and the parameters that give the lowest cost are selected. It is a process in which big mistakes are corrected by the supervisor. The learning process stops when the total cost falls below a predetermined and acceptable level. In this chapter, the process of incorporating product condition ratings generated by the Supervised Learning Model (SML) into the auction price models will be explained. The explanatory texts entered into the sales pages of the products carry valuable information about the real value of the product. Buyers can update their valuations for the product by analyzing this information. It is necessary to pull valuable information from these explanations and incorporate them into product price models. Otherwise, the models will have a crucial omitted variable bias. In the previous section, we grouped the most similar statements and added the number of clusters to the price model as a product condition rating. In that process, the actual price information of the products has never been used, only the similarity rates of the texts were taken into account. However, supervised learning models can be utilized and actual prices can be used as labels for the product descriptions. For this process sufficient number of samples are necessary to estimate the models. Technically, price of each product can not be used as a label for the product descriptions individually. On the other hand, there is no label for each product or product group in the dataset. For this reason, we need to group specific price ranges and product explanations in a virtual manner. When determining the number of groups, we will refer to the number of 4 groups used by some of the vendors as described in the previous section. The groups will represent “excellent”, “great”, “good” and “fair” product condition ratings respectively. From the economic theory and previous study the product condition ratings are linear and positive in relation to the prices. By taking this point of view, we have divided the prices of the products sold as BIN sales into 4 virtual groups between the minimum and maximum prices. We set the price ranges in such a way that there will be enough observations in each cluster. The following table and graph show the virtual product condition group created for 676 buy-it-now sales. The table shows price ranges for product groups and the number of observations in each group. The graph shows the change in the structure of groups over time. One can see from the chart that data is concentrated around $500 and the structure of the groups has not changed over time. The virtual grouping method might seem a subjective method. However, this is the easiest way for prices to be used as labels for product descriptions. The number of groups and product price ranges can be created in many different ways. However, one can assume that the effect of different methods on the price will be limited after experiencing a price model without using any price tags.

Table 5.12: Supervised Product Condition Rate Class Ranges

Min Price Boundary $	Max Price Boundary $	Product Condition Rate Class	Number of Observation
527.95	685.00	Excellent	145.00
490.00	525.00	Great	158.00
454.99	489.99	Good	172.00
249.99	450.00	Fair	201.00

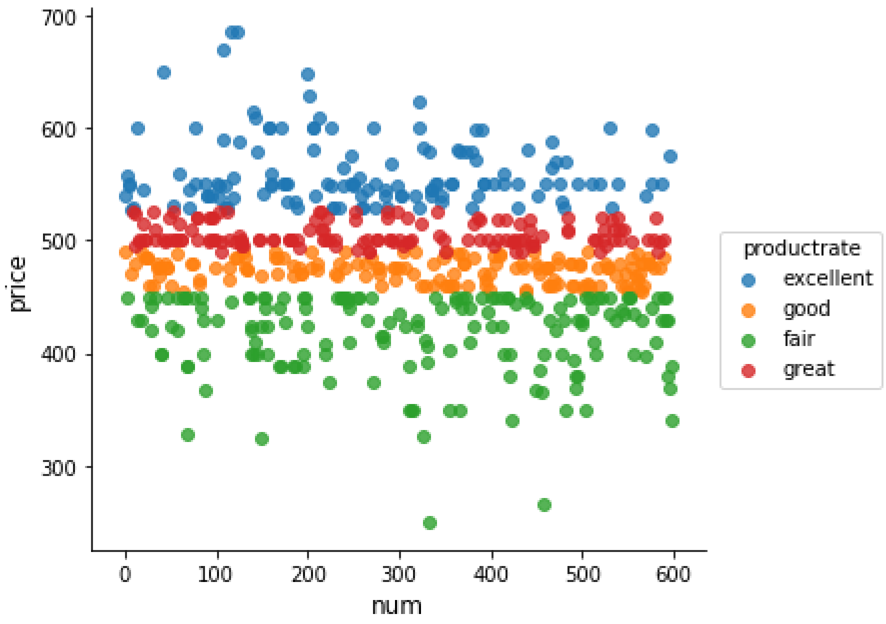


Figure 5.13: BIN Prices by Product Rates for Supervised Learning

The figure below describes the new process of incorporating the product condition ratings into the auction price models. As a first step, BIN sales were clustered and labeled according to the above mentioned virtual groups. In the labeled dataset, Natural Language Processing (NLP) methods were utilized to train various text classification models and their parameters were estimated. At this stage, through the models we have developed, vendor descriptions written for auction sales were analyzed and condition ratings for the products were estimated. The remaining processes are the same as in the previous section. That is, the product condition ratings were also included in the variable set. With the top-down approach, the independent variables were determined and the auction price model was estimated. Then the best features to be used for clustering were determined with certain feature selection algorithms. Subsequently, both all-sample and cluster models were estimated. The analysis will be completed by comparing both in-sample and out-of-sample performances of those models.

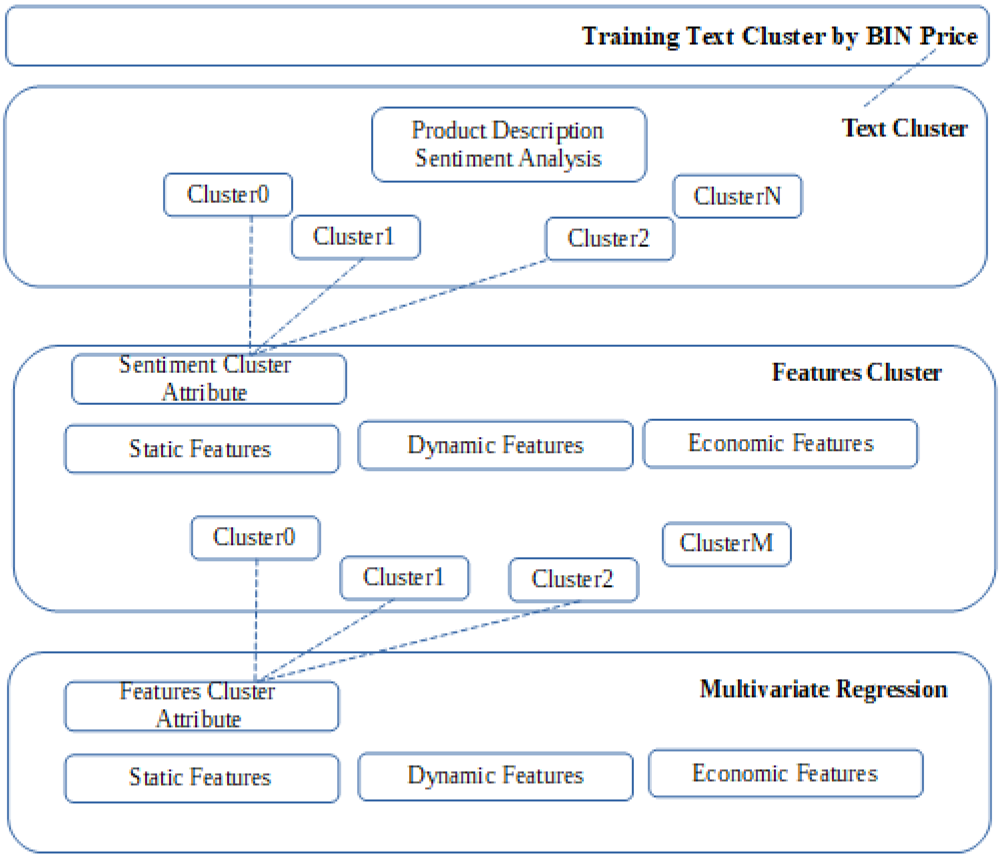


Figure 5.14: Supervised Machine Learning Process Map for Auction Prices

In this dissertation, scikit-learn module developed by Pedregosa et al. (2011) and supervised text classifier models will be used and they will be briefly explained below.

5.2.1 Multinomial Naive Bayes Text Clustering

5.2.1.1 Naive Bayes Model

The Naive Bayes (NB) classification uses the Bayes theorem rules. The assumption of conditional independence of all features adds a “naive” property to this method. Although this assumption in real life does not hold, NB classifiers give successful results. Zhang (2004) described why NB classifiers are successful. Briefly, the dependence between features loses effect when there is an equal distribution between classes. So they cancel out each other. The NB classifiers show fast performance compared to other models and can be trained with a small number of data.

According to Bayes' theorem, the relationship between and the class variable *y* and feature vector *x* can be shown below:

$$P(y \mid x_{1,......,}x_n) = \frac{P(y)P(x_{1,......,}x_n \mid y)}{P(x_{1,......,}x_n)} \tag{5.13}$$

with naive conditional independence assumption

$$P(x_i \mid y, x_{1,......,}x_{i-1,}x_{i+1,......,}x_n) = P(x_i \mid y) \tag{5.14}$$

this relationship can be written for all *i,*

$$P(y \mid x_{1,......,}x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_{1,......,}x_n)} \tag{5.15}$$

Given the input denominator is constant then classification rule can be simplified as

$$P(y \mid x_{1,......,}x_n) \propto P(y)\prod_{i=1}^{n} P(x_i \mid y) \tag{5.16}$$

Thus,

$$\hat{y} = \arg\max_{y} P(y)\prod_{i=1}^{n} P(x_i \mid y) \tag{5.17}$$

for *P(y)* and *P(x$_i$ | y)* estimation Maximum A Posteriori (MAP) can be used where *P(y)* is the relative frequency of class in the training dataset.

5.2.1.2 Multinomial Naive Bayes

Multinomial Naive Bayes (MNB) text classification model uses NB algorithm for multinomially distributed data. For the text classification process, data are represented as word of vectors and distribution is parameterized by vectors:

$\theta_y = (\theta_{y1}, ...\theta_{yn})$

y is the class, *n* is the number of features that is, size of vocabulary and θ_{yi} is the probability of feature *i* seen in sample belongs to class *y.*

By the relative frequency counting given below θ_y parameters can be estimated by a smoothed version of maximum likelihood model.

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \quad \textbf{(5.18)}$$

N_{yi} is frequency of feature *i* in a sample of class *y* in the training set and N_y is the total number of all features in class *y* and alpha is the smoothing factor.

5.2.2 Logistic Regression Text Clustering

Although the model name is regression, logistic regression can be used for classification purposes. In fact, it is an optimization problem that minimizes the following cost functions for L2 and L1 regularization respectively where regularization is optional. In this model, the probabilities of the outcomes of a single trial are modeled by a logistic function.

$$\min_{w,c} \| w \|_1 + C \sum_{i=1}^{n} \log(\exp(-y_i (X_i^T w + c)) + 1) \quad \textbf{(5.19)}$$

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i (X_i^T w + c)) + 1) \quad \textbf{(5.20)}$$

In the model, the observation y_i assumed to take -1, +1 values at trial *i.*

The above implementation provided by scikit-learn can be implemented for binary, One-vs-Rest, or Multinomial logistic regressions.

For some solvers, it is mentioned that by Bishop (2006), L2 penalization is found to be better in fast convergence. With multi-class labels, known as multinomials, the model learns true multinomial logistic regression model, meaning that its probability estimates are better than the default “one-vs-rest” setting. Thus, in this model L2 regularization is used as default penalization.

5.2.3 Linear Support Vector Classification (SVC) Text Clustering

LinearSVC is an application of Support Vector Classification with a linear kernel and this model uses "one-vs-rest" multi-class strategy for n class models. As stated in scikit-learn documentation, support vector machine builds hyper-planes in a high dimensional space for classification and regression purposes. A good separation can be achieved by the hyper-plane that has the largest distance to the nearest training data points of any class. This is called functional margin and in general the larger the margin means the lower the generalization error of the classifier. Mathematical formulation of the optimization problem and the decision function developed based on papers of Cortes and Vapnik (1995) and Guyon et al. (1993) are given below:

$$\begin{aligned}
&\textit{given training vectors } x_i \in R^p, i = 1,.....,n \textit{ and for two classes } y \in \{1,-1\}^n \\
&\textit{Support vector classifier solves below optimimization problem} \\
&\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i \qquad \textbf{(5.21)} \\
&s.to. \;\; y_i (w^T \phi(x_i) + b) \geq 1 - \zeta_i \\
&\zeta_i \geq 0, i = 1,.....,n
\end{aligned}$$

And its dual;

$$\begin{aligned}
&\min_{\alpha} \; (\frac{1}{2} \alpha^T Q \alpha - e^T \alpha) \\
&s.to. \;\; y^T \alpha = 0 \qquad \textbf{(5.22)} \\
&0 \leq \alpha_i \leq C, \;\; i = 1,.....,n
\end{aligned}$$

e is vector of all ones, *C* is upper bound, *Q* is nxn positive semi definite matrix

$$Q_{ij} = y_i y_j K(x_i, x_j), \;\; \textit{where } K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \textit{ is the kernel} \qquad \textbf{(5.23)}$$

and decision function:

$$\operatorname{sgn}(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + \rho) \qquad \textbf{(5.24)}$$

5.2.4 Random Forest Text Clustering

A random forest fits various decision tree (DT) classifiers on sub-samples of dataset and uses averaging method to improve the predictive performance and avoid over-fitting. Although, complex technical details are out of scope of this study, general representations for decision trees and random forest structure is given below.

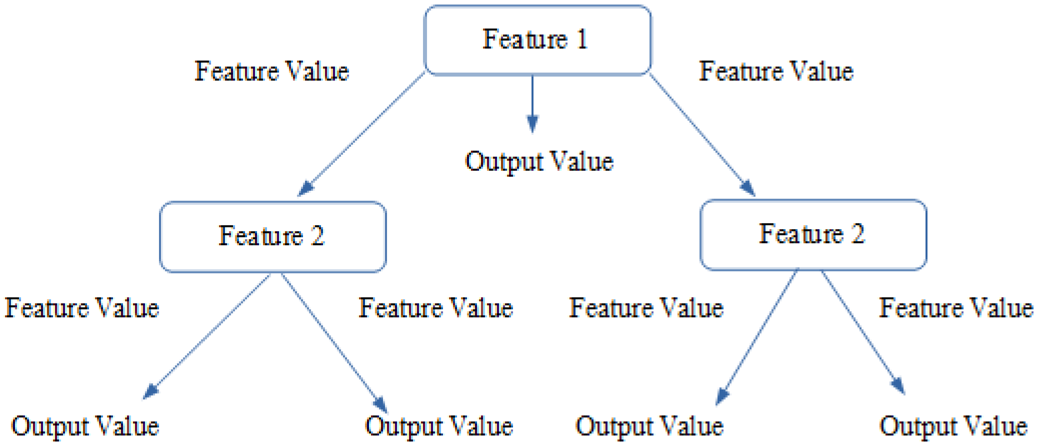


Figure 5.15: General Decision Tree Representation

DT is a non-parametric classification method. It uses a learning approach with simple if-else decision rules inferred from the data features.

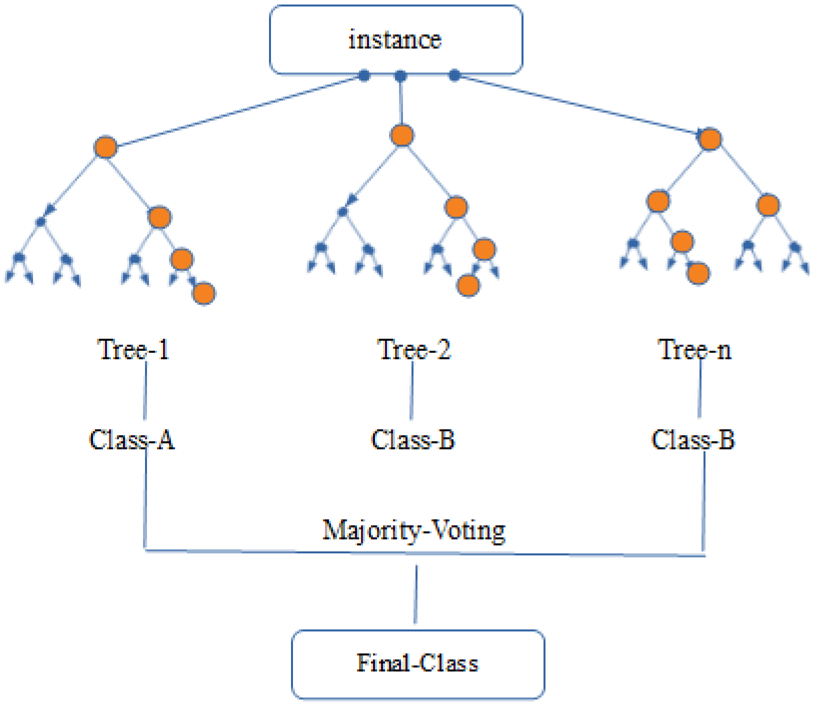


Figure 5.16: General Random Forest Structures Representation

Random forest classification method creates multiple decision trees and combines them together to get a more accurate and stable prediction. As its name implies, the algorithm creates a forest in some way and makes it random. It uses the techniques created by Breiman (1998) specifically designed for trees. That is, various classifiers are trained by introducing randomness in the model construction. The prediction is done by ensembling which means averaging the predictions of the individual classifiers.

5.2.5 Comparison of Text Clustering Algorithms

The performance of the above-mentioned text classification methods in the cross-validation stage are shown in the figure below. For the performance measurement, part of the training dataset is reserved for validation and the performance is calculated according to the "accuracy" metric measured at each step of validation. Accuracy metric is calculated as follows:

$$acc(y,\hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1} 1*(\hat{y}_i = y_i) \tag{5.25}$$

In this formula, *n* represents the number of samples, *y* is the actual class value and is $\hat{y}$ the estimated class value. The nominator in the accuracy metric is a binary variable which equals to 1 when two values are equal and 0 otherwise. *N* is the total number of observations in the denominator. In other words, it shows how much of the classification prediction for BIN sales is correct with the model developed. In the figure the success of Multinominal NB and Logistic Regression classifiers are the highest. At the cross-validation stage, we found that the success rate of the logistic classifier concentrate on a narrower area, but the Multinominal NB classifier can deliver a higher success rate. For this reason, we decided to use the Multinomial NB model for the classification of texts. The success of the classifier is around 44% maximum. One can normally notice the fact that classifiers do not have a very high degree of accuracy. This emerges from the fact that virtual text groups are created according to the sale prices in addition to lots of differences in each text. But what is important is not the success of the classification model, but the fact that the product condition information can be incorporated as a variable into the price models in a supervised learning manner. The Figure 5.18 below represents the number of observations in each labels:

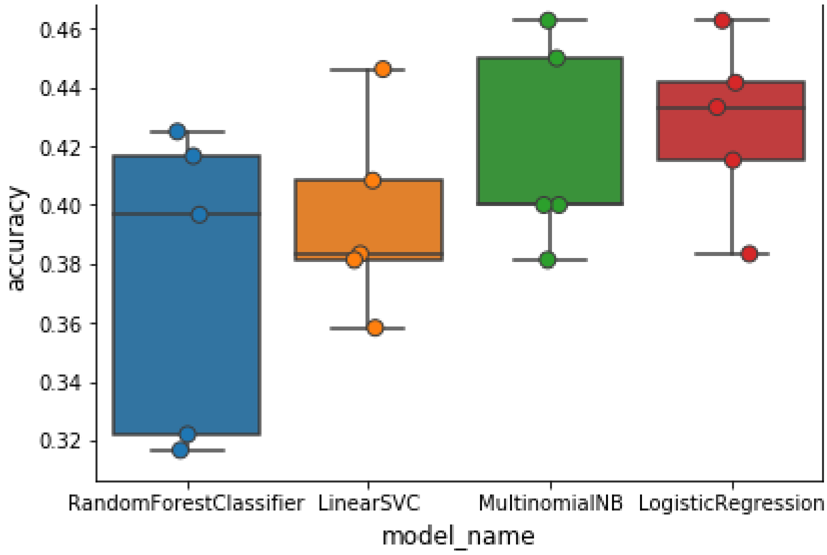


Figure 5.17: Cross Validation Accuracy Comparison for Text Clustering

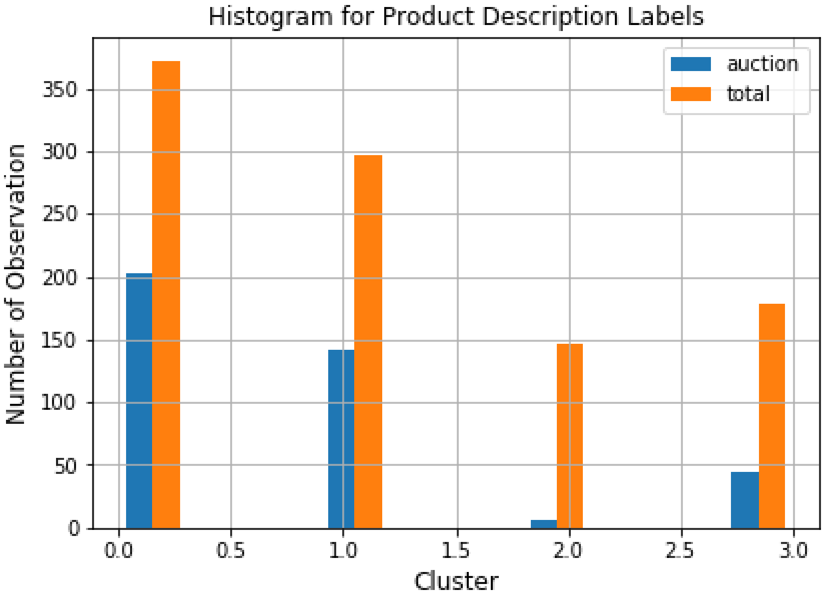


Figure 5.18: Histogram for Estimated Labels for Product Descriptions

The following table shows the 15 most important group of words in product descriptions for each cluster created with Multinominal NB classifier. The words that can give an idea about how the product was used and which can be a signal for product condition were marked bold and gray fill. The order of words in the same column shows the importance assigned by the model. So the words shown in the 1st row in the same column are the most important and frequent words in the product descriptions. The words in the 15th rank are less important and fewer in the texts.

Table 5.13: Top 15 Terms for Each Cluster by Multi nominal NB Algorithm

	excellent	**great**	**good**	**fair**
	excellent shape	gsm cdma	**small marks**	cleared original
	comes apple	**near perfect**	**phone need**	**dings screen**
	128gb rose	charging port	edge screen	comes gold
	100 guaranteed	**condition scratches**	**heavy wear**	phone cleared
	esn Verizon	**includes original**	pictures different	phone earbuds
	box brand	account iPhone	**judge overall**	**scratches signs**
	buy trusted	does include	view judge	settings comes
	shipping thank	box include	need don't	earbuds charger
	lightning fast feedback	phone shipped	**overall quality**	receive phone
	lightning	include charger	**quality phone**	earbuds used
	trusted seller	day purchase	don't hesitate	**condition scratches**
	showing little	cable charging	angles view	case earbuds
	save big	phone work	different angles	**scratches dents**
	unlocked sim	ship day	gold iPhone	item minor
	Verizon factory	used comes	tried pictures	protector case
Sentiment	Positive(+)	Neutral(+)	Neutral(0)	Negative(-)
Price Effect	High	Medium	Medium	Low

For the products with high price, which were labeled as “excellent” the text classification model determined the most import words as “excellent shape”. This word group is very compatible with the label name for that cluster. Also for this excellent label, phrases such as “buy trusted”, “trusted seller”, “save big” which highlight seller reliability and price versus quality performance have come to the forefront. Moreover, there is no word for the excellent label to imply any negative opinions about the condition of the product or the vendor. Considering all the words, one can foresee that product descriptions with the excellent label might create a positive sentiment in customers and this might have a high impact on the auction price. In other words, the coefficient in the price model for the product group labeled with “excellent” might be the highest compared to the other three variables.

In the product descriptions labeled as “great” the phrases “near perfect” are highlighted as the most important one. Besides, the words “condition scratches”, which indicates that there are some scratches on the products, came in second. Thirdly, the words “includes original” appear which has a positive meaning. By looking at these words, it can be said that the condition of the products is generally good, but some scratches on the products add a little bit negative sentiment. Thus, it can be foreseen that the statements written for the product will create a sentiment close to positive, but the price effect might not be as high as the excellent one.

The product description labeled as “good” include many words that can give an idea about the condition of the product. These are “small marks”, “phone need”, “heavy wear”, “judge overall”, “overall quality” and “quality phone” respectively. These words suggest that there might be some scratches on the product, and obsolescence from use and the current product quality might be in average level. For this reason, it can be predicted that products labeled as “good” might create a near-negative perception in the customers, which will have an impact on the price less than the others.

Finally, by looking at the list of the “fair” cluster, it can be observed that many words stand out with a negative perception about the condition of the product. These are the words “dings screen”, “scratches signs”, “condition scratches” and “scratches dents”. These words indicate plenty of scratches on the phones in this group. In addition, there are no additional words which can create a positive perspective on the product. Therefore, it can be argued that the product descriptions written for this cluster might create a negative feeling on the buyers resulting a drop in the price. That is, coefficient of this label might be the smallest in the price model.

5.2.6 Grouping Similar Products by Variable Clustering

In parallel with the previous study, the predictive performance of the price model might be increased by a clustering processes. K-Means and Hierarchical Variable Clustering methods, which were earlier explained in detail will be utilized. These techniques are unsupervised learning algorithms. Since there is no group label in the dataset, use of this kind of learning methods is necessary. In the K-Means method, initial centroids are assigned, the distance of each data point is calculated and the centroids are updated. This process is repeated until all centroids become stable. In agglomerative clustering, first step, each data is assumed to be a cluster, and the closest clusters are combined to each other until a single cluster remains. In order to use these approaches, it is necessary to determine the number of clusters in advance. There is no definitive method to determine the optimal number of clusters. Even so, information such as the elbow method, silhouette statistics can be used. In addition, the number of clusters can be defined by trying all the possibilities in a certain range to find out which combinations provide the best estimation performance. In this study, the latter method is used. There are a total of 50 features that can be used for clustering. We have identified the number of clusters in a way that can provide group diversity and contain a reasonable number of observations. To ensure this, it is considered that the number of clusters could be between 3 and 9. For each possible combination, the variables were determined by the feature selection algorithms, which were explained in detail in the previous section. For both of the clustering methods, in-sample and out of sample performances were calculated to choose the best ones. In order to find the best combination, the price model was estimated approximately 10,000 times. Supervised model performance of the feature selection methods for K-Means clustering and Hierarchical clustering methods are provided in Appendix G and Appendix H respectively. The following table summarizes each feature selection algorithm and the best performances of each clustering method. As it can be seen from this table, the best predictive performance is supported by clustering with 5 groups by K-Means method. The features used for the clustering were determined by the recursive feature elimination (RFE) method. These features are respectively “extras”, “neut”, “neg”, “k0”, “k1”, “k2”, “k3”, “percent”, “swednesday”, “sthursday”, “sfriday”, “ssaturday”, “ssunday”, “d3”, “d5”, “d7”, “d10”, “biderkrp”. The Figure 5.19 shows the number of observations assigned to each cluster and there is adequate data to estimate the price model for each cluster.

Table 5.14: Performance of Feature Selection Methods for Supervised Model

Feature Selection Method	Clustering Type	Number of Clusters	Number of Features	Insample MAPE	Weighted Insample MAPE	Out of Sample MAPE	Weighted Out of Sample MAPE
SelectKBest	K-Means	6.00	33.00	6.13	6.56	5.99	6.90
RFE	**K-Means**	**5.00**	**18.00**	**6.09**	**6.68**	**5.25**	**6.80**
PFA	K-Means	4.00	13.00	5.82	6.76	5.39	7.04
SelectKBest	Hierarchical	3.00	19.00	7.00	7.07	6.40	6.99
RFE	Hierarchical	4.00	15.00	6.40	6.88	6.00	7.15
PFA	Hierarchical	3.00	28.00	5.99	6.93	6.18	7.49
Grid Search	K-Means	3.00	3.00	5.46	7.09	4.30	7.01
Grid Search	Hierarchical	4.00	3.00	6.76	6.82	9.77	7.65

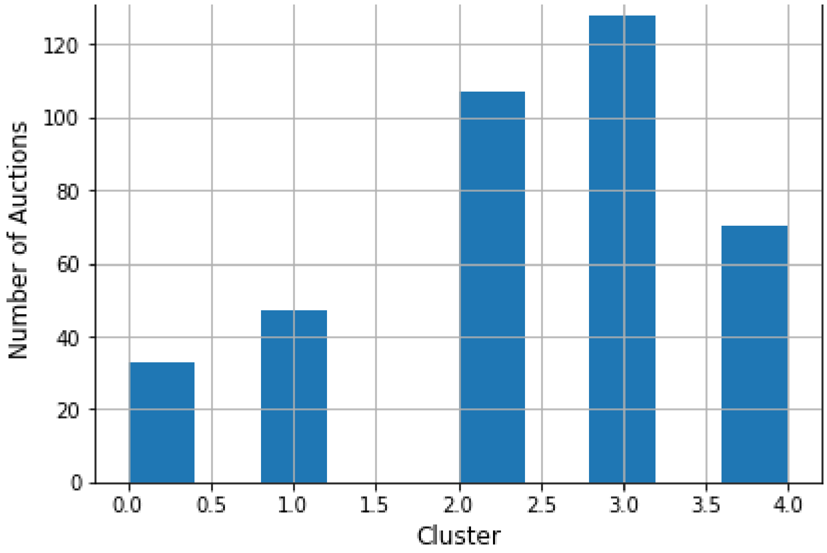


Figure 5.19: Number of Auctions for Each Cluster

5.2.7 Past Auction and BIN Price Clustering

There are two candidate methods which can be used to incorporate economic variables into the price model. The first is to determine past sales based merely on auction end time. We determined economic variables using this approach. Another method might be to cluster old sales and determine past prices according to clusters. This method is a challenging and time-consuming one. It also does not allow for sufficient data in each cluster. In computable models, the best performance can be achieved by 3 clusters with RFE method. The in-sample performance for the model is 5.56% and the out of sample performance is 6.84% in terms of MAPE. These statistics do not outperform the values which were found earlier. For this reason, it is a better strategy to determine economic variables only by time and this research will continue to do so in this study similar to the previous one.

