USING LOCAL LINEARIZATION TO IDENTIFY THE DYNAMICS OF FEATURES IN SECOND-HAND AUTOMOBILE PRICING

BERKAY BULUT

BOĞAZİÇİ UNIVERSITY

2019

USING LOCAL LINEARIZATION TO IDENTIFY THE DYNAMICS OF FEATURES IN SECOND-HAND AUTOMOBILE PRICING

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Master of Arts

in

Management

by

Berkay Bulut

Boğaziçi University

2019

DECLARATION OF ORIGINALITY

I, Berkay Bulut, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature.....

Date .....13.06.2019.....

ABSTRACT

Using Local Linearization to Identify the Dynamics of Features

in Second-Hand Automobile Pricing

In this study, nonlinear feature interactions are analyzed using application of local linearization on machine learning algorithms in second-hand automobile price modeling. To study the phenomenon in nonlinear models, a colossal amount of data collection is conducted for three years using advanced python scraping. The collected 313570 observations are from two main largest online listings website. The automobile prices were modeled as a function of an automobile's technical, visual, geospatial and macroeconomic features. The dataset is processed, transformed and modeled using basic linear least squares regression and advanced gradient boosting algorithm to compare linear and nonlinear modeling performance. Nonlinear modeling performed significantly better performance with 93% in R2 with 6529 TL mean absolute error compared to linear model performance at 58% in R2 with 23214 TL mean absolute error. In this research, we show that high-performance nonlinear models can be built to understand market dynamics while maintaining a similar level of interpretability as linear models. This research opens new opportunities in the application of advanced analytics in business and academic research in explaining how a specific prediction is made, understanding a very complex phenomenon and developing interpretable high-performance models in risk, insurance, and health care. Understanding why a model makes a prediction is essential for trust, actionability, accountability, debugging, and many other tasks.

ÖZET

İkinci El Otomobil Fiyatını Belirleyen Değişkenlerin Dinamiklerinin Lokal Doğrusallaştırma Yöntemi ile Belirleme

Bu çalışmada, doğrusal olmayan özellik etkileşimleri, ikinci el otomobil fiyat modellemesinde makine öğrenmesi algoritmalarına lokal doğru sallaştırma uygulaması kullanılarak analiz edilmiştir. Bu olguyu lineer olmayan modellerle incelemek için, üç yıl boyunca gelişmiş¸ Python web kazıma tekniği kullanılarak büyük miktarda veri toplanmıştır. Toplanan 313570 gözlem, en büyük iki çevrimiçi listeleme web sitesindendir. Otomobil fiyatları, otomobilin teknik, görsel, mekânsal ve makroekonomik özelliklerinin bir fonksiyonu olarak modellenmiştir. Veri seti, doğrusal ve doğrusal olmayan modelleme performansını karşılaştırmak için temel doğrusal en küçük kareler regresyonu ve gradyan güçlendirme algoritması kullanılarak işlenmiş, dönüştürülmüş ve modellenmiştir. Doğrusal olmayan modelleme, 23214 TL ortalama mutlak hata ve 58% R2'i olan doğrusal modellemeye kıyasla 6529 TL ortalama mutlak hata ve 93% R2 ile daha iyi performans göstermiştir. Bu çalışmada, lineer modellerle benzer bir yorumlana bilirlik seviyesini korurken, piyasa dinamiklerini anlamak için yüksek performanslı doğrusal olmayan modellerin kurulabileceği gösterilmiştir. Bu araştırma, özel bir tahminin nasıl yapıldığını açıklamak, çok¸ karmaşık bir olguyu anlamak ve risk, sigorta ve sağlık hizmetlerinde açıklanabilir yüksek performanslı modeller geliştirmek konusunda iş ve akademik araştırmalarda ileri analitiklerin uygulanmasında yeni fırsatlar sunmuştur.

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor Assist. Prof. Hüseyin Sami Karaca for the continuous support of my master study, research and professional career, for his patience, motivation, and immense knowledge. I would not have been where I am today without his contribution and guidance. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my graduate study.

I would like to thank my thesis committee member Assoc. Prof. Özlem Karaca and Prof. Kenan Aydin for their insightful comments.

I thank my fellow colleagues in for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last two years. Special thanks go to Feyzullah Alim Kalyoncu with whom we worked sleeplessly in mind-blowing data science projects.

I would like to thank Ayça for her kindness, love, and support through these hard-working years towards my goals.

Last but not the least, I would like to thank my family: my mother, my father and to my brother Berke for supporting me spiritually throughout writing this thesis and my life in general.

To my lovely parents and tonti

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1. Importance of used automobile marketplace

The automobile market has risen in the world and competitions are fierce. It has an essential standing in world economies due to its relationship with many macroeconomic factors such as employment and income. Understanding the right second-hand automobile price drivers is a catalyst for the healthy and stable development of the second-hand automobile market. There has been much research into what are the key factors that influence an automobile's price. In consumer-oriented marketing, consumer buying behavior and the process a consumer goes through have been a fascinating subject that a great deal of research been into. Understanding the insight into purchasing decisions is significant for developing marketing strategy (Louviere et al., 2000). The value of a product is the sum of all the values of its components or its parts in the customer's eyes. It is a known factor that the value of a second-hand automobile de-pends on a variety of factors. The most important ones are the age of the car, make, model, the origin of the car as in the country, its odometer in km and its horsepower. As information increases so do difficultly in understanding what the main effect is.

Data mining using linear methods applied to estimate the price of used cars exists in the existing literature. These explanatory models fall behind in performance and capturing nonlinearity in the used car pricing, however making up in interpretability. Furthermore, to study the phenomenon in nonlinear models, a colossal amount of data collection, processing, and computation power is required. In today's abundance of information, collecting big data has become more accessible,

developing advanced algorithms to generate insights have become computationally efficient and possible. As a data-centric approach, applied business analytic grows on improvement in database management and relies heavily on various data collection, extraction, and analysis technologies (Chaudhuri et al., 2011). The impact and insights that can be generated using advanced analytics to business problems are enormous.

Predicting the resale price of a second-hand automobile is not a simple task; however, it shows great potential. For customers, knowing the reasonable price of the car can make them buy or sell used car with no worries; for car rental companies, predicting the residual value is useful for the pricing of their rental service; for banks and other financial institutions, evaluating the price of a lender's car can help them control his or her loan quota. Understanding why a model made a prediction is essential for trust, actionability, accountability, debugging, and many other tasks.

Present research in understanding and predicting an automobile's price heavily depended on linear hedonic models. One of the reasons that are has been modeled in such a way is due to the limits in computational capacity, lacking in several observations and interpretability of linear hedonic models. However, there is a vast potential window with increased computational power, ability to collect big data and the ability to model nonlinear systems. Advantages arise such as the ease with which symbolic, nominal, or categorical variables can be that can be included; and the ability of these methods to deal with noisy data. Very precise pricing models can be developed, and insights can be generated with local linearization methods.

This research addresses the price structure of the used car market in the light of analytic methods developed to measure nonlinear effect using local linearization of predictive models. Most literature on nonlinear modeling would lead to the

conclusion that interpretation is impossible or very difficult and should be treated as a black box model. The proposed price model exhibits evidence in support of nonlinearity in the features for explaining price structure. Secondly, it exhibits how not easily quantifiable features come with. In this research, we examine to what extent the prices in the second-hand automobile market reflects the vehicle's attributes, how the characteristics affect prices using nonlinear modeling with application of local linearization method.

There is a lot of research in understanding the price determiners of new automobiles. However, this cannot be said for the second-hand vehicle car market. Therefore, this research contributes to the literature mainly by using second-hand market data, nonlinear, and local linearization. This is the first research study that is conducted to a unique and large dataset. The findings of this study are expected to add to the literature by filling the gap in the second-hand market and new areas that local linearization can be applied to such as risk modeling literature.

Most research on random forests, gradient boosting methods and nonlinear modeling would lead to the conclusion that interpretation is impossible or very difficult since these modeling techniques are typically treated as a black box. Indeed, a forest and gradient boosting consists of many deep decision trees, where each tree is trained on bagged data (sampled from the larger data set) while using a random selection of features. Understanding the decision process and the prediction process to its full extent by examining each individual tree is infeasible and impossible. To this end, various modeling techniques such as linear regression, decision trees, random forests and gradient boosting are presented in the existing literature, how they are constructed in functional form and intuition behind the models are introduced. Firstly, how the models are interpreted using statistical methods are

conceded, and lastly, local linearization literature is introduced. Both approaches are useful, however crude and static methods give little insight in understanding individual decisions and predictions on actual data. This is where local linearization brings novel insights.

The ability to interpret a nonlinear machine learning model opens many opportunities for business. First is that businesses can build very high accuracy models without the concern of losing interpretability. This interpretability is a similar level to linear models; however, in a sense that is dynamic, not static. Using local linearization, every prediction can be presented as a sum of each feature's contribution. With interpretability, analysts can interpret why a prediction is made; models can be debugged to understand why the predictions are weird or unexpected and heavily regulated industries such as insurance and credit banking can use nonlinear models that can be explained to regulatory bodies.

1.2. Manuscript Organization

This thesis is organized in five chapters, with specific content specified below. The thesis starts with an introduction, literature review, data explanation, analysis, and conclusion.

Chapter 1 is an introduction to the importance of understanding used automobile marketplace and the methodology used and proposed that enables to understand the dynamics of the pricing phenomenon. Chapter 2 reviews previous studies on pricing using information, pricing with hedonic models and theories, feature analysis using on second-hand automobiles, statistical methods that are considered in application to this research and cutting-edge analytic approaches used

to understand the phenomenon. Chapter 3 reviews the dataset that is used in modeling from the collection of data from online resources, processing the dataset and final dataset. Chapter 4 presents a summary of the theoretical models applied, empirical re-search findings and conclusions for the analysis conducted. Chapter 5 presents the results of this study, limitations and potential future rooms to build on the research.

CHAPTER 2

LITERATURE REVIEW

Predicting the price of second-hand automobiles has not received much attention in research despite its immense importance for the society and economy. The existing literature in used goods markets has so far focused on asymmetric information, quality perception, price-quality perception, dynamic pricing of used goods and hedonic modeling of price and independent features. This research aims to address the evaluations in the existing literature and deduct the theories in the focus of used automobile pricing. The automotive industry is an industry that manufactures road vehicles that will meet passenger and cargo transportation requirements (Eken and Çiçek, 2009). Among the goods and services produced by the industry, the automobile is a product with many features. Every automobile has observable and not observable qualities like size, weight, braking system, steering type, maximum speed, fuel consumption, brand, and so forth. Automobiles have a range of different qualities to attract the consumer's attention (Hadinejad and Shabgard, 2011).

2.1 Pricing Used Goods with Information

The existing literature has focused on general theories in asymmetric information, pricing, and quality perception but lacks field experiments to support evidentially. Turkish academics have attempted to create price forecasting models for used cars but have not utilized enough data (Balce, 2016) and were mostly on comparing the best model to forecast rather than examining the independent features (Ecer, 2013). Research shows that used goods market transactions have asymmetric information (Wilson, 1980). He stated that asymmetric information could be avoided on the

product quality when there is mutual trust between the transaction parties. The transaction parties are the seller, the buyer, and the intermediary and mutual trust of these par-ties and the information will eliminate the information asymmetry. In this paper, our research scope is to identify the main factors leading to a given sales price decision for used car models.

There is a lot of evidence that there is a relationship between a product's price and the product's perceived quality; however, there is insufficient evidence that this is sufficient (Gardner, 1971). The result of the study cast serious doubt on the possibility of a generalized price-quality relationship. He concluded that price does not influence the perception of product quality, yet this conclusion depends on the independent features. The price-quality research finding is that the judgment of quality by price can be confounded by non-price information about the product (Stafford and Enis, 1969). Sellers and marketers make use of the product features and the price to influence customers (Chang and Wildt, 1994). Thus, the product attribute information influences the perceived quality of the goods. Furthermore (Shapiro, 1968) states that a consumer often has more confidence in price as an indicator of quality than other cues. In our research, we will examine how indicators of quality in cars impact the price and whether these attributes of information are in line with the market data.

Consumers with vast information are knowledgeable of the configuration of attributes that comprise a product. Nevertheless, real-world consumer's perception of the quality of a product is an admixture of a variety of informational inputs concerning a set of criteria for judging the product (Tull et al., 1964). His findings suggest that consumers rely heavily upon price as a predictor of quality when there is substantial degree of uncertainty involved. He finds that for some products the price

and quantity sold relationship is kinked such that it bends backward after a certain price break point. These findings suggest that regardless of any market whether e-commerce or retail, there is a certain degree of pricing limit that a product can attribute.

Despite the expectation of a positive price-quality relationship, results of nearly 90 research studies in the past 30 years provide mixed evidence for this relationship (Zeithaml, 1988). He posits that most experimental studies related to quality assessment focused on price as the critical extrinsic quality signal, however price is but one of the several potential useful extrinsic cues. Brand name, packaging may be equally or more important. Both Gardner and Zeithaml have concluded that price-perceived quality relationship is inconclusive.

One of the objectives of price theory is to quantify the price change of a given good accurately. Changes in the quality of the goods should be considered to establish a basis for price assessments. Lancaster stated that each property has value for consumers and this value is defined as implicit price. Therefore, the hedonic price model refers to the price of the goods as the sum of the implied prices of the different properties of the goods (Abounoori and Rezvani, 2012), Lancaster's approach can be summarized as goods do not benefit the consumer alone, there are properties that belong to the good they increase the benefits, the goods as a whole may have different characteristics separately from other goods associated. The consumption of a single commodity or combination of goods is input, and the sum of the attributes of the goods is output. So, the benefit is obtained through the properties that the goods have (Lancaster, 1966). When the consumer balance is realized, the price that the consumer is willing to pay shall be equal to the implicit price of each property.

Therefore, implicit prices include the consumer's wishes (Hadinejad and Shabgard, 2011). Hedonic prices are defined as implicit prices of properties (Rosen, 1974).

2.2 Feature Analysis on Second-Hand Automobiles

The choice of attributes of any product reflects consumer preferences. There is a two-stage process of deciding to buy for second-hand automobiles. In the first step, the consumer chooses the type of automobile and limits age and kilometer range. The choice of automobile type is based on variables such as the size, design, technical equipment, engine power and brand image of the automobile. In the second step, if the automobile type was chosen, the consumer tries to find a campaign or a discount for the purchase. The consumer tries to select the automobile with the best price-quality level (Dexheimer, 2003). There are many factors that can influence the price of a second-hand automobile in technical details like the type of fuel it uses, the braking system, acceleration, the volume of its cylinders (measured in cc), weight of the car whether it has cruise control, whether it is automatic or manual transmission, its physical state, whether it is a sports car. Other attributes options such as air conditioner, sound system, power steering, cosmic wheels, GPS navigator influences the pricing of the vehicle. Some factors that are clustered as visual elements are the interior style, its size, number of doors, paint color and external information gathered on safety index, consumer reviews, prestigious awards won by the car manufacturer, whether it belonged to an individual or a company. Some unique factors which buyers attach importance are the local of previous owners, whether the car had been involved in severe accidents and whether it is a lady-driven car. The look and feel of the car positively contribute a lot to the price. As it can inevitably be observed, the price depends on many factors on environmental and internal factors. However, there

is a constant information asymmetry between the buyer and the seller. Information on all these factors is not always available, and the buyer must decide to purchase the vehicle at a price based on just several provided factors.

The seller knows unique and only several pieces of information about the car he is selling (Lewis, 2011). A seller may not be able to judge the mechanical condition of the car neither the buyer. He claims that there is undoubtedly information asymmetry in online markets which are because of the disclosure costs of listings. If there were to be contractual agreement on the information the asymmetric information is eliminated. Contrary to Lewis's findings, in our research, we are confident on the issue of information asymmetry since the transaction of used automobiles does not finalize until both parties meet face to face and there are no commitments or holding of cash or bank account balances until a final purchase is made.

Production year of automobiles is one of the principal characteristics influencing car prices and have identified fuel type and volume as a significant portion of the price of used car (Erdem and Sentürk, 2009). Their research involves a unique data set of 1074 cars with different origins.