5.2.8 Multivariate Linear Regression and Regularization

The OLS model, the cost function and the mathematical formulation of the estimators are presented below. Obviously, the models shown are the same as the models built in the previous section. Only the variables in the model will be different. In this study, variables were selected by the top-down method to refrain from the bias that may occur in regularized models thus no penalty has been added to the cost function. Multiple linear regression model represented below is used in the analysis.

$$Y_{nx1} = X_{nxm}\beta_{mx1} + u_{nx1} \qquad \textbf{(5.26)}$$

$$ap = sf * \beta + df * \theta + ef * \delta + u \qquad \textbf{(5.27)}$$

$$L(\beta) = \sum_{i=1}^{n} (Y_i - X_i\beta)^2 \qquad \textbf{(5.28)}$$

$$\beta^{ols} = \arg\min_{\beta \in R^p} \|Y - X\beta\| \qquad \textbf{(5.28)}$$

The proposed multivariate linear regression model, the names, coefficients, level of statistical significance and types of the independent variables are shown in the table below. For the training period of the model, all of the 380 observations in the dataset have been used. The adjusted R^2 of the model is calculated as 0.312.

Table 5.15: Results of Multivariate Regression for All Sample

Dependent Variables	**Auction Price**		
Number of Observations	380		
Independent Variable	**Coefficient**	**pValue**	**Feature Type**
auclast4	0.067	0.118	**economic**
binlast6	0.094	0.045	**economic**
binlast8	0.089	0.087	**economic**
extras	22.067	0.000	static
selrate	-0.014	0.122	static
neg	-5.791	0.038	static
views	0.020	0.001	static
k0 (fair)	89.358	0.000	sent.cluster atribute
k1 (good)	93.453	0.000	sent.cluster atribute
k2 (great)	98.990	0.000	sent.cluster atribute
k3 (excellent)	112.785	0.000	sent.cluster atribute
d1	77.519	0.000	static
d3	83.423	0.000	static
d5	90.540	0.000	static
d7	84.041	0.000	static
d10	59.062	0.002	static
smonday	59.624	0.000	static
stuesday	57.246	0.000	static
swednesday	52.449	0.000	static
sthursday	52.311	0.000	static
sfriday	60.733	0.000	static
ssaturday	48.356	0.000	static
ssunday	63.867	0.000	static
biddercrp	1.336	0.012	**dynamic**
Lbidcrp	0.329	0.000	**dynamic**
Lbidderratecrp	0.013	0.051	**dynamic**
Adj. R-square	0.312		

All variables except “auclast4” and “selrate” are statistically significant. According to the table, the coefficients of past BIN prices among economic variables are positive and significant. The positive coefficient shows the effect of market price changes on the auction price. That is, if the price increases in the BIN market for the product, it is likely to cause an increase in auction prices. As in the previous chapter, one can consider that buyers are able to follow the prices of earlier products considerably, rather than the price of the products closest to the end of the auction.

Extra items stated in the advertisements of the products, such as “charger”, “cable“, “case”, “box”, “warranty” etc. increase the auction price by $22. This is actually expected meaning that there is “no free lunch”. In other words, every component in the auction advertisement is priced.

Customer satisfaction is essential in all kinds of trade. Ebay serves a scoring system to represent the satisfaction of buyers and sellers. In this system, buyers and sellers can give each other a positive, negative and neutral score after any shopping activity. The coefficient of the variable indicating the total rating of the seller is a negative but close to zero, and not significant. It is seen that the negative ratings received by the seller have a significant effect on the price with a coefficient of -5.8$. Although a positive rating does not have any impact on the price, an additional negative rating from a buyer causes a fall of $6 on the product price. This result implies that the sellers need to prove a high level of customer satisfaction to make reasonable revenue. As a matter of fact, if the customer has received too many negative ratings, it might be useful to close the vendor profile and start a new one.

Another variable is the number of times the vendor web pages are visited by customers. In this model, similar to the previous one, it has a significant but close to zero coefficient. The number of customers following seller the web page may be inferred as customer loyalty which can cause customer segmentation and may lead to price increases.

The next group of variables is the product condition rating variables that were created for the product by NLP text classification algorithms. These variables are shown as “k0”, “k1”, “k2” and “k3” and stands for “fair”, “good”, “great” and “excellent” labels respectively. All of the product condition variables are statistically significant in the model. The coefficients are in ascending order as expected. To put it in more detail, we have argued that there might be a lot of scratches on “fair” labeled products and any word demonstrating a positive sentiment about the product do not exist in the list of important words. For this reason, buyers of “fair” labeled products offer less price. The product condition effect at the price with a fair label is $89.4. This value is $12 less on average from the other three labels.

On the other hand, in the product descriptions sold with the “excellent” label, as it was found earlier that the words that will create a positive perspective for the product are at the forefront. Moreover, there is no other word that will damage reputation of the product. One can see the effect of the product condition rating on the prices of the auctions by the k3 coefficient which is $112.8. This value is $18 higher than the average compared to others.

The variables used in the price model for labels “good” and “great” are k1 and k2, respectively. The product descriptions in these groups contain both positive and negative comments, so one may infer a neutral perspective on customers. The estimated coefficients in the model confirm this view. The coefficient calculated for k1 is $93.46 and the coefficient for k2 is around $99. These values are between the values specified for the other two labels. Since the positive word groups are more dominant for the “great” cluster, it is also a logical result that the coefficient in the price model is slightly higher. Looking at the coefficients in general, how the products were used has a significant effect on the auction prices. Of course, it is necessary to express this properly on the auction page. For this reason, it would be useful for the sellers to keep the quality of the product they want to sell above a certain level and express this in a positive way to the customers.

As emphasized earlier and it can be seen from the table below, product descriptions can be more easily internalized by experienced buyers. The following table shows the sets of product condition ratings developed by the SML model and the information for these clusters. The variables k0 and k1 represent more negative expressions. The coefficients of these variables and the price averages of these clusters and the ratings of the bidding customers are lower than the others. When the bidder ratings are examined, a similar situation can be observed as in the previous section. In other words, it is seen that less experienced buyers are bidding on products having a negative sentiment in product descriptions. In other words, in this model, inexperienced buyers do not hesitate to submit a bid in auctions and the experienced buyers do not immediately bid to negative sentiment products in the initial stage of auctions.

Table 5.16: Bidder rate and Product Condition Rating in SML Model

SuperVised		Price ($)		meanbidderrate		meanbidderratekrp	
Product Cond.	Coefficient	Mean	Median	Mean	Median	Mean	Median
k0	27.43	493.11	494.00	171.39	105.65	105.98	30.37
k1	31.52	495.56	497.00	111.90	78.98	68.81	28.06
k2	37.06	499.83	492.50	123.82	59.64	80.17	57.00
k3	50.85	527.53	526.00	119.48	66.36	108.16	35.46

The next group of variables in the model is the binary variable group created for the duration of the auction. These are “d1”, “d3”, “d5”, “d7” and “d10”. For example, if the auction period is 3 days, then the variable d3 takes 1 and 0 in all other cases. The effect of the auction time on the price seems to be in a non-linear manner similar to the model in the previous section. While the auction period is short, the effect on the price is low, it reaches the highest level in 5 days and decreases again for 7 or 10 days. From this point of view, one can conclude that the best length for the auction is 5 days. This period appears to be an optimum value as long as the customer is able to identify, evaluate and bid the product, as well as to obtain the product in a reasonable time. The duration variable of the 10-day auctions is estimated as the smallest factor in this model. This may be due to the fact that the length of the auction period is too long for the customers to wait for the delivery or to follow the auction and bear the competition. This creates waiting and hassle costs for the auction and reduces the competition leading to a price decrease. On the other hand, when the auction length is 1 day, the duration is so short for customers to identify the product and bid on it. Thus, it can be argued that the sellers need to specify the optimum length of the auctions to maximize profit and patient buyers can focus on longer auctions to maximize their surpluses.

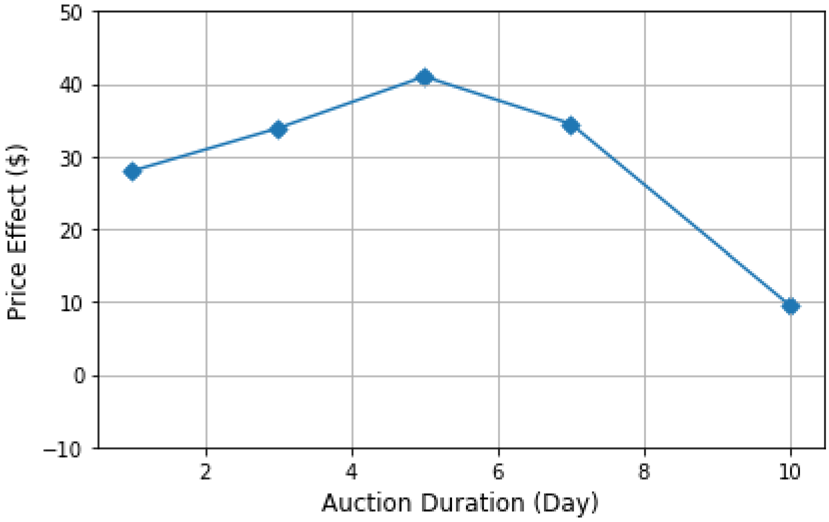


Figure 5.20: The Effect of Duration on Auction Price in Supervised Model

The next group of variables in the model is the days when the auction begins. These variables are binary variables and all of them are statistically significant. The highest coefficient values are on Sundays, Mondays, and Fridays. These findings are very similar to the findings in the previous section. First of all, it is proved that there is a lot of interest in the auctions at the initial stages. Some buyers strategically tend to bid as soon as the auctions start, while others discover the product as soon as it is auctioned, but they wait until the last minutes to bid on the product. It can be assumed that people prefer weekends for shopping or entertainment. For this reason, higher end price of Sunday auctioned products can be attributed to more buyers follow the auction and this causes more competition. On the other hand, the high coefficients of Monday and Friday can be linked to the fact that they are the first and last working day of the week. On the other hand, it can be argued that the auctions in the weekdays are more likely to be disregarded by customers because of other occupations which leads a decrease in the auction prices. There are important inferences that can be drawn by both buyers and sellers from these variables in general. For example, if the vendors want to pull more competition in their products, they might offer their products to the auction market on Sundays, Mondays or Fridays. Buyers might, on the other hand, be interested in products that are auctioned on weekdays if they are expecting a lower price.

The last variable group in the model is the dynamic features, namely, the number of the bidders, the last bid and the rating of the last bidder at the initial stage. The bids in this stage offer essential information for the valuation of the product. Besides, these variables can be called a proxy for competition. All variables in this group are statistically significant and the coefficients are positive. If the number of bidders in the initial stage as a proxy for the future competition in the auction, one can expect a higher final price. That is, it is logical that the coefficient of the variable is positive. The final bid in the initial stage is an important signal for product valuation. This signal is further strengthened if a buyer with a high rating gives a high bid. In other words, as soon as the auction starts, bidding by experienced buyers mean that the market value of the product will be high and the coefficients of these variables are positive.

The following representations show the in-sample performance of the proposed price model. In the graph on the top left, the estimated and the actual price go parallel to each other, and the estimation error terms move around zero. In the graph on the top right, the actual and estimated prices are seen around \$500. At the bottom, the actual price and errors are shown. It is seen that the model errors are largely within the range (-\$50, + \$50) and around 0 implying the proposed model is successful.

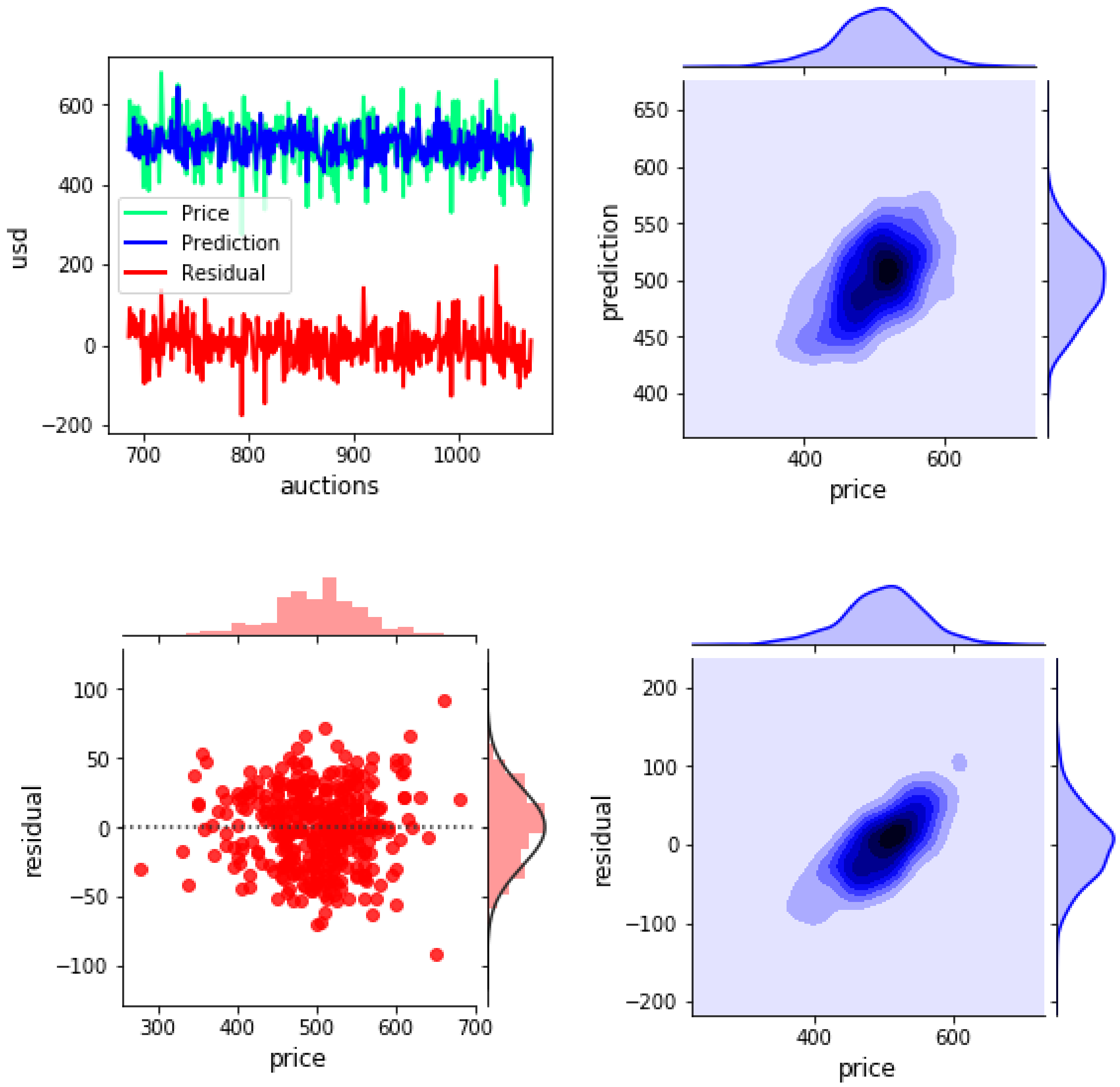


Figure 5.21: In Sample Prediction Performance for All Sample

The following graphs demonstrate the ability of the proposed price model to predict auction end price with the data that have never seen before. From the chart on the top left, it is seen that the variance of the actual prices is largely explained by the model. One can see from other charts that the actual and forecast prices are around $500 and that prediction errors occur in a line that is substantially near zero but is distributed nominally in the range of (-$50, + $50).

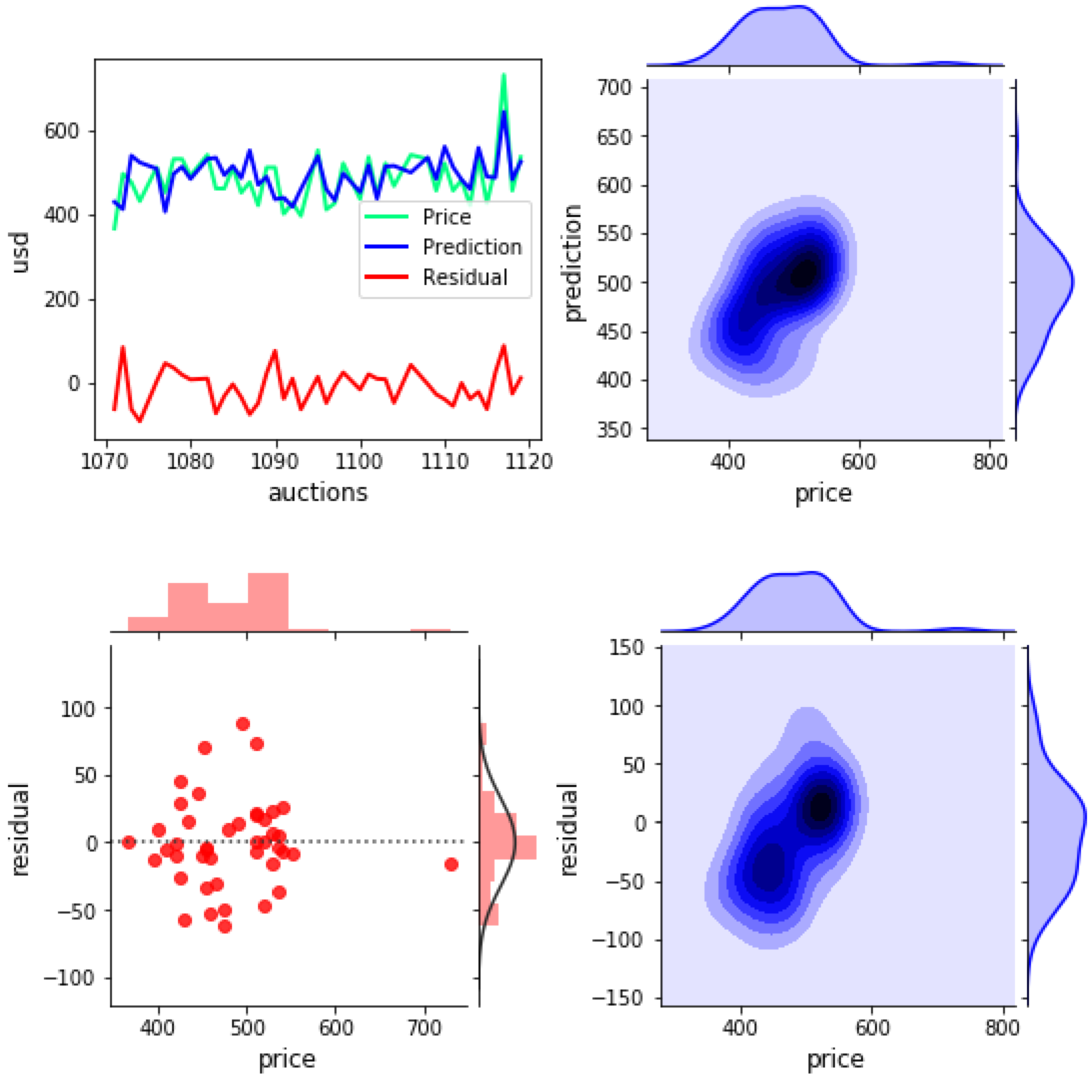


Figure 5.22: Out of Sample Prediction Performance for All Sample

5.2.9 Diagnostic Tests For the Multivariate Regression Model

In the previous chapter, all the necessary assumptions to create correct linear models were described. Important problems, such as model bias, inefficiency, inconsistency, might arise in case assumptions do not hold, The way how these assumptions can be tested were also emphasized. In this section, the assumptions of the model developed with supervised learning models have been tested and the results are given in the table below. There are many methods for testing the assumptions in the literature and the tests that give the best performance for each assumption may change. For this reason, it is desired to increase the reliability of the test results by using more than one test for the assumptions. Looking at the table, the model has passed both of the autocorrelation tests successfully. In other words, the null hypothesis that there is no autocorrelation in error terms could not be rejected. Similarly, according to two heteroscedasticity tests listed in the table, the hypothesis that the model has a homoscedastic variance is not rejected. As to normality tests of error terms below, the Anderson-Darling test supports the assumption that the error terms are normally distributed, on the other hand, the assumption of normality should be rejected according to the other two tests. However, taking into account the number of adequate observation and the proper variance-covariance structure, violation of this assumption might be ignored. Finally, linearity tests can be used to figure out whether the model is correctly specified. The proposed model has passed from one of the linearity tests and the failed the other. In addition, it is useful to remind that, to refrain from omitted variable bias, top-down variable selection method was used. Only one of the highly correlated variables was used to avoid multicollinearity and the end-of-auction decisions were not included to avoid the endogeneity problem.

The model in Study2 has passed one of the linearity test and failed in the other one. According to the literature such as Peeters and Tenev (2018), Larue et al. (2013), probability of winners curse increases as the number of bidder increases and players increase their bids at a lower rate. In this study, the relationship between auction prices and number of bidders may also be nonlinear. Therefore, square and root of the number of bidders at the initial stages are added into the model as separate variables. When the square of the number of bidders is added to the model, the residuals of the model become linear but the coefficient of the variable is statistically insignificant. When the square root of the number of bidders is added to the model, the coefficient is statistically significant, but the model does not pass one of the non-linearity test again. The results of these trials are given in Appendix I. To sum up, this model has passed one of linearity test and the model created in Study1 has passed both the linearity tests, so we could assume that the model is linear in this study.

Table 5.17: Results of Diagnostic Tests

OLS Diagnostic Test	Test Statistic	pValue	Results
Autocorrelation Tests			
Breush-Godfrey	15.17	0.51	Ho is not rejected. No AC
Ljung-Box	0.02	0.88	Ho is not rejected. No AC
Heteroscedasticity Tests			
Breusch-Pagan	22.73	0.59	Ho is not rejected. No HC
White	217.61	0.88	Ho is not rejected. No HC
Normality Tests			
Anderson-Darling	0.52	0.19	Ho is not rejected. Residuals are normal
Shapiro-Wilk	0.99	0.03	Ho is rejected. Residuals are not normal
Kolmogorov-Smirnov	0.49	0.00	Ho is rejected. Residuals are not normal
Non-Linearity Tests			
Utts Rainbow	1.30	0.04	Ho is rejected. Model is not Linear
Lagrange Multiplier Test	18.22	0.83	Ho is not rejected. Model is Linear

5.2.10 Multiple Linear Regression for Each Cluster

From the literature and the previous study, grouping similar auctions and re-estimating the parameters of the model might increase the in-sample and out of sample performance of the model. In the previous section, it is found that the K-Means clustering model using the features determined by the recursive feature elimination method (RFE) improves well the prediction performance of the supervised model. According to RFE, “extras”, “neut”, “neg”, “k0”, “k1”, “k2”, “k3”, “percent”, “swednesday”, “sthursday”, “sfriday”, “ssaturday”, “ssunday”, “d3”, “d5”, “d7”, “d10” and “biderkrp” variables are the most relevant set of variables in the clustering process.

The best performance is obtained when the number of clusters is 5. Regression results of the supervised models for each cluster are provided in Appendix J.

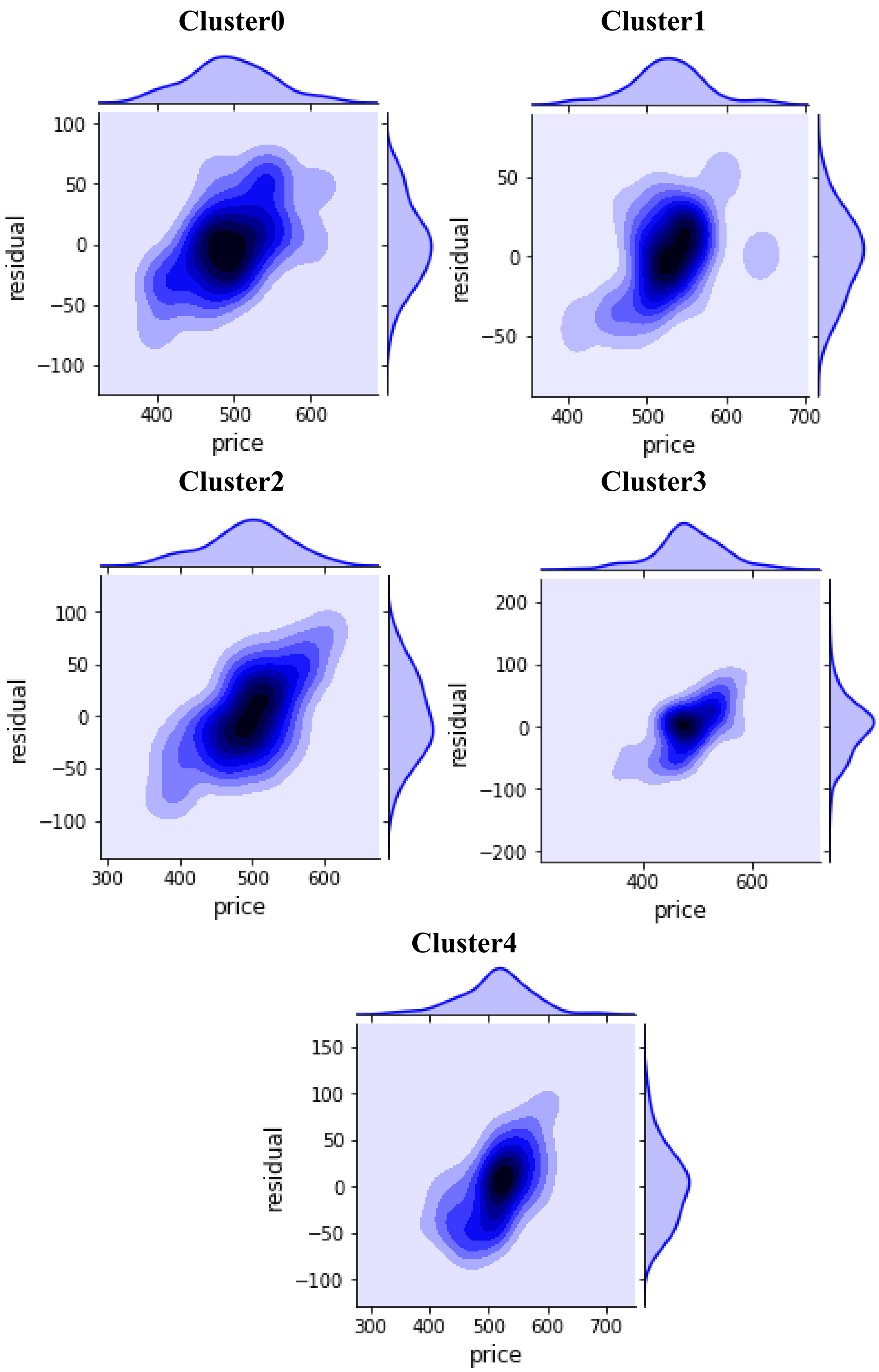
Table 5.18: Results of Multivariate Regression for Each Cluster

	Cluster0	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**
Number of Observations	33	47	105	128	67
Independent Variables	**pValue**	**pValue**	**pValue**	**pValue**	**pValue**
auclast4	0.670	0.465	0.351	0.421	0.874
binlast6	0.562	0.539	0.478	**0.090**	0.771
binlast8	0.906	0.795	0.939	0.162	0.295
extras	0.598	**0.074**	**0.005**	0.230	0.785
selrate	0.577	0.353	0.338	0.309	0.901
neg	0.641	0.371	0.534	0.478	0.455
views	0.813	**0.001**	0.217	0.546	0.467
k0 (fair)	0.114	**0.039**	0.688	0.275	**0.002**
k1 (good)	0.158	**0.018**	**0.001**	**0.061**	**0.004**
k2 (great)	0.308	**0.008**	**0.005**	**0.048**	0.401
k3 (excellent)	0.165	**0.009**	**0.000**	**0.044**	**0.006**
d1	0.162	**0.011**	**0.014**	0.108	**0.007**
d3	0.124	**0.002**	**0.001**	0.932	**0.002**
d5	0.356	**0.020**	**0.002**	**0.088**	**0.014**
d7	0.291	**0.002**	**0.000**	**0.036**	**0.005**
d10	0.273	0.203	0.571	0.236	**0.081**
smonday	0.340	**0.006**	**0.010**	0.146	**0.042**
stuesday	0.886	**0.012**	**0.023**	0.124	**0.015**
swednesday	0.485	0.106	0.656	0.323	**0.018**
sthursday	0.988	**0.018**	0.268	0.220	0.166
sfriday	0.992	0.449	**0.087**	**0.020**	**0.019**
ssaturday	0.195	**0.002**	**0.019**	0.452	0.891
ssunday	0.490	**0.003**	**0.011**	0.181	**0.015**
biddercrp	0.871	**0.085**	0.398	**0.030**	0.306
Lbidcrp	**0.032**	**0.056**	**0.000**	**0.000**	**0.001**
Lbidderratecrp	0.837	**0.003**	**0.084**	0.666	0.892
Adj. R-square	0.163	0.482	0.383	0.192	0.264

The Table 5.18 summarizes the statistical significance values of the independent variables, the number of observations used and the adjusted R^2 of each model.

The statistically significant variables are marked as bold. Most of the model parameters do not appear statistically significant in cluster models. This is a natural result of clustering processing since each cluster has different segments of variables. In Cluster0, initial stage variable is statistically significant. The majority of the variables were statistically insignificant. Therefore, adjusted R^2 is low. In Cluster1, many variables are statistically significant. These can be listed as extras, views, product condition ratings, variables related to the length of the auction, the days when the auction commences and all of the dynamic variables. The list of variables that are statistically significant in Cluster2 is similar to Cluster1. In Cluster3, dynamic variables, the length of the auction, the starting day and product condition ratings, the previous prices of BIN sales appear effective in the price model. Cluster4 has also a similar model structure.

In the table, product condition ratings are significant in 4 of the 5 clusters means that these variables un-disregardable. In each cluster, at least one auction duration, the starting day of the auction and the dynamic variable are statistically significant. Therefore, we have tested the effect of variables on price once again by determining significant sets of variables in clusters. Adjusted R^2 of the models are higher in 2 of the 5 clusters than the all-sample one and lower in 3 of them. Even if the model compliance is low, we can continue to apply clustering since the first performance criterion is based on price estimation and we will mention the forecast performance of the model below. The following two figures show in and out-of-sample performances for 5 clusters, respectively. One can infer from the dark blue colors that the actual price and errors in each cluster are concentrated around \$500 and \$0, respectively. The images show the model is perfectly compatible with the data. In the out-of-sample performance graphs, the errors are mostly close to zero, except for a cluster, and the other four have a prediction error between (-\$20, +\$20). This is actually better result than in-sample and previous studies performances. These values show that the model can predict successfully an auction price it has never seen before.