The variability of used car prices was investigated by (Asilkan, 2011) of the used cars in Europe. Although Asilkan's approach has significant potential of having high sample size by using data mining techniques from several websites, he has contributed to literature in investigating the regression analysis and artificial neural network and the methodology comparison to use in forecasting used automobile prices.

2.3 Price modelling and hedonic models

Present literature predicts prices of products using some previous data and so did (Pudaruth, 2014) who predicted prices of second-hand cars in a specific location. Using multiple linear regression, k-nearest neighbors, Naive Bayes and decision trees algorithm to predict the car prices, models achieved some degree of accuracy. Comparing the prediction results with these techniques has shown that the values of these methods are strictly comparable. However, it was found that the decision tree algorithm and the Naive Bayes method could not classify and predict numerical values. Pudaruth's research also concluded that the limited number of cases in a dataset does not provide high predictability (Pudaruth, 2014). Many instances give better statistical power than a low number of instances, allowing the data mining experiment to converge and give significant results. (Wu et al., 2009) used fuzzy neural information-based system to estimate the price of used cars. There are only three factors in this study: the brand of the car, the year in which it was produced, and the style of the engine are discussed in this study. Wu's data lacks in-vehicle variability features to capture many factors contributing to the price of a used car.

The hedonic price theory starts with the assumption that the goods are heterogeneous and are considered as a combination of the individual qualities or characteristics of each commodity. Each quality feature is treated as property or service of its own and in this way has its price is found. These features distinguish between different car models and thus represent the quality of each vehicle (Murray and Sarantis, 1999).

The first implementer of the hedonic price model is G. C. Haas in 1922. In his study for agricultural land pricing, G. C. Haas has built a simple hedonic price model for farm area by using distance from the city center and city size variables.

Following Haas, American car industry expert A. T. Court, who used the word hedonic for the first time in the study, predicted the hedonic price model as a function of the properties of the good (Colwell and Dilmore, 1999).

A. T. Court, for the period 1925-1935, described the automobile's price as a function of the various features of the automobile and analyzed the hedonic price index of the automobile, which is a different commodity. In the study, he observed that the average price of the automobile has increased as of the said period, but in fact considering the properties of an automobile such as weight, length, horsepower, and so forth. Price has decreased. In the meantime, approximately ten years before A. T. Court's work, Waugh in 1929 in his study, appears as the first researcher to perform a systematic analysis of the impact of quality on the price of a good. Waugh defined quality using various observable features and estimated the quality as the implicit price of each of these features (Sheppard, 1999).

Thereafter, Lancaster in 1966 and Rosen in 1974 formed the theoretical bases of the hedonic price approach. In the case of heterogeneous products, the demand for the product depends not on the product itself, but its properties. Each feature is a benefit to the consumer and the level of benefit that consumers receive from any product depends on the different qualities in these products. When it is decided not to buy or buy a product, the consumer compares the total benefit to the cost. Therefore, the cost of any product in equilibrium is equal to the sum of the values of the properties in this product (Ayan and Erkin, 2014). The hedonic method uses multiple regression techniques with an extensive data set and requires a structure based on microeconomics analysis. The hedonic method is mainly used for market values that include the benefits and impact attributes of goods (Selim, 2011).

In the model developed by Rosen, the goods (Z) are the sum of their n characteristics Zi where i includes n characteristic and shows the amount of each characteristic. Rosen proposed the general form of hedonic price function as P (Z) = Z(Z1, Z2, Z3, ..., Zn) where P (Z) is the observed market price of the product. Z (Z1, Z2, Z3, n., Zn) is the vector of objectively measured properties. Z elements are measured objectively in terms of perceptions of all consumers (Rosen, 1974). Rosen developed the hedonic model based on product characteristics within the scope of Lancaster's work and based on this price function, the effect of each feature on the price can be expressed by taking the partial derivatives of equation (Baldemir et al., 2007).

In the hedonic method, price is used as a dependent variable, and product properties are used as independent variables (Court,1939:109). In this study, a model was designed based on Lancaster's hedonic hypothesis based on consumer behavior theory. The consumer aims to maximize the benefits of the automobile with its features and qualities. The developed hedonic model is investigating automobile demand in this respect.

2.4 Statistical methods

To understand effect of automobile pricing factors, the big data is analyzed using statistical methods. Statistical analysis methods apply hypothesis testing on data set features to score and then select a set number of highest scoring features as an input to a descriptive or predictive model (Wang et al., 2013). Therefore, univariate and multivariate statistics that can provide a systematic road to quantify phenomenon tested. The method goal can be two folds; to test for hypothesis such as significant

changes in the market or to understand and reduce the size of the feature set leading to the optimum modeling performance.

Feature statistical analysis methods can also be divided into two classes: univariate and multivariate. Although automobile pricing factors can be analysed using both techniques by going over one to one relations, multivariate analysis has many ad-vantages of the former. Univariate methods use univariate statistics to test features and can be used to score the features. However, univariate methods unavoidably discard features that, when taken in aggregate, would have provided useful information about the experimental conditions (Norman et al., 2006). By contrast, multivariate feature selection methods can avoid this problem by computing multivariate statistics for feature ranking because they consider the dependencies between the features when calculating scores for features (Wang et al., 2013).

To quantify and understand social research statistical data analysis procedures are performed. With increasing numbers of features, statistical modeling becomes necessary and more complex. In univariate statistics, just one feature is of interest while in bivariate statistics the relationship between two features is analyzed (Baur and Lamnek, 2007). Multiple variable analysis consists of more than two characteristics in the relaxed sense of the term and at least two dependent and two independent characteristics in the narrow sense.

Complex and multifactorial models are more appropriate to social sciences because many pieces of social reality are tangled and seldom there is a single concentration. In automobile multivariate analysis satisfy the multivariate analysis requirements are valid, standardized dataset, set minimum number of observations, random sampling, specified variable scale types, certain distribution of features and residuals and a minimum variance of features.

The statistical methodology to test for significance varies depending on dataset and feature types. In the second-hand automobile pricing, the data follows a tabular format with nominal, ordinal and continuous variables. Nominal variables require the use of non-parametric tests, and three commonly used significance tests can be used for this nominal datatype (McHugh, 2013). The first and most common is the chi-square. The second is the final proof of Fisher research, which is a bit more accurate than the chi-square but is used only for 2 x 2 Table. The third test is the chi-square test, which is the most common usage ratio when the data set is too small to meet the size assay of the square test sample. (Scott et al., 2013).

2.4.1 Pearson correlation

The purpose of the Pearson correlation coefficient (Pearson's r) analysis is to measure and interpret the strength of a linear or nonlinear (e.g., exponential, polynomial, and logistic) relationship between two continuous variables. When conducting a correlation analysis, we use the term association to mean "linear association." In second hand automobile dataset there are, continuous and categorical features. The features linear association is interpreted using Pearson correlation. The score of Pear-son correlation analysis lies between +1 and −1, a range which is defined as negatively correlated (−1) to positively correlated (+1). The sign of the coefficient exhibits the direction of the correlation whereas the value exhibits the power of the correlation. The Pearson correlation coefficient is also known as the sample correlation coefficient (r), product-moment correlation coefficient, or the coefficient of correlation (Neter et al., 1990) measures the linear relationship between two random variables. For ex-ample, when the value of the km feature of an

automobile is manipulated (increased or decreased) by a fixed amount, the outcome variable price of the vehicle changes proportionally (linearly).

2.4.2 Regression

In the existing literature, mostly hedonic models have been used for price forecasting models. This is understandable since housing takes one of the significant portions of the disposable income of individuals. Regression based hedonic models have been used in (Jim and Chen, 2009) research in house price modeling in Hong Kong and it has been used by (Fletcher et al., 2004) in comparing hedonic modeling of housing. (Coulson and McMillen, 2008) estimated time, age and vintage effects in housing prices using hedonic regression. He posits that hedonic modeling in the pricing of housing is overcoming the difficulty of separating the effects of age and cohort on house prices. He finds some of the features to be highly correlated with perfect collinearity; thus, when applying such methods; it is advisable to keep in mind the collinearity of features in used car pricing.

It seems appropriate and beneficial that the use of regression modeling will bring new insights into the used car market and to the literature. Both statistical methods of classification and regression can be used for relevance analysis which attempts to identify attributes that are significantly relevant to the objective information (variable) that is being modeled (Han et al., 2011).

The primary goal of linear regression is to fit a straight line through the data that predicts Y based on X. To estimate the intercept and slope regression parameters that determine this line, the least squares method is commonly used. This regression method, a set of regression parameters are found such that the sum of squared

residuals (i.e., the differences between the observed price value of automobiles and the fitted values derived from listings) are minimized. The fitted price value of second-hand automobile is then computed as a function of the given x values (i.e., the features of the listing such as city, color, km, seller type) and the estimated intercept and slope regression parameter. It is meaningful to interpret the value of the Pearson correlation coefficient r by squaring it; hence, the term R-square (R2) or the coefficient of determination. R2 is a good estimate of how the model fits and is calculated using a formula which consists of one minus sum of squared errors of the model divided with sum of squared errors from the mean of actual observations. This measure (with a range of 0–1) is the fraction of the variability in price of automobiles predicted that can be explained by the variability in big data of second-hand automobiles through their linear relationship or vice versa. R2 is a measure of how well the price prediction model performs in comparison to having a simple mean line across all price points of the online listing.

The second measure of success is generally visually determined or measuring how random the residuals are distributed by observing how the residuals are plotted across each actual value. The residual plot should show Gaussian (random) behavior. The erratic behavior can be observed by detecting a Gaussian distribution or a helicopter drop of all residuals across the scatter plot. If there is a specific pattern to the plot of the residuals than that could be interpreted in three ways. Firstly, the model does not have enough observations to capture the phenomenon. Secondly, the model does not have enough attributes to capture the phenomenon. Lastly, the model structure is not fit for the specific problem at hand. The model structure should be changed from a linear to nonlinear or vice versa.

2.4.3 Lasso regression

In second hand automobile pricing regression analysis there are a lot of features that can be considered. Most regression problems fall short in predictive power usually calculated with R2. One of the biggest challenges in using linear regression is in how to choose predictors for output. However, using all the features can lead to misleading linear regression output due to shared information in two or more features. These highly colinear features should be removed which can cause multicollinearity in the trained predictive model and disturb the prediction outcomes. Secondly, some features can have high noise which can significantly decrease predictive power in out-of-bag samples of the dataset. Feature selection plays a major role in how a good price predictive model for second-hand automobile market can be built.

Lasso regression's advantage is that it has a regularization which punishes feature's coefficient towards zero. The lasso does this by imposing a constraining on the model parameters that push the coefficients of the predictors to zero after the shrinkage process is excluded from the model. Variables that do not have a zero coefficient are variables that are strongly associated with the response variable. For example, a feature such as city information can have coefficient very small or zero thereby reducing the effect of this feature on the price of an automobile. If a feature's coefficient is punished to zero, it is eliminated and if the feature's coefficient decreases to very low values the feature's effect on predictor is suppressed. Therefore, lasso regression not only helps in decreasing the over-fitting of the model to the training dataset, but it also helps to select features by using regularization.

2.5 Advanced data mining

Increased computer and processing power of the society has enhanced capabilities to generate and collect data from diverse resources. The exponential growth in saved and transferred data has created opportunities to analyze the economic and behavioral dynamics. Data mining can be considered a combination of selection, exploration, and modeling of large databases. Unknown correlations, patterns, and trends can be found by using data mining tools. In other words, data mining can be used to extract knowledge from any data (Giudici and Castelo, 2003). It can find close relationships that could not have been obtained by observing, recording individual data items alone. Data mining is an interdisciplinary subject, and it is also called knowledge discovery from data. The data mining process as can be seen in Figure 1 has an iterative sequence of steps.

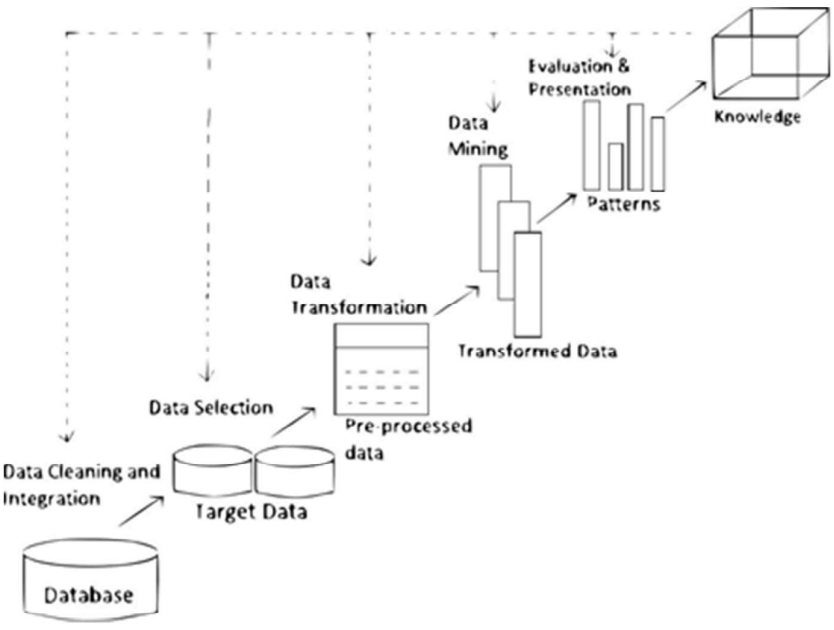

Figure 1. Data mining process.

The steps from data collection to data transformation are a collection of different forms of data preprocessing for the data that will be applied to knowledge extraction. First data is collected through a variety of sources. The collected data is called raw data. The raw data is stored in its original format to be utilized in other methods before application of any transformation. The second step is data cleaning where stored raw data is transformed for the objective of the study. Unnecessary knowledge and private information are filtered out. Data integration is a step where data from a variety of sources are combined in a structured and organized format. This enables the researcher to apply methods on a consistent basis. For some research requirements, data may need to be transformed. For example, a time component such as the month of the year may be needed in an ordinary variable for prediction purposes, converting text to ordinary numbers is called data transformation.

Data mining algorithms are commonly used in several fields. In the banking sector, data mining applications are widely used for credit applications, fraud detection, and stock markets. Data mining algorithms are also used in other fields such as medicine, telecommunications, physics and biology (Alpaydın, 2010).

2.5.1. Decision trees

Decision tree classifiers are used successfully in a lot of diverse areas such as radar signal classification, character recognition, remote sensing, medical diagnosis, expert systems, and speech recognition, to name only a few (Safavian and Landgrebe, 1991). One of the most important aspect of a decision tree classifiers is that they are powerful to break down a complex decision-making process so that it is composed of simpler decision towards a goal. The breaking down aspect makes the model is easy

to interpret. The main objectives of decision trees are to classify correctly as much of the training sample as possible and generalize beyond the training sample so that unseen samples could be classified with a high of an accuracy as possible. The decision trees can be easy to update as more training sample becomes available. The design of a decision tree can be decomposed into the following tasks (Kulkarni and Kanal, 1976). Firstly, the appropriate choice of the tree structure. Secondly, the choice of feature subsets to be used at each internal node and lastly the choice of the decision rule or strategy to be used at each node are critical.

Decision trees are an exceptional form of understanding of how the features interact with each other to explain the target variable. The performance of the decision tree is due to how it can interact with several features with each other in a nonlinear methodology. The power of the decision trees come from its ability to create hyper-planes in the solution space that cannot be executed with linear models. Secondly, the interpretation of a decision tree is easy. By observing the decision tree structure, inferential can be made to how the phenomenon is structured.

The decision trees lack the ability to find the optimal splitting hyper-plane. This is the major drawback of using a decision tree. The resulting tree structure does not have an optimum global guarantee. The reason behind the sub-optimal is the greedy tree splitting structure. A decision tree creates new splits where the entropy is minimized without considering the next step. This causes the decision tree to stay in sub-optimal local points in the hyperspace. Secondly, if the decision tree is not correctly tuned with pruning and the number of points in the leaf parameter, the algorithm falls back in

2.5.2 Ensemble models

In recent years, there has been a lot of interest in ensemble modeling approaches. The most prominent industrial usage example is the Netflix Prize Challenge in 2006 (Hafner, 2006). The goal of the competition was to achieve a 10% reduction in Root Mean Squared Error (RMSE) on a dataset of 100 million customer generated movie ratings. The result of the challenge was that no single model had the best performance metric. Instead, the best performance came from combining predictions of models that complement each other. It was found that each separate model addresses a variety of structure in the data set (Bell and Koren, 2007).