Cluster0

Cluster1

Cluster2

Cluster3

Cluster4

Figure 5.23: In Sample Prediction Performance for Clusters

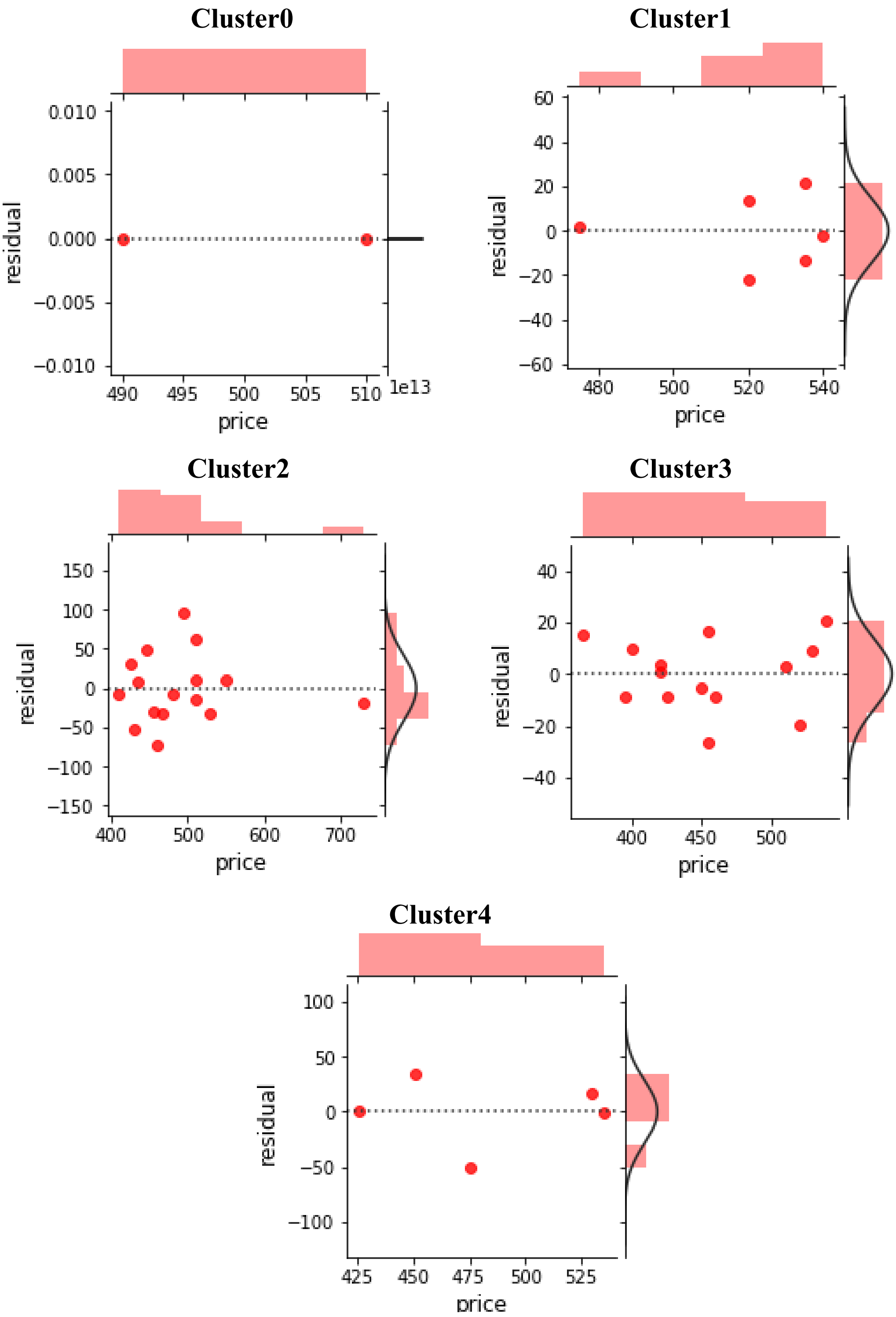


Figure 5.24: Out of Sample Prediction Performance for Clusters

5.2.11 Comparison of Regression Results

The performance of the developed model is shown in the following tables with Mean Absolute Percentage Error (MAPE) values. First of all, according to in-sample performance result, it can be claimed that all-sample model can accurately estimate the actual prices with a 7.57% estimation error. Indeed, the price prediction patterns of the clusters are more accurate. By MAPE levels, the average of the errors in cluster models is 6.09% and the error rate weighted by the number of observations in the clusters is 6.68%. In both cases, it is seen that the prediction performance of the models has increased significantly by the clustering process.

Second performance improvement with the clustering process can be seen in the standard deviation and maximum values of errors. In other words, error rates in the cluster models have a lower standard deviation. Also, the maximum values of error rates are significantly reduced, which was a very high rate, 60% in the all-sample model, it falls to 27% in the cluster models. In fact, one can also suggest applying the clustering process by considering only the improvement in the maximum values of errors. Finally, the in-sample performance of the model is better than the one in the previous section.

Table 5.19: OLS In Sample Performance for All Sample and Each Cluster

In Sample	AllSample	Cluster0	Cluster1	Cluster2	Cluster3	Cluster4	Average	Weighted Av
count	380.00	33.00	47.00	105.00	128.00	67.00	76.00	
mean	**7.57**	**5.27**	**3.93**	**6.68**	**8.22**	**6.35**	**6.09**	**6.68**
std	6.86	4.48	3.05	4.83	8.47	4.89	5.14	5.82
min	0.01	0.00	0.00	0.07	0.00	0.00	0.01	0.02
25.00%	2.85	1.81	1.36	3.08	2.50	2.05	2.16	2.38
50.00%	6.14	4.41	3.26	5.93	5.64	5.64	4.98	5.32
75.00%	10.08	7.73	5.80	9.39	11.83	9.89	8.93	9.71
max	63.26	18.71	12.40	23.51	59.47	20.88	27.00	33.37

The most important performance criterion for predictive models is accuracy of the out of sample forecasts. That is, forecasting the data, never seen in the training period. The following table shows the MAPE values of the specified model. First of all, all-sample model could forecast the end price for 43 auctions reserved before, with a 7.64% prediction error. This value is very close to the performance of the unsupervised model. If the table is examined in detail, it can be seen that the clustering process significantly increases the price estimation performance. The average estimate error of the cluster models is a good ratio of 4.65%. The weighted error rate is 6.80% by weighting the errors with the number of observations in the clusters. In either case, it indicates that cluster models are more successful in price forecasts. In fact, this ratio is lower than the unsupervised model's results. In addition, standard deviation and maximum values of the errors are lower in cluster models. This means that the cluster models can largely minimize the errors and keep the prediction success rate high.

Table 5.20: OLS Out of Sample Performance for All Sample and Each Cluster

Out of Sample	AllSample	Cluster0	Cluster1	Cluster2	Cluster3	Cluster4	Average	Weighted Av
count	43.00	2.00	6.00	16.00	14.00	5.00	8.60	
mean	**7.64**	**0.95**	**4.08**	**8.41**	**7.50**	**5.32**	**5.25**	**6.80**
std	5.70	0.97	4.09	6.09	4.47	5.51	4.23	4.98
min	0.23	0.27	0.08	0.30	0.04	0.79	0.29	0.24
25.00%	2.39	0.61	0.93	3.90	4.93	2.67	2.61	3.52
50.00%	6.65	0.95	3.21	6.91	8.04	2.92	4.40	6.02
75.00%	11.90	1.30	6.08	11.04	9.75	5.52	6.74	8.83
max	21.39	1.64	10.74	19.29	15.03	14.70	12.28	15.36

5.3 Image Classification For Auction Prices

The photos of the products can express the meanings that thousands of words cannot. Therefore, in the current literature, price estimation models by the product photographs have been developed. These studies generally can be divided into two types. The first one is the models developed by using the direct photographs of the products, and the second one is the models based on the time series product price graphs. Technically, estimation models uses image classification techniques. In our age, thousands of products are sold through e-commerce and millions of product photographs are created in the heavy internet traffic. While the product information is usually written by the vendors, the information must be checked and categorized by the electronic trading companies. Zahavy et. al (2018) mentioned the need to develop product classification methods and they proposed a decision level fusion approach for multi-modal product classification. In the mean time, Chen et. al (2018) have suggested that product image classifications could be used for product price estimation and they prepared an algorithm using bicycle and car photographs. The authors trained several deep CNNs using both transfer learning and their architectures, for both regression and classification and found that with deep CNNs the model significantly outperformed others in a variety of metrics. Similarly, Maurya (2016) proposed an approach classifying the products according to their photographs (eg, towels, shoes) and developed a price estimation model according to the average of that class. In addition to these, in some studies such as Limsombunchai (2004), You et al. (2017), Bency et al. (2017, and Law and Russel (2018) house price prediction models have been developed with satellite image information of houses and streets.

On the other hand, some researchers found that the time series graphs of the product prices can deliver essential information about the price estimation and they proposed the models showing the price change over time in a graph and estimating future prices by image classification techniques. For example, Siripurapu (2014) explored a particular application of CNNs: namely, using convolutional networks to predict movements in stock prices from a picture of a time series of past price fluctuations, with the ultimate goal of using them to buy and sell shares of stock in order to make a profit. Similarly, Akhtar (2018) developed a price prediction model using bitcoin time series price figures and Groß, et al (2017) showed how multivariate time series can be interpreted as space-time pictures, thus expanding the applicability of the tricks-of-the-trade for CNNs to this important domain. To conclude, we can infer that the images of auction price paths can provide useful information about end price prediction. Nevertheless, there is no such study as far as our literature review and we aim to fill this gap by this study.

In the previous studies, the price models for all-sample and clusters have been specified by using many features gathered from auction and BIN sales. However, accurately extracting many features and adding them to the models incur some risks and costs. To exemplify, there is always a risk that some of the related features can be missing in the sales information. In this case, either it is necessary to estimate the missing data with some assumptions or to remove that sales information completely from the dataset. In other words, the modeler has to make a trade off: either losing meaningful sales information or making an incorrect assumption. In addition, extracting data from the database, cleaning and editing have the cost of time and effort. To remove the cost and risks mentioned above, this chapter, will propose a method which is illustrated below to estimate the auction prices by using simpler and less diverse data.

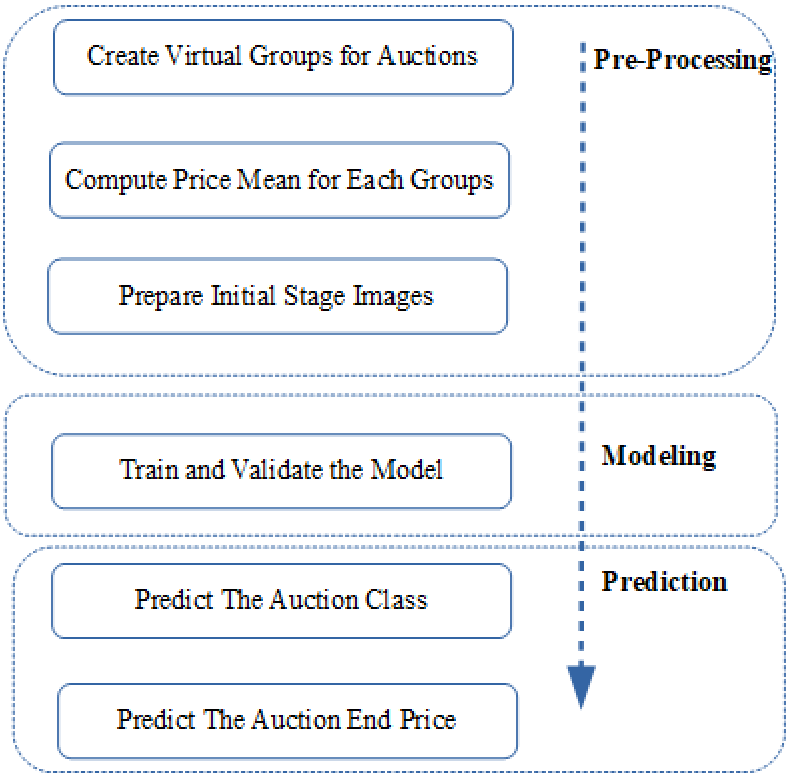

Create Virtual Groups for Auctions
Compute Price Mean for Each Groups
Prepare Initial Stage Images
Pre-Processing
Train and Validate the Model
Modeling
Predict The Auction Class
Predict The Auction End Price
Prediction

Figure 5.25: Process Map for Price Prediction by Image Classification

The graphical representations of the auctions contain essential information about how the auction will close. As mentioned earlier, auctions can be divided into three stages according to the distribution of bids within the bidding period. These are the initial stage, intermediate stage, and final stage. Initial and final stages are very intense in bids i.e. “auction fever” is high. When it comes to the final stage, there is not enough time to make a price estimate and take an action correspondingly at this point. In other words, as soon as the initial stage ends, one needs to estimate the closing price with a sufficient amount of time to make strategic decisions. The research study will use less data for the price model in this section. In other words, the study will not include the information defined as static and economic features in the model. I will use only the in-auction variables which are called dynamic features.

The purpose of using a small variety of data is to be able to show the data on the chart and use it in image classification models.

The figure below shows the distribution of the bids given in the 22nd auction by time. The horizontal axis on the figure shows the bidding period. 0 represents the time when the auction commences, and 1 represents the time when it ends. This auction was also realized in 3 stages. As soon as the auction commences, a lot of bids have been submitted and the price has increased rapidly but similar to a concave function. In the interim period, few offers were submitted and the price change was almost non-existent. When the auction entered to the last period, it was bid intensively causing a price increase in a convex manner resulting the auction to close with a high price. By looking at the intensity of the bids, it can be claimed that the competition in this type of auction is very high and the bids follow an auction path similar to the horizontal inverse S path.

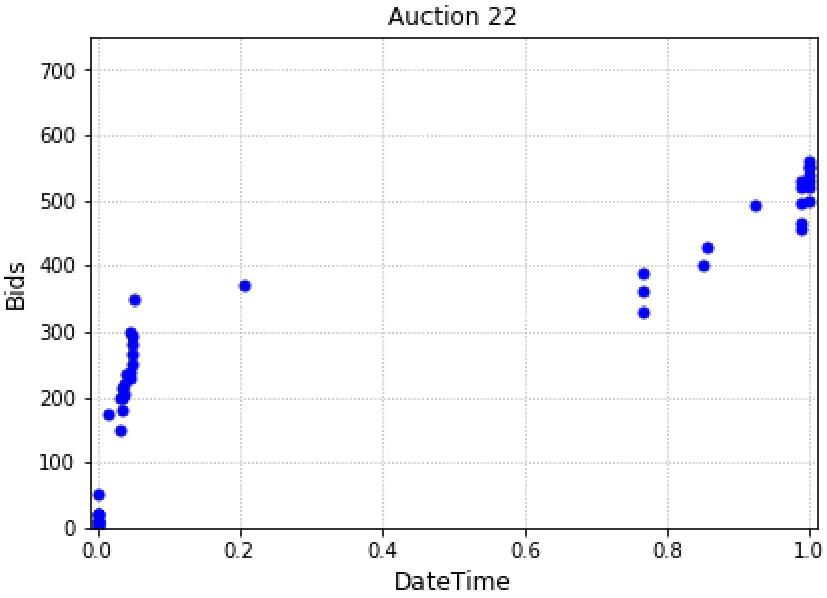


Figure 5.26: Bids During Auction 22 by Time

5.3.1 Information Aggregation for Auctions

Each bid in the initial stage can give some signals about the value of the product, possible competition and the end price of the auction. Therefore, the initial period of the auction can be called as an "information aggregation" stage. The bidders can use the information and the data collected at this stage in their pricing strategies. I have included this aggregate information in the previous models using the nominal values of data. On the other hand, the data can also be illustrated on a graph and image classification models can be utilized. In other words, the developed classification model can be used to forecast the closing price of the auction by giving it the image of first stage information. Actually, this can be compared to the process of an expert guessing the artist after seeing a piece of an art work. In order to achieve high success, the expert must be experienced enough for each artist and has to see a lot of her works.

To give a little bit detail, the minimum auction end price in the dataset is \$276. Considering this, it can be assumed that the bids, which have passed the minimum price as soon as the auction starts, will provide an important signal about the value of the product. For this reason, bids which are above a certain value (here \$300) are pointed in red color. The following graph shows the bids given in the first stage of the "Auction22". A lot of inferences can be drawn from this graph. First of all, the auction starts at \$0 and continues with very small bids, but in a short time the price increases to over \$400 with high and intense bids. By looking at the graph carefully, the majority of the blue points can indicate that the bids and competition in the auction will be high in the future. The high number of red dots indicates that the product has a high rating. The image classification method that will developed will take into account the intensity of the blue and red dots. Taking into account the path followed, and it will forecast the final price of the auction by estimating the value of the product and the possible competition in the auction.

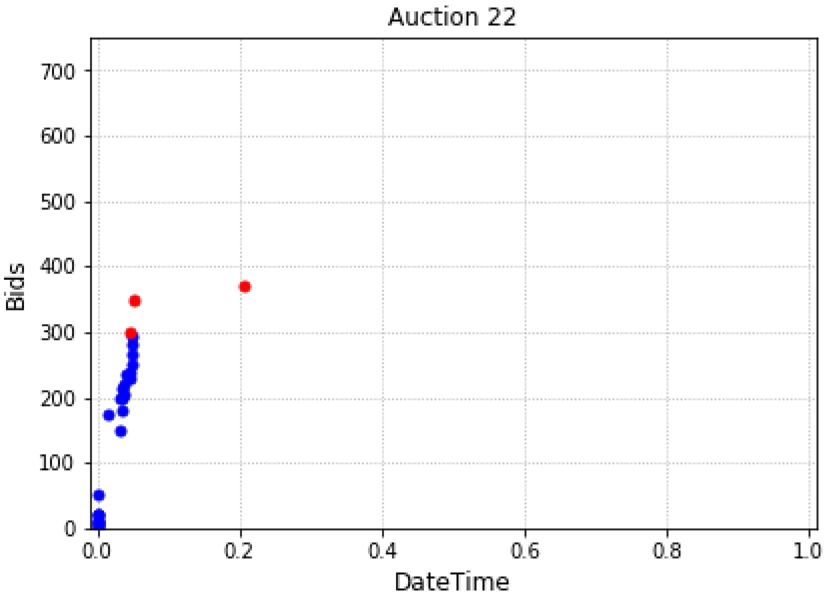


Figure 5.27: Bids at the Beginning of Auction 22

5.3.2 Image Processing and Image Classification

The classification models will be trained by processing the initial auction images as mentioned above. There will just be axis, blue and red points on the images to process. There is need for another information for classification purposes, the labels. The dataset used in the analysis do not include any label to classify the initial stages of the auctions. For this reason, the auction data needs to be grouped virtually for the training process of the model. As conducted in the previous section, the virtual auction groups can be set according to the price ranges between the minimum and maximum end prices. The following table shows the price ranges created for virtual group labels. For example, the products have been labeled as Cluster0 with an auction end price between $230- $450. Other products were also classified according to the cluster in which the end price was included.

Table 5.21: Virtual Price Ranges and Cluster Labels

Label Name	Min Price	Max Price
Cluster0	$230	$450
Cluster1	$450	$490
Cluster2	$490	$525
Cluster3	$525	$750

First, some pre-processing operations were carried out for images. The first of these is the virtual grouping process described above. Secondly, auction figures up to 20% of the bidding time were cut and saved in a folder by the auction number. In these graphs, only the blue and red bidding information and the axes of the graphics are used. Apart from this information, no markings, a name or any other information is left in the graphs. The purpose of this exercise is to ensure that the model to be trained to take into account the red dots representing the valuation, blue dots representing the competition. The proposed model will classify the auctions which were seen in neither training nor validation period for the out of sample performance prediction. The average end price calculated for the training and validation period of that class will be the price estimate of the auction. There are several image processing methods and one of the most prominent methods, convolutional neural network approach was used as a model in this study, which will be elaborated in the following sections.

5.3.3 Convolutional Neural Network

Artificial Neural Network (ANN) is a machine learning method inspired by the human brain and nervous system. It is frequently used in both industry and literature because it provides successful results in natural language processing and image processing. Artificial Neural Network consists of neurons and layers. The first layer is called the input layer and the last one is the output layer. There may be hidden layers between them. Single layer networks are called the “perceptron”. In the following figure, the general structure of a perceptron is shown which can be used to solve simple problems.

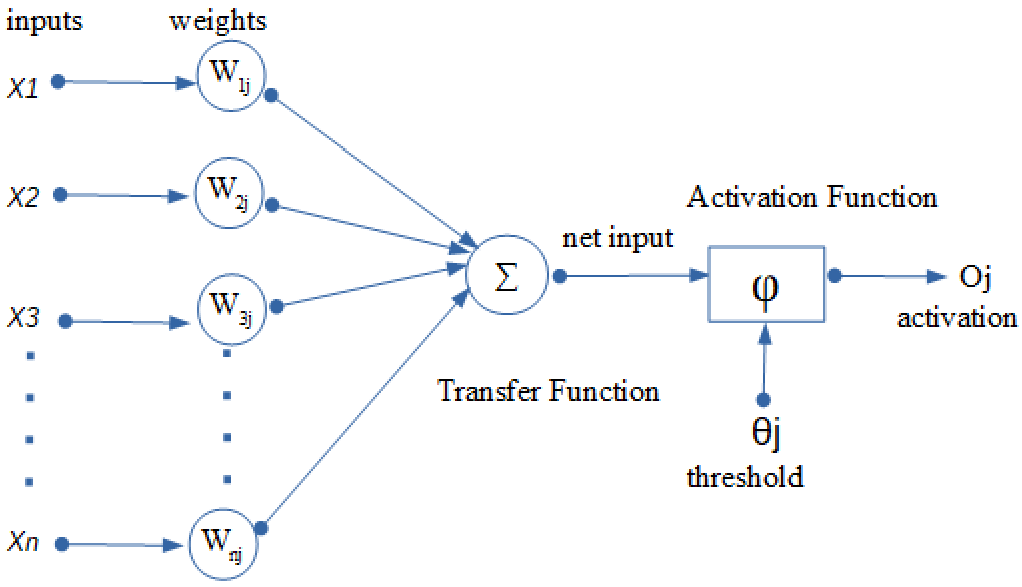


Figure 5.28: General Artificial Neural Network Structure

For more complex problems, thousands of perceptrons should be used. In this case, the neurons are connected to each other by synapse, i.e by weights of coefficients. Real neurons in the human nervous system do not react immediately to each signal and wait for a certain threshold to pass. Artificial neural network processing is also based on this logic. All incoming input values are converted to the sum of net inputs by means of a weight transfer function. The net input value is also evaluated by a function called the activation function. If the function value exceeds a certain threshold, an output from the neuron can be obtained. Otherwise, the neuron does not produce any output signal. To express it simply, weak signals cannot produce any output from the neuron. By this technique, the linear and nonlinear relationship between the input and output can be modeled by the activation function. Activation functions can be a simple step function (1 if input is greater than zero, -1 if it is less than zero) or slightly more complex sigmoid, Relu and Tanh functions.

The first and simplest type of artificial neural networks, which have multiple neurons and layers, is called feedforward neural network. This structure contains multiple neurons in layers which are weighted according to predetermined objective function. No calculation is performed on the input layer and the incoming data is transferred to the hidden layer. The hidden layer is an intermediate layer where the incoming input is calculated and sent to the output layer and is not connected to the outside world. The information layer is called the output layer. One-way information flow exists in this kind of learning. Although there is one layer of input and output layer, there can be many hidden layers. Another method is the Back Propagation Algorithm. This method learns from model errors and optimizes the weight coefficients. In other words, there can be a return from the output layer or hidden layers to the previous layers Thus, the information flow is twofold. Thus, in this type learning method, the current output takes into account the current and previous inputs.

ANN method can be applied in many areas, such as classification, clustering, estimation, and optimization operations. In order to use ANN within the scope of image classification, some image processing operations must be done first. Therefore, this process is also called Convolutional Neural Network (CNN). Basically, CNN uses standard Neural Network models to solve the classification problem, but it uses other layers to determine the required information and to extract certain features from the images. The basic steps in this process are shown below in the form of CNN general structure. These are the steps to take input, convolution, pooling, fully connection and the estimation of the output. The first 4 steps are called feature learning and the last one is classification.

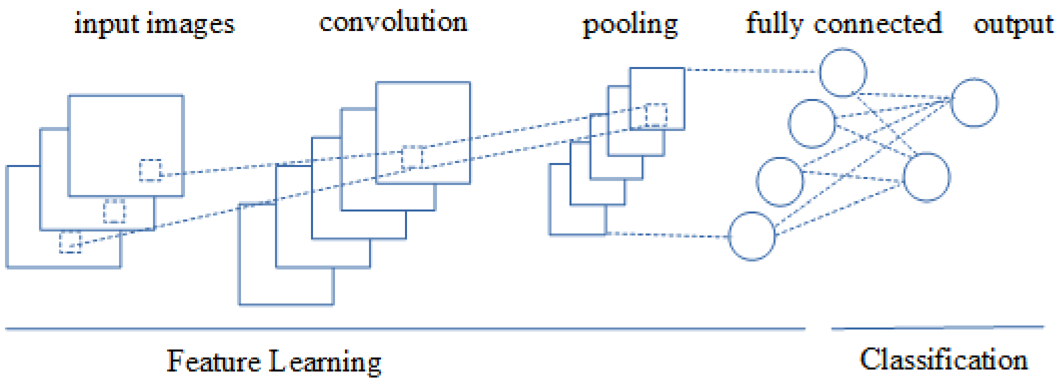


Figure 5.29: General Convolutional Neural Network Structure

First of all, the images in the database are uploaded to the system. “Convolutional Layer” is the main building block of CNN and is the stage in which the features are extracted. One or more filters are applied to remove the unnecessary properties of the image in this layer. Technically, the entire matrix structure of the image is converted by a filter matrix. To do this, dot product calculation is made and a smaller size of the output matrix is created. This matrix is also called the feature map since it shows the location of the desired feature in the photo. Next stage is pooling, also known as “downsampling”. The task of this layer is to reduce the processing time by decreasing the parameters in the neural network and the number of calculations. Filtering is done in a similar way, but the protection of the most important features is ensured. A method commonly used is the max pooling. In this way, the highest values of the feature matrix are selected so that less important features are filtered. The matrix structure is then made into a one-dimensional array, a structure in which the artificial neural network algorithm can be used. This process is also called flattening. By using data taken from this point, the classification model is developed with artificial neural network models in a fully connected phase. At this stage, if all the nodes in a layer are connected to the next layer, then it is fully connected. In the final stage, the classes are estimated with the developed model. For the purpose of image processing and classification “keras” and “tensorflow” modules were used in python. The data is divided into three sets namely train, validation and test. Most of the data was used for the training period. The model, in parallel with the previous studies, trained and validated with 4 classes to select the best weights for neural networks. In this process, the technical inputs were defined as follows: epochs are 20, batch size is 16, sample per epoch is 1500 and the validation step is 400.

5.3.4 Auction Price Prediction with Image Classification

The following tables show the model's training, validation and out of sample performance according to MAPE statistics, respectively. The developed model produced its best performance in the training period. When we look at the MAPE weighted by the number of observations in the clusters, the error margin of the model is calculated as 9.53%. This value is lower than the ones in the validation and out of sample performance. These values are not better than the performance of the models developed in the previous sections. However, it is a very serious achievement to estimate the end-price of the auction only by interpreting the information in the initial stage graph without using any economic or static features that affect the price of the auction directly. This shows us that the information aggregated in the first period of the auction can give crucial insights about the auction result. In other words, the bids submitted in this period can give serious clues about the value of the product and the future competition that will be in the auction.

Table 5.22: Price Performance in Training for Image Classification

	Cluster0	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster Mean**	**Weighted Average**
count	44.00	42.00	72.00	73.00	57.75	
mean	17.79	10.92	6.54	6.68	10.48	9.53
std	12.73	6.98	6.03	6.11	7.96	7.51
min	0.67	0.15	0.07	0.40	0.32	0.30
25.00%	6.17	5.97	2.18	2.32	4.16	3.67
50.00%	17.69	9.87	3.04	5.80	9.10	7.95
75.00%	25.60	16.94	10.34	8.18	15.27	13.77
max	52.92	23.68	21.40	29.53	31.88	30.39

Table 5.23: Price Performance in Validation for Image Classification

	Cluster0	Cluster1	Cluster2	Cluster3	Cluster Mean	Weighted Average
count	29.00	45.00	45.00	45.00	41.00	
mean	15.26	13.82	8.07	8.48	11.41	11.03
std	12.41	5.80	6.36	8.41	8.24	7.84
min	0.31	1.93	0.26	0.04	0.64	0.67
25.00%	6.04	9.39	2.18	2.32	4.98	4.88
50.00%	13.04	15.72	8.22	6.18	10.79	10.57
75.00%	22.57	18.23	13.00	12.30	16.52	15.93
max	50.07	23.41	22.15	38.07	33.42	31.80

Table 5.24: Out of Sample Price Performance for Image Classification

	Cluster0	Cluster1	Cluster2	Cluster3	Cluster Mean	Weighted Average
count	16.00	12.00	11.00	10.00	12.25	
mean	17.57	14.51	7.43	6.67	11.55	12.32
std	14.97	5.43	5.28	5.83	7.88	8.59
min	0.31	7.09	0.26	2.13	2.45	2.33
25.00%	4.00	10.28	2.99	4.39	5.41	5.39
50.00%	15.54	12.96	8.22	5.19	10.48	11.15
75.00%	29.44	18.48	10.34	6.09	16.09	17.70
max	47.44	23.68	17.43	22.91	27.86	29.88

5.3.5 Image Classification with Auction Path Models

In the previous section, the price ranges were used for the image classification labels. In this study, the auction path information will be used as a label. At the beginning of the dissertation, we mentioned Auction Path Models (APM) which are the paths that bids follow in general and their performances to estimate the auction closing prices. Since most of the auctions are composed of 3 stages such as initial, interim and final, the piecewise linear function of 3rd degree and polynomial 3rd degree functions can be the best match to the auction paths. In addition, some of auction paths are grouped in order to reach a sufficient number of observations for each cluster. This is not exactly a virtual grouping since the end price prediction performance of the auction path is used. Yet, it might be called a semi-virtual grouping. We have labeled the auctions which can be predicted best (according to price prediction error (PPE) performance) by linear, sigmoid functions and only two-point auctions as Cluster0. The auctions including multiple bids in which the piecewise linear 2nd degree function achieved best PPE was classified as Cluster1. Piecewise linear function 3rd degree model was classified as Cluster2. Finally, the auctions in which the highest performance is achieved by a polynomial 2nd degree or 3rd degree were labeled as Cluster3. Thus, the whole dataset was divided into 4 classes in total and each auction was labeled. Due to the fact that most of the auction structures are composed of 3 stages, the much of the observations exist in Cluster2 and Cluster3.