One of the earliest powerful ensemble models is boosting. The main idea of the algorithm is to maintain a distribution or set of weights over the training set. The weight of this distribution on training example i on around t is denoted Dt(i). Initially, all weights are set equally, but on each round, the weights of incorrectly classified examples are increased so that the weak learner is forced to focus on the hard examples in the training set. The weak learner's job is to and a weak hypothesis $h_t : X \rightarrow \{+1, -1\}$ appropriate for the distribution Dt. The goodness of a weak hypothesis is measured by its error (Schapire, 1999).

Bagging predictor is one of the early ensemble-learning models. The main idea of bagging is to generate multiple versions of a predictor model and utilize these learning models to get an aggregated predictor (Breiman, 1996). In bagging the learning set of Θ consists of data {(yn, xn), n = 1, ..., N } where the y's are either class labels or a numerical response feature with N independent observations. The algorithm creates a k number of bootstrapped samples Θk from the original data set with the same underlying distribution as Θ. On each of the Θk a single learning

predictor is used to get predictions. The class label y is majority voted across all the predictor's results to form the final prediction y.

A common element of all the ensemble methods is that for each tree a random vector of Θk is generated with same distribution from the learning data independent of previous Θ1,...,Θk−1. A more powerful modeling method, Random Forests are a combination of decision tree predictors where each individual tree derives on a random vector bootstrapped sampled independently with the same distribution for each of the tree in the forests (Breiman, 2001). Introducing random split selection at each level of the tree was able to create uncorrelated decision trees formed by selecting at random at each decision node on a bootstrapped sample of input training data. Injecting the right kind of randomness makes them accurate classifiers and regressors. Furthermore, the framework in terms of strength of the individual predictors and their correlations gives insight into the ability of the random forest to predict (Breiman, 2001). Empirical improvement in classification error and mean error have been achieved from growing ensembles of decision trees letting them vote for the most popular class or averaging continuous results. Besides, it is very user-friendly in the sense that it has only two parameters (the number of features in the random subset at each node and the number of trees in the forest) and is usually not very sensitive to their values.

Gradient boosting model is a greedy algorithm of the ensemble models and can over-fit the data very effortlessly. Gradient Boosting builds an additive model using a forward fashion. At each iteration for subsequent trees, decision trees or decision stumps which are called weak learners are built and added to the existing weak learners to minimize the loss of the model (Friedman, 2001). A gradient boosting has three main elements: A loss function to optimize for, weak learners to

make predictions and an additive model to add weak learners to minimize the loss function. Therefore, there are several hyperparameters (number of trees, random sampling, penalized learning) to tune which can decrease the variance of the model created. The number of trees is an essential parameter to tune to increase accuracy and decrease variance. Random sampling of the data is a significant insight into ensemble models. Another approach to decreasing variance is stochastic gradient boosting (random sampling) (Friedman, 2001). At each iteration, a sub-sample of the training data is drawn at random (without replacement) from the full training dataset. The randomly selected sub-sample is then used, instead of the full sample, to fit the base learner.

2.5.3 Entropy with bagging tree methods

Many factors influence the success of machine learning in a given task. The quality of data is such a factor; if the information is irrelevant or unnecessary or if the data is noisy and unreliable, the discovery of information is more difficult during training. The feature subset selection is the process of identifying and removing as much as possible, unrelated and extra information. Machine learning algorithms vary in the amount of importance they give to feature selection. On one side, there are algorithms like the nearest neighbor modeling which classifies new samples by taking the nearest stored training instance using all available properties in the distance calculations. The other extreme pseudo-algorithms are the algorithms that try to focus clearly on the relevant features and ignore the irrelevant ones.

Decision tree modeling is an example of extracting relevant feature and ignoring the irrelevant ones using entropy. By experimenting with the values of

specific features, decision tree algorithms try to divide the training data into subgroups that contain a substantial majority of a class. The increase in the majority of one class leads to a decrease in entropy in that subgroup. This is accomplished by first selecting a small number of highly predictive features (significant in entropy loss) to avoid over-use of training data.

2.6 Feature analysis in nonlinear modelling

One of the most significant setbacks when it comes to the usage of nonlinear models occurs where there is a loss of capacity in understanding the prediction outputs in the model. Due to higher capacity and higher accuracy, researchers have since then given up on the interpretability. One of the most, if not the most, vital abilities is to be able to understand the prediction models feature effect onto the output correctly. With the availability of interpretability, researchers are given trust and understanding of the phenomenon mentioned above. During many applications, researchers have started to apply simple models. The simple models can include the following: linear regression, ANOVA, as well as decision trees. They have done this even though they run the possibility of being less accurate than the more complex models. We have witnessed a truly significant increase in big data and computation power. Parallel to this, we have also seen a notable increase in the use of far more advanced nonlinear models to detail various phenomena. However, there has been a drawback regarding the loss in feature understanding. When observing current research, researchers have used linear regression in attempts to quickly understand and deduce the relationship between features and output value with the use of simple statistical analysis. In addition to this, a more sophisticated analysis was conducted. These were performed by decision trees to inspect and understand the important features visually. Lastly,

when the researchers are using nonlinear models, the examples could be gradient boosting and random forests researchers used entropy gains and number splits to understand the behavior of the model. They do this by taking very high-level pictures of millions of trees that come together to predict one output.

The nonlinear models are very complex however it is easy to approximate a linear functional form around the vicinity of an instance. The model is treated as a black box, however the instance that needs to be explained is perturbed and a sparse linear model is learned around it to provide an explanation. As can be seen in Figure 2. local linearization procedure has a structured approach to be applied. We sample instances around X and weight them according to their proximity to X (weight here is indicated by size).

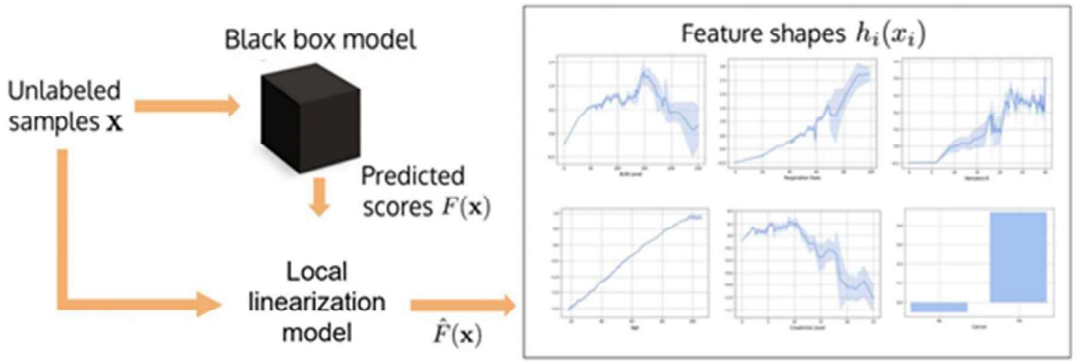

Figure 2. Given a black box model and unlabeled samples, local linearization approach leverages model distillation to learn feature shapes that describe the relationship between input features and model predictions.

In contrast, there are some authors that consider tendency to explain such com-plex models in a narrow point of view approach to simply be far too optimistic (Alvarez-

Melis and Jaakkola, 2018), to successfully counter this, it can be said that the simple approximations do in fact give important, valuable awareness into understanding the causality features that drive the model's prediction as well as its accuracy. They also give opportunities for vigorous validation procedures and improvement of the model (Fong and Vedaldi, 2017). Another way of saying this is the following; these interpretation methods were established to help illuminate scientific problems where the human intuition and domain knowledge are limited often (Montavon et al., 2017). Contrary to certain beliefs, the purpose of the existing interpretation method is not to explain the sensible idea behind the black box but to give realistic answers for the choice in an instance. (Guidotti et al., 2018). The current interpretation framework has led to inconsistent results and often opposing results for machine learning algorithms. The SHAP framework has shown an original solution in explanation models intended for post-hoc interpreting machine learning methods where this is more in line with human intuition. Dissimilar from attempts that provide a specific global predictor, the SHAP framework explains the tree ensemble's overall behavior in the method of specific feature contributions. In line with other methods for interpreting machine learning predictions, SHAP is continually and with great vigor becoming a prevalent tool in the forecasting of natural and social phenomena.

For example, the current study of Janizek put both SHAP and an XGBoost tree-based method together to predict successfully and explain the synergy of novel drug combinations for a much more accurate cancer treatment process (Janizek et al., 2018). The idea of the polynomial time algorithm for SHAP values as an alternative exponential time algorithm is to repetitively follow the amount of all possible subsets as they settle down into each tree leaf. The rapid reduction in difficulty gives options to the traditional partial dependence, and feature importance plots (Friedman, 2001),

this has been termed as SHAP dependence and SHAP summary plots, individually. Because they are feature attributes that have been individualized exclusively to each and every prediction, the SHAP values do allow for better captures of interaction effects. Different from partial dependence plots that give representation to the dependency of a model on a subset of features that have all the other features fixed, the SHAP dependency plots do in fact capture the features importance - as well as the changes as the value of the features varies. Furthermore, differing from the protocol partial dependence plots that produce lines, SHAP dependence plots collect interaction effects in the model, represented as vertical distribution. By combining the SHAP dependence plot and SHAP plots, we can see true global interaction patterns which were not able to be identified previously. We have used supervised clustering based upon SHAP feature attributions, SHAP summary plots, and partial SHAP dependency and interaction plots to explore the effects of interaction between the relevant factors.

In second-hand automobile pricing, by utilizing a local linearization modeling, the phenomenon behind how a second-hand automobile price is nonlinearly constructed is interpreted. These methods enable to represent a consistent and locally accurate additive feature attribution method of the phenomenon, and thus the pricing of second-hand automobiles can be interpreted in novel methods.

CHAPTER 3

DATA AND PREPROCESSING

Data set used in experimentation is secondary data type from a highly reliable online listing web page. Due to its nature of being a secondary data of the used automobile cars and along with the data collection and cleaning methods, the data does not need to be processed for validation and verification to check reliability and consistency. The marginal automobiles are excluded from the sample to prevent outliers.

Data used approximately consist of 313570 number of listings in the online car listing websites from the primary listing source, sahibinden.com. Each listing unit has dependent and independent features such as price, brand, model, series, color, sales location, milometer or odometer, warranty.

3.1 Advanced modelling strategy

A predefined strategy is used to incorporate the collected data into a structure, build a model, analyses performance, and apply local linearization on the predictive model. The advanced modelling strategy as can be seen in Figure 3. an advanced predictive model is built from start to end with a structured approach.

In the initial step, data is collected from comprehensive online web sources using advanced scraping algorithms. In the first phase, the dataset is refined using the information on existing literature to identify the most influencing factors that should be included in the dataset. Two separate datasets are formed, one for encoded variables that are used in linear regression and the second dataset as the features encoded with a categorical formation that is used in the gradient boosting algorithm.

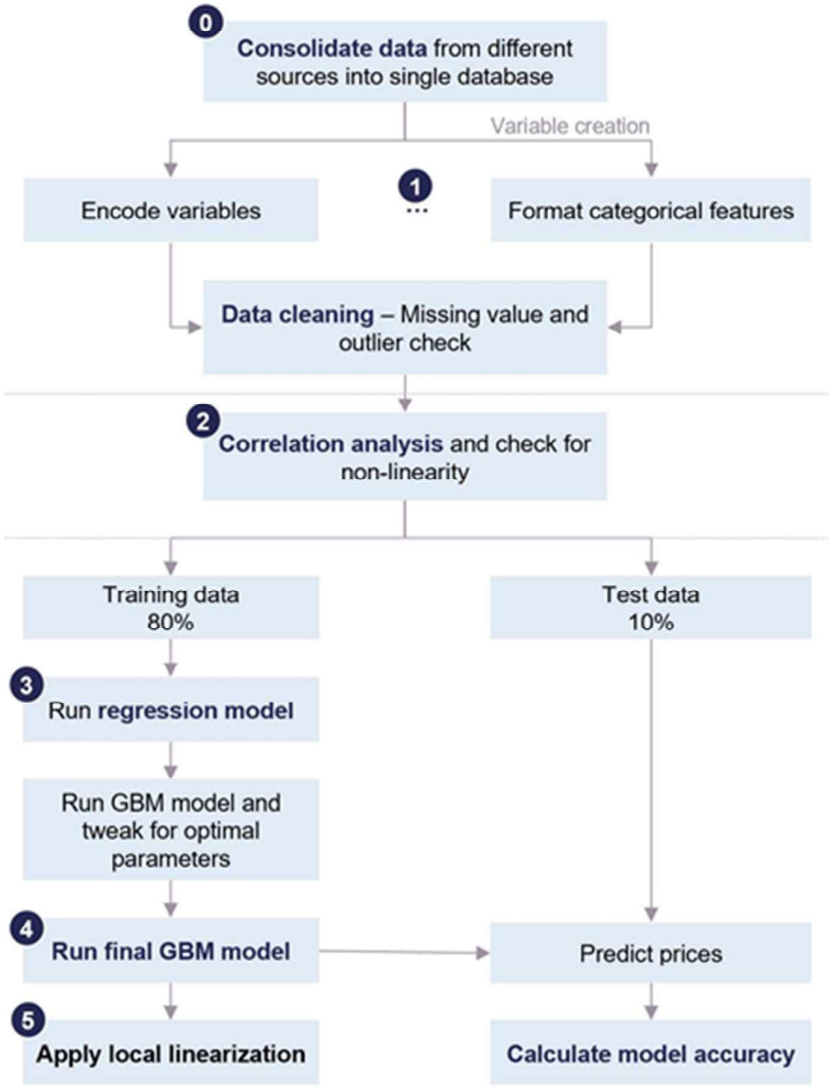

Figure 3. The strategy developed to build the model

In the second phase, a correlation analysis of the most influencing factors is analyzed. Moreover, visual analysis of the dataset is performed to generate fundamental statistical analysis.

In the third phase, the linear regression model is trained on the training data, and baseline performance is obtained. In the fourth phase gradient boosting algorithm is trained a tuned to obtain a robust predictive model. In the final step,

local linearization is applied to understand the effect of features in a nonlinear modeling context. The process is built using python 3.6 environment on an Intel XEON 8-Core CPU with 64GB RAM. The used python libraries are NumPy (Van Der Walt et al., 2011), SciPy (Oliphant, 2007), Pandas (McKinney, 2015), Jupyter (Kluyver et al., 2016), SkLearn (Pedregosa et al., 2011), Plotly (Sievert et al., 2017), LightGBM (Ke et al., 2017) and Shap (Lundberg et al., 2018)

3.2 Exploratory data analysis

After preprocessing the data, it is analyzed through the visual exploration to gather insights about the model that can be applied to the data, understand the diversity in the data and the range of every field. In the analysis, a bar chart, box plot and distribution graph are used to explore each feature varies and its relationship with other features including the target feature.

3.3 Data collection

The internet contains all kinds of information of different origins; some of those are social, financial, security and academic. Most people access information through the internet for educational purposes. Information on the web is available in different formats and through different access interfaces. Therefore, indexing or semantic processing of the data through websites could be cumbersome.

Web scraping is the technique which aims to address this issue. Web scraping is used to transform unstructured data on the web into structured data that can be stored and analyzed in a central local database or spreadsheet (Sirisuriya et al., 2015). Traditional copy and paste are the basic and tiresome web scraping technique

where people need to scrap lots of datasets. Web scraping software is the easiest scraping technique since all the other techniques except traditional copy and pastes require some form of technical expertise.

A web scraper and spider are used to collect relevant information on used automobile listings. The task of capturing and structuring data mined from the web has two distinct phases: crawling and scraping. Second-hand automobile data collection was completed in the run-time of 6-hours uninterrupted using Python 3.6 with Scrapy library from the most extensive online second-hand market sahibinden.com and arabam.com.

Scrapy is an open-source web crawling framework written in Python. Its primary purpose is to provide support for web scraping, compatible with Python 3.6 Scrapy is an integrated system that includes an engine for controlling the data flow between all components, a scheduler for receiving requests, a downloader for fetching web pages and custom classes which are called spiders written by users to parse responses and extract items from the responses. The scrapers are working as human reader would go through webpages. First it loads the list of links and goes through each link one by one to extract automobile data. In each link it loads only text information to save on data while loading which increases speed and also reduces cost of data extraction. If the parsing of list of links have completes, the scraper automatically starts parsing a different brand. The complete parsing and extraction of information completes in seven hours. The website example can be observed in Figure 4. and Figure 5. These are random examples from the two main webpages that were scraped in data collection.

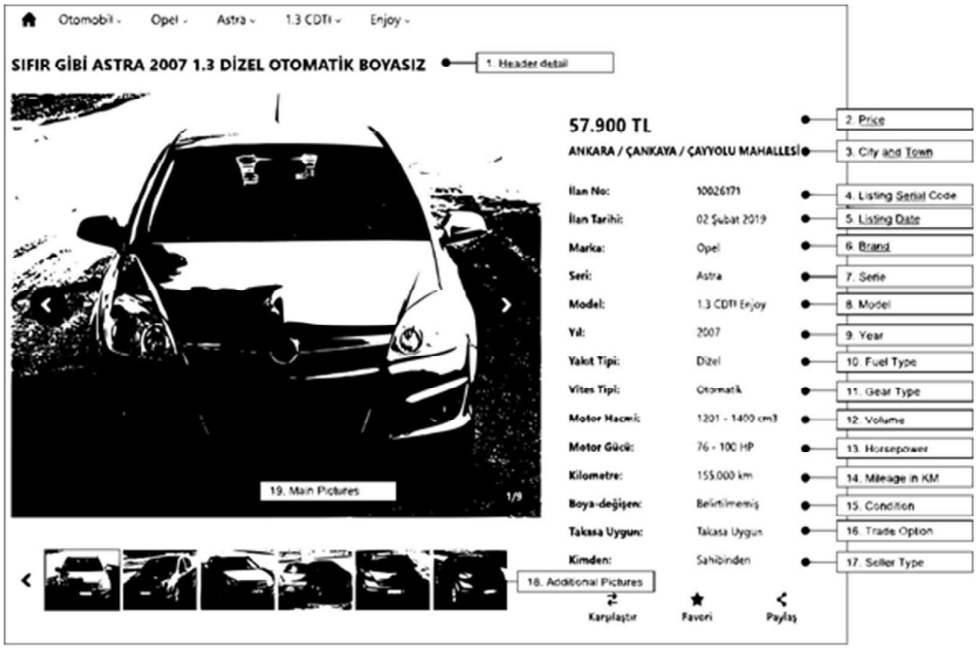

Figure 4. Example webpage of arabam.com listing page where the scraper targets for information

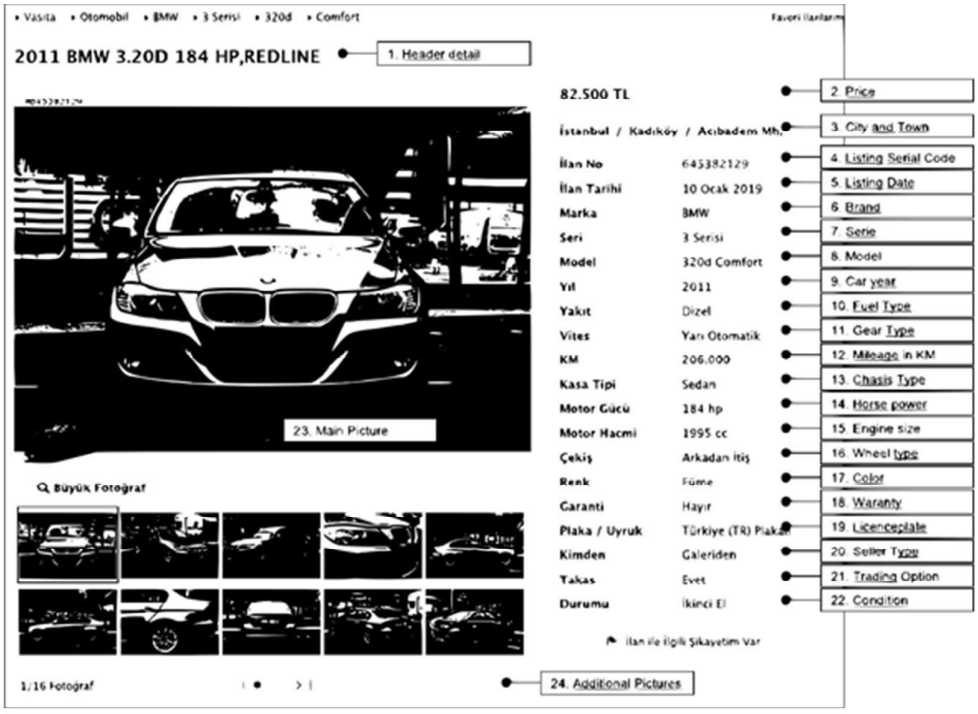

Figure 5. Example webpage of sahibinden.com listing page where the scraper targets for information.

3.4 Data preprocessing

The data has various categorizations of categorical and ratio data types. Most data will be of ratio and categorical such as brand, model, location as nominal but there will also be ratio data such as age and mileage of the car.

1. Keep only listings for cars sold by private owners and filter out those sold by dealerships
2. Keep only listings for cars being sold from sellers and dealerships
3. Filter out cars manufactured before 1900 and after 2018, and derive the car's age
4. Filter out all cars with unrealistic power values
5. Filter out listings which don't have an associated price
6. Filter out all cars listed as unavailable
7. Filter out invalid registration dates
8. Convert Boolean (true/false) fields to numeric (0/1) based
9. Filter out all data with value as 'NA' (Not Available)

Feature price and kilometers are in thousands. Categorical features such as location, region, warranty, fuel type, gear type, features are recorded in binary format. Continuous features such as engine volume, horsepower, price, kilometers (mileage), model, age recalculated and if necessary transformed.

CHAPTER 4

COMPUTATION AND ANALYSIS

In this section, the main analysis is conducted on visual exploration, model building in linear and nonlinear functional forms are built using the cleaned dataset, application of sensitivity analysis using partial dependency plots and application of local linearization using shap method.

The data is retrieved from two independent sources to over two years at varying seasonal intervals. The collected data comes from the largest two major online automobile listing source; arabam.com and sahibinden.com. After the collection of data for two years, each source is randomly sampled by 160,000 listings per source.

The random sampling in a high quantity of observations as can be seen in Figure 6 allows building models that are not biased towards one source of data but a representation of the second-hand automobile vehicle market. It can be observed that the distributions are very close when data is randomly sampled at high quantities, the median is very close, and the probability distributions of the prices are similar at each randomly sampled source.

A descriptive analysis of post data cleaning is conducted to understand the underlying distributions that the model learns. The main highlighting results are gathered using pandas describes function. In total there are 313570 observations from two large resources as can be seen in Table 1. The cars on have a maximum of 7 seats, the year-make of the vehicles range from 1980 to 2019. This means that the models will learn a model age of 40 to brand new cars.

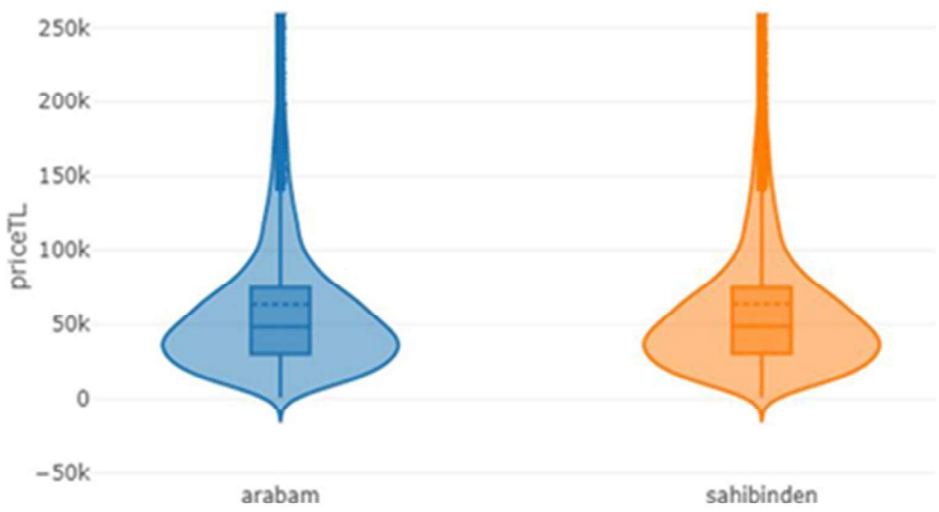

Figure 6. Comparison of price to data sources using a box-plot, probability distribution estimation.

The currency has a high variance however this is due to the fact of increasing currency rate in the period that the data was collected. Although this feature may not have a direct price effect, it can be a useful predictor in explaining the price difference of the same car in different years at local linearization. The most important is the target feature variable which is the price of vehicles. This is the data that is being described using all the available features. The maximum price of a vehicle is 800000 Turkish Lira with a minimum price of 1000 Turkish Lira. The most occurring (median) price is 48500 TL and most of the vehicles at 80th percentile are 80000 TL.

The model will learn a very general price distribution that encapsulates a variety of customer segments. The number of drivers in the city data shows how the licensed drivers in the city changes. This is a proxy feature for demand and supply of

second-hand automobiles. It shows that the lowest number of the licensed city has 7000 and the maximum licensed city has 5.8 million licensed drivers.

Table 1. Descriptive analysis of numerical features.

	count_seats	year	priceTL	horsepower
count	313570	313570	313570	313570
mean	0.000159454	2017.630386	63889.36525	111.559476
std	0.030094457	1.450724363	60128.77919	44.8527798
min	0	1980	1000	39
25%	0	2018	29750	84
50%	0	2018	48500	102
75%	0	2018	75000	125
max	7	2019	799000	700
	volume	tryusd	tryeur	
count	313570	313570	313570	
mean	1531.118273	4.691573065	5.33995238	
std	395.8305181	0.593619526	0.708851871	
min	57	3.4069	3.684	
25%	1390	4.429	5.1748	
50%	1499	4.5996	5.3482	
75%	1598	5.3863	6.0319	
max	7011	5.3863	6.113	
	detail_length	number_of_drivers_2017		
count	313570	313570		
mean	87.22695092	2164494.202		
std	235.1729479	2302052.403		
min	0	7044		
25%	25	384867		
50%	40	828041		
75%	59	5871653		
max	16167	5871653		

As a key step in data analysis, observing how the individual model is represented in each random sample of the data source is important to validate that there are enough variations in the dataset.

It can be observed in Figure 7 that both sources resemble very similar distributions in quantity of brand listings and average prices for the vehicles.

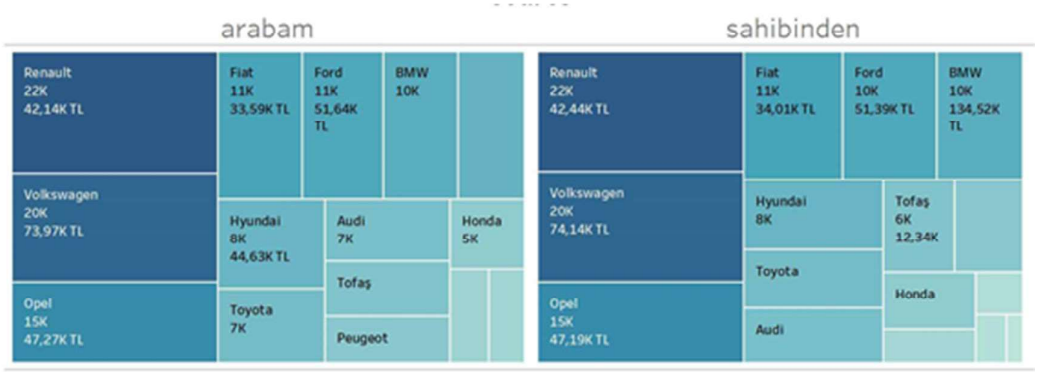

Figure 7. Descriptive analysis of continuous features.

One of the most important factors is how the population resembles the market. It can be observed in Figure 8 that most vehicle types are from Renault with 43000 observations, followed by Volkswagen with 39000 observations and Opel ranking as third most observations with 29000. High-end vehicles such as BMW and Mercedes-Benz have 19000 and 17000 observations respectively.

Color is one of the essential features in second-hand automobile literature. Analysis of the color breakdown using group by shows that white cars are in abundance in the market with 37.9% of the total second-hand market as can be seen in Figure 9, followed by black cars at 13.4% and Silver-Gray vehicles with 11.06% of the entire market.

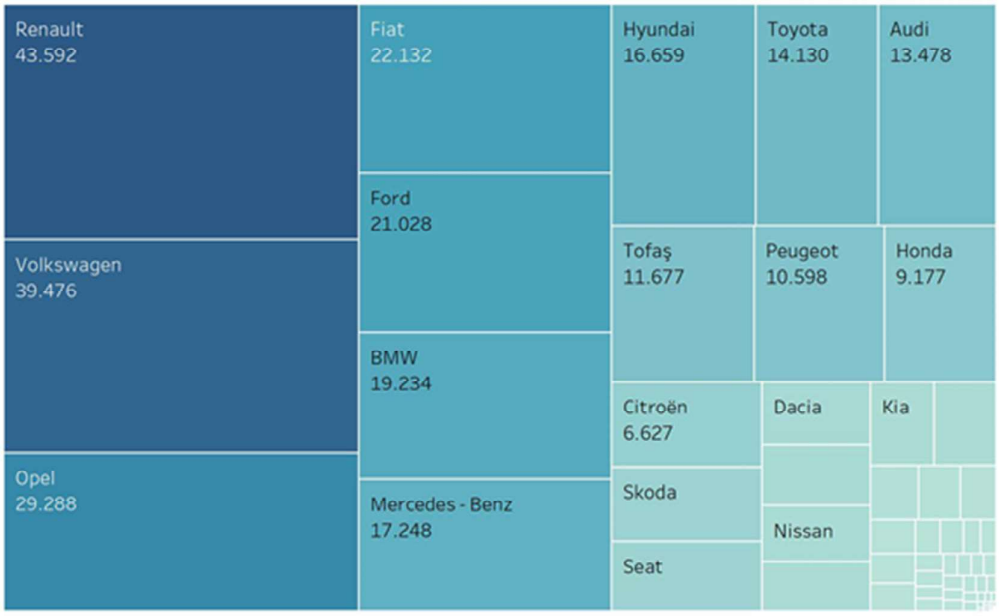

Figure 8. Breakdown of brands in counts of listings

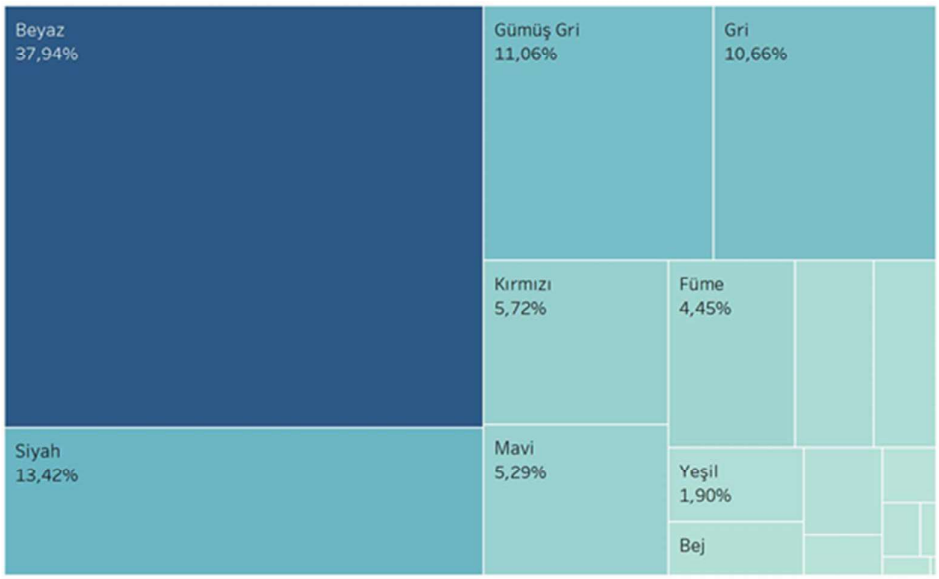

Figure 9. Breakdown using boxing for color features in the percentage of total listings

The listings are analyzed by the gear type which can be a significant factor influencing the price. It can be observed in Figure 10 that manual-type gear is the most listed with 190542 listings but with lowest average price and semi-automatic cars have the highest average price over 110000 TL with 83449 listings. We recorded each gear type to observe its effect on the price, having manual gear type as a base.