In this study, the dataset is divided into three sections as a training, validation, and testing set. CNN image classification model was developed during the training and validation period by using the graphs created for the initial stage of the auctions. Then the model was tested for the out of sample performance. In this process, firstly, the cluster of the input image is estimated and auction end price forecast is determined by the average price of that cluster.

The following tables show the performance of the classifier in the training, validation and testing periods. In the first table, the error rate in terms of MAPE in the training period is 8.97%. When it is weighted by the number of observations in the cluster, the error rate becomes 9.02%. These values are better than the performance of the model we developed previously. In the validation and testing period, the error rate of the developed model has slightly increased but is calculated very close to the result of the previous study. As a result, although the price prediction performance of the developed model is not very high, it has been seen once again that the predetermined auction paths labels and only the information obtained from the graph formed in the auction initial stage provide very useful information to estimate the auction price.

Table 5.25: Performance in Training for Image Classification with APM

	Cluster0	Cluster1	Cluster2	Cluster3	Cluster Mean	Weighted Average
count	45.00	50.00	111.00	63.00	67.25	
mean	8.53	9.69	9.28	8.40	8.97	9.02
std	6.50	9.62	10.34	8.92	8.84	9.23
min	1.35	0.43	0.34	0.02	0.54	0.45
25.00%	4.48	3.03	2.83	1.96	3.07	2.94
50.00%	7.00	6.08	7.15	6.68	6.73	6.82
75.00%	9.79	12.02	11.37	9.95	10.78	10.89
max	31.62	42.34	83.60	42.86	50.10	57.69

Table 5.26: Performance in Validation for Image Classification with APM

	Cluster0	Cluster1	Cluster2	Cluster3	Cluster Mean	Weighted Average
count	20.00	25.00	50.00	31.00	31.50	
mean	13.33	10.95	10.32	11.53	11.53	11.22
std	10.08	8.25	9.59	12.89	10.21	10.22
min	1.35	0.14	0.18	0.43	0.52	0.42
25.00%	4.24	4.39	3.48	3.48	3.90	3.78
50.00%	12.61	10.16	6.08	5.57	8.60	7.80
75.00%	16.90	15.55	13.62	11.99	14.51	14.12
max	40.82	36.70	44.78	53.10	43.85	44.60

Table 5.27: Out of Sample Performance for Image Classification with APM

	Cluster0	Cluster1	Cluster2	Cluster3	Cluster Mean	Weighted Average
count	14.00	7.00	21.00	7.00	12.25	
mean	14.19	18.30	8.02	15.56	14.02	12.33
std	11.05	15.10	7.68	16.65	12.62	10.98
min	0.64	2.37	0.64	0.64	1.07	0.89
25.00%	3.98	9.72	3.76	5.96	5.86	4.99
50.00%	11.99	14.62	6.16	8.03	10.20	9.30
75.00%	20.30	22.39	9.29	21.75	18.43	16.08
max	38.83	46.88	32.48	44.78	40.74	38.11

CHAPTER 6

COMPARISON OF STUDIES

The in-sample and out-of-sample performances of the models developed within the scope of this dissertation are summarized in the tables below.

Table 6.1: In Sample Price Prediction Performance Comparison

	Unsupervised		**Supervised**		**Image Train**	**Image APM Train**
	AllSample	**Cluster Weighted Av.**	**AllSample**	**Cluster Weighted Av.**	**Cluster Weighted Av.**	**Cluster Weighted Av.**
count	380.00	95.00	380.00	76.00	58.00	67.00
mean	7.64	6.94	7.57	**6.68**	9.53	9.02
std	6.84	6.20	6.86	5.82	7.51	9.23
min	0.00	0.03	0.01	0.02	0.30	0.45
25.00%	2.93	2.48	2.85	2.38	3.67	2.94
50.00%	5.91	5.65	6.14	5.32	7.95	6.82
75.00%	10.56	9.62	10.08	9.71	13.77	10.89
max	61.49	41.64	63.26	33.37	30.39	57.69

According to in-sample prediction performance, the supervised model is the best, the unsupervised model is the second and the third is image classification method with APM. Clustering approach has increased performance of the models. For the supervised model, clustering process has increased sample prediction performance by 1%.

Table 6.2: Out of Sample Price Prediction Performance Comparison

	Unsupervised		Supervised		Image	Image APM
	AllSample	**Cluster Weigthted Av.**	**AllSample**	**Cluster Weigthted Av.**	**Cluster Weigthted Av.**	**Cluster Weigthted Av.**
count	43.00	10.75	43.00	8.60	12.25	12.25
mean	7.59	6.91	7.64	**6.80**	11.55	12.33
std	5.55	4.51	5.70	4.98	7.88	10.98
min	0.23	0.68	0.23	0.24	2.33	0.89
25.00%	2.43	3.17	2.39	3.52	5.39	4.99
50.00%	7.58	7.02	6.65	6.02	11.15	9.30
75.00%	11.93	9.45	11.90	8.83	17.70	16.08
max	21.46	14.74	21.39	15.36	29.88	38.11

As to out-of-sample forecast performance, the supervised model is the best again and unsupervised model is the second and the third is the image classification method. Clustering approach again contributes well to the performance of the models by fundamentally reducing the maximum level of forecast error and variance.

CHAPTER 7

CONCLUSIONS

In this dissertation, the structure of fixed-time auctions and the factors affecting the auction end price were examined. In this context, firstly, the auction paths followed by the bids and the relationship between the BIN sales and the auctions were analyzed. In addition to the static, dynamic and economic features, product descriptions have been incorporated in the auction price models by the virtue of machine learning algorithms. Finally, image classification models that can estimate the auction end price from the graphs in the initial stage of the auctions were developed. Fundamental findings are summarized briefly as follows below.

7.1 Auction Paths

It can be concluded that auction paths that the bids follow can be determined best by either piecewise 3rd degree or polynomial 3rd degree functional models by curve fitting applications. Together with this conclusion, it can be claimed that there are 3 phases or stages for the auctions. The initial stage, the middle stage and the final stage of the auction. From the data analysis section, it is observed that there are lots of bids at the initial and final stages but this is not typically the case for the middle stage: There are a few bids in the middle stage. Then, one can think that buyers evaluate the product at the beginning of the auction and bid several times at the initial stage. However, they do not bid at the middle stage to refrain from the competition and do not open up their private valuation. Therefore one can focus on the initial stage of the auctions and retrieve valuable information to forecast the end price. This study proves that the initial stage demonstrates essential inputs for the price models.

7.2 Average Auction Price

There is a high positive correlation between average auction prices with the auction price lags (APL) and BIN prices. As time passes, the effect of the previous auction and BIN prices on the auctions also increases. Moreover, the impact of the BIN prices is higher than the impact of APL on average auction prices. This analysis proves that, both BIN and previous auction prices are important variables to in auction pricing models.

7.3 Unsupervised Model

APL and BIN price have a positive relationship with the auction price. Accessories sold with the iPhone brings about approximately $20 in addition to the auction price. It means that if an extra item such as case, screen protector, charger etc. are included in sales, it might increase the auction price by $20 on the average. From this, it can be inferred that if the buyer needs only the iPhone, then s/he should focus on the auctions that do not include accessories. Also, one can infer from this conclusion that, the seller can add accessories with the product to increase the price of the auction.

Although total seller rate has a significant impact, the coefficient is close to zero. Contrary to the findings of Lackes et al. (2013) and Li (2017), the total seller rate does not have a substantial impact on the auction price. At the same time, it has been determined that the number of negative ratings of the seller has a significant and negative effect on the prices. Additional one negative rating that the seller receives causes more than a price drop of $6 in the auction for that product. This is a solid and compatible result with literature by Melnik and Alm (2002), Bajari and Hortacsu (2003), Cabral and Hortacsu (2010). We might infer that the sellers should avoid getting a negative rating from the buyers and they could be advised to increase or at least try to keep customer satisfaction at some level. The number of views of the seller profile is positive but has a lower impact on auction price.

Bajari and Hortacsu (2003) did just cover the “blemishes” on the products in their price model and they found that blemishes decrease auction price but they did not incorporate a product condition rating for multiple conditions. They just used dummy variables for blemishes only. The effects of the product descriptions written for the products have been analyzed for the first time in literature by this research. Product description has valuable information for the product and it has an important and positive effect on the prices. Expressions such as “excellent”, “new”, “original” are crucial words and have positive outcomes on the auction price of the products and they have the biggest coefficient in the price model. Expressions such as “scratches”, “dings”, “wear” are very important and have the lowest coefficient in the auction price model in similar way to the literature. In addition we have found that, inexperienced bidders do not internalize product descriptions well and they do not refrain from bidding to the worn products. This analysis shows that the phone condition is a crucial feature and must be included in the price models.

The auction duration exists in the price model as well. The effect of the auction duration is positive but it is very interesting. If the auction period is 10 days, the duration coefficient is very low. This is contrary to the result of the research done by Lucking-Reiley et al. (2007). The authors mentioned that as the auction length increases the auction price also increases. But this is not the case in our study. In other words, planning a long auction may not be a good option for the vendors. The duration coefficient is the greatest for 5-day auctions. It can be concluded that the optimum auction length might be 5 days for this type of product and market. If the auction length lower or higher than 5 days, the coefficient of duration variables decreases. Then, one can propose sellers that, it is optimal to plan 5-day auctions for maximizing auction price and propose patient buyers to buy from 10-day auctions since the duration coefficient is the lowest for 10 days.

Various papers studied the effects of auction end day such as Bajari and Hortacsu (2003), Hou (2007). On the other hand auction start day is disregarded. We found that, there is a positive and significant effect of the day auction begins. Buyers notice the products on the first days of auction and they follow the auction quietly by and large. The price coefficient of an auction that starts on Sunday is high. Since Sunday is a non-working day for many buyers, an auction is more likely to draw followers’ attention. Again one can also propose sellers to start the auction at weekends for price maximization, on the other hand, one can propose buyers to focus on auctions started on Wednesday or Thursday.

In literature, bidder experience was not studied much. There are a few researches namely Bajari and Hortacsu (2003) and Houser and Wooders (2006) and they found that bidder experience was not statistically significant. Nevertheless, it is found in this research that, information retrieved from the initial stage of the auction, initial bids, bidders and bidder ratings (experience) have a positive and significant effect on the end price. In addition to this, when the bids given in the initial stage of the auction are included in the price models, starting price becomes insignificant.

Similar to the studies Kaur et al. (2012) and Kaur et al. (2014), We have found that, clustering auctions by certain characteristics increases the performance of the in-sample predictions and the out of sample price forecast performance. Moreover, the auction clustering decreases the maximum prediction error for the forecast period. That is, clustering is also a helpful methodology for the price prediction analysis.

7.4 Supervised Model

All of the above results and implications are also valid for the supervised models. In fact, the supervised model outperforms the unsupervised one.

Clustering has been carried out from the product descriptions under BIN prices and it is observed that product description has a positive and significant effect on the auction prices. Similarly, descriptions including such as “excellent shape” have the greatest effect on price and the effects of words such as “scratches”, “dings”, “dents” are the least. From this section of the study, BIN prices are found to have valuable product information to predict the auction prices.

7.5 Image Classification

Bids, number of bidders and intensity of the bids are important signals at the initial stage of auctions. This information can be pictured by a graph and a price model can be developed with these figures by image classification tools. Although their performance is slightly lower than the previous models, the pricing models by image classification can be used for a quick and preliminary evaluation for the auctions.

7.6 Future Studies

Online auction prices are a very fertile area for research and many new studies can be conducted. Models can be generalized by specifying additional price models for other products. Bids values can be weighted by bidder rates as a signal for experience and valuation and the new models can be specified. In addition, the auctions can actually be defined in a game theoretic framework and strategic decisions can be included in price models of Machine Learning applications.

REFERENCES

Akhtar, A. (2018). Machine learning for market trend prediction in bitcoin. On Internet: *http://a.web.umkc.edu/aa95b/doc7.pdf*

Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268*),* 765-769.

Ariely, D., Ockenfels, A., & Roth, A. E. (2005). An experimental analysis of ending rules in internet auctions. *RAND Journal of Economics,* 890-907.

Athey, S. (2017). Beyond prediction: using big data for policy problems. *Science,* 355(6324), 483-485.

Athey, S., & Imbens, G. (2015). Machine learning methods for causal effects. On Internet: *www. nasonline. org/programs/sackler-colloquia/documents/athey. Pdf.*

Bajari, P., & Hortacsu, A. (2003). The winner's curse, reserve prices, and endogenous entry: Empirical insights from eBay auctions. *RAND Journal of Economics,* 329- 355.

Bajari, P., & Hortacsu, A. (2004). Economic insights from internet auctions. *Journal of Economic Literature,* 42(2), 457-486.

Bajari, P., & Hortacsu, A. (2005). Are structural estimates of auction models reasonable? evidence from experimental data. *Journal of Political Economy,* 113(4), 703-741.

Bency, A. J., Rallapalli, S., Ganti, R. K., Srivatsa, M., & Manjunath, B. S. (2017). Beyond spatial auto-regressive models: predicting housing prices with satellite imagery. *In Applications of Computer Vision (WACV), 2017 IEEE Winter Conference* (pp. 320-329). IEEE.

Benkard, C. L., & Bajari, P. (2005). Hedonic price indexes with unobserved product characteristics, and application to personal computers. *Journal of Business & Economic Statistics,* 23(1), 61-75.

Bishop, C. M. (2006). Pattern recognition and machine learning. *SpringerLink*, Chapter 4.3.4

Bose, S., & Daripa, A. (2017). Shills and snipes. *Games and Economic Behavior,* 104, 507-516.

Breiman, L. (1998). Arcing classifier. *The Annals of Statistics,* 26(3), 801-849.

Breusch, T. S. (1978). Testing for autocorrelation in dynamic linear models. *Australian Economic Papers,* 17(31), 334-355.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society,* 1287-1294.

Cabral, L., & Hortacsu, A. (2010). The dynamics of seller reputation: evidence from eBay. *The Journal of Industrial Economics,* 58(1), 54-78.

Chan, N. H., & Liu, W. W. (2017). Modeling and forecasting online auction prices: a semi-parametric regression analysis. *Journal of Forecasting,* 36(2), 156-164.

Chen, S., Chou, E., & Yang, R. R. (2018). The price is right: predicting prices with product images. *arXiv preprint arXiv:* 1803.11227.

Chow, V. (2017). Predicting auction price of vehicle license plate with deep recurrent neural network. *arXiv preprint arXiv:* 1701.08711.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning,* 20(3), 273-297.

D’Agostino, R. & Pearson, E. S. (1973). Tests for departure from normality. *Biometrika,* 60, 613-622

Einav, L., Farronato, C., Levin, J., & Sundaresan, N. (2018). Auctions versus posted prices in online markets. *Journal of Political Economy,* 126(1), 178-215.

Einav, L., Kuchler, T., Levin, J., & Sundaresan, N. (2015). Assessing sale strategies in online markets using matched listings. *American Economic Journal: Microeconomics,* 7(2), 215-47.

Frongillo, R. M. (2015). Machine learning and microeconomics. *Transactions on Embedded Computing Systems,* 9(4).

Ghani, R., & Simmons, H. (2004). Predicting the end-price of online auctions. *In International workshop on data mining and adaptive modeling methods for economics and management.*

Godfrey, L. G. (1978). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica: Journal of the Econometric Society,* 1303-1310.

Groß, W., Lange, S., Bödecker, J., & Blum, M. (2017). Predicting time series with space-time convolutional and recurrent neural networks. *Proc. of the 25th ESANN,* 71-76.

Guyon, I., Boser, B., & Vapnik, V. (1993). Automatic capacity tuning of very large VC-dimension classifiers. *In Advances in Neural Information Processing Systems* (pp. 147-155).

Grossman, J. (2013). Predicting-ebay-auction-sales-with-machine-learning. *Working Paper, Published on Web.Research* *http://jaygrossman.com/post/2013/06/10/ Predicting-eBay-Auction-Sales-with-Machine-Learning.aspx*

Gupta, R., & Pathak, C. (2014). A machine learning framework for predicting purchase by online customers based on dynamic pricing. *Procedia Computer Science,* 36, 599-605.

Hortacsu, A., Jerez, A., M., & Douglas, J. (2009). The geography of trade on eBay and mercado libre. *American Economic Journal: Microeconomics,* v. 1, no.1, February 2009, p. 53-74.

Hou, J. (2007). Price determinants in online auctions: a comparative study of eBay China and US. *Journal of Electronic Commerce Research,* 8(3).

Houser D, Wooders J. (2006). Reputation in auctions: theory, and evidence from eBay. *Journal of Economics and Management Strategy* 15: 354–369.

Jank, W., Shmueli, G., & Zhang, S. (2010). A flexible model for estimating price dynamics in on-line auctions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(5), 781-804.

Johnson, S., C. (1967). Hierarchical clustering schemes. *Psychometrika*, 2:241-254.

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis *(Vol. 344). John Wiley & Sons.*

Kaur, P. (2014). Development of automated dynamic bidding agents for final price prediction in online auctions *(Doctoral Dissertation)*.

Kaur, P., Goyal, M. L., & Lu, J. (2017). A comparison of bidding strategies for online auctions using fuzzy reasoning and negotiation decision functions. *IEEE Trans. Fuzzy Systems,* 25(2), 425-438.

Kaur, P., Goyal, M., & Lu, J. (2011). Pricing analysis in online auctions using clustering and regression tree approach. *In International Workshop on Agents and Data Mining Interaction(pp. 248-257). Springer, Berlin, Heidelberg.*

Kaur, P., Goyal, M., & Lu, J. (2012a). Price forecasting using dynamic assessment of market conditions and agent's bidding behavior. *In International Conference on neural Information Processing (pp. 100-108). Springer, Berlin, Heidelberg.*

Kaur, P., Goyal, M., & Lu, J. (2012b). An integrated model for a price forecasting agent in online auctions. *Journal of Internet Commerce,* 11(3), 208-225.

Kaur, P., Goyal, M., & Lu, J. (2014). A price prediction model for online auctions using fuzzy reasoning techniques. *In Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on (pp. 1311-1318).*

Kolmogorov, A. (1933). Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.,* 4, 83-91.

Kumar, S., & Rishi, R. (2017). Hybrid dynamic price prediction model in online auctions. *International Journal of Applied Engineering Research,* 12(5), 598-604.

Lackes, R.,Börgermann, C.,& Frank, E. (2013). Which price you will get in your eBay-auction a data mining analysis to identify price determining factors in online-auctions. *International Journal of Computer Science and Electronics Engineering (IJCSEE) Volume 1, Issue* 4 (2013) ISSN 2320-401X; EISSN 2320- 4028

Law, S., Paige, B., & Russell, C. (2018). Take a look around: using street view and satellite images to estimate house prices. *arXiv preprint arXiv:1807.07155.*

Lawrence, R. D. (2003). A machine-learning approach to optimal bid pricing. In Computational modeling and problem solving in the networked world. (pp. 97- 118). *Springer, Boston, MA.*

Larue, B., Jeddy, M., & Pouliot, S. (2013). On the number of bidders and auction performance: when more means less. *Working paper*

Levin, J. (2004). Auction theory. On Internet *www. stanford. edu/jdlevin/Econ, 20286.*

Li, Z. (2017). Effects of last-minute bidding behavior and seller reputation on online auctions. *Journal of Marketing Management,* 5(1), 12-20.

Limsombunchai, V. (2004). House price prediction: hedonic price model vs. artificial neural network. *In New Zealand Agricultural and Resource Economics Society Conference* (pp. 25-26).

Lin, C. S., Chou, S., Chen, C. H., Ho, T. R., & Hsieh, Y. C. (2006). A final price prediction model for online english auctions-a neuro fuzzy approach. *In JCIS.*

Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika,* 65(2), 297-303.

Lu, Y., Cohen, I., Zhou, X. S., & Tian, Q. (2007). Feature selection using principal feature analysis. *In Proceedings of the 15th ACM international conference on Multimedia* (pp. 301-304). ACM.

Lucking-Reiley, D., Bryan, D., Prasad, N., & Reeves, D. (2007). Pennies from eBay: The determinants of price in online auctions. *The Journal of Industrial Economics,* 55(2), 223-233.

Maurya, A. (2016). Clicktoprice: incorporating visual features of product images in price prediction. *In INFORMS*

Mary, S., Kolhe, L., N., & Pathari, R., D. (2014). A study on different methods and algorithms to predict end prices in online auctions. *International Journal of Engineering Research & Technology (IJERT)*

Melnik, M. I., & Alm, J. (2002). Does a seller's ecommerce reputation matter? evidence from eBay auctions. *The journal of Industrial Economics,* 50(3), 337- 349.

Milgrom, P., & Segal, I. (2002). Envelope theorems for arbitrary choice sets. *Econometrica,* 70(2), 583-601.

Milgrom, P. R., & Tadelis, S. (2018). How artificial intelligence and machine learning can impact market design (No. w24282). *National Bureau of Economic Research.*

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied Linear Statistical Models. (*Vol. 4, p. 318, pp.289). Chicago: Irwin.

Nicholson, D., & Paranjpe, R. (2013). A novel method for predicting the end-price of eBay auctions. *Research Paper, Stanford.*

O'Regan, R., T.,(2005). A look at the game theory of online auctions the choice between end- time formats on yahoo! Auctions. *Boston College Electronic Thesis or Dissertation*

Ockenfels, A., & Roth, A. E. (2006). Late and multiple bidding in second price Internet auctions: theory and evidence concerning different rules for ending an auction. *Games and Economic Behavior,* 55(2), 297-320.

Ockenfels, A., Reiley, D., & Sadrieh, A. (2006). Online auctions. *(No. w12785). National Bureau of Economic Research.*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research,* 12(Oct), 2825-2830.

Peeters, R., & Tenev, A. (2018). Number of bidders and the winner's curse, *University of Otago Economics Discussion Papers No.* 1802

Raykhel, I., & Ventura, D. (2009). Real-time automatic price prediction for eBay online trading. *In IAAI.*

Rezende, L. (2008). Econometrics of auctions by least squares. *Journal of Applied Econometrics,* 23(7), 925-948.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika,* 52(3/4), 591-611.

Siripurapu, A. (2014). Convolutional networks for stock trading. *Stanford University Department of Computer Science.*

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics,* 19(2), 279-281.

Storch, D. (2013). Towards an intelligent bidding agent in quibids penny auctions. *Research Paper Brown University Department of Computer Science.*

Tseng, K. K., Lin, R. F. Y., Zhou, H., Kurniajaya, K. J., & Li, Q. (2018). Price prediction of e-commerce products through Internet sentiment analysis. *Electronic Commerce Research,* 18(1), 65-88.

Utts, J. M. (1982). The rainbow test for lack of fit in regression. *Communications in Statistics-Theory and Methods,* 11(24), 2801-2815.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance,* 16(1), 8-37.

Wang, S., Jank, W., & Shmueli, G. (2004). Forecasting eBay's online auction prices using functional data analysis. *University of Maryland, College Park, MD, 20742.*

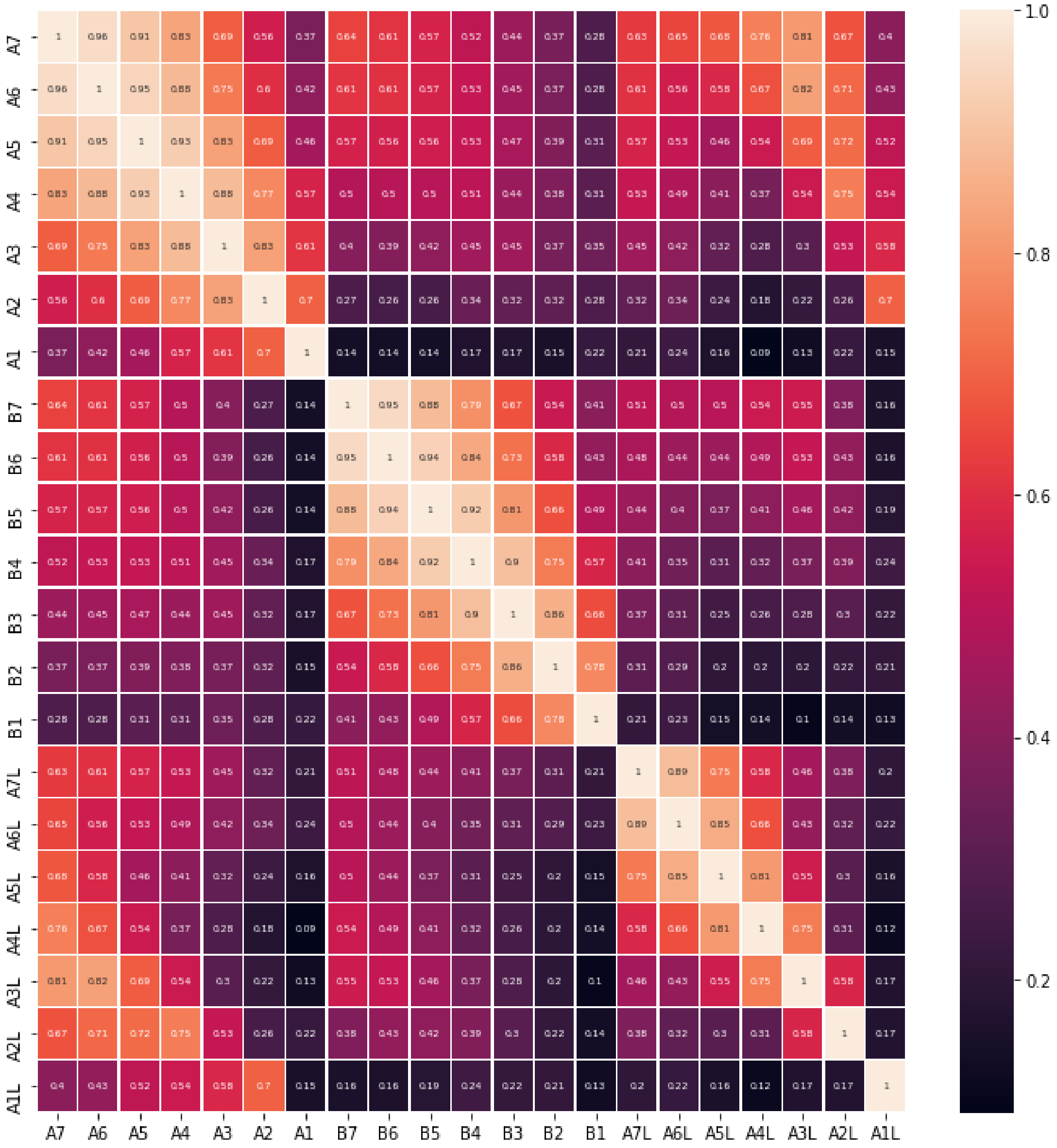
White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society,* 817-838.

You, Q., Pang, R., Cao, L., & Luo, J. (2017). Image-based appraisal of real estate properties. *IEEE Transactions on Multimedia,* 19(12), 2751-2759.

Zahavy, T., Krishnan, A., Magnani, A., & Mannor, S. (2018). Is a picture worth a thousand words? a deep multi-modal architecture for product classification in e-commerce. *In AAAI.*

Zhang, H. (2004). The optimality of naive bayes. *AA*, 1(2), 3.