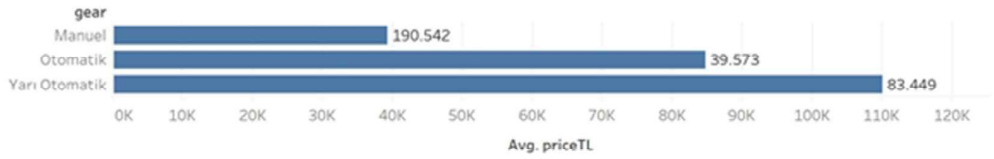

Figure 10. Breakdown of gear types in counts and average listing price.

The listings are analyzed by the fuel type which can be a significant factor influencing the price. It can be observed in Figure 11 that Diesel (Dizel) type of fuel is the most listed with 85157 listings. LPG kit automobiles have the lowest average price, and hybrid cars have the highest average price over 160000 TL with 76 listings. We recorded each fuel type to observe its effect on the price, having gasoline fuel type as a base to see the effect of having an LPG kit or Diesel Engine.

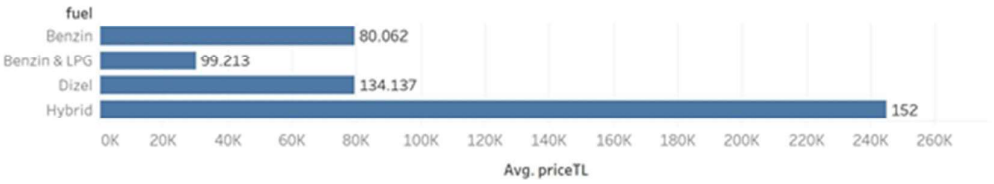

Figure 11. Breakdown of fuel types in counts and average listing price.

In Turkey, cars have the horsepower in the range of 60 to 140. There is a relation between a car's horsepower and its price as can be seed in Figure 12. As the horsepower increases above the market, the prices increase remarkably. This is related to the volume of the engine because cars are taxed per their engine volumes in Turkey. The taxation increase is reflected in the prices. It can be observed that in each level of taxation brackets (1300, 1600, 2000, 2500, 3000, 3500, 4000 Plus for volumes the average prices increase. In the analysis, we included both features because the price is not solely based on taxation on engine volume but also the prices are affected by the preference for more powerful cars.

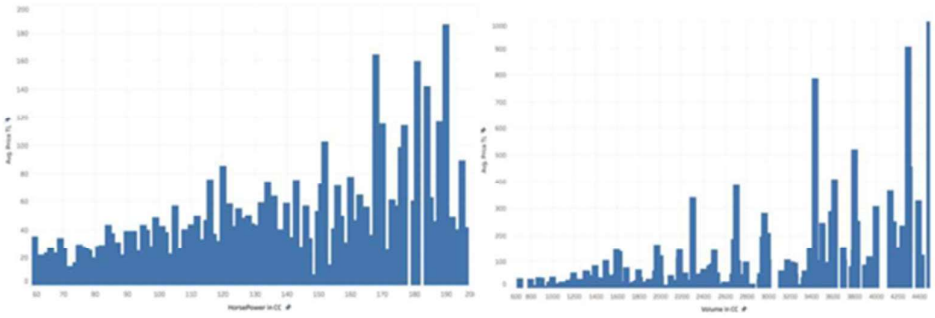

Figure 12. Left: Average listing prices plot with respect to horsepower. Right: Average listing prices plot with respect to vehicle engine volume.

Using the model year of the car, the age can be calculated for the observations. As the car ages the average price of the car decreases as can be seen in Figure 13 but as the car ages above a certain level, it becomes antique models; thus, this can be observed as the age increases price decreases to a certain level and then a slight bumped shape occurs.

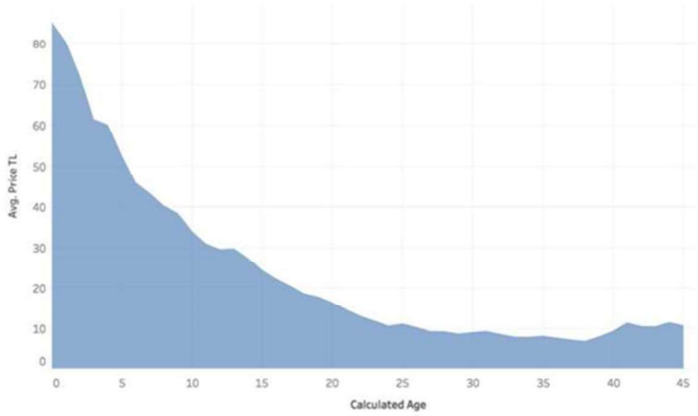

Figure 13. Average price change with car age as a calculated variable from model year

In analyzing how the mileage distributes across the country, it can be observed in Figure 14 that Istanbul has the most used cars followed by the Southern Eastern region of Turkey. This separation can be a significant factor to utilize city information as an explanatory feature for understanding how prices change within different regions.

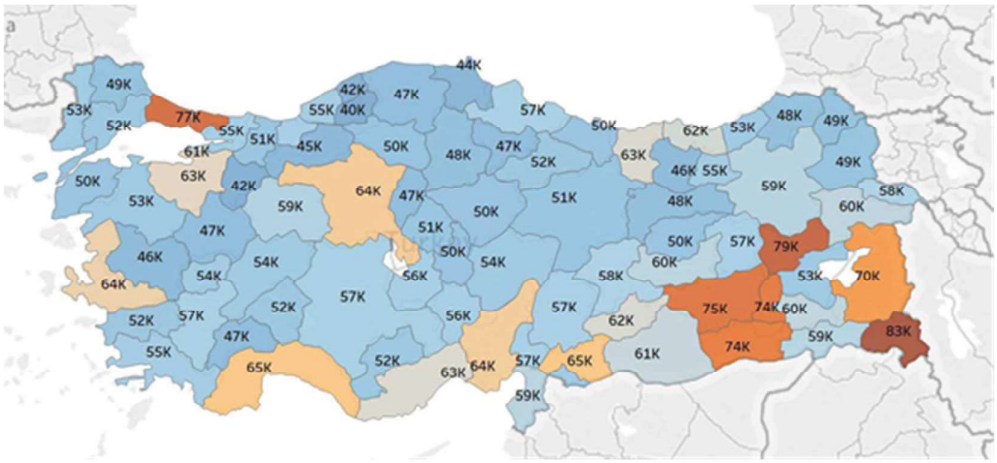

Figure 14. Geographic price plotting of vehicles. Listing prices are averaged by city aggregation of Turkey.

One of the key features of a listing is the horsepower of a vehicle. Horsepower of a vehicle indicates how powerful that car is in acceleration, top speed, and fuel consumption. It is critical to visualize how average horsepower distributes across the country in cities. It can be observed in Figure 15 that major cities such as Istanbul and Southern Eastern region have the highest average horsepower vehicles. In a business perspective, this makes sense due to two main reasons; firstly, economic size of a region is correlated with vehicle types because richer people will tend to purchase, own and sell more powerful vehicles and geographic region type can be correlated with horsepower because due to business nature of southern, eastern regions in agriculture, farming and steep hilled nature the vehicles need to be more powerful with higher horsepower.

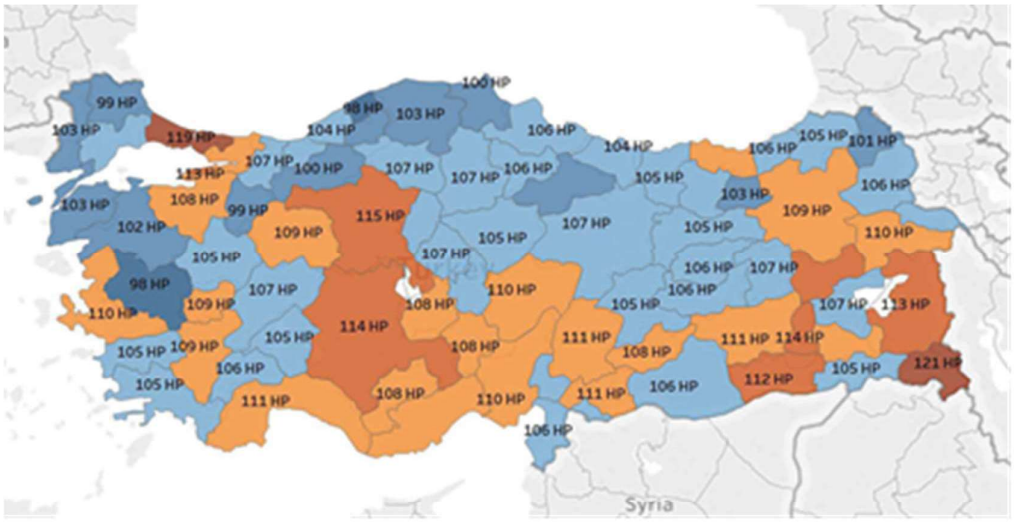

Figure 15. Geographic horsepower plotting of vehicles. Listing horsepower are averaged by city aggregation of Turkey.

In the dataset collected there are a total of 313570 listings. A high-level descriptive analysis validates the diversity of the market information that is included in the

research. The target variable which is the price ranges from 1000 TL to 800000 TL with a 75th percentile at 80000 TL. It can be observed in Table 2 that in the car listing dataset, the range of year of the car is between 1980 to 2019 with an average model year of 2017. This is because most of the listings have been collected during the years of 2016, 2017, 2018 and the beginning of 2019. The listing's horsepower ranges from 40hp to 700 HP indicating that in the analysis a full range of vehicle types are covered. The currencies are retrieved individually for listings dates and TRY to USD range from 3.4 to 5.4 in exchange rated where TRY to EUR range from 3.6 to 6.04 on these dates. All listings have details in the web pages; it can be observed that details have been written from no detail to texts up to 16 thousand words. Lastly, we can observe that the number of licensed drivers for the city listing is from can range from seven thousand to six million. It can be concluded from the listings that the data in the research is a diverse set with information ranging from low segment to high segment vehicles.

Table 2. Descriptive analysis of continuous features of the dataset. The analysis contains count, mean, standard deviation, minimum value, maximum value, 25th percentile, 50th percentile and 75th percentile.

	count_seats	year	priceTL	horsepower	km	volume	tryusd	tryeur	detail_length	number_of_drivers_2017
count	313570	313570	313570	313570	313570	313570	313570	313570	313570	313570
mean	0.000159454	2017.630386	63889.36525	111.5594764	40261.38909	1531.118273	4.691573065	5.33995238	87.22695092	2164494.202
std	0.030094457	1.450724363	60128.77919	44.85277981	99827.77564	395.8305181	0.593619526	0.708851871	235.1729479	2302052.403
min	0	1980	1000	39	0	57	3.4069	3.684	0	7044
25%	0	2018	29750	84	870	1390	4.429	5.1748	25	384867
50%	0	2018	48500	102	1680	1499	4.5996	5.3482	40	828041
75%	0	2018	75000	125	2900	1598	5.3863	6.0319	59	5871653
max	7	2019	799000	700	499990	7011	5.3863	6.113	16167	5871653

As mileage of the vehicle increases prices tends to show a decline which can be seen in Figure 16. This means that there is a negative relationship between the price of the

vehicles and the mileage. This can be considered as a factor for depreciation of the priced good. Consumers will be willing to pay less to goods due to the depreciation.

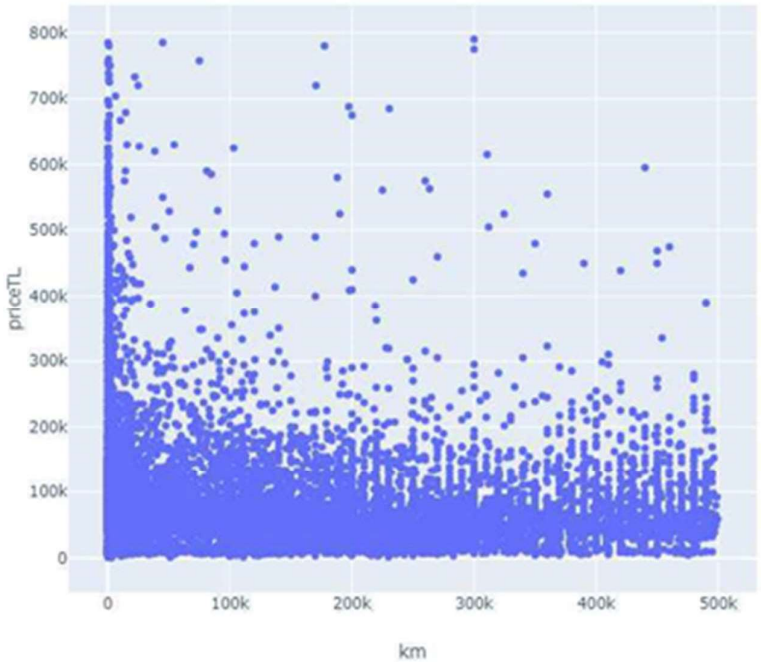

Figure 16. Scatter plot of odometer in kilometers to price.

Moreover, the price scatter-plot with respect to km exhibits a significant logarithmic relationship as can be seen in Figure 17. This means that the price and mileage have not only a negative relationship but also a nonlinear relationship. As for mileage increase the price de-creases at a decreasing rate.

The logarithmic nonlinear relationship can be validated by transforming the dependent variable relationship with log-transformation. The linear scale shows the absolute price of vehicles over mileage while the logarithmic scale shows the change of the price of vehicles over mileage. In the log-transformed figure, it can be observed that the logarithmic relationship is flattened out. This validates that the

relationship is logarithmic. Modeling the phenomenon with non-transformed linear regression models can yield to poor, learned relationship models.

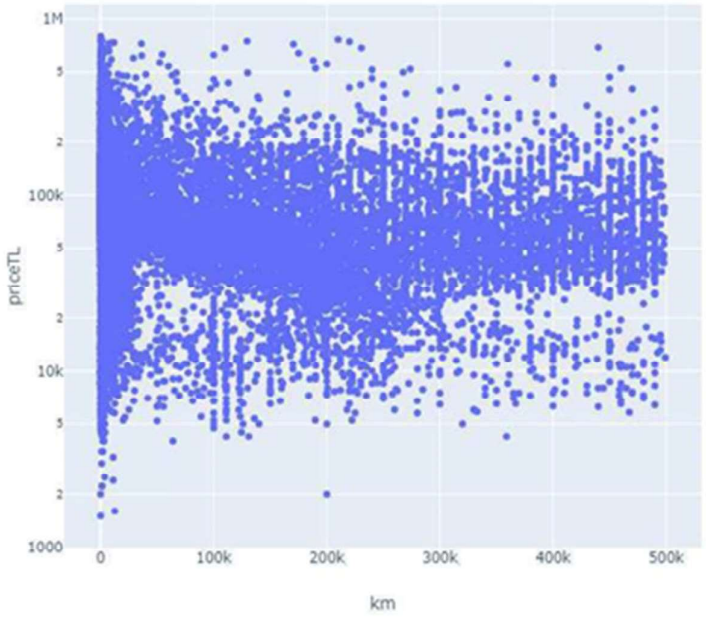

Figure 17. Scatter plot of odometer in kilometers to logarithm of price.