APPENDIX A. AUCTION AND BIN PRICE DAY AVERAGE HEAT MAP



APPENDIX B. PERFORMANCE OF UNSUPERVISED MODEL BY KMEANS AND FSA

Table B.1 Performance of Unsupervised Model by KMeans with SelectKBest

Cluster	3		4		5		6		7		8		9	
Features	in	out	in	out	in	out	in	out	in	out	in	out	in	out
11	7.0	10.3	6.7	14.8	6.6	15.1	6.7	13.6	6.6	13.1	5.9	15.4	5.7	13.5
12	7.0	10.5	6.9	11.1	6.5	10.2	6.4	12.2	6.1	16.2	nan	nan	nan	23.3
13	6.7	9.8	6.7	10.1	6.1	10.8	6.0	10.6	5.8	11.9	nan	12.9	nan	10.9
14	6.7	9.8	6.7	9.5	6.7	10.4	6.6	11.2	6.3	14.2	6.0	13.7	5.6	11.0
15	6.6	11.3	7.0	10.1	6.5	10.2	6.2	10.4	5.6	11.6	5.8	12.9	4.8	18.2
16	7.2	9.5	6.4	11.4	6.1	10.1	5.8	16.1	5.7	14.5	5.2	17.2	5.2	16.0
17	7.1	8.6	6.8	9.8	6.4	10.4	6.1	10.6	6.0	nan	5.6	nan	5.2	nan
18	7.0	9.1	6.8	8.7	6.5	10.0	6.0	13.0	6.1	12.3	5.7	nan	5.3	11.1
19	7.0	9.1	6.8	8.7	6.4	10.7	5.9	13.2	6.1	12.6	5.7	12.5	5.4	nan
20	6.9	7.8	6.5	8.0	5.9	11.5	5.9	12.2	5.9	11.7	5.5	24.7	nan	30.2
21	6.9	9.7	6.6	9.2	6.4	10.6	6.0	11.6	5.8	nan	5.8	10.3	5.7	nan
22	7.0	8.4	6.7	9.3	6.4	11.6	6.4	nan	6.2	nan	5.7	nan	5.4	nan
23	7.0	8.4	6.6	8.6	6.3	9.3	6.4	10.3	6.0	9.5	5.4	nan	5.2	nan
24	6.8	8.1	6.7	9.3	6.5	8.6	6.4	9.8	6.0	nan	nan	nan	nan	nan
25	6.4	12.8	6.5	11.9	6.4	12.0	6.4	nan	5.7	nan	nan	nan	nan	nan
26	6.4	12.8	6.5	12.3	6.3	nan	5.8	nan	5.8	nan	nan	nan	nan	nan
27	6.4	12.8	6.5	12.3	6.3	nan	5.8	nan	5.3	nan	6.1	nan	5.8	nan
28	6.4	12.8	6.5	12.3	6.4	12.0	6.2	nan	5.8	nan	nan	nan	nan	nan
29	6.4	12.8	6.5	12.3	6.4	12.0	6.4	13.8	6.2	nan	5.7	nan	nan	nan
30	6.3	nan	6.1	nan	6.1	nan	6.0	nan	5.6	nan	nan	nan	5.4	nan
31	7.0	8.5	7.0	10.0	6.5	11.8	6.1	10.5	5.5	nan	nan	nan	nan	nan
32	7.0	8.5	7.0	10.0	6.4	12.6	6.1	10.4	5.4	nan	4.7	nan	nan	nan
33	7.0	8.5	6.7	7.5	6.5	11.7	6.1	10.5	5.5	nan	5.8	nan	5.7	nan
34	7.0	8.5	7.0	10.0	6.5	11.7	6.1	10.3	5.4	nan	5.1	nan	4.8	nan
35	6.9	10.8	6.9	11.0	6.4	10.2	6.3	13.8	6.1	13.5	nan	14.9	5.1	nan
36	6.9	10.8	6.7	8.3	6.4	10.2	6.2	9.9	6.0	14.0	5.6	nan	nan	nan
37	6.9	10.8	6.7	8.3	6.4	10.2	6.3	13.4	6.1	13.5	5.6	nan	nan	nan
38	6.9	8.1	6.5	14.0	6.6	9.3	6.4	9.1	6.3	13.3	5.2	nan	4.9	nan
39	6.9	8.1	6.5	14.0	6.6	9.3	6.4	9.1	6.3	13.3	5.3	nan	5.3	nan
40	6.9	10.8	6.5	11.3	6.5	11.6	5.6	nan	5.4	nan	5.4	14.8	5.5	nan
41	6.9	10.8	6.5	11.3	6.4	11.7	5.5	nan	5.5	nan	5.2	17.8	5.3	nan
42	6.9	8.1	6.4	7.6	6.6	8.6	6.2	11.0	5.9	10.3	5.7	14.7	5.4	14.4
43	6.9	8.1	6.4	7.6	6.6	8.6	6.2	10.8	5.9	10.3	5.7	14.0	5.4	14.4
44	7.0	8.5	6.6	9.2	6.6	12.4	6.2	12.9	6.1	12.9	5.8	14.9	5.6	14.7
45	7.0	8.5	6.6	9.2	6.6	12.4	6.2	12.9	6.0	11.6	5.8	14.7	5.6	14.7

Table B.2 Performance of Unsupervised Model by KMeans with RFE

Cluster	3		4		5		6		7		8		9	
Features	in	out	in	out	in	out	in	out	in	out	in	out	in	out
11	7.2	9.5	6.3	9.4	5.9	8.7	5.7	9.1	5.8	8.7	4.9	11.3	nan	nan
12	7.2	9.5	7.2	9.7	6.3	10.8	6.3	10.8	5.9	9.6	5.7	9.6	nan	nan
13	7.1	9.0	7.2	9.7	6.7	11.7	6.4	11.7	5.7	nan	nan	nan	nan	nan
14	6.8	9.0	6.7	9.2	6.1	10.4	6.2	10.7	6.2	12.7	5.0	nan	5.0	nan
15	6.7	9.5	6.7	8.9	6.4	9.5	6.3	10.0	6.0	11.8	5.4	nan	nan	nan
16	6.6	9.3	6.9	10.7	6.6	11.6	6.3	11.1	6.0	13.3	5.7	nan	5.3	nan
17	6.9	10.4	6.6	12.7	6.4	11.0	6.3	15.0	5.9	13.2	5.8	14.8	5.6	13.4
18	6.9	10.4	6.6	12.7	6.4	11.0	6.3	15.0	5.9	13.2	5.4	nan	5.1	nan
19	7.1	9.1	6.6	9.6	6.6	9.6	6.1	10.7	5.7	13.3	5.5	13.9	5.6	nan
20	6.9	10.4	7.0	11.4	6.6	11.8	6.5	13.1	6.1	12.5	5.9	nan	5.7	nan
21	7.1	9.1	6.6	13.5	6.7	9.1	6.3	16.1	6.1	14.0	5.7	12.5	5.3	nan
22	7.1	9.1	6.8	10.6	6.3	10.6	6.4	12.4	6.2	13.2	5.5	nan	5.4	12.4
23	7.1	9.1	6.7	9.7	6.7	11.6	6.4	13.5	6.3	13.1	5.5	17.0	5.4	9.9
24	7.1	9.1	6.7	9.7	6.7	11.6	6.4	10.7	6.0	10.8	5.2	14.4	4.6	16.0
25	7.1	9.2	7.1	9.5	6.7	9.1	6.3	10.7	6.3	11.1	5.5	nan	5.4	nan
26	6.9	10.4	6.8	10.5	6.6	12.5	6.2	11.6	6.2	12.2	6.1	nan	5.4	nan
27	6.9	10.4	6.8	10.5	6.6	12.5	6.2	11.6	6.2	12.2	6.1	nan	5.4	nan
28	6.8	9.8	6.9	10.2	6.7	10.4	6.1	12.6	5.9	nan	6.0	10.3	5.7	11.9
29	6.8	9.8	6.9	9.1	6.5	9.4	6.4	12.1	6.2	12.6	6.0	12.0	5.7	13.5
30	7.1	9.0	6.3	10.9	6.1	9.9	5.8	11.2	5.5	nan	5.4	13.7	5.2	21.4
31	6.9	9.7	7.2	10.3	6.7	9.8	6.4	10.3	6.2	12.0	5.6	nan	5.5	nan
32	6.9	9.7	7.2	10.3	6.7	9.8	6.4	10.2	6.0	13.2	5.8	nan	5.8	nan
33	6.9	9.7	6.5	8.8	6.6	9.1	6.3	12.2	6.2	12.1	5.7	12.4	5.8	nan
34	6.9	9.7	7.2	10.3	6.6	10.1	6.3	10.8	6.0	13.3	5.6	11.9	5.7	12.8
35	6.9	9.7	7.2	10.3	6.7	10.2	6.3	10.3	6.2	13.0	5.6	11.9	5.7	12.8
36	7.0	8.5	6.7	9.5	6.5	12.8	6.3	12.4	5.7	11.2	6.2	nan	5.6	nan
37	7.0	8.5	6.7	9.5	6.6	12.0	6.4	11.9	6.1	9.9	5.3	nan	5.5	nan
38	7.0	8.5	6.7	9.5	6.6	12.0	6.2	14.8	6.0	14.8	5.5	nan	5.4	nan
39	7.0	8.5	6.6	9.2	6.6	12.0	6.2	14.8	6.0	14.8	5.5	nan	5.4	nan
40	7.0	8.5	7.0	11.4	6.6	11.7	6.2	14.8	6.0	14.8	5.8	11.7	5.8	13.0
41	6.8	9.8	6.9	9.8	6.7	10.7	6.4	12.9	6.0	16.2	6.1	13.3	5.8	12.0
42	7.1	8.5	6.5	9.9	6.3	10.0	5.8	15.6	5.9	13.7	5.8	nan	5.3	nan
43	7.0	8.5	6.6	9.2	6.7	12.2	6.2	14.8	5.9	11.1	5.5	12.1	5.7	nan
44	7.0	8.5	6.6	9.0	6.7	12.2	6.5	11.6	6.2	13.6	6.1	13.9	5.7	13.5
45	7.0	8.5	6.6	9.2	6.7	12.2	6.2	12.9	6.1	12.0	5.8	11.4	5.6	nan

Table B.3 Performance of Unsupervised Model by KMeans with PFA

Cluster	3		4		5		6		7		8		9	
Features	**in**	**out**	**in**	**out**	**in**	**out**	**in**	**out**	**in**	**out**	**in**	**out**	**in**	**out**
11	6.9	10.4	6.8	9.8	6.5	11.6	6.1	11.1	5.3	13.4	5.5	10.4	5.5	11.6
12	7.1	9.0	6.6	10.0	6.6	11.6	6.3	11.7	5.5	26.7	5.8	11.4	nan	11.9
13	7.1	9.1	6.7	12.0	6.5	13.7	6.5	10.8	6.0	14.2	5.2	nan	nan	nan
14	7.1	9.2	7.2	10.5	6.6	8.7	6.3	8.5	5.9	9.3	5.6	nan	nan	nan
15	6.7	9.6	6.9	10.9	6.6	12.2	6.5	12.1	5.6	10.4	5.6	nan	nan	15.5
16	6.9	9.4	6.6	10.0	6.6	11.3	6.4	10.4	6.2	10.5	6.0	12.3	5.7	13.5
17	6.7	9.1	6.6	8.9	6.4	10.8	6.2	10.8	5.9	12.3	5.8	17.1	5.6	nan
18	6.7	9.1	6.3	8.2	6.3	8.4	6.2	15.3	6.0	16.6	5.5	nan	5.5	nan
19	7.0	9.7	6.5	9.0	5.8	10.5	5.4	18.9	6.2	14.6	5.2	nan	5.3	nan
20	7.0	9.1	6.7	9.6	6.8	10.0	6.6	10.2	6.3	8.4	5.4	nan	5.1	nan
21	7.0	9.1	6.8	9.3	6.7	9.0	6.2	10.3	6.1	13.7	5.5	16.0	5.2	nan
22	6.9	11.4	6.3	10.3	6.5	15.6	6.5	9.8	5.9	nan	6.0	12.5	5.7	12.1
23	6.9	11.4	6.3	10.3	6.6	12.8	6.4	11.8	6.1	16.8	5.9	15.5	5.9	15.1
24	6.9	10.4	7.1	11.0	6.5	12.1	6.3	14.6	6.1	17.6	5.7	13.3	5.6	12.8
25	6.8	9.8	6.9	9.8	6.7	10.8	6.4	11.6	6.1	16.1	5.8	12.6	5.8	12.4
26	6.9	10.4	6.9	10.9	6.6	11.7	6.2	11.6	6.0	nan	5.8	12.9	5.5	15.0
27	6.9	10.4	7.0	10.4	6.5	12.8	6.0	14.5	6.1	14.4	5.4	nan	5.0	nan
28	6.9	10.4	7.2	9.7	6.7	10.9	6.4	10.6	6.0	nan	5.3	19.5	5.2	nan
29	6.9	10.4	6.8	9.0	6.9	11.6	6.4	12.4	6.1	12.6	5.5	24.4	5.5	24.6
30	6.9	10.4	7.0	11.4	6.7	12.4	6.4	12.4	6.1	13.0	5.9	15.9	5.4	28.0
31	7.1	8.4	6.6	10.1	6.6	11.9	6.2	14.1	5.9	16.2	5.6	nan	5.5	15.2
32	7.1	8.4	7.0	11.4	6.5	11.7	6.5	9.7	5.4	nan	5.3	nan	5.4	nan
33	7.1	8.4	6.6	10.1	6.6	11.9	6.2	13.4	5.9	16.2	5.8	19.2	5.4	64.8
34	7.1	8.5	6.5	10.7	6.3	13.3	6.1	nan	6.3	nan	5.9	10.1	5.6	14.4
35	6.8	9.3	6.4	8.5	6.1	10.2	6.3	12.9	6.2	nan	5.5	15.4	5.4	13.5
36	7.1	8.4	6.6	10.1	6.6	11.9	6.3	13.2	6.1	12.7	5.9	14.7	5.4	64.8
37	7.0	8.5	6.6	9.2	6.6	11.7	6.2	14.8	6.2	12.4	5.4	nan	5.0	nan
38	7.0	8.5	6.6	9.1	6.6	12.0	6.2	14.8	5.9	11.1	5.6	12.7	5.7	10.5
39	7.2	9.0	6.6	10.1	6.6	12.0	6.3	13.2	6.1	12.7	5.8	15.4	5.6	20.3
40	7.0	8.5	6.6	9.1	6.6	11.7	6.2	12.9	6.2	13.0	5.4	nan	5.5	nan
41	7.0	8.5	6.6	9.1	6.6	12.0	6.2	12.9	5.9	11.1	5.9	11.0	5.6	14.7
42	7.0	8.5	6.6	9.1	6.6	12.0	6.2	12.9	6.2	12.9	5.9	10.7	5.6	15.5
43	7.0	8.5	6.6	9.2	6.6	12.4	6.2	12.9	6.0	12.1	5.8	12.0	5.6	14.4
44	7.0	8.5	6.6	9.2	6.6	12.4	6.2	12.9	6.1	12.9	5.9	13.8	5.8	14.5
45	7.0	8.5	6.6	9.2	6.6	12.4	6.2	12.9	6.1	12.9	5.6	15.2	5.2	15.8

Table B.4 Performance of Unsupervised Model by KMeans with Grid Search

		Clusters	3		4	
Static Feature	Dynamic Feature	Economic Feature	in	out	in	out
Monday	biderkrp	last1	6.66	7.38	6.52	7.70
Monday	biderkrp	last2	6.66	7.64	6.07	8.00
Monday	biderkrp	last3	6.66	7.40	4.89	6.85
Monday	biderkrp	last4	6.69	7.75	5.00	14.38
Monday	biderkrp	last5	6.69	7.98	6.72	6.88
Monday	biderkrp	last6	6.69	7.75	6.71	8.57
Monday	biderkrp	last7	6.69	7.75	4.86	7.20
Monday	biderkrp	last8	6.69	7.75	6.60	8.52
Monday	biderkrp	last9	6.69	7.75	6.59	8.59
Monday	biderkrp	last10	6.69	7.75	6.59	9.15
Monday	biderkrp	binlast1	6.63	7.64	6.58	8.23
Monday	biderkrp	binlast2	6.68	7.81	6.70	7.45
Monday	biderkrp	binlast3	6.69	7.75	6.52	7.81
Monday	biderkrp	binlast4	6.66	7.51	6.70	7.36
Monday	biderkrp	binlast5	6.69	7.75	6.74	8.22
Monday	biderkrp	binlast6	6.69	7.75	5.86	8.65
Monday	biderkrp	binlast7	6.70	7.96	6.33	7.91
Monday	biderkrp	binlast8	6.69	7.75	6.61	8.08
Monday	biderkrp	binlast9	6.67	7.74	6.59	7.47
Monday	biderkrp	binlast10	6.69	7.75	6.38	7.64
Monday	lastbidkrp	last1	6.87	8.69	6.63	9.25
Monday	lastbidkrp	last2	6.68	8.17	6.68	9.40
Monday	lastbidkrp	last3	4.33	18.96	4.97	16.58
Monday	lastbidkrp	last4	3.83	18.93	4.63	17.39
Monday	lastbidkrp	last5	6.84	7.95	6.49	9.00
Monday	lastbidkrp	last6	6.72	8.05	4.70	9.75
Monday	lastbidkrp	last7	6.69	8.32	5.16	9.32
Monday	lastbidkrp	last8	6.75	8.45	6.63	11.20
Monday	lastbidkrp	last9	6.71	7.98	6.70	9.39
Monday	lastbidkrp	last10	4.42	9.44	5.02	9.57
Monday	lastbidkrp	binlast1	6.86	8.06	6.71	8.14
Monday	lastbidkrp	binlast2	6.87	7.75	6.72	11.74
Monday	lastbidkrp	binlast3	6.64	8.70	6.28	9.62
Monday	lastbidkrp	binlast4	6.89	7.98	6.71	9.67
…	…	…	…	…	…	…

APPENDIX C. PERFORMANCE OF UNSUPERVISED MODEL BY HIERARCHICAL CLUSTERING AND FSA

Table C.1 Performance of Unsupervised Model by Hierarchical and SelKBest

Cluster	3		4		5		6		7		8		9	
Features	in	out	in	out	in	out	in	out	in	out	in	out	in	out
11	6.7	12.1	6.5	13.8	6.7	17.2	nan	19.9	nan	19.3	nan	23.6	nan	22.6
12	6.7	14.7	6.4	13.2	6.7	13.9	6.5	17.8	nan	17.5	nan	37.1	nan	38.5
13	6.5	13.0	6.4	19.8	6.7	19.7	6.3	18.4	6.2	16.3	nan	19.2	nan	20.1
14	6.5	13.1	6.4	18.6	6.7	15.2	6.3	18.1	6.2	16.2	nan	17.8	nan	20.8
15	6.7	14.7	6.5	17.7	6.8	15.6	6.7	16.0	6.4	21.2	nan	25.1	nan	37.7
16	7.4	12.5	7.2	15.4	6.7	14.9	6.3	20.0	6.1	20.1	6.0	18.5	nan	21.4
17	6.8	13.1	6.5	17.0	6.8	18.0	6.8	15.9	6.4	18.0	6.3	19.3	6.1	24.9
18	6.9	9.6	6.7	12.9	6.5	15.6	6.1	14.6	6.1	13.4	5.9	14.3	5.9	16.7
19	6.9	8.6	6.7	13.0	6.5	13.3	6.1	17.6	6.1	17.6	5.9	16.1	5.9	16.8
20	6.7	16.0	6.3	15.4	6.6	15.0	6.5	15.8	6.4	18.6	6.1	7.5	5.3	15.4
21	6.6	9.9	6.5	9.1	6.3	9.8	6.3	12.3	6.3	14.4	5.9	14.2	5.2	17.2
22	6.8	14.7	6.3	17.9	6.4	17.5	6.1	18.6	6.1	18.9	nan	19.0	nan	20.6
23	6.8	11.9	6.3	16.2	6.4	16.0	6.1	16.2	6.1	15.0	nan	14.9	nan	16.1
24	6.8	11.9	6.3	12.5	6.4	11.7	6.1	13.6	6.1	13.1	nan	14.0	nan	15.0
25	6.8	14.6	6.3	17.7	6.4	17.0	6.3	15.7	nan	14.9	nan	15.8	nan	19.6
26	6.8	13.8	6.3	17.3	6.4	16.9	6.3	16.4	nan	17.1	nan	19.0	nan	19.0
27	6.8	13.8	6.3	17.3	6.4	16.9	6.3	16.8	nan	19.3	nan	17.6	nan	16.2
28	6.8	13.8	6.3	17.3	6.4	16.9	6.3	16.8	nan	19.3	nan	19.8	nan	18.2
29	6.8	13.8	6.3	17.3	6.4	16.9	6.3	16.8	nan	19.3	nan	19.8	nan	18.2
30	6.8	13.8	6.3	17.3	6.4	16.9	6.3	16.8	nan	19.3	nan	19.8	nan	18.2
31	6.8	11.7	6.3	17.2	6.4	16.4	6.2	18.7	nan	20.8	nan	38.4	nan	36.1
32	6.8	11.7	6.3	17.2	6.4	19.8	6.0	19.3	nan	15.3	nan	22.6	nan	25.6
33	6.8	11.8	6.3	17.9	6.3	18.2	5.9	15.7	5.9	17.0	nan	17.1	nan	19.8
34	6.7	11.0	6.3	14.3	6.3	15.2	6.3	12.0	6.0	18.9	nan	17.5	nan	17.4
35	6.7	12.0	6.3	12.2	6.4	12.8	5.9	12.8	6.0	14.8	nan	16.4	nan	22.4
36	6.8	10.2	6.3	16.4	6.4	15.6	5.9	14.4	6.0	16.1	nan	19.8	nan	19.1
37	6.8	11.9	6.3	14.0	6.3	13.4	6.3	15.7	6.2	16.8	nan	15.4	nan	32.6
38	6.8	12.7	6.3	14.3	6.4	15.0	6.3	14.8	6.2	14.2	nan	15.1	nan	17.3
39	6.7	9.6	6.2	12.7	6.1	15.5	6.2	15.6	6.1	18.7	nan	20.3	nan	21.1
40	6.8	14.4	6.3	17.5	6.4	15.9	6.4	16.8	6.3	18.6	nan	27.9	nan	27.5
41	6.7	9.4	6.1	10.5	6.2	10.7	6.3	12.3	6.0	19.9	nan	22.1	5.7	23.4
42	7.0	10.9	6.9	15.2	6.4	20.1	6.1	16.2	5.9	19.3	5.4	16.9	5.4	20.2
43	7.0	9.7	6.8	11.2	6.4	12.6	6.3	16.4	5.9	22.6	5.7	21.6	5.4	23.8
44	7.0	15.0	6.7	12.8	6.4	13.7	6.1	19.9	5.7	24.0	5.3	24.1	5.2	22.7
45	7.1	9.2	6.7	10.5	6.4	14.0	6.0	12.4	6.1	18.0	5.8	20.5	5.4	18.3

Table C.2 Performance of Unsupervised Model by Hierarchical with RFE

Cluster	3		4		5		6		7		8		9	
Features	in	out	in	out	in	out	in	out	in	out	in	out	in	out
11	7.0	10.2	6.5	7.7	6.2	16.9	5.9	17.2	5.7	18.6	5.5	18.2	5.4	20.4
12	6.9	15.3	6.4	15.0	6.3	21.0	6.2	21.6	6.1	23.4	6.1	23.0	5.7	23.9
13	6.9	7.7	7.0	12.6	6.7	12.8	6.4	12.3	5.8	16.1	5.6	18.3	5.1	24.3
14	6.9	9.9	6.6	12.9	6.2	12.0	5.6	28.7	5.3	27.7	5.3	25.9	5.3	24.6
15	7.0	9.2	6.5	13.4	5.9	17.1	5.5	17.4	5.5	17.7	5.6	16.4	5.4	17.2
16	6.7	12.5	6.9	11.8	6.8	14.5	6.4	18.2	6.0	15.7	nan	14.7	nan	13.2
17	6.7	13.1	6.9	14.6	6.5	17.4	6.5	18.2	6.1	13.8	nan	14.6	nan	18.1
18	6.7	9.3	6.9	8.6	6.5	15.9	6.5	16.1	6.1	19.8	nan	22.0	nan	25.4
19	6.8	9.3	6.9	8.9	6.5	11.7	6.3	19.8	nan	19.2	nan	14.5	nan	19.2
20	6.7	8.8	6.3	13.1	6.3	13.2	6.3	21.4	6.2	20.9	5.8	25.5	5.6	24.2
21	6.7	9.6	6.3	14.1	6.4	14.0	6.3	20.3	6.2	19.7	5.8	26.4	5.5	26.7
22	6.7	8.4	6.3	12.9	6.4	13.4	6.3	22.9	6.2	23.0	5.8	28.7	5.6	29.4
23	6.8	13.5	6.4	12.3	6.3	13.9	5.8	16.3	5.8	12.8	5.5	12.2	5.4	14.7
24	7.0	7.9	6.7	11.9	6.3	14.3	6.0	15.2	6.1	12.0	5.8	12.1	5.6	12.8
25	6.4	12.9	6.1	15.5	6.2	15.9	6.3	13.7	6.0	17.7	6.0	19.0	5.9	17.5
26	6.9	7.8	6.4	9.2	6.5	12.1	6.3	14.1	6.1	15.2	6.1	18.5	5.8	19.7
27	6.7	9.1	6.3	10.2	6.1	9.2	6.2	11.7	6.0	13.8	5.8	15.6	5.6	16.5
28	6.7	9.2	6.4	11.2	6.2	15.0	6.3	12.7	6.1	13.2	5.9	11.9	5.9	12.4
29	6.6	10.7	6.1	16.2	6.1	24.4	5.8	27.2	5.8	25.9	6.0	23.6	5.8	21.3
30	7.1	9.4	6.9	14.0	6.6	14.7	6.2	19.7	6.0	14.4	5.8	14.7	5.3	31.8
31	7.1	9.4	6.9	14.7	6.6	14.6	6.2	20.4	6.0	18.8	5.8	18.7	5.3	25.5
32	7.0	9.5	6.8	8.9	6.2	14.4	6.3	17.7	6.0	17.6	5.6	20.0	5.2	30.8
33	7.1	10.8	6.7	10.1	6.2	14.1	6.1	17.6	5.7	18.1	5.6	22.6	5.1	31.9
34	7.1	9.6	6.8	11.6	6.4	14.8	5.9	24.3	5.9	26.4	5.8	25.6	5.2	33.2
35	7.1	9.6	6.8	11.6	6.4	14.8	5.9	24.3	5.9	24.3	5.8	23.7	5.2	31.5
36	7.0	9.3	6.8	11.6	6.5	13.7	6.0	23.7	5.9	25.0	5.5	26.4	5.0	34.6
37	7.1	10.2	6.8	11.0	6.6	16.7	6.0	30.6	5.8	30.9	5.3	35.0	4.9	32.6
38	7.1	8.8	6.8	9.7	6.3	16.7	6.0	15.3	5.9	14.8	5.7	18.2	5.4	16.0
39	7.1	8.8	6.9	9.2	6.5	13.0	6.1	20.7	5.9	21.8	5.5	21.9	5.3	22.0
40	7.1	8.8	6.9	9.2	6.5	13.0	6.1	20.7	5.9	21.8	5.5	21.9	5.3	22.0
41	7.1	8.4	6.9	11.9	6.5	12.7	6.1	22.1	5.9	24.3	5.6	24.0	5.3	18.4
42	7.1	8.3	6.9	10.8	6.5	13.6	6.1	17.8	6.0	22.3	5.6	20.7	5.3	18.6
43	7.1	10.7	6.8	14.7	6.6	13.8	6.2	15.6	5.9	22.2	5.6	25.8	5.2	25.6
44	7.1	8.0	6.7	10.1	6.2	16.4	6.1	20.2	5.7	23.4	5.4	21.9	5.0	39.8
45	7.1	8.7	6.7	13.8	6.3	17.0	6.1	15.8	5.7	17.4	5.3	18.7	5.1	19.7

Table C.3 Performance of Unsupervised Model by Hierarchical with PFA

Cluster	3		4		5		6		7		8		9	
Features	in	out	in	out	in	out	in	out	in	out	in	out	in	out
11	6.8	11.7	6.8	12.2	6.2	14.7	6.2	13.0	6.0	15.0	nan	31.7	4.7	35.2
12	7.0	8.9	6.9	9.6	6.8	13.0	6.2	17.0	5.9	27.8	5.8	26.2	5.5	25.1
13	6.9	9.5	6.8	9.3	6.6	15.8	6.2	14.5	6.0	12.4	5.9	14.5	5.7	12.5
14	7.1	9.9	6.8	11.1	6.4	13.0	6.4	15.2	5.9	19.4	5.7	21.2	5.4	18.3
15	7.1	8.9	6.7	10.5	6.3	8.2	6.3	13.7	5.9	15.0	5.7	17.3	5.7	20.3
16	7.0	9.6	6.8	13.7	6.3	12.2	6.2	12.5	5.9	15.9	5.5	19.0	5.1	44.2
17	7.1	8.9	6.3	18.4	6.4	19.7	6.1	25.6	5.9	22.5	5.5	26.9	5.3	24.9
18	7.0	11.9	7.0	12.8	6.4	19.0	6.2	19.5	5.7	21.0	5.5	21.0	5.0	39.2
19	7.1	13.1	6.7	12.5	6.1	18.8	6.0	18.5	5.5	23.9	5.2	25.7	4.8	45.7
20	7.1	12.8	6.7	13.1	6.1	12.3	6.0	12.3	5.7	13.3	5.3	15.3	5.2	17.9
21	7.1	12.0	6.7	10.4	6.4	11.2	5.9	25.1	5.7	24.9	5.4	25.0	5.2	21.9
22	6.8	12.9	6.7	13.6	6.3	13.4	5.8	23.5	5.7	22.5	5.5	20.7	5.0	20.6
23	7.1	8.9	6.7	12.1	6.3	20.8	6.2	17.9	5.8	24.1	5.6	25.8	5.1	48.6
24	6.9	8.9	6.9	16.0	6.4	20.4	6.2	20.3	6.1	16.5	5.6	17.2	5.1	39.8
25	7.1	13.3	6.8	12.9	6.5	15.7	6.3	17.2	5.8	17.1	5.5	23.2	5.3	22.2
26	6.9	11.8	6.8	13.9	6.4	13.3	6.0	16.4	5.8	16.9	5.7	16.3	5.6	19.4
27	7.0	8.7	6.9	13.0	6.6	14.5	6.4	18.3	6.1	15.1	5.9	19.0	5.6	26.2
28	7.0	9.9	6.6	15.8	6.1	19.1	6.2	18.4	5.9	16.2	5.7	15.4	5.3	16.8
29	7.1	9.9	6.7	12.2	6.3	13.1	6.0	19.5	6.0	18.3	5.8	19.5	5.2	31.2
30	7.0	9.5	6.9	11.4	6.1	30.3	5.8	27.0	5.5	30.0	5.5	24.8	5.4	31.7
31	7.0	11.8	6.9	12.8	6.3	16.3	5.9	18.4	5.6	19.5	5.5	20.1	5.4	20.2
32	7.1	13.9	6.7	16.3	6.2	19.3	5.9	16.9	5.3	17.1	5.1	14.9	4.6	38.4
33	6.9	11.5	6.2	13.2	5.9	16.4	5.0	24.0	5.3	24.4	5.1	24.4	5.1	26.5
34	7.1	14.5	6.4	13.0	6.3	17.4	6.0	17.1	5.9	17.4	5.8	21.0	5.5	19.1
35	7.2	13.3	6.8	11.4	6.2	14.5	5.9	16.2	5.8	15.7	5.7	19.0	5.4	20.2
36	6.8	13.7	6.7	13.3	5.9	33.2	5.6	27.7	5.8	25.6	5.5	28.8	5.5	22.3
37	7.0	12.5	6.8	11.5	6.6	10.5	6.1	19.8	5.7	16.7	5.4	22.7	5.3	21.1
38	7.1	12.6	6.8	10.5	6.3	10.1	6.0	12.6	5.7	22.3	5.6	12.0	5.2	14.7
39	6.9	9.6	5.9	30.5	6.0	24.1	5.7	24.9	5.1	37.2	5.1	27.6	5.1	27.9
40	6.9	14.4	6.9	13.9	6.3	15.4	6.2	13.3	6.0	16.4	5.9	19.3	5.7	18.0
41	6.9	9.0	6.9	9.7	6.3	15.5	6.1	15.3	6.0	14.5	5.8	18.9	5.6	19.1
42	6.9	8.0	6.9	8.5	6.3	20.6	6.1	18.0	6.0	20.1	5.8	19.1	5.6	17.0
43	7.0	8.4	6.7	9.7	6.2	18.9	6.0	14.5	5.9	13.5	5.9	18.4	5.5	19.2
44	7.0	8.6	6.7	13.3	6.3	14.5	6.0	12.7	5.9	17.4	5.9	20.4	5.6	20.8
45	7.0	9.2	6.8	9.8	6.5	13.7	6.1	19.3	5.9	18.8	5.5	22.0	5.2	21.4