Violin plots of Mercedes, Audi, and BMW, in Figure 18 indicate that the prices have a similar distribution with close means and percentiles. Statistically if categories have similar median and mean with similar distribution, this indicates that cars in similar segments have similar distributions. The similar behavior is not observed for vehicles that have a different segment attribute.

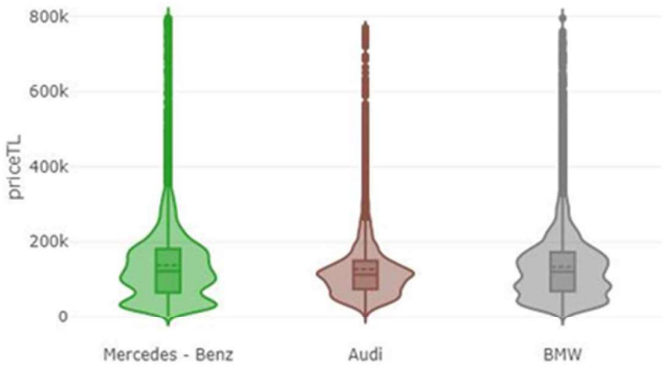

Figure 18. Comparison of price distribution to Mercedes-Benz, Audi and BMW using a box-plot, probability distribution estimation.

Similarly, Renault, Opel, and Honda show similar price distributions in Figure 19. This is because brands in the same segment exhibit very close means and distributions, however different distribution than other segments.

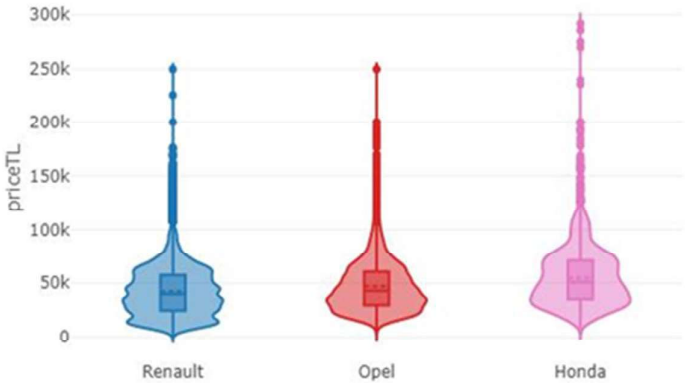

Figure 19. Comparison of price distribution to Renault, Opel and Honda using a box-plot, probability distribution estimation.

It can be observed that the probability distributions of three types of sellers are variant in Figure 20. The data exhibits that prices are lower when the seller is the direct owner of the car, secondly on average prices are higher when the seller is second-hand vehicle merchant and thirdly if the vehicle is being sold from a verified channel the prices are highest. This relationship has two outcomes in understanding how the features' effect on vehicle prices. First is that from a statistical modeling perspective, a model will be able to find separation planes because of different probability distributions exhibited by the feature. The different distributions will help the model to learn using separate distributions by giving splits on three different categories. Second is that from business perspective reduction in risk of purchasing a second-hand vehicle is added to the price of the car and additional operating costs of having a business are reflected on the automobile prices.

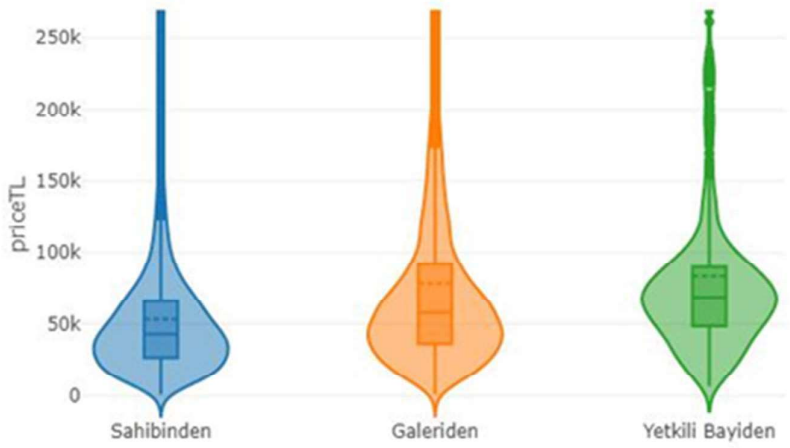

Figure 20. Comparison of price distribution to ownership by Sahibinden (Personal Owner), Galeriden (Car Distributor), Yetkili Bayiden (Certified Distributor) using a box-plot, probability distribution estimation.

The probability distributions and the box-plots of categorical variables contain different information at each category level. The lowest median and distribution range is with vehicles that have an LPG gas system installed as can be seen in Figure 21. The cars that only have gasoline has a higher median than with cars that have the LPG system installed, and the distribution has higher prices. Thirdly, diesel cars exhibit similar distribution with only gasoline system vehicles with similar median and distribution of prices. Lastly, electric vehicles show a significant price distribution shift compared to other categories. Like seller type feature, fuel-type can be a significant feature that can help model the price of second-hand automobiles. Firstly, the significant shifts in probability distributions are good explanatory categories to describe a phenomenon.

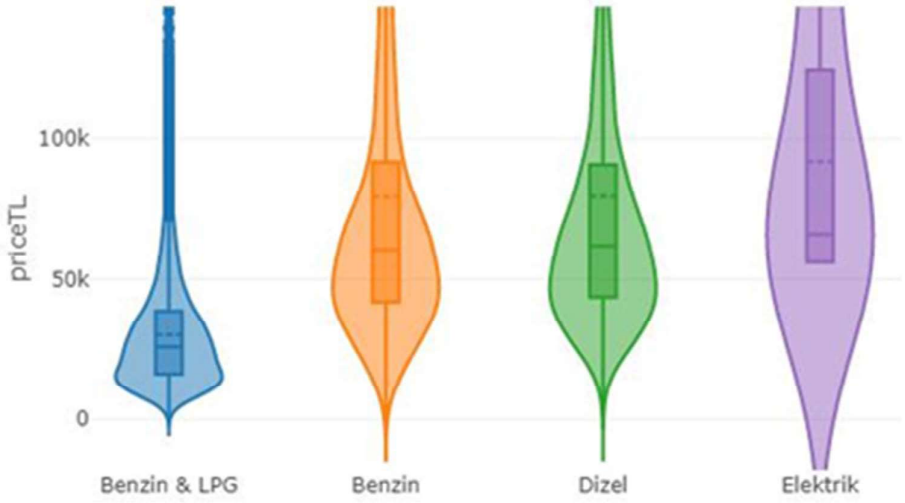

Figure 21. Comparison of price distribution to fuel type of vehicle in Benzene-LPG mix, Benzene, Diesel and Electric using a box-plot, probability distribution estimation.

The dataset shows that black cars (tagged as siyah) have the highest mean price and range of prices, followed by white cars (tagged beyaz) have the second highest prices in Figure 22. The distribution of car prices in different colors is also varying. It can be observed that black cars are more accumulated around below 50000 TL price whereas white cars are more accumulated around above 50000 price TL. It can be observed that colors and prices have different distributions in probability density curves. This shows that the color of an automobile can be a perfect predictor of prices for second-hand automobiles.

Figure 22. Comparison of price distribution to color types using a box-plot, probability distribution estimation.

One of the critical features in second-hand prices is whether the automobiles have warranties in listings. The warranties are a mean of security of consumers and are a Premium for any second-hand automobile as can be seen in Figure 23

The dataset's results show that have a warranty increases the median price of automobiles and now having a warranty decreases the median prices. Secondly, a

clear differentiation in probability density curves can be seen. The density differentiation is a good predictor that can be used in regression of second-hand automobile prices.

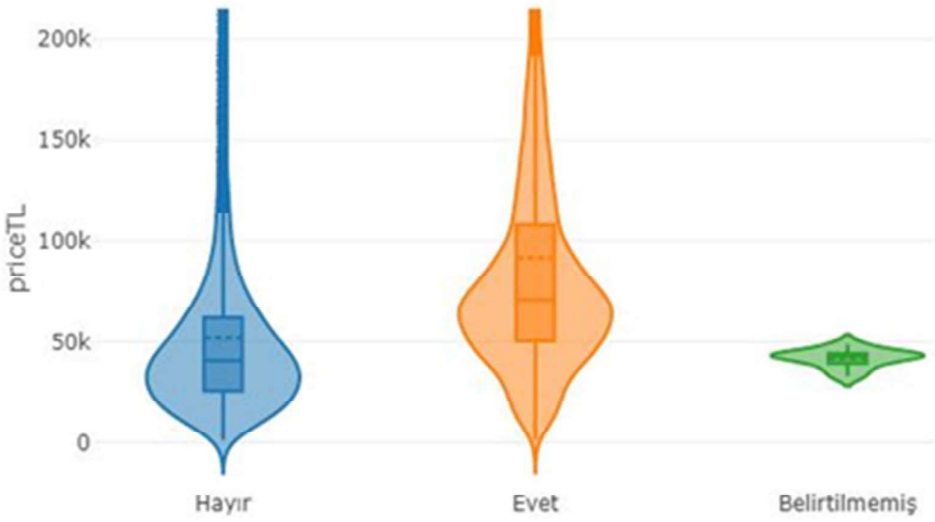

Figure 23. Comparison of price distribution to whether warrant exists using a box-plot, probability distribution estimation.

In second-hand automobile listings, the dataset exhibits that if a car has some crash, it decreases the median list price of the vehicles. If the vehicle has been tagged as no crash the median value increases as can be seen in Figure 24. Interestingly, if a vehicle has not been given any information in crash conditions the overall median price increases, second-hand automobile market has a vehicle in diverse conditions.

The most important factor for buyers is the condition of a vehicle in the cosmetic and internal engine. If a vehicle has a crash, it would have a lower value whereas if a vehicle has no crash, it is value will increase. Lastly, the crash condition is a factor that is critical for consumers and is a significant factor that should be

modeled in the second-hand automobile price listings. Although the probability distributions do not exhibit wide discrepancy for each category when inspected from the local linearization approach the predictor can gain importance in describing the phenomenon.

Figure 24. Comparison of price distribution to whether there is no crash, unstated or crash registered using a box-plot, probability distribution estimation.

A correlation matrix is a table that shows the correlation coefficients between each independent feature that is used in the modeling. At each box for a pair of the feature, the value is the degree of correlation which ranges from negatively correlated (-1) to positively correlated (+1) if the feature does not correlate then it means that the two features exhibit no behavior that moves together.

In the correlation analysis, it can be observed in Figure 25 that several factors move together with price and features that move together which signals collinearity.

This gives insight into what feature to use in the model or generate insights on the feature dynamics of the market. The most positively correlated feature with the price of a second-hand vehicle is the horsepower and gear of the vehicle. The most negatively correlated feature with the price is the mileage (km) in the odometer and detail length of the listing. This has two main outcomes from a business perspective; firstly, the technical specifications of a vehicle are critical in determining the prices of second-hand automobiles and secondly the usage of the vehicle which is signaled in mileage (km) of the vehicle shows a value depreciation in the negative direction.

4.1 Predictive models

In understanding how the features affect price in second-hand automobiles it is essential to build a predictive model that can capture the whole behavior with a reasonably high R2 value and low mean absolute errors. If the model has a low R2 than this would prevent to make reliable conclusions on how the model features describing the phenomenon are affecting the price for predictions.

4.1.1 OLS regression

Ordinary least squares (OLS) is a powerful regression method that has been used in literature for second-hand goods market and specifically for second-hand automobiles. The phenomenon can be described with an additive formula from which a holistic effect can be concluded. The second-hand automobile dataset was trained using OLS with an 80% split on training data with 313570 observations and 20% split into testing data with 24 variables.

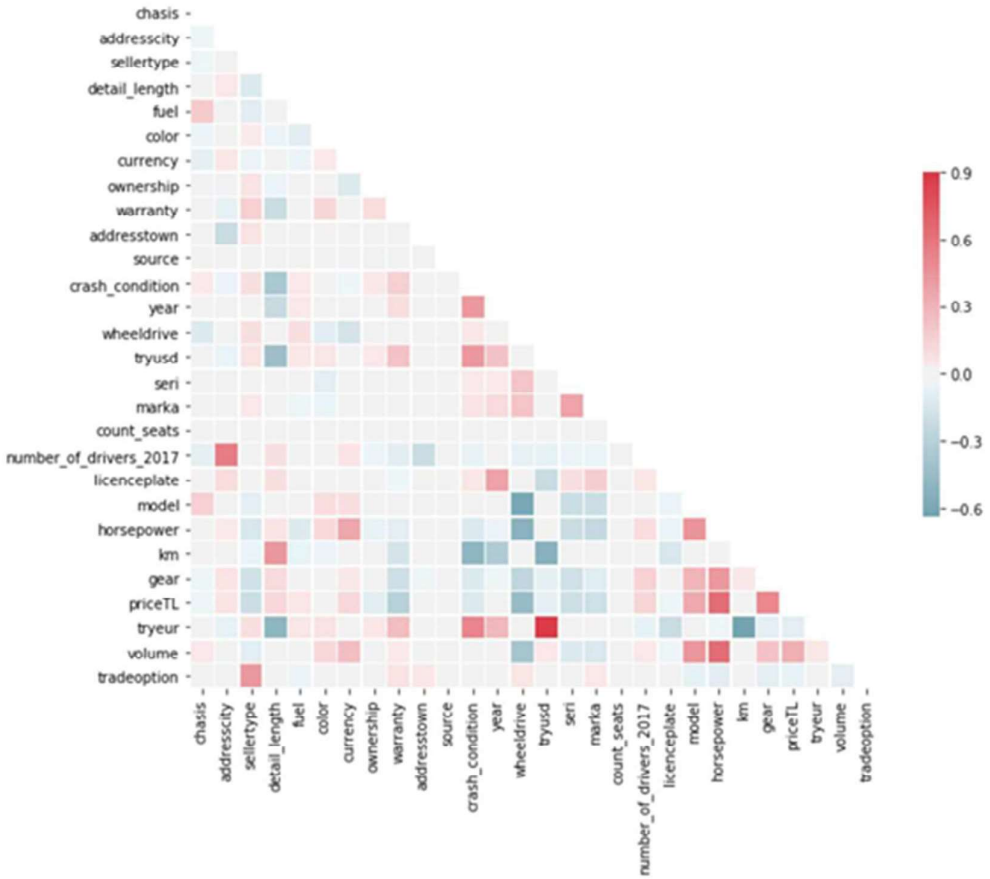

Figure 25. Correlation matrix of features in continuous with continuous transformation.

We check the model's p-value to check whether our sample provides enough evidence that we can reject null and we can observe that our p-value that can be seen in Table 3 is close to 0 at 2.2e −16 and we can reject the null hypothesis and conclude that our sample and features provide enough evidence to describe the used car prices with the influencing factors.

Table 3. Ordinary Least Squares model results.

OLS Regression Results			
Dep. Variable:	priceTL	R-squared:	0.58
Model:	OLS	Adj. R-squared:	0.58
Method:	Least Squares	F-statistic:	18160.00
Date:	Sun, 31 Mar	Prob (F-statistic):	0.00
Time:	14:25:26	Log-Likelihood:	-3758900.00
No. Observations:	313570	AIC:	7518000.00
Df Residuals:	313545	BIC:	7518000.00
Df Model:	24		
Covariance Type:	nonrobust		

The model's F-statistics is below 0.05 which means that the model is significant enough. In Table 4 it can be observed that the R2 is 0.58 which means that the model's variation in the sum of squared errors has been captured at 58% with the predictive model.

Table 4. OLS model performance on training and testing datasets.

Dataset	Metric	Performance
Train	R2	0.58
	Mean Absolute Error	23092.46
Test	R2	0.58
	Mean Absolute Error	23214.44

The dataset performance with train dataset is 0.58 R2 and the test dataset's performance is like train R2. This indicates that the model has not over-fit to the

training dataset as can be seen in Figure 26. However, mean absolute error for the OLS is around 23000 TL which is a high price error considering the mean of the dataset is only 63000 TL.

Figure 26. Ordinary Least Squares model scatter plot of predictions of test dataset to real value

It can be observed in Figure 4.27 that the model fails to predict very low values and very high values but performing good in mid-range values. Because regression is a methodology which the model learns a mean when modeled with a linear function the model is not able to capture the full dynamics of the phenomenon for second-hand automobiles.

A residual plot should have a random distribution of residuals. The model's residuals show a significant pattern in Figure 27 which indicates it has not been able to learn the full pattern of the phenomenon which is due to the linearity of OLS and the additive functional form it takes. The main drawback with this approach is that ordinary least squares have a very low R2 compared to gradient boosting. Thus, the

insights generated from an OLS regression will not be valid for specific and local predictions. The model performs statistically high R2 at 0.58; however, due to nonlinear factors in the phenomenon, a better descriptive model using nonlinear approaches is built with gradient boosting.

Figure 27. Ordinary Least Squares model residual plot

The ordinary least squares regression model performs suboptimal at 0.58 R2 on both train and test dataset. The first reason an ordinary least squares regression model performs low is that the dynamics of the phenomenon described cannot be captured using a linear form equation derived from the features of the dataset. This means that the interactions of the features in describing the second-hand automobile prices are not describing enough when modeled in a linear additive form and without capturing quadratic forms. Third reasons are that in an OLS regression the categorical features must be encoded numerically in two alternative ways; ordinary numeric encoding or one-hot encoding. Both encoding methods create over-fitting or information

distortion in the regression model thus causing a reduced test R2. Ordinary numeric encoding forces the feature to be numerically ordered, and when there is no order in the cate-gory, this distorts information thereby forcing the model to learn a non-representative functional linear form. One-hot encoding is the best method to convert categories to a numeric form; however, this data transformation increases the dimensionality creating the model to over-fit. The over-fitting due to high dimensionality is the curse of dimensionality (Friedman, 1997). The third reason is that linear models cannot capture enough variance of the second-hand automobile market. The dataset has a high variance with information from low segment vehicles to luxury segment vehicles. The high variance cannot be encapsulated with one model that has a singular coefficient per feature, where the model lacks the flexibility to capture a variety of behaviors.

4.1.2 Gradient boosting regression

Using gradient boosting, a predictive model that describes how features of second-hand automobile listings describe prices is trained. A gradient boosting model is a non-linear and nondifferentiable model structure. This model structure makes it impossible to generalize coefficients of the features in the dataset which explain the pricing dynamics, whereas using an ordinary least squares method, conclusions are difficult to make due to low model performance. The developed gradient boosting nonlinear model has a very high performance; therefore, conclusions can be made however there is still no single formula that can be derived from the nonlinear structure. To overcome the challenge of understanding which features affect the pricing of second-hand automobiles, in the research we used a local linearization

approach for random predictions to understand how feature's effects play a role in the model structure.

The model building consists of three datasets. The first dataset has 80% of the data, and it is called train dataset. The second dataset consists of 10% of the data, and it is called development dataset. The last dataset consists of 10% of the data, and it is called test dataset. The purpose of splitting the dataset is two folds; first reasons is to build an unbiased model on a training dataset and optimize the model parameters on a development dataset. The second reasons are to keep an unbiased dataset where the model performance can be observed and validated. This is because the model performance can only be trusted if the performance measured is derived from an unbiased dataset.

Training the model shows R2 of 98% in train dataset, development performance shows an R2 of 95%, and lastly, test performance shows a significant performance of 94% as can be seen in Table 5 The three-performance metrics show that the model is very powerful in explaining the structure of the car price predictions. In the second figure, the prediction plot can be observed.

It can be observed in Figure 28 that the model can predict prices at all ranges with similar performance. Although the performance of high-value predictions slightly askew, the overall performance is very high because of the small number of high-priced vehicles in the listings.

Table 5. Gradient boosting decision tree model performance on training, development and testing dataset

Dataset	Metric	Performance
Train	R2	0.98
	Mean Absolute Error	5438.2
Development	R2	0.95
	Mean Absolute Error	6556.06
Test	R2	0.94
	Mean Absolute Error	6579.22

To evaluate model performance, the residuals are also analyzed in Figure 30 to search for any skewness and non-normal distribution plots. In the event of a pattern in the residual plots, the model could have been concluded that it still needs more features to learn from. However, in the residual plots, it can be observed that the distribution is normal with a slight skewness in loss of prediction performance in high values. The skewness could derive from two main reasons; observations and features. Due to the low number of high prices vehicles, there are not enough observations that the model can train on.

Figure 28. Gradient boosting decision tree model plot of real value and prediction

From a business perspective, in second-hand automobiles, high prices vehicles are not listed as much as average prices vehicles due to customer segment. Secondly, the skewness is due to not enough features that explain high prices vehicles. If the information is not embedded such as luxury factors, the predictive model cannot learn characteristics enough that it can describe the phenomenon.

Figure 29. Gradient boosting decision tree model residuals plot

There are three main reasons for the gradient boosting to have high predictive performance. First, the gradient boosting model structure has a nonlinear functional form. The model learns a nonlinear representation of the features by interacting features using decision trees. The decision tree form enables to capture of a sub-optimal feature space split. The second reason is that the gradient boosting model combines hundreds of decision trees thereby enabling to learn a high variance representation without falling to over-fitting. Combination of many decision trees

enables to capture variance in the phenomenon being described. Third, the gradient boosting models and decision models that are under the hood can handle categorical feature set without the need for a feature transformation. A decision tree can split the space using a categorical feature without a need for one-hot encoding thereby the feature space is not increased by one-hot encoded features, and secondly, without a need for ordinality for numeric values, the model can learn the best representation.

As a result, having seen high R2 metrics and mean absolute error and Gaussian distribution of the residuals, this nonlinear gradient boosting model can be used to understand how the predictors influence the pricing of second-hand automobiles.

4.2 Feature dynamics analysis

4.2.1 Entropy analysis

One of the key steps in understanding how listing features exhibit an effect on second-hand price automobile have used a nonlinear predictive model is to compare how the decision tree splits using the features. Gradient boosting splits each node with respect to that feature's entropy minimization effect. If a feature has high minimization for entropy, it means that at a global level that features is highly important to explain the target which in this research is the price of a second-hand automobile.

The results in Table 5 show that most entropy is gained with using continuous feature horse-power, followed by gear type as a categorical variable and third with the model as another categorical variable and the 4th important variable is the km in the odometer. Secondly, it can be observed that the description length of the listing

has a very low effect on the model understanding of how prices are determined for second-hand vehicles. The reasons that brand and series is very low prioritized is that the nonlinear model has already learned the vehicle's brand attributes from a more specific feature which is model of the car and it does not require any more information to generalize from. This shows that in explaining how the prices are modeled concerning features the vehicles feature such as technical is prioritized over more perception features such as the brand of the vehicle.

Table 6. Gradient boosting decision tree model feature importance chart model output based on entropy gain.

Feature	Feature Importance
horsepower	0.23
gear	0.19
model	0.11
km	0.11
fuel	0.07
warranty	0.06
wheeldrive	0.03
volume	0.03
currency	0.02
seri	0.02
brand	0.01
detail_length	0.01

4.2.2 Local linearization

In this section of the results, the local level predictions are explored using a local linearization methodology. At each level of prediction, random samples are taken around the instance, and the features are weighted using a linearization method. This enables to create an additive model structure specific to the instance. At each instance, the weights change therefore the exact effect of features can be approximated.

In the research out of 30,000 test instances that the model has not observed before 5 random instances are selected at random to which local linearization is applied. For each of the 5 instances, the prediction to actual values is compared, how the features are interacting to output the predicted price is analyzed, and lastly insights in both modeling the second-hand automobile and business impact is concluded.

The gradient boosting model predicts a Volvo car with a value of 56000 TL with only 3400 TL of error which is 7% of error in a prediction. It can be observed in Figure 31 that for this prediction four features have been playing an important role in predicting the price; fuel type, brand, gear type, and chassis type. The fact that the car is a Volvo with a Diesel engine has increased the prices whereas the fact that the car is a three-door hatchback with a manual gear type has lowered the price. Secondly, although the global feature importance for the Gradient Boosting model showed horsepower as the most important predictor, it can be observed that it plays an insignificant role when linearized to a local level prediction.

Figure 30. Random listing sample which is a Volvo brand in Bursa city with an actual price of 56000 TL. Predicted price is 52652 TL with an error of 3400 TL.

The predictive model predicted the price of a Seat in Ankara at 48000 TL whereas the actual price is 46500 TL with an error of 1500 TL. It can be observed in Figure

32 that for a Seat car, three main factors played a significant in predicting the price; mileage, fuel type and gear of the vehicle. It shows that low odometer value has increased the price however fuel type as benzene and LPG and manual gear type lowered the price of the vehicle in the dataset.

Figure 31. Random listing sample which is a Seat brand in Ankara city with an actual price of 46500 TL. Predicted price is 48024 TL with an error of 1500 TL.

The nonlinear gradient boosting model shows that four main features are the main drivers of price in the second-hand automobile listing in Figure 33. The price increasing variables are the fact that the fuel type is Diesel and the car is a brand-new car with only 400 km in the odometer. The price lowering variables are the fact that the brand is a Renault and the car is equipped with a manual gear type. The other features are very insignificant in pushing the features to higher and lower directions in this prediction at a local level.

Figure 32. Random listing sample which is a Renault brand in Izmir city with an actual price of 69750 TL. Predicted price is 76706 TL with an error of 6956 TL.

The nonlinear model predicts an Audi car in Istanbul at 113000 TL where the actual price is 121000 TL with an error of 8000 TL as can be seen in Figure 34. Four main features are the main influences in the price of the second-hand automobile listing which are gear type, fuel type, brand and horsepower of the vehicle. The fuel type diesel and brand Audi increase the price where the horsepower 110 lowers the price in the prediction at a local level as can be observed in Figure 34. Although the decrease in price due to horsepower look counter-intuitive at first because of the general perception that a higher horsepower will increase the price, however, in a local space where the brand is a premium brand such as Audi, a lower horsepower range would indicate a lower segment car on average and the prices would be lowered.

Figure 33. Random listing sample which is an Audi brand in Istanbul city with an actual price of 121900 TL. Predicted price is 113521 TL with an error of 8379 TL.

In the random example in Figure 35, it can be observed that horsepower increases the price in the local domain. This is because that higher range horsepower with a premium brand such as BMW would mean a higher priced vehicle. Secondly, it can be observed that configurations that are preferred such as an automatic car, and a diesel fuel type increases the prices in general.

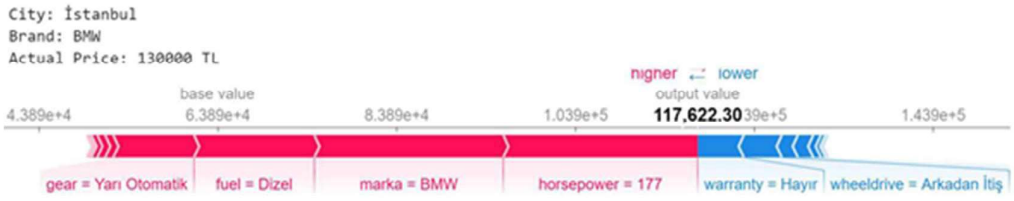

Figure 34. Random listing sample which is a BMW brand in Istanbul city with an actual price of 130000 TL. Predicted price is 117622 TL with an error of 12378 TL.

In the analysis, five samples were taken at random to build local linearization around the prediction to transform the nonlinear model to a linearized form of an equation. Firstly, it can be observed that premium brands with a high price perception exhibit a price increasing behavior whereas brands with a low-price perception exhibit a price lowering behavior. There is no one significant coefficient in a linear regression model that can capture the brand effect. Features should not always have a negative or positive definite coefficient like in linear regression; they can differ at local level predictions. Secondly, feature importance can change at local predictions. Although the general model exhibit that horsepower is the most important factor with the highest entropy, at local predictions horsepower fall into less important positions. From a business point, this means that for a specific brand or model, horsepower loses its importance and other factors emerge as the main driver of prices.

4.2.3 Partial dependency plots

Partial Dependence Plots (PDP) were introduced by Friedman with the intention to interpret complex modeling, specifically in nonlinear domain (Friedman, 2001). Interpreting a linear regression model is not complicated because it provides an additive formula that can be understood; however, interpreting nonlinear methods

such as Gradient Boosting Machine models are very complex and almost impossible. In the partial dependency plots if there is a significant variation for any given feature that is used in training that means the value of that feature affects the nonlinear model significantly quite a lot but if the plot is constantly near zero it means that the feature has almost no effect on the target feature which is the price of second-hand automobiles. Partial dependence plots are a powerful method to extract insights from complex nonlinear models. They can generate powerful insights into how to improve describing the phenomenon using the nonlinear model as well as into the business aspect of the problem.

The SHAP partial dependence does not show the predicted value of the nonlinear model; instead, it exhibits how the value is changing with the change in the given feature that explains the phenomenon. Secondly, it shows each instance that is in the test dataset as node thereby allowing to visualize how the effect is in the full second-hand vehicle domain. SHAP dependence chart plots the value of the independent feature on the x-axis and the SHAP value of the same feature on the y-axis. This shows how the model depends on the given feature and is like a richer extension of the classical partial dependence plots.

The Figure 36 exhibits three main insights in trend with increasing values of km, at low values of km and when the value of km is zero. Firstly, observing the trend in general as the km in odometer increases, the feature has an effect of lowering the target value price. Secondly, when the value of odometer is zero, it exhibits high increase impacts on prices. Thirdly when the km is at low boundaries below 80000 odometers, it continues to have an increasing impact on the price of vehicles. From a statistical modeling point of view, a phenomenon such complex

with quantitative and qualitative understanding should be modeled using nonlinear approaches that can capture the nonlinearity in a feature.

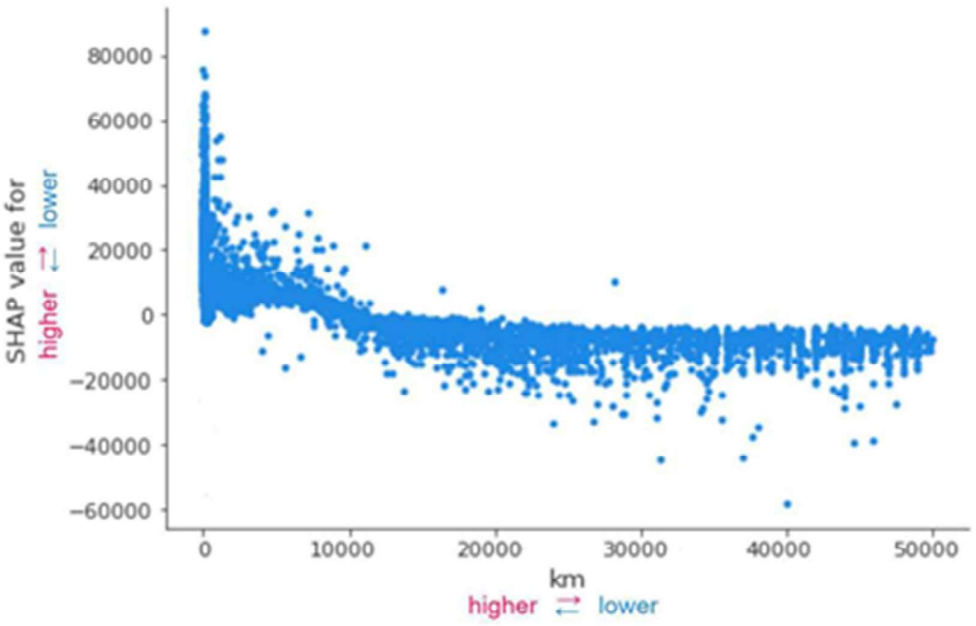

Figure 35. Dependency plot for kilometers in odometer to price of listings, exhibiting nonlinear behavior.

Modeling the second-hand automobile prices with linear regression methods falls behind in capturing the behavior. From a business point of view, the market tends to price low odometer vehicles at relatively higher values until a threshold and passing the threshold starts to have an increased effect of decreasing the value of the vehicle.