APPENDIX D. REGRESSION RESULTS OF UNSUPERVISED MODEL BY NONLINEAR DURATION VARIABLE

OLS Regression Results

Dep. Variable:	price	R-squared:	0.348
Model:	OLS	Adj. R-squared:	0.312
Method:	Least Squares	F-statistic:	9.575
Date:	31 Jan 2019	Prob (F-statistic):	0.00
Time:	08:09:02 PM	Log-Likelihood:	-2006
No. Observations:	380	AIC:	4054
Df Residuals:	359	BIC:	4137
Df Model:	20		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
extras	19.89	5.97	3.33	0.00	8.15	31.63
biderkrp	1.44	0.53	2.70	0.01	0.39	2.49
last4	0.08	0.04	1.75	0.08	-0.01	0.16
binlast6	0.09	0.05	1.88	0.06	0.00	0.18
binlast8	0.10	0.05	1.99	0.05	0.00	0.20
selrate	-0.02	0.01	-1.74	0.08	-0.03	0.00
neg	-6.83	2.80	-2.44	0.02	-12.34	-1.32
views	0.02	0.01	3.62	0.00	0.01	0.03
k0	127.87	28.54	4.48	0.00	71.74	184.01
k1	120.32	27.74	4.34	0.00	65.77	174.88
k2	127.86	28.82	4.44	0.00	71.18	184.55
k3	135.46	27.68	4.89	0.00	81.02	189.89
durat2	5.57	2.47	2.26	0.03	0.71	10.42
smonday	77.57	16.78	4.62	0.00	44.57	110.57
stuesday	74.62	17.07	4.37	0.00	41.05	108.19
swednesday	70.39	17.07	4.12	0.00	36.82	103.97
sthursday	67.84	17.58	3.86	0.00	33.27	102.41
sfriday	76.81	17.29	4.44	0.00	42.82	110.81
ssaturday	65.44	16.63	3.94	0.00	32.73	98.16
ssunday	78.84	16.91	4.66	0.00	45.58	112.10
lastbidkrp	0.34	0.03	10.98	0.00	0.28	0.40
lastbidderratekrp	0.01	0.01	1.72	0.09	0.00	0.03

APPENDIX E. IMPACT NUMBER OF ALTERNATIVE PRODUCTS

Table E.1 Regression Results of the Model by Number of Alternative Products When the Auction Starts

OLS Regression Results			
Dep. Variable:	price	R-squared:	0.353
Model:	OLS	Adj. R-squared:	0.309
Method:	Least Squares	F-statistic:	8.061
Date:	31 Jan 2019	Prob (F-statistic):	0.00
Time:	09:06:03 PM	Log-Likelihood:	-2004.6
No. Observations:	380	AIC:	4059
Df Residuals:	355	BIC:	4158
Df Model:	24		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
quantity	0.71	0.52	1.38	0.17	-0.3	1.7
extras	19.92	6.00	3.32	0.00	8.12	31.72
biderkrp	1.40	0.54	2.62	0.01	0.35	2.46
last4	0.07	0.04	1.60	0.11	-0.02	0.15
binlast6	0.08	0.05	1.71	0.09	-0.01	0.17
binlast8	0.09	0.05	1.76	0.08	-0.01	0.19
selrate	-0.02	0.01	-1.71	0.09	-0.03	0.00
neg	-6.61	2.82	-2.34	0.02	-12.15	-1.06
views	0.02	0.01	3.57	0.00	0.01	0.03
k0	94.27	18.90	4.99	0.00	57.10	131.45
k1	85.57	18.12	4.72	0.00	49.94	121.21
k2	92.52	19.31	4.79	0.00	54.54	130.51
k3	101.31	18.16	5.58	0.00	65.59	137.02
d1	72.99	15.66	4.66	0.00	42.19	103.78
d3	78.42	15.33	5.12	0.00	48.28	108.56
d5	87.52	15.78	5.55	0.00	56.48	118.56
d7	78.85	14.41	5.47	0.00	50.50	107.19
d10	55.90	18.54	3.02	0.00	19.45	92.35
smonday	57.60	11.90	4.84	0.00	34.20	81.01
stuesday	53.73	11.92	4.51	0.00	30.28	77.18
swednesday	51.22	11.98	4.28	0.00	27.66	74.79
sthursday	47.67	12.55	3.80	0.00	23.00	72.35
sfriday	58.05	12.40	4.68	0.00	33.68	82.43
ssaturday	46.27	11.70	3.96	0.00	23.26	69.28
ssunday	59.12	11.73	5.04	0.00	36.04	82.19
lastbidkrp	0.33	0.03	10.58	0.00	0.3	0.4

Table E.2 Regression Results of the Model by Number of Alternative Products When the Auction Initial Stage Ends

OLS Regression Results			
Dep. Variable:	price	R-squared:	0.351
Model:	OLS	Adj. R-squared:	0.307
Method:	Least Squares	F-statistic:	8.011
Date:	31 Jan 2019	Prob (F-statistic):	0.00
Time:	09:22:07 PM	Log-Likelihood:	-2005.1
No. Observations:	380	AIC:	4060
Df Residuals:	355	BIC:	4159
Df Model:	24		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
quantity2	0.58	0.55	1.06	0.29	-0.50	1.67
extras	20.04	6.01	3.34	0.00	8.23	31.86
biderkrp	1.43	0.54	2.66	0.01	0.37	2.49
last4	0.07	0.04	1.65	0.10	-0.01	0.16
binlast6	0.08	0.05	1.70	0.09	-0.01	0.17
binlast8	0.09	0.05	1.76	0.08	-0.01	0.20
selrate	-0.02	0.01	-1.70	0.09	-0.03	0.00
neg	-6.67	2.82	-2.36	0.02	-12.22	-1.12
views	0.02	0.01	3.55	0.00	0.01	0.03
k0	93.69	18.92	4.95	0.00	56.48	130.91
k1	85.60	18.14	4.72	0.00	49.93	121.28
k2	92.35	19.35	4.77	0.00	54.29	130.41
k3	101.24	18.18	5.57	0.00	65.48	137.00
d1	73.26	15.69	4.67	0.00	42.40	104.12
d3	78.80	15.34	5.14	0.00	48.63	108.96
d5	87.46	15.80	5.54	0.00	56.39	118.53
d7	78.87	14.43	5.47	0.00	50.49	107.25
d10	54.50	18.50	2.95	0.00	18.11	90.89
smonday	57.87	11.91	4.86	0.00	34.45	81.30
stuesday	53.67	11.97	4.49	0.00	30.14	77.20
swednesday	50.48	12.01	4.20	0.00	26.86	74.10
sthursday	47.59	12.57	3.79	0.00	22.87	72.30
sfriday	58.05	12.41	4.68	0.00	33.65	82.46
ssaturday	46.11	11.72	3.94	0.00	23.07	69.15
ssunday	59.11	11.75	5.03	0.00	36.00	82.21
lastbidkrp	0.33	0.03	10.62	0.00	0.27	0.39
lastbidderratekrp	0.01	0.01	1.66	0.10	0.00	0.03

APPENDIX F. REGRESSION RESULTS OF UNSUPERVISED MODEL FOR EACH CLUSTER

Table F.1 Regression Results of Unsupervised Model for Cluster0

OLS Regression Results						
Dep. Variable:	price	R-squared:	0.664			
Model:	OLS	R-squared:	0.238			
Method:	Least Squares	F-statistic:	1.559			
Date:	Tue 29 Jan 2019	Prob (F-statistic):	0.194			
Time:	10:05:06 PM	Log-Likelihood:	-162.76			
No. Observations:	35	AIC:	365.5			
Df Residuals:	15	BIC:	396.6			
Df Model:	19					
Covariance Type:	nonrobust					
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
extras	38.33	29.50	1.30	0.21	-24.5	101.2
biderkrp	-1.51	2.50	-0.61	0.55	-6.84	3.81
last4	0.04	0.14	0.26	0.80	-0.27	0.34
binlast6	0.24	0.16	1.46	0.16	-0.11	0.59
binlast8	-0.05	0.22	-0.25	0.81	-0.51	0.41
selrate	-0.02	0.06	-0.28	0.78	-0.13	0.10
neg	-13.47	15.97	-0.84	0.41	-47.51	20.58
views	0.02	0.03	0.87	0.40	-0.04	0.08
k0	41.40	26.75	1.55	0.14	-15.61	98.41
k1	51.55	26.49	1.95	0.07	-4.92	108.02
k2	68.59	54.90	1.25	0.23	-48.44	185.61
k3	55.86	27.69	2.02	0.06	-3.15	114.87
d1	0.00	0.00	2.07	0.06	0.00	0.00
d3	217.40	103.75	2.10	0.05	-3.73	438.52
d5	0.00	0.00	-2.17	0.05	0.00	0.00
d7	0.00	0.00	-0.72	0.48	0.00	0.00
d10	0.00	0.00	-1.58	0.14	0.00	0.00
smonday	38.39	40.68	0.94	0.36	-48.31	125.09
stuesday	46.34	28.15	1.65	0.12	-13.66	106.33
swednesday	26.63	22.61	1.18	0.26	-21.56	74.81
sthursday	25.68	23.65	1.09	0.30	-24.73	76.09
sfriday	24.30	28.96	0.84	0.42	-37.43	86.03
ssaturday	6.82	46.50	0.15	0.89	-92.28	105.92
ssunday	49.24	28.05	1.76	0.10	-10.54	109.02
lastbidkrp	0.17	0.14	1.21	0.25	-0.13	0.48
lastbidderratekrp	0.03	0.03	1.32	0.21	0.0	0.1

Table F.2 Regression Results of Unsupervised Model for Cluster1

OLS Regression Results			
Dep. Variable:	price	R-squared:	0.374
Model:	OLS	R-squared:	0.306
Method:	Least Squares	F-statistic:	5.44
Date:	Tue 29 Jan 2019	Prob (F-statistic):	0.00
Time:	10:06:49 PM	Log-Likelihood:	-1178.4
No. Observations:	223	AIC:	2403
Df Residuals:	200	BIC:	2481
Df Model:	22		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
extras	11.37	8.44	1.35	0.18	-5.3	28.0
biderkrp	1.93	0.78	2.49	0.01	0.41	3.46
last4	0.08	0.06	1.46	0.15	-0.03	0.20
binlast6	0.06	0.06	0.99	0.32	-0.06	0.18
binlast8	0.08	0.07	1.15	0.25	-0.06	0.23
selrate	-0.02	0.02	-1.62	0.11	-0.05	0.01
neg	-6.36	5.47	-1.16	0.25	-17.15	4.42
views	0.01	0.02	0.38	0.70	-0.03	0.04
k0	88.48	23.43	3.78	0.00	42.27	134.68
k1	74.75	22.89	3.27	0.00	29.61	119.90
k2	100.24	24.29	4.13	0.00	52.34	148.15
k3	94.25	22.31	4.23	0.00	50.26	138.23
d1	88.85	23.57	3.77	0.00	42.37	135.33
d3	0.00	0.00	4.54	0.00	0.00	0.00
d5	105.21	23.60	4.46	0.00	58.68	151.74
d7	96.86	21.95	4.41	0.00	53.58	140.14
d10	66.80	27.31	2.45	0.02	12.95	120.65
smonday	53.73	14.80	3.63	0.00	24.55	82.91
stuesday	53.02	14.85	3.57	0.00	23.73	82.32
swednesday	45.86	15.45	2.97	0.00	15.40	76.33
sthursday	53.18	16.69	3.19	0.00	20.27	86.09
sfriday	56.34	15.78	3.57	0.00	25.22	87.47
ssaturday	35.63	14.91	2.39	0.02	6.23	65.04
ssunday	59.95	14.69	4.08	0.00	30.99	88.91
lastbidkrp	0.38	0.04	8.75	0.00	0.30	0.47
lastbidderratekrp	0.02	0.01	2.01	0.05	0.0	0.0

Table F.3 Regression Results of Unsupervised Model for Cluster2

OLS Regression Results			
Dep. Variable:	price	R-squared:	0.741
Model:	OLS	R-squared:	0.33
Method:	Least Squares	F-statistic:	1.805
Date:	Tue 29 Jan 2019	Prob (F-statistic):	0.148
Time:	10:08:31 PM	Log-Likelihood:	-156.83
No. Observations:	32	AIC:	353.7
Df Residuals:	12	BIC:	383
Df Model:	19		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
extras	22.02	39.61	0.56	0.59	-64.3	108.3
biderkrp	3.24	3.04	1.07	0.31	-3.39	9.87
last4	0.22	0.43	0.50	0.63	-0.73	1.16
binlast6	-0.08	0.42	-0.20	0.85	-1.00	0.84
binlast8	-0.20	0.27	-0.73	0.48	-0.79	0.39
selrate	-0.02	0.05	-0.33	0.75	-0.13	0.10
neg	35.98	31.41	1.15	0.27	-32.45	104.41
views	0.05	0.15	0.35	0.74	-0.28	0.38
k0	49.25	73.35	0.67	0.52	-110.58	209.08
k1	22.93	88.70	0.26	0.80	-170.33	216.19
k2	55.72	86.08	0.65	0.53	-131.83	243.26
k3	89.55	88.17	1.02	0.33	-102.56	281.66
d1	0.00	0.00	0.61	0.55	0.00	0.00
d3	217.45	313.57	0.69	0.50	-465.76	900.66
d5	0.00	0.00	0.73	0.48	0.00	0.00
d7	0.00	0.00	-0.63	0.54	0.00	0.00
d10	0.00	0.00	0.63	0.54	0.00	0.00
smonday	70.73	31.74	2.23	0.05	1.57	139.88
stuesday	-20.57	59.78	-0.34	0.74	-150.82	109.67
swednesday	46.61	49.25	0.95	0.36	-60.69	153.91
sthursday	45.88	77.10	0.60	0.56	-122.11	213.87
sfriday	8.31	69.05	0.12	0.91	-142.13	158.75
ssaturday	53.03	52.98	1.00	0.34	-62.41	168.47
ssunday	13.47	75.47	0.18	0.86	-150.96	177.89
lastbidkrp	0.54	0.13	4.21	0.00	0.26	0.82
lastbidderratekrp	-0.08	0.12	-0.66	0.52	-0.3	0.2

Table F.4 Regression Results of Unsupervised Model for Cluster3

OLS Regression Results

Dep. Variable:	price	R-squared:	0.418
Model:	OLS	R-squared:	0.226
Method:	Least Squares	F-statistic:	2.183
Date:	Tue 29 Jan 2019	Prob (F-statistic):	0.00774
Time:	10:10:16 PM	Log-Likelihood:	-468.19
No. Observations:	90	AIC:	982.4
Df Residuals:	67	BIC:	1040
Df Model:	22		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
extras	11.67	13.72	0.85	0.40	-15.7	39.1
biderkrp	2.62	1.55	1.69	0.10	-0.47	5.72
last4	0.04	0.11	0.39	0.70	-0.17	0.26
binlast6	0.26	0.13	2.03	0.05	0.01	0.52
binlast8	0.18	0.12	1.51	0.14	-0.06	0.41
selrate	-0.02	0.02	-1.24	0.22	-0.06	0.02
neg	-2.72	4.91	-0.55	0.58	-12.52	7.08
views	0.01	0.01	0.98	0.33	-0.01	0.03
k0	56.28	45.50	1.24	0.22	-34.53	147.08
k1	46.73	42.37	1.10	0.27	-37.85	131.31
k2	59.93	43.96	1.36	0.18	-27.81	147.68
k3	65.91	45.10	1.46	0.15	-24.11	155.94
d1	78.09	41.88	1.87	0.07	-5.49	161.67
d3	0.00	0.00	-1.30	0.20	0.00	0.00
d5	63.14	46.17	1.37	0.18	-29.02	155.30
d7	60.80	41.76	1.46	0.15	-22.55	144.16
d10	26.82	48.51	0.55	0.58	-70.01	123.64
smonday	30.64	29.61	1.04	0.30	-28.46	89.75
stuesday	35.98	29.15	1.23	0.22	-22.20	94.16
swednesday	37.10	31.35	1.18	0.24	-25.48	99.67
sthursday	2.86	30.08	0.10	0.92	-57.18	62.91
sfriday	57.19	28.82	1.98	0.05	-0.34	114.72
ssaturday	32.08	25.92	1.24	0.22	-19.66	83.82
ssunday	33.00	28.64	1.15	0.25	-24.17	90.17
lastbidkrp	0.26	0.08	3.37	0.00	0.11	0.41
lastbidderratekrp	0.00	0.01	0.10	0.92	0.0	0.0

APPENDIX G. PERFORMANCE OF SUPERVISED MODEL BY KMEANS AND FSA

Table G.1 Performance of Supervised Model by KMeans with SelecKBest

Cluster	3		4		5		6		7		8		9	
Features	**in**	**out**	**in**	**out**	**in**	**out**	**in**	**out**	**in**	**out**	**in**	**out**	**in**	**out**
11	6.6	9.7	6.8	11.0	6.6	14.2	6.5	14.3	nan	14.6	6.0	12.5	5.6	14.4
12	6.6	9.7	6.8	11.0	6.7	10.4	6.4	11.0	6.3	12.4	5.8	12.2	nan	nan
13	6.6	9.7	6.8	11.0	6.6	10.7	6.2	11.2	6.0	13.8	6.2	15.4	nan	nan
14	7.2	9.5	7.0	10.5	6.4	10.1	6.4	11.4	6.1	13.7	6.2	11.5	5.3	14.8
15	6.9	9.8	6.8	9.3	6.8	10.7	6.5	13.6	6.1	12.4	6.0	14.8	5.7	14.0
16	6.9	9.8	6.8	9.3	6.5	9.6	6.5	13.6	6.0	16.4	5.6	13.5	4.9	nan
17	6.9	9.8	6.8	9.3	6.5	9.6	6.5	13.6	6.0	16.3	5.6	13.5	5.1	nan
18	6.9	7.5	7.0	8.2	6.4	8.6	6.2	10.4	5.7	18.0	nan	nan	nan	nan
19	7.1	8.0	6.6	9.5	6.3	10.0	6.1	10.8	6.0	11.2	5.8	10.8	nan	8.2
20	7.0	7.9	6.7	10.2	6.5	10.9	6.4	12.8	6.0	13.0	5.3	nan	nan	nan
21	7.0	7.9	7.0	9.7	6.9	9.5	6.5	11.0	6.2	13.0	nan	15.2	nan	15.3
22	6.2	11.2	5.4	14.1	5.0	22.8	5.3	15.5	5.4	nan	nan	nan	nan	nan
23	6.7	8.4	6.6	7.7	6.3	8.4	6.1	14.3	6.3	13.8	nan	21.6	nan	30.6
24	6.8	7.5	6.0	11.0	6.2	6.7	6.1	11.7	5.4	14.7	nan	nan	nan	nan
25	7.0	7.9	6.5	7.2	5.4	9.9	5.9	8.8	5.6	9.0	6.1	15.1	nan	24.0
26	6.8	7.5	6.6	7.6	6.0	10.7	6.1	6.8	6.1	10.8	5.7	8.6	nan	15.6
27	6.8	7.5	6.6	7.4	6.4	7.5	6.2	nan	nan	14.6	5.4	13.8	nan	13.3
28	6.8	7.5	6.6	7.4	6.4	7.5	6.2	nan	nan	14.6	5.1	9.2	nan	10.9
29	7.1	8.1	6.8	8.4	6.4	9.0	6.4	8.6	6.2	8.3	5.7	11.1	5.5	10.5
30	6.8	7.5	6.3	7.6	6.5	8.7	6.2	7.6	5.7	10.8	6.0	11.9	nan	nan
31	6.8	7.5	6.3	7.6	5.7	7.7	5.8	15.8	5.7	15.3	5.4	12.7	nan	10.8
32	6.9	6.8	6.8	7.6	6.7	8.5	6.2	8.3	6.1	8.7	5.4	15.0	5.5	12.6
33	7.0	7.5	6.3	6.7	6.3	6.7	6.1	6.0	6.1	9.2	5.0	10.5	nan	nan
34	7.0	7.5	6.3	6.7	6.3	6.7	5.8	7.4	5.8	7.4	5.6	8.9	5.4	nan
35	7.0	7.5	6.3	6.7	6.1	9.5	6.1	8.6	5.8	12.1	5.9	7.6	nan	nan
36	7.0	7.5	6.3	6.7	6.2	9.5	6.1	8.6	5.6	9.1	5.9	14.3	5.7	9.5
37	7.0	7.5	6.4	6.7	6.1	9.5	6.0	8.7	5.1	8.2	5.7	9.2	5.7	7.9
38	7.0	7.5	6.4	6.7	6.1	9.5	6.2	8.2	5.4	6.7	5.7	9.3	5.8	7.9
39	7.0	7.5	6.4	6.7	6.1	9.5	6.3	8.6	5.4	6.7	5.9	11.1	5.7	9.3
40	7.0	7.5	6.4	6.7	6.1	9.5	6.3	8.4	5.4	6.7	5.3	nan	5.2	nan
41	7.0	7.5	6.2	7.2	6.4	8.3	5.4	7.7	5.6	nan	6.1	10.3	5.9	10.5
42	7.0	7.5	6.2	7.2	6.4	8.3	5.8	8.2	5.5	nan	5.4	7.3	5.4	9.2
43	7.0	7.5	6.2	7.2	6.4	8.3	5.8	8.2	5.5	nan	5.7	7.3	5.4	8.2
44	6.9	7.7	6.9	8.5	6.7	7.8	6.4	6.7	6.4	9.5	6.4	7.8	5.8	8.8
45	6.9	7.7	6.9	7.5	6.6	9.9	6.1	7.8	5.7	7.7	6.1	9.4	5.8	9.8

Table G.2 Performance of Supervised Model by KMeans with RFE

Cluster	3		4		5		6		7		8		9	
Features	in	out	in	out	in	out	in	out	in	out	in	out	in	out
11	6.5	12.0	5.9	10.2	5.9	9.6	nan	8.5	nan	nan	nan	nan	nan	nan
12	6.5	12.0	5.9	10.2	6.1	9.2	nan	8.2	nan	11.4	nan	13.6	nan	212.1
13	6.2	6.7	6.0	10.1	6.0	9.5	5.7	9.4	5.4	8.7	5.2	nan	4.7	nan
14	6.4	6.9	6.2	6.6	5.9	8.7	6.1	8.6	6.0	8.5	5.6	7.2	5.5	8.0
15	7.0	7.5	6.3	6.7	6.2	7.1	6.1	8.9	6.0	8.0	5.8	8.3	5.5	8.2
16	7.0	7.5	6.3	6.8	6.3	7.3	5.9	6.6	6.1	6.1	5.9	7.5	5.3	11.9
17	7.0	7.5	6.2	6.6	6.3	6.7	6.0	8.0	5.9	8.7	5.8	10.9	5.9	8.7
18	7.0	7.5	6.3	6.8	6.1	5.3	6.4	5.6	5.7	9.1	5.8	10.9	5.9	8.7
19	7.0	7.5	6.9	7.5	6.6	7.1	6.1	7.6	5.9	8.4	5.7	9.5	5.7	8.8
20	7.0	7.5	6.2	6.6	6.3	6.7	6.2	8.3	5.9	8.6	5.6	19.3	5.7	17.8
21	7.0	7.5	6.2	6.6	6.3	6.7	6.2	8.3	5.9	8.7	5.4	10.8	5.2	10.3
22	7.0	7.5	6.2	6.6	6.3	6.7	6.2	8.3	5.1	6.9	5.3	11.1	4.8	13.3
23	7.0	7.5	6.2	6.6	6.3	6.7	6.2	8.3	5.1	6.9	5.7	10.8	5.4	11.4
24	6.9	6.8	6.9	7.0	6.4	6.9	6.2	6.6	6.0	6.5	5.4	22.1	nan	nan
25	6.9	6.8	6.9	7.0	6.4	6.9	5.8	6.9	6.0	7.4	5.5	10.6	5.3	12.2
26	6.9	6.8	6.9	7.7	6.3	6.7	6.0	8.3	5.8	7.4	5.8	8.4	5.7	9.1
27	7.0	7.5	6.2	5.9	6.3	6.7	6.1	6.2	5.2	10.4	5.6	20.1	5.7	18.7
28	7.0	7.5	6.4	6.7	6.3	7.3	5.6	11.0	6.0	8.3	6.0	7.0	5.5	9.2
29	7.0	7.5	6.3	6.7	6.3	6.7	6.2	7.6	5.9	9.0	5.5	nan	5.4	nan
30	7.0	7.7	6.9	7.7	6.4	8.7	6.4	7.1	5.7	9.1	5.6	6.6	5.4	7.9
31	7.0	7.7	6.9	7.7	6.3	6.7	5.8	7.8	5.1	12.2	5.8	8.5	5.8	8.2
32	7.0	7.7	6.8	8.7	6.2	7.1	6.3	8.9	5.7	10.3	5.0	nan	5.0	nan
33	7.1	6.9	6.9	8.7	6.0	7.1	6.1	6.6	5.6	7.8	5.7	8.8	5.4	10.5
34	6.7	9.0	6.6	8.6	6.5	9.0	5.9	11.2	5.5	11.7	5.7	9.4	nan	10.3
35	6.4	6.9	6.5	6.6	6.4	7.1	6.4	6.8	6.0	6.9	5.7	9.1	5.8	11.9
36	7.2	7.9	6.4	6.7	6.0	7.6	5.6	12.7	5.5	9.4	5.7	9.1	nan	nan
37	6.4	6.9	6.0	7.3	6.4	7.1	6.3	7.5	6.0	nan	5.6	19.9	5.4	11.5
38	6.7	9.0	6.5	6.6	6.2	7.1	6.4	9.4	5.5	11.7	5.7	7.6	5.2	10.0
39	6.7	9.0	6.5	6.6	6.2	7.1	6.4	9.4	5.7	10.1	5.8	7.6	5.9	6.9
40	6.7	9.0	6.5	6.6	6.3	9.2	6.1	10.4	6.2	11.9	5.7	7.1	5.6	8.0
41	6.8	7.5	6.8	7.6	6.4	6.7	6.2	6.9	6.5	7.1	5.7	10.9	5.3	13.6
42	6.8	7.5	6.7	7.0	6.6	8.1	6.4	8.1	6.2	11.8	5.6	7.7	5.5	9.8
43	6.8	7.5	6.7	7.0	6.5	8.4	6.1	6.2	5.7	6.2	5.6	7.7	5.5	9.8
44	6.8	7.5	6.7	7.0	6.6	8.1	6.1	6.2	5.7	6.2	6.1	10.8	5.9	11.1
45	7.0	7.5	6.3	6.8	6.3	7.3	6.4	7.2	6.4	7.7	5.5	15.9	nan	14.0

Table G.3 Performance of Supervised Model by KMeans with PFA

Cluster	3		4		5		6		7		8		9	
Features	in	out	in	out	in	out	in	out	in	out	in	out	in	out
11	6.7	8.2	6.6	7.5	6.5	7.2	6.4	9.7	6.1	9.7	5.6	11.3	5.5	27.5
12	7.0	7.5	6.8	8.1	6.5	7.8	5.9	12.0	5.9	14.6	5.0	11.2	5.0	11.4
13	7.0	7.5	5.8	5.4	5.9	6.4	5.1	7.2	nan	9.7	nan	16.7	nan	14.4
14	6.8	7.5	6.4	6.6	6.6	6.8	6.2	7.3	5.9	9.4	5.7	11.2	5.3	12.0
15	7.0	7.5	6.9	7.5	6.3	9.8	6.0	9.8	5.2	11.4	5.4	10.7	4.6	nan
16	6.1	6.8	6.3	6.8	6.1	5.8	6.1	6.4	6.2	6.2	5.4	7.6	5.0	8.9
17	6.9	7.7	6.9	7.7	6.3	7.3	6.4	7.8	6.5	7.4	5.8	8.1	6.0	9.8
18	7.0	7.5	6.3	6.8	6.0	7.5	5.9	8.1	5.9	8.2	6.2	9.5	5.8	11.9
19	6.8	7.5	6.8	6.3	6.3	6.7	5.9	7.6	5.8	8.0	5.7	8.0	5.6	8.9
20	7.0	7.5	6.3	6.8	6.3	7.3	6.3	7.4	6.0	7.5	5.8	7.7	5.5	8.8
21	6.9	6.8	6.3	6.8	6.6	6.1	6.2	9.1	6.2	8.8	5.2	10.0	5.6	9.5
22	6.9	7.7	6.8	7.0	6.3	6.1	5.9	7.6	5.7	7.8	5.9	nan	5.5	11.1
23	7.0	7.5	6.3	6.8	6.3	7.3	6.3	7.2	5.9	8.6	5.6	8.9	5.4	10.2
24	7.0	7.5	6.3	6.8	6.3	7.3	5.8	10.3	5.8	10.0	5.7	9.0	5.5	10.0
25	7.0	7.5	6.6	7.3	5.9	8.7	5.8	8.7	5.8	10.3	5.6	8.7	5.2	nan
26	7.0	7.5	6.6	7.3	5.9	8.7	6.2	7.7	6.0	9.0	5.8	11.2	5.5	9.8
27	7.0	7.5	6.6	7.3	5.9	8.9	5.9	8.9	5.8	10.3	6.1	9.4	5.5	11.1
28	7.0	7.5	6.3	6.8	6.3	7.3	6.1	6.2	5.6	8.9	5.5	9.4	5.3	8.8
29	7.1	7.3	6.3	6.8	5.9	8.7	6.1	9.0	5.9	10.0	5.8	9.6	5.2	14.4
30	6.9	7.7	6.5	6.6	6.3	7.3	6.2	7.0	6.2	7.0	5.9	7.9	5.4	11.0
31	6.9	6.8	6.1	7.1	5.9	7.6	5.3	9.0	5.6	8.9	6.0	7.9	5.8	nan
32	6.9	7.7	6.9	7.7	6.8	6.9	6.6	7.6	6.2	8.9	6.2	8.1	5.9	9.9
33	6.8	8.1	6.9	8.5	6.8	6.9	6.8	7.8	6.3	8.6	5.8	10.9	5.3	13.2
34	6.9	6.8	6.6	7.6	6.5	6.2	5.9	7.5	5.8	10.4	5.8	7.1	5.4	8.5
35	6.9	8.0	6.8	7.0	6.3	7.0	6.4	7.2	5.8	7.8	5.8	15.5	5.7	nan
36	6.7	7.9	6.7	6.8	6.6	8.9	6.5	9.5	5.7	8.0	6.0	10.7	5.8	12.2
37	6.7	7.9	6.5	6.6	6.3	7.0	6.3	7.4	6.2	10.9	5.8	8.6	5.7	10.5
38	6.4	6.9	6.3	6.6	6.0	7.5	6.3	7.4	6.1	10.8	5.5	7.7	5.3	14.2
39	7.2	7.9	6.6	6.3	6.3	8.4	6.1	7.4	5.9	10.1	5.4	13.7	5.3	14.3
40	6.6	9.3	6.7	7.0	6.7	7.4	6.5	8.5	6.0	8.5	6.3	8.7	6.1	8.8
41	6.6	8.3	6.7	7.0	6.7	7.6	6.5	8.5	6.1	8.8	5.4	13.7	5.3	14.3
42	6.9	8.0	6.3	6.8	6.2	9.5	6.0	8.0	5.7	7.6	5.8	12.3	5.3	15.0
43	6.5	7.6	6.3	6.6	6.3	7.0	6.4	7.0	6.0	8.4	5.7	10.8	nan	9.1
44	7.0	7.5	6.3	6.8	6.3	7.3	6.6	7.2	6.2	7.6	nan	8.3	5.1	13.2
45	6.9	7.7	6.9	7.5	6.6	9.9	6.1	7.8	5.7	7.7	5.6	9.3	5.2	11.3

Table G.4 Performance of Supervised Model by KMeans with Grid Search

		Clusters	3		4	
Static Feature	Dynamic Feature	Economic Feature	in	out	in	out
Monday	biderkrp	last1	6.43	8.15	6.22	8.27
Monday	biderkrp	last2	6.43	8.40	5.90	9.06
Monday	biderkrp	last3	6.46	8.06	4.77	8.07
Monday	biderkrp	last4	6.52	8.47	4.88	15.32
Monday	biderkrp	last5	6.52	8.44	6.49	7.59
Monday	biderkrp	last6	6.52	8.47	6.50	9.43
Monday	biderkrp	last7	6.52	8.47	4.78	8.65
Monday	biderkrp	last8	6.52	8.47	6.49	8.78
Monday	biderkrp	last9	6.52	8.47	6.49	8.99
Monday	biderkrp	last10	6.52	8.47	6.45	8.86
Monday	biderkrp	binlast1	6.42	8.45	6.43	9.57
Monday	biderkrp	binlast2	6.52	8.28	6.62	8.17
Monday	biderkrp	binlast3	6.52	8.47	6.41	8.46
Monday	biderkrp	binlast4	6.45	8.02	6.43	8.56
Monday	biderkrp	binlast5	6.52	8.47	6.56	9.23
Monday	biderkrp	binlast6	6.52	8.47	5.74	7.32
Monday	biderkrp	binlast7	6.51	8.44	5.90	8.48
Monday	biderkrp	binlast8	6.52	8.47	6.39	9.05
Monday	biderkrp	binlast9	6.52	8.56	6.56	8.41
Monday	biderkrp	binlast10	6.52	8.47	6.20	7.69
Monday	lastbidkrp	last1	6.59	8.97	6.47	8.79
Monday	lastbidkrp	last2	6.47	8.56	6.44	9.62
Monday	lastbidkrp	last3	4.29	17.96	4.93	16.00
Monday	lastbidkrp	last4	3.72	22.53	4.50	19.53
Monday	lastbidkrp	last5	6.63	8.31	6.41	8.38
Monday	lastbidkrp	last6	6.48	9.08	4.60	11.83
Monday	lastbidkrp	last7	6.49	9.15	5.17	9.85
Monday	lastbidkrp	last8	6.64	8.21	6.45	10.47
Monday	lastbidkrp	last9	6.54	8.58	6.59	8.70
Monday	lastbidkrp	last10	4.39	9.72	4.97	10.18
Monday	lastbidkrp	binlast1	6.67	8.23	6.55	8.22
Monday	lastbidkrp	binlast2	6.68	8.07	6.47	10.70
Monday	lastbidkrp	binlast3	6.46	9.38	6.28	9.43
Monday	lastbidkrp	binlast4	6.70	8.25	6.49	9.43
…	…	…	…	…	…	…

APPENDIX H. PERFORMANCE OF SUPERVISED MODEL BY HIERARCHICALCLUSTERING AND FSA

Table H.1 Performance of Supervised Model by Hierarchical with SelecKBest

Cluster	3		4		5		6		7		8		9	
Features	in	out	in	out	in	out	in	out	in	out	in	out	in	out
11	6.5	16.8	6.3	15.1	6.6	19.7	nan	17.0	nan	16.3	nan	17.7	nan	14.8
12	6.5	15.7	6.3	13.9	6.6	14.7	nan	20.0	nan	16.8	nan	19.0	nan	21.0
13	6.7	15.6	6.9	15.4	6.6	17.5	6.6	15.3	nan	15.2	nan	15.7	nan	16.6
14	6.8	12.4	6.7	11.1	6.7	13.8	6.4	12.6	nan	14.5	nan	19.5	nan	23.6
15	6.6	7.0	6.6	12.6	6.8	13.2	6.5	17.1	6.3	17.2	6.0	19.5	5.6	18.1
16	6.7	13.4	6.4	12.3	6.6	12.2	6.4	18.0	6.2	21.5	5.9	17.3	5.5	16.8
17	6.4	8.5	6.5	13.4	6.7	13.2	6.4	17.7	6.2	22.2	5.9	20.3	5.5	19.8
18	6.7	8.5	6.6	14.4	6.4	14.5	6.2	14.3	6.3	17.4	6.1	12.9	5.8	11.9
19	7.0	6.4	6.9	16.1	6.7	14.4	6.2	19.8	6.0	15.1	5.7	17.2	5.9	18.3
20	6.8	8.6	6.6	12.7	6.8	16.8	6.6	16.1	6.3	16.6	6.2	24.8	nan	19.5
21	6.8	8.6	6.6	12.7	6.4	14.6	6.6	17.6	6.3	16.9	5.4	18.0	nan	20.3
22	6.8	8.6	6.6	11.5	6.8	15.1	6.6	17.5	6.3	17.6	6.2	19.6	nan	19.0
23	6.8	8.6	6.6	11.5	6.6	12.6	6.3	14.1	6.3	17.6	nan	18.6	nan	22.4
24	6.8	8.6	6.6	11.5	6.6	12.8	6.3	14.1	6.3	17.6	nan	18.6	nan	23.9
25	6.8	8.6	6.6	11.5	6.6	11.2	6.6	17.2	nan	17.8	nan	18.6	nan	19.3
26	6.8	8.6	6.6	8.4	6.6	11.3	6.6	10.3	nan	12.2	nan	13.8	nan	15.4
27	6.8	8.6	6.6	8.4	6.6	11.3	6.6	10.3	nan	12.2	nan	12.4	nan	14.2
28	6.8	8.6	6.6	8.4	6.6	11.3	6.6	10.3	nan	12.2	nan	12.4	nan	14.2
29	6.8	8.6	6.6	8.4	6.6	11.3	6.6	10.3	nan	16.8	nan	18.6	nan	19.7
30	6.7	12.9	6.5	15.4	6.6	14.2	6.4	16.9	nan	17.8	nan	14.9	nan	16.0
31	6.7	14.7	6.5	17.7	6.6	16.2	6.4	15.2	nan	14.9	nan	13.1	nan	13.8
32	6.7	12.0	6.5	14.5	6.6	13.5	6.4	13.3	nan	13.3	nan	13.5	nan	15.2
33	6.1	11.1	6.4	10.2	6.3	12.7	6.1	15.5	5.6	14.9	5.1	15.7	nan	26.6
34	7.0	10.6	6.7	11.1	6.2	14.7	6.2	18.7	5.7	21.6	5.1	22.5	4.6	20.5
35	7.0	10.6	6.7	11.1	6.2	15.9	6.2	13.1	5.7	17.3	5.0	24.4	4.6	25.6
36	7.0	10.6	6.7	11.1	6.2	15.9	6.2	13.1	5.7	16.7	5.0	22.5	4.6	26.2
37	7.0	10.8	6.7	11.2	6.2	16.2	6.2	15.1	5.7	19.0	5.0	23.2	4.7	33.4
38	7.0	11.9	6.7	13.4	6.6	11.3	6.1	14.9	5.7	14.2	5.0	13.0	4.6	19.7
39	7.0	10.1	6.7	10.0	6.2	11.2	6.2	14.2	5.7	19.0	5.1	18.2	4.7	18.7
40	7.0	10.1	6.7	10.0	6.2	11.2	6.2	14.2	5.7	23.8	5.1	25.0	4.7	30.4
41	7.2	14.6	6.9	10.4	6.6	12.2	6.3	13.6	5.8	17.5	5.1	26.9	4.8	38.3
42	7.2	12.9	6.9	17.2	6.6	17.0	6.3	14.5	5.8	11.9	5.1	20.0	4.8	19.4
43	7.2	14.6	6.9	10.4	6.6	12.3	6.3	18.1	5.8	19.2	5.1	19.4	4.8	20.7
44	5.7	13.9	5.8	13.7	5.9	22.2	5.8	23.4	6.0	22.1	5.7	16.0	5.5	21.0
45	5.7	10.9	5.8	12.2	5.9	18.6	5.8	21.1	6.0	23.9	5.7	21.2	5.5	29.0

Table H.2 Performance of Supervised Model by Hierarchical with RFE

Cluster	3		4		5		6		7		8		9	
Features	**in**	**out**	**in**	**out**	**in**	**out**	**in**	**out**	**in**	**out**	**in**	**out**	**in**	**out**
11	6.2	7.1	6.3	8.0	6.0	13.8	5.9	23.8	5.7	22.7	5.6	23.9	5.5	22.8
12	6.4	6.4	6.6	11.3	6.2	12.8	6.1	14.7	6.0	15.6	5.8	19.8	5.7	23.3
13	6.2	10.6	6.4	12.0	6.5	13.9	6.1	19.3	5.9	17.5	5.8	17.0	5.6	22.1
14	6.2	10.6	6.4	7.1	6.1	8.0	5.8	9.2	5.7	14.0	5.7	16.3	5.3	21.7
15	6.2	6.2	6.4	6.0	6.3	6.3	6.2	17.6	5.9	18.4	5.7	21.6	5.6	21.0
16	6.0	7.3	6.4	10.2	6.3	12.2	6.3	14.6	5.8	20.9	5.4	28.8	5.1	34.8
17	6.0	7.3	6.4	10.0	6.3	11.5	6.3	14.2	5.8	18.6	5.4	23.5	5.1	23.5
18	6.0	7.3	6.4	10.0	6.3	11.5	6.3	9.9	5.8	17.0	5.4	23.8	5.1	28.8
19	6.0	7.3	6.4	9.5	6.3	9.4	6.2	7.4	5.7	16.9	5.7	24.0	5.7	21.9
20	6.0	7.3	6.1	9.6	6.0	10.9	6.0	16.9	5.6	22.3	5.6	18.6	5.2	20.7
21	6.0	7.3	6.1	9.7	6.0	15.6	6.0	12.1	5.6	18.2	5.6	18.9	5.2	21.0
22	6.0	7.3	6.1	9.1	6.0	13.9	6.0	11.9	5.6	18.3	5.7	14.3	5.3	14.5
23	6.2	10.3	6.1	9.4	6.0	11.7	6.2	12.8	5.7	17.1	5.5	22.7	5.1	27.1
24	6.1	6.4	6.4	10.5	6.4	11.7	6.4	11.4	5.9	20.3	5.4	23.4	5.1	21.1
25	5.9	7.5	6.3	11.3	6.2	13.1	5.7	15.7	5.7	17.0	5.3	17.9	5.1	20.9
26	6.1	6.4	6.4	10.5	6.4	12.1	6.3	12.1	5.8	12.8	5.4	16.5	5.2	19.9
27	6.2	7.2	6.4	7.2	6.3	12.8	6.3	20.1	5.8	28.7	5.4	29.6	5.4	29.9
28	6.2	16.1	6.6	15.7	6.4	18.5	6.2	17.7	5.6	19.3	5.3	25.0	5.2	24.5
29	6.2	7.6	6.6	10.8	6.4	13.2	6.2	13.5	5.7	15.8	5.4	21.5	5.3	28.2
30	6.2	7.6	6.6	10.8	6.4	13.2	6.2	13.5	5.7	15.8	5.4	21.5	5.3	28.2
31	6.2	7.6	6.6	10.8	6.4	13.2	6.2	13.5	5.7	15.8	5.4	21.5	5.3	28.2
32	6.2	7.6	6.5	7.2	6.4	8.9	5.8	18.9	5.8	17.8	5.4	23.9	5.4	24.4
33	6.2	8.1	6.3	17.7	6.2	24.7	6.2	22.7	5.9	25.2	5.8	25.1	5.8	23.5
34	6.2	8.1	6.3	17.7	6.2	24.7	6.2	21.5	6.1	21.5	6.1	26.6	5.9	30.2
35	6.2	11.7	6.3	17.2	6.2	22.6	6.2	22.8	6.1	21.2	6.1	24.1	5.9	26.2
36	6.2	11.5	6.2	22.3	6.1	23.1	6.1	22.0	6.1	19.8	5.9	24.6	5.9	22.2
37	6.2	11.5	6.2	17.3	6.1	19.1	6.1	19.8	6.1	23.3	5.9	21.2	5.9	25.8
38	6.0	10.8	6.0	15.0	6.1	15.6	6.1	22.6	5.7	22.4	5.6	20.4	5.6	24.8
39	5.8	10.7	5.9	21.7	5.8	20.3	5.9	21.0	5.9	19.0	5.8	24.7	5.8	22.8
40	5.8	17.7	5.9	26.5	5.8	25.0	5.9	24.8	5.9	22.8	5.8	21.9	5.8	24.7
41	6.9	6.9	6.0	11.2	6.2	12.2	6.2	14.1	6.0	12.8	5.8	16.3	5.8	15.9
42	7.2	11.1	6.3	12.6	6.1	12.2	5.8	18.2	5.7	21.0	5.3	21.6	5.1	23.6
43	6.0	8.9	6.1	8.2	6.0	10.3	5.9	12.5	5.8	13.4	5.5	17.9	5.3	17.6
44	5.9	10.0	6.1	10.8	5.9	13.3	5.9	17.4	5.8	18.3	5.5	19.1	5.3	21.6
45	6.5	12.3	6.4	19.5	6.4	16.0	6.4	16.9	6.5	18.2	6.1	17.9	5.4	18.7

Table H.3 Performance of Supervised Model by Hierarchical with PFA

Cluster	3		4		5		6		7		8		9	
Features	in	out	in	out	in	out	in	out	in	out	in	out	in	out
11	6.8	11.0	6.6	11.9	6.0	15.4	6.1	20.5	5.9	19.7	5.9	22.5	5.7	23.4
12	6.7	14.3	6.8	9.7	6.8	15.0	6.5	17.0	6.2	20.4	5.8	16.3	5.6	19.8
13	7.0	7.6	6.7	10.3	6.5	15.4	6.0	19.1	5.5	26.7	5.4	26.3	5.5	31.2
14	7.1	12.6	6.8	14.0	6.5	18.2	5.7	16.0	5.3	19.0	5.3	19.0	5.3	24.9
15	7.1	7.2	6.8	11.7	6.5	13.5	6.4	13.2	6.3	18.7	5.8	18.5	5.4	23.8
16	6.1	6.3	6.5	9.3	6.3	10.5	5.8	13.0	5.8	13.6	5.7	20.4	5.3	25.4
17	6.1	17.0	6.0	19.8	6.0	18.6	5.9	28.4	5.8	28.2	5.8	29.8	5.6	33.6
18	6.7	9.6	5.5	11.5	5.6	12.1	5.7	12.7	5.7	12.9	5.6	10.9	5.7	15.1
19	6.6	7.2	6.4	11.9	6.3	15.9	6.5	15.8	6.0	14.8	5.7	14.0	5.5	20.4
20	6.5	7.2	6.4	11.9	6.3	15.8	6.5	15.9	6.0	14.9	5.7	14.0	5.5	20.4
21	6.8	13.8	6.1	15.4	6.2	14.1	6.0	21.3	5.9	22.2	5.7	19.6	5.6	26.3
22	7.0	11.4	6.1	7.5	6.1	12.3	6.0	15.0	5.6	15.6	5.4	13.0	5.3	21.6
23	6.7	13.3	6.1	14.4	6.3	15.5	6.1	16.9	5.9	18.4	5.7	19.2	5.4	18.7
24	6.1	6.3	5.8	10.3	5.9	9.9	5.9	18.4	5.9	21.3	6.0	20.2	5.7	19.7
25	6.1	8.5	6.2	10.1	6.4	11.9	6.0	15.8	6.0	18.1	5.6	21.8	5.7	23.7
26	6.9	11.7	5.9	15.5	5.7	13.5	5.9	14.4	5.5	17.2	5.4	19.6	5.3	23.7
27	6.9	9.7	5.9	11.1	5.7	10.0	5.9	11.7	5.5	13.9	5.4	17.7	5.3	21.3
28	6.0	6.2	6.1	9.0	6.2	7.7	6.4	8.8	6.2	15.5	5.9	21.7	5.8	23.2
29	5.8	11.6	6.1	10.7	6.1	11.3	6.2	13.6	5.5	12.2	5.3	13.0	5.4	16.0
30	5.9	9.8	6.0	9.8	6.2	12.0	5.9	8.5	6.0	12.6	5.9	15.1	5.7	17.6
31	5.9	7.7	6.0	10.7	5.7	8.6	5.7	15.0	5.8	12.0	5.8	17.8	5.8	22.0
32	5.9	6.7	6.0	10.1	5.6	8.9	5.7	17.0	5.7	16.7	5.7	17.4	5.8	16.1
33	6.8	10.7	5.9	9.5	6.0	14.7	6.1	18.2	5.8	20.1	5.8	21.5	5.6	21.5
34	6.8	8.5	5.9	12.0	6.1	14.2	6.1	13.7	5.8	15.3	5.3	19.0	5.3	19.3
35	6.9	11.3	6.0	9.9	5.5	18.1	5.8	14.3	5.8	22.0	5.6	20.1	5.6	18.1
36	6.8	9.5	6.7	11.0	5.9	14.7	5.9	15.2	5.5	27.7	5.5	27.5	5.5	26.2
37	5.9	13.2	6.2	8.9	6.2	11.4	6.1	13.0	5.3	36.3	5.4	36.0	5.0	35.9
38	6.8	7.7	6.7	9.3	5.9	11.8	5.9	14.9	5.5	26.4	5.5	27.1	5.5	25.2
39	6.9	15.0	5.9	10.0	5.6	13.7	5.9	16.8	5.8	13.3	5.5	19.5	5.3	23.1
40	6.8	10.3	6.3	16.5	6.5	16.7	6.3	18.2	5.9	26.4	5.7	27.5	5.0	26.9
41	7.1	10.0	6.6	10.4	6.3	15.1	6.5	18.3	6.2	21.9	5.3	20.4	4.8	25.7
42	5.7	10.1	5.8	11.6	6.1	11.6	5.9	13.5	5.6	15.7	5.2	19.3	5.3	27.2
43	5.7	14.0	5.8	16.2	6.1	15.8	5.9	13.0	5.6	19.3	5.2	22.2	5.3	20.6
44	5.7	14.0	6.1	12.8	6.0	14.8	6.1	15.8	5.8	16.4	5.7	13.4	5.2	16.8
45	6.0	10.9	6.0	11.5	6.0	20.4	5.9	21.3	6.1	21.6	5.8	17.0	5.6	17.4

APPENDIX I. NON-LINEARITY IN NUMBER OF BIDDERS

Table I.1 Regression Result of Supervised Model with Square of Number of Bidders

OLS Regression Results

Dep. Variable:	price	R-squared:	0.345
Model:	OLS	Adj. R-squared:	0.302
Method:	Least Squares	F-statistic:	8.14
Date:	31 Jan 2019	Prob (F-statistic):	0.00
Time:	07:45:35 PM	Log-Likelihood:	-2007
No. Observations:	380	AIC:	4062
Df Residuals:	356	BIC:	4157
Df Model:	23		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
extras	23.30	5.80	4.02	0.00	11.90	34.70
biderkrp2	0.03	0.03	1.14	0.26	-0.03	0.10
last4	0.07	0.04	1.62	0.11	-0.02	0.15
binlast6	0.09	0.05	1.97	0.05	0.00	0.19
binlast8	0.09	0.05	1.68	0.09	-0.02	0.19
selrate	-0.01	0.01	-1.62	0.11	-0.03	0.00
neg	-5.36	2.79	-1.92	0.06	-10.85	0.13
views	0.02	0.01	3.25	0.00	0.01	0.03
k0	91.37	18.42	4.96	0.00	55.15	127.58
k1	96.20	18.54	5.19	0.00	59.75	132.65
k2	101.97	24.68	4.13	0.00	53.42	150.51
k3	115.45	19.69	5.86	0.00	76.72	154.18
d1	80.24	15.44	5.20	0.00	49.88	110.61
d3	85.94	15.43	5.57	0.00	55.60	116.27
d5	92.70	15.95	5.81	0.00	61.34	124.07
d7	86.11	14.48	5.95	0.00	57.63	114.58
d10	59.99	18.63	3.22	0.00	23.35	96.64
smonday	60.99	11.95	5.11	0.00	37.50	84.49
stuesday	58.97	11.95	4.94	0.00	35.47	82.46
swednesday	53.56	12.04	4.45	0.00	29.87	77.24
sthursday	54.04	12.61	4.29	0.00	29.24	78.83
sfriday	61.63	12.51	4.93	0.00	37.03	86.22
ssaturday	50.61	11.68	4.33	0.00	27.64	73.58
ssunday	65.20	11.81	5.52	0.00	41.98	88.42
lastbidkrp	0.32	0.03	10.34	0.00	0.26	0.38
lastbidderratekrp	0.01	0.01	2.05	0.04	0.00	0.03

Table I.2 Regression Result of Supervised Model with Square Root of Number of Bidders

OLS Regression Results

Dep. Variable:	price	R-squared:	0.367
Model:	OLS	Adj. R-squared:	0.326
Method:	Least Squares	F-statistic:	8.973
Date:	31 Jan 2019	Prob (F-statistic):	0.00
Time:	07:52:38 PM	Log-Likelihood:	-2000.4
No. Observations:	380	AIC:	4049
Df Residuals:	356	BIC:	4143
Df Model:	23		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
extras	20.80	5.74	3.63	0.00	9.52	32.09
biderkrp2	7.90	2.12	3.73	0.00	3.73	12.07
last4	0.06	0.04	1.41	0.16	-0.02	0.14
binlast6	0.09	0.05	1.96	0.05	0.00	0.18
binlast8	0.09	0.05	1.85	0.07	-0.01	0.20
selrate	-0.01	0.01	-1.48	0.14	-0.03	0.00
neg	-6.02	2.74	-2.19	0.03	-11.41	-0.62
views	0.02	0.01	3.29	0.00	0.01	0.03
k0	86.58	18.14	4.77	0.00	50.91	122.25
k1	89.29	18.31	4.88	0.00	53.28	125.31
k2	94.59	24.31	3.89	0.00	46.78	142.39
k3	109.62	19.42	5.65	0.00	71.44	147.80
d1	73.85	15.27	4.84	0.00	43.81	103.89
d3	80.24	15.21	5.27	0.00	50.32	110.16
d5	87.78	15.71	5.59	0.00	56.88	118.67
d7	80.79	14.30	5.65	0.00	52.66	108.91
d10	57.43	18.32	3.13	0.00	21.40	93.46
smonday	57.13	11.79	4.85	0.00	33.95	80.31
stuesday	55.26	11.77	4.70	0.00	32.13	78.40
swednesday	50.27	11.87	4.23	0.00	26.92	73.61
sthursday	50.31	12.42	4.05	0.00	25.90	74.73
sfriday	59.45	12.30	4.83	0.00	35.26	83.65
ssaturday	45.85	11.54	3.97	0.00	23.15	68.54
ssunday	61.81	11.64	5.31	0.00	38.92	84.70
lastbidkrp	0.34	0.03	10.96	0.00	0.28	0.40
lastbidderratekrp	0.01	0.01	1.79	0.07	0.00	0.03

	P-Value	**Linearity**
Non-Linearity Tests-Linear Rainbow	0.03	Model is Not Linear
Non-Linearity Tests-LM	0.88	Model is Linear

APPENDIX J. REGRESSION RESULTS OF SUPERVISED MODEL FOR EACH CLUSTER

Table J.1 Regression Result of Supervised Model for Cluster0

OLS Regression Results			
Dep. Variable:	price	R-squared:	0.608
Model:	OLS	R-squared:	0.163
Method:	Least Squares	F-statistic:	1.366
Date:	Tue 29 Jan 2019	Prob (F-statistic):	0.274
Time:	09:47:18 PM	Log-Likelihood:	-161.8
No. Observations:	33	AIC:	359.6
Df Residuals:	15	BIC:	386.5
Df Model:	17		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
extras	-15.22	28.22	-0.54	0.60	-75.36	44.93
biderkrp	0.33	2.00	0.17	0.87	-3.93	4.59
last4	0.09	0.21	0.44	0.67	-0.36	0.54
binlast6	0.10	0.17	0.59	0.56	-0.26	0.47
binlast8	0.03	0.23	0.12	0.91	-0.46	0.52
selrate	0.04	0.07	0.57	0.58	-0.11	0.18
neg	9.32	19.61	0.48	0.64	-32.48	51.12
views	-0.01	0.04	-0.24	0.81	-0.09	0.07
k0	76.96	45.90	1.68	0.11	-20.88	174.79
k1	0.00	0.00	-1.49	0.16	0.00	0.00
k2	59.13	56.07	1.06	0.31	-60.38	178.64
k3	0.00	0.00	1.46	0.17	0.00	0.00
d1	0.00	0.00	-1.47	0.16	0.00	0.00
d3	136.08	83.45	1.63	0.12	-41.79	313.96
d5	0.00	0.00	-0.95	0.36	0.00	0.00
d7	0.00	0.00	1.10	0.29	0.00	0.00
d10	0.00	0.00	-1.14	0.27	0.00	0.00
smonday	46.59	47.28	0.99	0.34	-54.18	147.36
stuesday	4.03	27.77	0.15	0.89	-55.15	63.22
swednesday	15.24	21.28	0.72	0.49	-30.12	60.61
sthursday	-0.42	28.06	-0.02	0.99	-60.24	59.40
sfriday	0.42	39.54	0.01	0.99	-83.85	84.70
ssaturday	90.55	66.79	1.36	0.20	-51.81	232.90
ssunday	-20.33	28.70	-0.71	0.49	-81.49	40.84
lastbidkrp	0.46	0.20	2.36	0.03	0.05	0.88
lastbidderratekrp	0.01	0.03	0.21	0.84	-0.06	0.07

Table J.2 Regression Result of Supervised Model for Cluster1

OLS Regression Results

Dep. Variable:	price	R-squared:	0.719
Model:	OLS	R-squared:	0.482
Method:	Least Squares	F-statistic:	3.041
Date:	Tue 29 Jan 2019	Prob (F-statistic):	0.00443
Time:	09:50:32 PM	Log-Likelihood:	-217.46
No. Observations:	47	AIC:	478.9
Df Residuals:	25	BIC:	519.6
Df Model:	21		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
extras	-37.82	20.31	-1.86	0.07	-79.7	4.0
biderkrp	2.66	1.49	1.79	0.09	-0.40	5.72
last4	0.07	0.10	0.74	0.47	-0.13	0.27
binlast6	0.09	0.15	0.62	0.54	-0.22	0.41
binlast8	-0.04	0.15	-0.26	0.80	-0.34	0.26
selrate	-0.01	0.01	-0.95	0.35	-0.04	0.02
neg	-19.18	21.05	-0.91	0.37	-62.54	24.17
views	0.04	0.01	3.75	0.00	0.02	0.06
k0	0.00	0.00	2.18	0.04	0.00	0.00
k1	0.00	0.00	-2.54	0.02	0.00	0.00
k2	246.59	85.03	2.90	0.01	71.47	421.71
k3	240.81	84.38	2.85	0.01	67.02	414.59
d1	82.41	30.02	2.75	0.01	20.60	144.23
d3	118.74	34.69	3.42	0.00	47.31	190.18
d5	100.94	40.62	2.49	0.02	17.28	184.60
d7	109.51	31.07	3.52	0.00	45.52	173.50
d10	75.79	57.99	1.31	0.20	-43.64	195.22
smonday	76.13	25.54	2.98	0.01	23.52	128.74
stuesday	85.08	31.28	2.72	0.01	20.66	149.50
swednesday	53.13	31.69	1.68	0.11	-12.13	118.39
sthursday	80.69	31.75	2.54	0.02	15.29	146.08
sfriday	24.39	31.70	0.77	0.45	-40.89	89.68
ssaturday	93.18	26.95	3.46	0.00	37.67	148.68
ssunday	74.80	22.91	3.27	0.00	27.61	121.98
lastbidkrp	0.16	0.08	2.00	0.06	-0.01	0.33
lastbidderratekrp	0.04	0.01	3.34	0.00	0.0	0.1

Table J.3 Regression Result of Supervised Model for Cluster2

OLS Regression Results			
Dep. Variable:	price	R-squared:	0.496
Model:	OLS	R-squared:	0.383
Method:	Least Squares	F-statistic:	4.4
Date:	Tue 29 Jan 2019	Prob (F-statistic):	1.03E-06
Time:	09:52:29 PM	Log-Likelihood:	-534.67
No. Observations:	105	AIC:	1109
Df Residuals:	85	BIC:	1162
Df Model:	19		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
extras	28.81	10.08	2.86	0.01	8.8	48.8
biderkrp	1.01	1.19	0.85	0.40	-1.35	3.37
last4	0.07	0.08	0.94	0.35	-0.08	0.22
binlast6	0.06	0.08	0.71	0.48	-0.10	0.22
binlast8	-0.01	0.09	-0.08	0.94	-0.19	0.18
selrate	-0.01	0.01	-0.96	0.34	-0.04	0.01
neg	-7.12	11.41	-0.62	0.53	-29.80	15.56
views	0.01	0.01	1.24	0.22	-0.01	0.03
k0	0.00	0.00	0.40	0.69	0.00	0.00
k1	180.97	54.49	3.32	0.00	72.64	289.31
k2	0.00	0.00	-2.89	0.01	0.00	0.00
k3	0.00	0.00	-3.66	0.00	0.00	0.00
d1	45.96	18.28	2.52	0.01	9.62	82.31
d3	58.31	17.03	3.42	0.00	24.45	92.18
d5	0.00	0.00	3.22	0.00	0.00	0.00
d7	60.19	15.84	3.80	0.00	28.69	91.69
d10	16.50	29.02	0.57	0.57	-41.19	74.20
smonday	33.17	12.50	2.65	0.01	8.31	58.02
stuesday	34.95	15.09	2.32	0.02	4.95	64.95
swednesday	5.77	12.91	0.45	0.66	-19.89	31.43
sthursday	15.64	14.03	1.12	0.27	-12.24	43.53
sfriday	26.96	15.57	1.73	0.09	-4.00	57.92
ssaturday	30.43	12.73	2.39	0.02	5.12	55.74
ssunday	34.05	13.12	2.60	0.01	7.96	60.14
lastbidkrp	0.41	0.06	7.01	0.00	0.29	0.53
lastbidderratekrp	0.03	0.02	1.75	0.08	0.0	0.1

Table J.4 Regression Result of Supervised Model for Cluster3

OLS Regression Results

Dep. Variable:	price	R-squared:	0.313
Model:	OLS	R-squared:	0.192
Method:	Least Squares	F-statistic:	2.592
Date:	Tue 29 Jan 2019	Prob (F-statistic):	0.00107
Time:	09:53:48 PM	Log-Likelihood:	-687.49
No. Observations:	128	AIC:	1415
Df Residuals:	108	BIC:	1472
Df Model:	19		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
extras	15.05	12.46	1.21	0.23	-9.6	39.8
biderkrp	2.75	1.25	2.20	0.03	0.27	5.22
last4	0.08	0.10	0.81	0.42	-0.11	0.27
binlast6	0.18	0.10	1.71	0.09	-0.03	0.38
binlast8	0.15	0.11	1.41	0.16	-0.06	0.36
selrate	-0.04	0.04	-1.02	0.31	-0.11	0.03
neg	-7.99	11.22	-0.71	0.48	-30.23	14.25
views	0.02	0.04	0.61	0.55	-0.06	0.11
k0	57.73	52.56	1.10	0.28	-46.46	161.91
k1	0.00	0.00	1.90	0.06	0.00	0.00
k2	137.69	68.75	2.00	0.05	1.41	273.97
k3	0.00	0.00	2.04	0.04	0.00	0.00
d1	61.08	37.72	1.62	0.11	-13.68	135.84
d3	0.00	0.00	-0.09	0.93	0.00	0.00
d5	0.00	0.00	-1.72	0.09	0.00	0.00
d7	80.68	37.91	2.13	0.04	5.53	155.82
d10	53.66	45.02	1.19	0.24	-35.57	142.89
smonday	32.08	21.91	1.46	0.15	-11.36	75.52
stuesday	29.77	19.18	1.55	0.12	-8.24	67.78
swednesday	21.64	21.81	0.99	0.32	-21.58	64.87
sthursday	26.81	21.75	1.23	0.22	-16.30	69.92
sfriday	44.47	18.86	2.36	0.02	7.08	81.86
ssaturday	13.45	17.82	0.76	0.45	-21.88	48.78
ssunday	27.20	20.21	1.35	0.18	-12.85	67.26
lastbidkrp	0.30	0.06	4.67	0.00	0.17	0.43
lastbidderratekrp	0.01	0.03	0.43	0.67	0.0	0.1

Table J.5 Regression Result of Supervised Model for Cluster4

OLS Regression Results

Dep. Variable:	price	R-squared:	0.487
Model:	OLS	R-squared:	0.264
Method:	Least Squares	F-statistic:	2.185
Date:	Tue 29 Jan 2019	Prob (F-statistic):	0.0146
Time:	09:55:58 PM	Log-Likelihood:	-343.72
No. Observations:	67	AIC:	729.4
Df Residuals:	46	BIC:	775.7
Df Model:	20		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
extras	4.71	17.18	0.27	0.79	-29.9	39.3
biderkrp	1.66	1.61	1.04	0.31	-1.57	4.89
last4	-0.02	0.13	-0.16	0.87	-0.27	0.23
binlast6	-0.03	0.12	-0.29	0.77	-0.28	0.21
binlast8	0.19	0.18	1.06	0.30	-0.17	0.54
selrate	-0.01	0.04	-0.13	0.90	-0.09	0.08
neg	-13.26	17.60	-0.75	0.46	-48.69	22.17
views	-0.03	0.04	-0.73	0.47	-0.11	0.05
k0	150.14	46.25	3.25	0.00	57.05	243.24
k1	145.24	48.20	3.01	0.00	48.22	242.27
k2	60.16	70.99	0.85	0.40	-82.74	203.06
k3	0.00	0.00	-2.89	0.01	0.00	0.00
d1	142.19	50.49	2.82	0.01	40.57	243.81
d3	0.00	0.00	-3.33	0.00	0.00	0.00
d5	120.08	46.96	2.56	0.01	25.55	214.61
d7	0.00	0.00	2.98	0.01	0.00	0.00
d10	93.28	52.36	1.78	0.08	-12.13	198.68
smonday	55.83	26.72	2.09	0.04	2.05	109.62
stuesday	58.93	23.33	2.53	0.02	11.98	105.89
swednesday	65.49	26.58	2.46	0.02	11.98	118.99
sthursday	38.66	27.45	1.41	0.17	-16.59	93.92
sfriday	66.36	27.26	2.44	0.02	11.49	121.22
ssaturday	5.67	41.32	0.14	0.89	-77.51	88.85
ssunday	64.60	25.48	2.54	0.02	13.33	115.88
lastbidkrp	0.32	0.09	3.64	0.00	0.14	0.50
lastbidderratekrp	0.00	0.01	-0.14	0.89	0.0	0.0

APPENDIX K. CURRICULUM VITAE

EMRAH ÖZ

Ph.D in Economics (2019)

Dissertation Title: Empirical Studies on Price Determinants of Online Auctions with Machine Learning Applications, Middle East Technical University, Ankara, Turkey

M.Sc in Economics (2010)

Thesis Title: Can Relative Yield Curves Predict Exchange Rate Movements? Example From Turkish Financial Market, Middle East Technical University, Ankara, Turkey

B.S in Industrial Engineering (2007)

Honor Degree (3.43 / 4.00), Middle East Technical University, Ankara, Turkey

Süleyman Demirel Anatolian High School (2002)

Ranked the 1st among 100 students with Degree (5.00 / 5.00)

Cost and Pricing Engineer

With broad cost and pricing experience from component level to large project level in automotive and electronic industries.

Skills

Advanced technical modeling skills in machine learning.

APPENDIX L. TURKISH SUMMARY / TÜRKÇE ÖZET

Elektronik ticaret günümüzde konvansiyonel ticaretin yerini hızla almaktadır. Yaygın, hızlı internet, güvenli elektronik ödeme sistemleri ve hızlı kargo sistemleri bu değişimi etkileyen başlıca faktörlerdir. Satıcılar internet üzerinden ürünlerini kolaylıkla pazarlayabilmekte, alıcılar ise birçok alternatif ürünü zahmetsizce karşılaştırabilmekte ve satın alma tercihinde bulunabilmektedir. Bu nedenle e-ticaret firmaları gün geçtikçe daha da değerlenmektedir. Hatta bazı e-ticaret firmalarının piyasa değeri 1 trilyon doları bile geçmiştir. İstatistiklere göre global e-ticaret son yıllarda daha hızlı artmakta bu değişim doğal olarak Türkiye'yi de etkilemektedir. Türkiye'de son 5 yıl içinde e-ticaretin toplam perakende satışlara oranı %1.9'dan %4.1'e çıkmıştır. Bu oranın gelecekte daha da artması da beklenmektedir.

Elektronik ticaret olanağı sağlayan firmalar genellikle iki farklı fiyatlandırma stratejisi kullanırlar. Bunlardan birincisi fiyatların ihale ortamında belli olması, ikincisi ise "şimdi satın al" fiyatlandırmasıdır. Ihale fiyatlandırması ilginç dinamiklerden oluşmaktadır. İhale ile satılan ürünlerin fiyatları alıcılar tarafından ihale içi rekabet ortamında oluşur. Şimdi satın al satışlarında ise satıcılar ürünlerine bir etiket fiyatı belirler ve belirlenen fiyattan müşterilerin gelmesini beklerler. Satıcılar benzer ürünler için iki farklı satış stratejisini de kullanabilir hatta aynı gün içinde benzer ürünler için farklı fiyatlar da ortaya çıkabilir. Bu nedenle satış fiyatını etkileyen faktörleri belirlemek ve fiyatı önceden tahmin etmek önemli bir çalışma alanı olarak ortaya çıkmaktadır. Günümüzde bu alandaki açık kaynak makine öğrenme algortimaları ve veri bolluğu hem profesyonellerin hem de akademisyenlerin dikkatini çekmektedir.

Satıcılar fiyatı etkileyen faktörleri önceden bilmesi ile ürünlerin satışlarını arttırabilir ve karlarını maksimize edebilirler. Alıcılar ise istedikleri bir ürünü daha uygun bir fiyata satın alabilirler. Ayrıca arbitraj fırsatı varsa aracı kişiler piyasadaki düzensizlikleri ortadan kaldırabilirler. Ayrıca alıcı ve satıcılara danışmanlık hizmeti sağlanabilir. Bütün bunlar ve daha fazlası ihale fiyatlarını etkileyen faktörleri belirlemek, fiyatları önceden tahmin edebilen modeller geliştirmek üzerine yapılacak çalışmalar için bir motivasyon oluşturmaktadır. Bu çalışmanın ana amacı aşağıdaki sorulara akademik cevaplar bulmaktır.

- Ebay gibi sabit bitiş zamanlı ihalalerin fiyatını etkileyen faktörler nelerdir?
- İhale fiyatları önceden tahmin edilebilir mi?
- Makine öğrenmesi yöntemleri ihale fiyat modellerine ne katabilir?

Bütün bu sorulara cevap ararken ihale başladığında belli olan ve zaman içinde değişmeyen özellikler, ihale zamanı içinde belli olan özellikler ve benzer ürünlerin geçmiş dönemdeki fiyatları kullanılacaktır. Ayrıca, ürün ihale sayfasına girilmiş, ürün hakkındaki açıklamaların ürünün ihale bitiş fiyatına etkisi, ihalelerin ilk döneminde verilmiş teklifler de analiz edilecektir.

Bu çalışmanın literatüre katkısı şöyle özetlenebilir. Benzer ürünlerin geçmiş dönemdeki fiyatları bu tez ile detaylı bir şekilde analiz edilmiştir. İhalede verilen tekliflerin oluşturduğu ihale yolları modellenmiştir. İhale sayfasına girilmiş ürün açıklamaları metin sınıflandırma yöntemleri ile ilk defa derecelendirilmiş ve ihale fiyat modellerine bildiğimiz kadarı ile ilk defa dahil edilmiştir. Doğrudan satış sayfalarında girilen ürün açıklamaları ve fiyat bilgisi ile oluşturulan metin sınıflandırma modeli kullanılarak ilk defa ihale edilen ürünlerin açıklama bilgileri derecelendirilmiştir. Literatürde daha çok ihalenin bittiği günün fiyata etkisi incelenmiştir. Bu tezde ise ihalenin başladığı günün fiyata etkisi analiz edilmiştir. İhalede verilen tekliflerin zamana göre yoğunluğu incelendiğinde ihalenin süresi ne kadar uzun olursa olsun, ihalenin başında ve sonunda yoğun bir şekilde teklif verildiği ancak ara dönemde ise çok teklif verilmediği tespit edilmiştir. İhalenin son döneminde verilen teklifler ihale bitiş fiyatını oluştursa da ihale bitmeden bu bilgilere ulaşmak kolay değildir. Öte yandan ihalenin başında verilen tekliflerin de ihale bitiş fiyatını önemli bir şekilde etkilediği gösterilmiştir. İhalenin ilk döneminde verilen teklifleri veren kişilerin deneyim seviyesinin de ihale fiyatında önemli bir yeri olduğu bulunmuştur. İhale edilen ürünün yanında aksesuar olarak verilen şarj aleti, kutu, kablo, garanti vb. gibi extra parçaların da ihale içinde fiyatlandığı, hiç bir şeyin bedava olmadığı da gösterilmiştir. Ayrıca klasik fiyat modellerine ilave olarak kümeleme işleminin modellerinin tahmin performansını arttırdığı bulunmuştur. Konvansiyonel modellerin yanında ihalede tekliflerin zamana göre oluşturduğu ihale yol grafiklerinden de fiyat tahmin modelinin geliştirilebileceği gösterilmiştir.

Tezde öncelikle ihale ve şimdi satın al satış stratejilerinin benzerlikleri ve farklılıkları ortaya konulmuş daha sonra literatürde bulunan teorik ve pratik çalışmalar anlatılmıştır. Tezde kullanılan veriler 4. bölümde detaylı bir şekilde anlatılmıştır. 5. bölümde ise ihale bitiş fiyatını etkileyen faktörleri tespit etmek için deneysel çalışmalar yapılmıştır. Bu bölüm 3 farklı çalışmadan oluşmaktadır. Model sonuçları 6. bölümde karşılaştırılmış, 7. bölümde ise tezin sonuçları ve gelecekte yapılabilecek çalışmalar özetlenmiştir.

Literatüre göre deneyimli satıcılar şimdi satın al satışlarına, deneyimli alıcılar ise daha çok ihale ile satılan ürünlere yönelmektedir. İhale satışları bazen müşteriler tarafından bir eğlence aracı olarak da görülmektedir. Genellikle fiyatları satıcılar tarafından tam olarak belirlenemeyen ürünler ihale ile satılmaktadır. İkinci el ürünler de bu sınıfa girebilir. Bazı alıcılar sabırsız davranıp doğrudan satış yapılan ürünleri tercih etmektedir. Bütün bunlar ihale ve şimdi satın al satışlarında bir pazar farklılaşması olduğunu göstermektedir. Literatüre göre ihale ile satışlarda fiyatlar ortalama olarak daha düşük ancak satış gerçekleşme oranı ise daha yüksek olmaktadır. Ebay tipi ihalede ihalenin süresi sabittir ve değişmez. Bazı ihale özellikleri satıcı tarafından belirlenir ve ihale boyunca bu özellikler sabit kalır değişmez. Literatürde ihalelerin bu özelliklerine statik özellikler denir. Örneğin ürün sayfasına girilmiş açıklamalar, ihalenin başladığı gün, ihalenin süresi, tekliflerdeki artış oranı gibi. Bazı özellikler ise ihale başlamadan bilinmeyen ve ancak ihale süresi içinde belli olan özelliklerdir. İhaleye katılan kişi sayısı, verilen teklif sayısı, tekliflerin sıklığı gibi. Literatürde bu bilgilere dinamik özellikler denilmektedir. Ebay ihalesinde politika gereği ihaleye katılan bütün oyuncular lider kişinin teklifi hariç bütün teklifleri ve tekliflerin verildiği tarihi görebilmektedir. Lider kişinin verdiği gerçek teklif ihale sayfasında gösterilmez. Lider kişinin teklifi en yüksek ikinci kişinin teklifinin bir miktar üstünde gösterilir. Sistemde görünen liderin teklifi daha yüksek bir teklif geldiğinde otomatik olarak gerçek teklifine kadar yavaş yavaş yükseltir. Bir rakip tarafından liderin gerçek teklifinden daha yüksek bir teklif verilmediği sürece liderin gerçek teklifi sistemde saklı tutulur. Rakip oyuncu liderin gerçek teklifinden daha yüksek bir teklif verdiğinde yeni lider belli olur ve yukarıdaki süreç yeni lider için uygulanır. Ebay ihale sistemi aslında ikinci fiyat ihalesine benzemektedir. Yani en yüksek teklifi veren kişi ihaleyi kazanır ancak en yüksek ikinci fiyat kadar bir ödeme yapar. Ebay sisteminde ihale bitişinde görünen lider kişinin teklifi aslında liderin o ürün için ödeyeceği fiyattır.

Literatürde ihale üzerine yapılan çalışmalar genellikle oyun teorik denge mekanizmaları ve oyuncuların stratejileri üzerine kurulmuş ya da bol veri imkanıyla birlikte ihale fiyatlarını etkileyen faktörleri belirlemek üzerine yoğunlaşmıştır. Ancak, yeni makine öğrenme yöntemleri ile birlikte bu alanda hala yapılabilecek bir çok yeni çalışma imkanı da vardır. İhale fiyat modelleri literatürde yapısal ve deneysel modeller olarak ikiye ayrılabilir. Yapısal modeller genellikle oyuncuların ürün değerlemesi üzerinden oluşturulan denge modelleridir. Ancak bu modeller ihale fiyatını önceden tahmin etme amacına uygun değildir. Deneysel modeller ise uygulama kolaylığı ve fiyatları önceden tahmin edebilme yeteneğiyle oldukça popülerdir. Bu tez metin sınıflandırma, kümeleme, grafik sınıflandırma gibi yeni nesil makine öğrenme yöntemleriyle ihale fiyat modellemesi çalışmalarına yeni bir bakış açısı getirecektir.

Tezde 2 Mart 2018 ile 2 Temmuz 2018 arasında gerçekleşen iPhone 7Plus ürünü için ihale ve şimdi satın al satış bilgileri kullanılmıştır. Bu kapsamda 444 ihale ve 676 şimdi satın al satış bilgileri eBay websitesinden çeşitli web makroları yardımıyla toplanmıştır. Fiyat bilgileri karşılaştırıldığında ihale ve şimdi satın al satışlarının birbirine yakın sonuçlar oluşturduğu gözlemlenmiştir. Verileri modellerde kullanılabilir bir hale getirmek için bazı ön işlemlerden geçirilmiştir. Bu kapsamda python ve libreoffice calc programları kullanılmıştır.

İhale verileri incelendiğinde göze ilk çarpan şey ihalenin başında ve sonunda bir çok teklifin verildiğidir. Ebay ihalelerinde genellikle oyuncular kendi ürün değerlemesini belli etmemek adına ihale bitiş zamanına kadar teklif vermezler ve son saniyelerde tekliflerini verirler. Ancak ihalenin ilk döneminde verilen tekifler bize ürünün değeri hakkında önemli sinyaller sağlamaktadır.

Modellerde sıradışı verilerin etkisini azaltmak için genel veri dağılımına uymayan veriler veri setinden çıkarılmıştır. Ayrıca kümeleme işlemlerinde veriler belli bir formüle göre normalize edilmiştir.

İhalede verilen tekliflerin zamana göre değişiminin iyi anlaşılması gerekmektedir. Zamana göre tekliflerin değişimine bu tezde “ihale yolu” denilmiştir. 444 ihalede oluşan tekliflerin tamamı incelendiğinde en sık kullanılan ihale yolları şöyle isimlendirilmiştir: Lineer, Çeyrek Çember, Gamma, Ters L, Yatay Ters S ve İki Nokta ihale yollarıdır. Bu ihale yolları Lineer, 2. ve 3. derece polinom, 2. ve 3. derece parçalı lineer ve 4 değişkenli lojistik fonksiyon ile modellenebilmektedir. R^2 ve fiyat tahmin performansına bakıldığında en iyi performansı 3. derece polinom ve parçalı fonkisyonların verdiği görülmüştür. Bu sonuç da ihalelerin 3. bölümden oluştuğu görüşünü desteklemektedir.

Modellerde hangi değişkenlerin kullanılacağını önceden tespit etmek önemlidir. Makine öğrenmesi literatüründe buna değişken mühendisliği denilmektedir. Regresyon modellerinde eksik değişken hatasından kaçınmak için yukarıdan-aşağı değişken seçim yöntemi kullanılmıştır. Yani en olası bütün değişkenler modele dahil edilmiş ve istatistiksel olarak en anlamsız olan değişkenler sırayla tek tek modelden çıkarılmıştır. Kümeleme işlemi için kullanılacak değişkenler ise Python Sklearn uygulaması ile geliştirilmiş değişken seçim algoritmaları ile bulunmuştur. Bunlar SelectKBest, RFE, PFA algoritmalarıdır. Bunların yanında her değişken kümesinden bir değişken alarak en iyi alt kümenin bulunması yöntemi olan Grid Search yöntemi de kullanılmıştır.

Online ihale fiyatlarının bitiş fiyatlarını etkileyen faktörleri belirlemek için öncelikle bütün değişkenlerin tespit edilmesi gerekmektedir. Ürün sayfasına girilmiş ürün açıklamaları metin olarak girilmiş bilgilerdir. Bu bilgilerin sınıflandırılması ve rakamsal bir ürün durum derecesi değişkeni haline getirilmesi gerekmektedir. Bu işlem için bu tezde iki farklı yöntem kullanılmıştır. Birinci çalışmada ürün açıklamaları öğreticisiz öğrenme yöntemi ile kümelenmiş ve ürün durum derecesi değişkeni oluşturulmuştur. OLS modeli ile ihale fiyat modeli kurulmuş ve modelin performansı kümeleme yöntemiyle arttırılmıştır. İkinci çalışmada ise şimdi satın al satış yöntemiyle satılan ürünlerin fiyatları ve açıklamaları öğreticili öğrenme yöntemiyle modellenmiş ve ihale satışlarındaki açıklamalar bu modeller ile sınıflandırılmıştır. Benzer şekilde bütün olası değişkenler kullanılarak OLS modeli ile ihale fiyat modeli kurgulanmış ve modelin performansı kümeleme yöntemleri ile arttırılmıştır.

Öğreticisiz öğrenme, veri setinde herhangi bir beklenen sonuç etiketinin olmaması durumunda geliştirilen modelleme yöntemidir. Öğreticili öğrenme ise beklenen sonuç bilgisinin de veri setinde bulunması durumunda kurgulanan modelleme işlemine denmektedir.

Birinci çalışmada ürün açıklamaları KMeans metin sınıflandırma yöntemleri ile ürün durum derecesine çevrilmiş ve değişken setine eklenmiştir. Daha sonra doğrusal regresyon yöntemiyle fiyat modeli oluşturulmuş ve modelin performansı kümeleme algortimaları ile arttırılmıştır. Bu çalışmada ürün açıklamalarının bazı satıcıların yaptığı gibi 4 sınıfta toplanabileceği varsayılmıştır. Ürün açıklamaları sınıflandırıldığında Küme3'te gruplanan ürünlerin açıklamalarında en sık olarak "original", "excellent" ve "new" kelimeleri ön plana çıkmıştır. Küme1'de ise "straches", "wear" kelimeleri görülmüştür. Buradan yola çıkarak ürünün kullanım durumu en iyi olan ürünlerin Küme3'te, en kötü olanların ise Küme1'de olduğu iddaa edilebilir. Küme0 ve Küme2'de ise kullanım durumları diğer ikisinin arasında olan ürünler yer almaktadır. Fiyat modelinde de beklendiği üzere Küme3'te bulunan ürünlerin ürün kullanım derecesinin katsayısı en yüksek çıkmış, Küme1'de ise en düşük çıkmıştır. Küme0 ve Küme2'nin katsayıları ise arada bir değerde çıkmıştır. Değişken seçim algoritmalarından Grid Search yöntemiyle seçilen değişkenlerin (3 adet) en iyi performansı 4 kümede verdiği tespit edilmiştir. Belirlenen küme sayısı ellbow eğrisi ve silhoutte katsayısı ile de desteklenmiştir. İhale fiyatlarını etkileyen faktörlere gelince, geçmiş dönemde oluşan satış fiyatlarının ihalenin bitiş fiyatını pozitif etkilediği ortaya çıkmıştır. Ürünlerle birlikte satılan aksesuarların ihale bitiş fiyatını arttırdığı ve aksesuarların müşteriler tarafından gözardı edilmediği bulunmuştur. Satıcının deneyim seviyesi, özellikle müşterilerinden aldığı negatif puanların satış fiyatlarını düşürdüğü tespit edilmiştir. Satıcı profilinin müşteriler tarafından ziyaret edilme sayısı da ürünlerin fiyatlarında pozitif bir etki oluşturduğu ortaya çıkmıştır. Ürünün kullanım durumu derecesine gelince bütün değişkenler istatistiksel olarak anlamlı çıkmıştır. Beklendiği üzere en yüksek katsayı Küme3'te en düşük katsayı ise Küme1'de oluşmuştur. İhale süresinin bitiş fiyatı üzerindeki etkisi istatiktiksel olarak anlamlı olmasının yanında bu ilişki detaylı bir şekilde incelemeye değerdir. İhale süresi 1 günden 5 güne kadar arttıkça bitiş fiyatı üzerindeki etki de artmakta, 5 günden sonra ise bu etki azalmaktadır. Buradan eBay tipi ihalede satıcılar açısından optimum sürenin 5 gün olduğu vurgulanabilir. 5 günlük bir süre ürünün alıcılar tarafından keşfedilmesi için yeterli bir süre iken alıcılar tarafından da çok beklemeden ürüne ulaşabileceği kadar kısa bir süredir. İhalenin başladığı günün fiyat üzerindeki etkisi ise şöyledir. Pazar, Pazartesi ve Cuma günü başlayan ihalelerde bitiş fiyatları daha yüksek olmaktadır. Alıcıların bugünlerde daha çok ürün araştırmalarına vakit ayırabildiklerini, belirledikleri ürünü takip edip teklif verdiklerini ve yoğun rekabet nedeniyle belirtilen günlerde ihaleye çıkan ürünlerin fiyatlarının da yüksek olduğunu söyleyebiliriz. Daha önce belirttiğimiz gibi ihale başlar başlamaz verilen tekliflerin ürünün piyasa değeri hakkında önemli bir fikir verebileceği modelin sonucu ile tespit edilmiştir. İhalenin ilk başladığında teklif veren kişi sayısı, bu dönemde verilen son teklif ve son teklifi veren kişin deneyim seviyesi ürün fiyatında pozitif bir etki yaptığı bulunmuştur. Geliştirilen modelin istatiksel olarak doğruluk testleri yapılmış ve model bütün testlerden başarı ile geçmiştir. Tezde her küme için fiyat modeli tekrar tahmin edilmiş ve kümeleme işlemi ile fiyat tahmin performası daha da arttırılmıştır.

İkinci çalışmada ise şimdi satın al satışı şeklinde satılan ürünlerin fiyatları ve ürün açıklamalarından yola çıkarak metin sınıflandırma modelleri oluşturulmuştur. Öğreticili öğrenme modellerinde öncelikle verilerde beklenen sonuç etiketinin olması gerekmektedir. Modellerin tahmin edilebilmesi için her sınıfta yeterince veri olması da gerekmektedir. Şimdi satın al satışlarında ürün açıklamaları için bir sonuç etiket bilgisi bulunmamaktadır. Bu nedenle fiyat aralıklarına göre yapay olarak ürün durum derecesi etiketi oluşturulmuştur. Fiyat gruplarına göre de metin sınıflandırma modelleri tahmin edilmiştir. Bu çalışmada öğreticili öğrenme yöntemlerinden Multinomial Naive Bayes, Logistic Regression, Linear Support Vector Classification, Random Forest metin sınıflandırma yöntemleri kullanılmış ve en iyi performansı veren Multinomial Naive Bayes yöntemi bu tez için seçilmiştir. Bu çalışmada da bir öncekine göre bazı satıcıların yaptığı gibi ürün kullanım dereceleri 4 sınıfta toplanmış ve ihale satışlarında girilen açıklamalar seçilen model ile derecelendirilmiştir. Her küme için en sık geçen ve en önemli kelime grupları belirlenmiştir. Benzer şekilde ürün durum derecesi değişken setine eklenmiş ve doğrusal regresyon yöntemiyle fiyat modeli oluşturulmuş, modelin performansı KMeans ve Hiearchial kümeleme yöntemleri ile de arttırılmıştır. Öğreticili öğrenme yöntemiyle oluşturulan fiyat modeli ile ihale fiyatını etkileyen faktörler bir önceki çalışmayı destekler nitelikte çıkmış modelin performansı bir önceki çalışmaya göre daha başarılı bulunmuştur.

Üçüncü çalışmada ise sadece ihalenin ilk aşamasında oluşan ihale yolunun grafikleri grafik sınıflandırma yöntemleri ile modellenmiştir. Bu çalışmada etiket olarak fiyat grupları ve ihale yolu bilgileri kullanılmıştır. Bu yöntemde Convolutional Neural Network (CNN) algoritmaları kullanılmıştır. Fiyat tahmini problemine hızlı bir çözüm olarak oluşturulan bu model için sadece ihale grafiklerinin ilk aşaması tamamlanması yeterlidir. Bu yöntemin fiyat tahmin performansı makul seviyede olsa da ilk 2 modelden biraz düşüktür.

Sonuç olarak bu tezde eBay tipi bir ihalede, ihale fiyatını etkileyen faktörler tespit edilmiş, fiyatı önceden tahmin edebilen modeller geliştirilmiştir. Gelcekte bu yöntem diğer ürünler için de gelecekte genelleştirilebilir, verilen teklifler teklifi veren müşterinin deneyimi ile ağırlıklandırılabilir, ihaleler oyun teorik çerçevede değerlendirilip makine öğrenme uygulamalarının getirdiği yenilikler bu alana uygulanabilir.

APPENDIX M. TEZ İZIN FORMU / THESIS PERMISSION FORM

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences □

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences □

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematic □

Enformatik Enstitüsü / Graduate School of Informatics □

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences □

YAZARIN / AUTHOR

Soyadı / Surname : ÖZ
Adı / Name : EMRAH
Bölümü / Department : İKTİSAT

TEZİN ADI / TITLE OF THE THESIS (İngilizce / English) :

Empirical Studies on Price Determinants of Online Auctions with Machine Learning Applications

TEZİN TÜRÜ / DEGREE: Yüksek Lisans / Master □ Doktora / PhD □

1. **Tezin tamamı dünya çapında erişime açılacaktır.** / Release the entire work immediately for access worldwide. □

2. **Tez iki yıl süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of **two year.** * □

3. **Tez altı ay süreyle erişime kapalı olacaktır.** / Secure the entire work for period of **six months.** * □

** Enstitü Yönetim Kurulu Kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir. A copy of the Decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.*

Yazarın imzası / Signature............................ Tarih / Date.....................