Horsepower is one of the best explanatory features that has the most entropy gain in the gradient boosting model. The Figure 37 exhibits two main insights where horsepower is below and above 100 HP. First is that when the horsepower is below 100, the horsepower is a feature that has a negative direction impact on the price. Secondly, when the price is above 100, the horsepower features starts to have a

positive direction impact on the second-hand automobile prices. As a result, at the local level, the second-hand automobile market shows separate behaviors with horsepower. Compared to a linear regression model where the horsepower has only a positive coefficient cannot be applied at local level feature understanding and will cause low-performance explanatory power. From a business point, the horsepower can be clustered to two segments; automobiles with below 100 horsepower that are economic cars and seek low prices and automobiles with powerful engines that are in an upper segment that are more expensive than others.

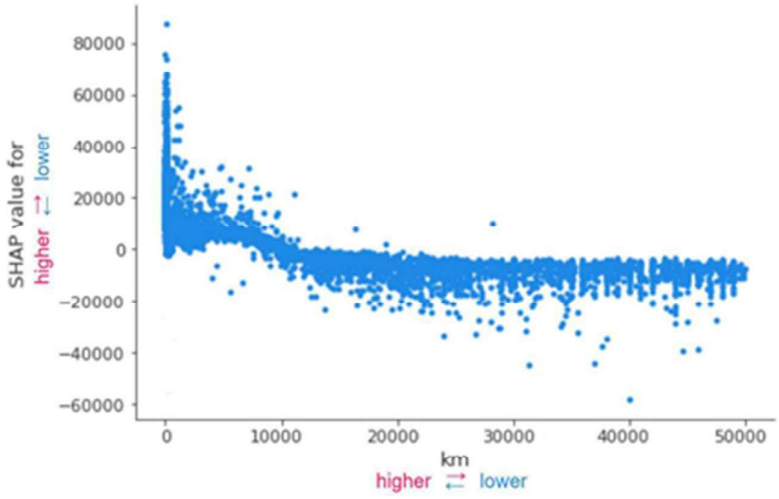

Figure 36. Dependency plot for horsepower to price of listings, exhibiting nonlinear behavior.

Capturing nonlinear relationships with linear models is impossible. Linear models fall behind in explaining local level phenomenon due to lack of capacity in capturing high variance datasets. The generalization over the sampled population is not enough to make conclusions in a linear setting. From a business point of view, the most important feature which is horsepower has an increasing and a decreasing effect of

the price at a variety of locals in data. The oscillation horsepower effect can only be observed with linear localization methods. Lastly, like horsepower, odometer (km) cannot have a single coefficient, it exhibits nonlinear behavior with respect to price. Low values of mileage have a significant increase in effect whereas the effect loses its power as the km increases, the effect is not stationary. General linear models and correlation matrix exhibit a negative coefficient with odometer km; however, when the km is low it has a positive effect on the price, and this behavior can only be captured with nonlinear methods.

CHAPTER 5

CONCLUSION

In this research, we propose three new improvements over existing research in automobile pricing; modeling, opening-up nonlinear models for explanation and feature dynamics on prices. Firstly, the research shows how performance over linear regression models can be improved using nonlinear regression methods to develop a model that explains second-hand automobile pricing. Secondly, exhibit how to keep interpretability of models using local linearization method without a trade-off in losing performance by confronting to linear regressions. Lastly, how the second-hand automobile features have a non-stationary effect on prices with proofs and experiments showing the changing dynamics affect predictions.

Having a high-performance model is a must first step to make significant conclusions on feature dynamics to automobile prices. In the research, we proved that linear regression performs lower in R2 and mean absolute error than a nonlinear regression with gradient boosting decision trees. The performance is attributed to nonlinear modeling structure, performance in variance to bias trade-off and handling of features without a need for transformation. Secondly the phenomenon described has very high nonlinear attributes that can only be captures with nonlinear advanced modelling methods such as gradient boosting decision trees.

Local linearization for interpretation is a very recent research area that has never been utilized in automobile pricing literature. Highest model accuracy is achieved by using complex models that even domain experts struggle to make meaningful insights from. Employing local linearization on the second-hand automobile dataset, we exhibited insights into what are the drivers and dynamics of

automobile pricing models. Thirdly, the dataset used in price modeling is collected over two years from two different online auction sources. The dataset has a high variance with a diversity of information on a listed automobile. Furthermore, in the dataset, we incorporated external information such as daily exchange rates and the number of drivers per city.

The analysis is conducted in three phases; first is analyzing the two most important continuous feature in the modeling. This analysis resulted that horsepower and km are the two most continuous important features followed by car model and gear type as the two most categorical important features of cars in terms of entropy gain. In the second analysis, randomly selected 5 listings were analyzed to compare how the feature's importance changes when predicting a price using local linearization. This showed that feature rankings significantly change in different predictions, exhibiting clusters of similar feature attributes to price. In the third analysis, using dependence plots the two most continuous important feature is analyzed to check for dynamics with respect to change in the independent features. This showed a significant nonlinear relationship with the price and that there is no single coefficient that can be attributed to continuous independent features of second-hand automobiles.

This research highlights several crucial issues in a managerial and academic context. The predictive models that can be developed and explained using interaction plots have implications for consumers, sellers and the marketplace. Firstly, the model and data can help the public in verifying and comparing the price of a used car whether it is overpriced or underpriced. Secondly, by utilizing a nonlinear approach, purchasers can understand what the main drivers of the price for the online listing are and utilize this information in their purchasing decisions. The price analysis can be

beneficial for consumers to easily spot high priced vehicles to increase bargain or seek alternative solutions.

The car sellers and dealers can set the pricing of the vehicle by utilizing advanced analytic methods. It has managerial benefits such that the results can be used by institutions or car dealers when setting the price for the used cars and furthermore, determine the most critical factors and focus the listings on the significant characteristic of the used cars listed. Expertise business can utilize advanced analytical tools to suggest optimal pricing of a vehicle by collecting sensory data of the vehicle. Models build using a detailed data scraping combined with sensory data can result in perfect descriptive pricing engines.

Application of explanation tools for nonlinear modeling in second-hand automobile marketplace open opportunities and promote to apply a variety of analytical solutions. As for the usage of such tools increase, information discrepancy, the advantage of information power will decrease, and the market can reach a price equilibrium in the second-hand automobile market. Secondly, as the demand for nonlinear modeling with big data increase, the data requirement from sellers of vehicles will increase. Increase in information on automobiles will require for a better priced second-hand automobiles market. As data on second-hand automobiles increase, nonlinear models built will increase in prediction power, decrease in error. This will result in better explanations on feature dynamics of second-hand automobiles. Nonlinear and complex models can be explained using local linearization. As a result, big data and advanced analytic methods will improve decision making in business, remove ambiguity in un-certainty and information asymmetry.

CHAPTER 6

LIMITATIONS AND FUTURE WORK

6.1 Limitations

The developed model's R2 is 93%; the performance can be improved by incorporating additional features that are not provided in an online listing. For example, some of the features that can be additionally incorporated can be the previous owner's occupation, the brand of an internal sound system and information from other websites. Although the model's performance is very significant compared to a linear model, these types of additional information account for limitation in performance.

In an online auction setting, the pricing of a good is a result of the "value construction" mechanism where the starting price of the good is an informative indicator of the good's quality and thus increases the quality perception of the item (Haubl and Leszczyc, 2003). Higher prices have a positive effect on valuations. Our research is only limited to determining price characteristics; price-quality perception is not included.

In an online auction, markets state that there is a positive relationship between the starting bid price of a good and the final auction price is valid for items where the market price of the good is difficult to determine (Brint, 2003). In the online used car market, there is no bidding system but a direct purchase, therefore in our research our assumption does not include the high pricing of goods and its effects on the purchase price.

A primary factor affecting the price is the demand for that good or service. Measuring the effect of demand on price was out of this research's scope. There is a potential for further research to include additional environmental and external features to describe the price of the used cars better.

6.2 Future research

In this research, the brand value or brand perception of the used automobiles is not included, further research which can improve the R2 of the model meanwhile accounting for the models and their perceptions and values will improve insights. More-over, including environmental and macroeconomic features into the model such as demand for the brands or models and car usage by region can be further researched to improve on the present literature and findings.

In auctions, picture posting provides an effective way to inform buyers (Spence, 1974). Picture posting is an effective way in today's electronic commerce. The more the pictures are, the more informed the buyers. In recent researches for eBay auctions, it is stated that increased and specific product information reduces quality uncertainty and thereby gives flexibility to increase buy-it-now prices (Yin, 2006). In this light of information, our research's data had initially had collected such information, but due to technical complications, we will run the model in our next project with updated information and analyze the effect of detailed information such as text detail given, and picture detail given in further research.

The ability to interpret nonlinear models to open a mammoth of opportunities for business modeling. Ability to conduct research nonlinear risk models for credit

scoring, loan, credit default models to understand how the risk changes using local linearization opens new potential areas for research and business applications.

References

Abounoori, E., & Rezvani, A. (2012). Using Hedonic Prices to Estimate Quality Changes concerning Iranian Automobile Market. *Iranian Journal of Economic Studies, 1*, 1-12.

Alpaydın, E. (2010). *Introduction to Machine Learning* (Vol. 3). Massachusetts: The MIT Press.

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *ArXiv, 40*, 85-104.

Asilkan, Ö. (2011). İkinci El Otomobillerin Güncel Pazar Fiyatlarının Veri Madenciliği Yöntemleriyle Modellenmesi. *Akademik Bakış Dergisi, 24*, 1-19.

Ayan, E., & Erkin, H. C. (2014). Hedonic modeling for a growing housing market: valuation of apartments in complexes. *International Journal of Economics and Finance, 6*, 188-199.

Balce, A. O. (2016). Factors affecting prices in an used car e-market. *Journal of Internet Applications & Management, 7*, 5-20.

Baldemir, E., Kesbiç, C. Y., & İnci, M. (2007, 9). Emlak piyasasında hedonik talep parametrelerinin tahminlenmesi. *Türkiye Ekonometri ve İstatistik Kongresi*, 24-25.

Bapna, R., Jank, W., & Shmueli, G. (2008). Price formation and its dynamics in online auctions. *Decision Support Systems, 44*, 641-656.

Baur, N., & Lamnek, S. (2007). Multivariate analysis. *The Blackwell Encyclopedia of Sociology*, 1-3.

Bell, R. M., & Koren, Y. (2007). Lessons from the Netflix prize challenge. *SiGKDD Explorations, 9*, 75-79.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123-140.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5-32.

Brint, A. T. (2003). Investigating buyer and seller strategies in online auctions. *Journal of the Operational Research Society, 54*, 1177-1188.

Case, B., & Quigley, J. M. (1991). The dynamics of real estate prices. *The Review of Economics and Statistics, 73*, 50-58.

Cason, T. N., & Friedman, D. (1996). Price formation in double auction markets. *Journal of Economic Dynamics and Control, 20*, 1307-1337.

Chang, T.-Z., & Wildt, A. R. (1994). Price, product information, and purchase intention: An empirical study. *Journal of the Academy of Marketing Science, 22*, 16-27.

Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. *Communications of the ACM, 54*, 88-98.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*, 1165.

Colwell, P. F., & Dilmore, G. (1999). Who was first? An examination of an early hedonic study. *Land Economics, 75*, 620-626.

Coulson, N. E., & McMillen, D. P. (2008). Estimating time, age and vintage effects in housing prices. *Journal of Housing Economics, 17*, 138-151.

Cover, T. M., & Thomas, J. A. (1991). Entropy, relative entropy and mutual information. *Elements of Information Theory, 2*, 1-55.

Dexheimer, V. (2003). Hedonic methods of price measurement for used cars. *Statistisches Bundesamt (Destatis), 18*, 1011-1029.

Ecer, F. (2013, 4 07). Türkiyede 2. El otomobil fiyatlarının tahmini ve fiyat belirleyicilerinin tespiti. *Anadolu Üniversitesi Sosyal Bilimler Dergisi, 3*, 219-229. Retrieved from https://hdl.handle.net/11421/43

Eken, M. H., & Çiçek, M. (2009). Türkiye'de Otomotiv Sektöründeki Ürünlerin Kredilerle Finansmaninin Satişlara Etkisi. *Maliye ve Finans Yazıları, 1*(84), 61-77.

Erdem, C., & Sentürk, I. (2009). A hedonic analysis of used car prices in Turkey. *International Journal of Economic Perspectives, 3*, 141.

Erdogdu, E. (2014). Motor fuel prices in Turkey. *Energy Policy, 69*, 143-153.

Fletcher, M., Mangan, J., & Raeburn, E. (2004). Comparing hedonic models for estimating and forecasting house prices. *Property Management, 22*, 189-200.

Fong, R. C., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. *Proceedings of the IEEE International Conference on Computer Vision*, (pp. 3429-3437).

Friedman, J. H. (1997). On bias, variance, 0/1---loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery, 1*, 55-77.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics, 38*, 1189-1232.

Gardner, D. M. (1971). Is there a generalized price-quality relationship? *Journal of Marketing Research, 8*, 241-243.

Giudici, P., & Castelo, R. (2003). Improving Markov chain Monte Carlo model search for data mining. *Machine Learning, 50*, 127-158.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*, 93.

Hadinejad, M., & Shabgard, B. (2011). Hedonic Price For Car in Iran. *Sosyal Bilimler Dergisi*(2), 118-127.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining Trends and Research Frontiers.* Elsevier.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering, 9*, 90.

Janizek, J. D., Celik, S., & Lee, S.-I. (2018). Explainable machine learning prediction of synergistic drug combinations for precision cancer medicine. *BioRxiv*, 331769.

Jank, W., & Shmueli, G. (2006). *Studying Heterogeneity of Price Evolution in Ebay Auctions via Functional Clustering.* Elsevier.

Jim, C. Y., & Chen, W. Y. (2009). Value of scenic views: Hedonic assessment of private housing in Hong Kong. *Landscape and Urban Planning, 91*, 226-234.

Kaun, D. E., & Spence, A. M. (1975). *Marketing Signaling: Informational Transfer in Hiring and Related Screening Processes.* (Vol. 29). Harvard University Press.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, (pp. 3146-3154).

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., . . . others. (2016). *Jupyter Notebooks - a publishing format for reproducible computational workflows.* Elpub.

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review, 69*, 066138.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy, 74*, 132-157.

Lewis, G. (2011). Asymmetric information, adverse selection and online disclosure: The case of eBay motors. *American Economic Review, 101*, 1535-46.

Louviere, J. J., Hensher, D. A., Swait, J. D., & Adamowicz, W. (2000). *Stated Choice Methods.* Cambridge University Press.

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *ArXiv, 1802*(03888).

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia Medica, 23*, 143-149.

McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing, 14*.

Miller, J., & Haden, P. (2006). Statistical analysis with the general linear model. *Creative Commons Attribution*.

Monson, M. (2009). Valuation using hedonic pricing models. *Cornell Real Estate Review, 7*, 10.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition, 65*, 211-222.

Murray, J., & Sarantis, N. (1999). Price-quality relations and hedonic price indexes for cars in the United Kingdom. *International Journal of the Economics of Business, 6*, 5-27.

Neter, J., Wasserman, W., & Kutner, M. H. (1990). Regression, analysis of variance, and experimental design. *Applied Statistical Models*, 614-619.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences, 10*, 424-430.

Oliphant, T. E. (2007). Python for scientific computing. *Computing in Science and Engineering, 9*, 10-20.

Olson, J. C. (1978). Inferential belief formation in the cue utilization process. *ACR North American Advances*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . others. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 8*.

Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. *International Journal of Information & Computation Technology, 4*(7), 753-764.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy, 82*, 34-55.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics, 21*, 660-674.

Scott, M., Flaherty, D., & Currall, J. (2013). Statistics: Dealing with categorical data. *Journal of Small Animal Practice, 54*, 3-8.

Selim, S. (2011). Determinants of house prices in Turkey: A hedonic regression model. *Doğuş Üniversitesi Dergisi, 9*, 65-76.

Shapiro, B. P. (1968). The psychology of pricing. *Harvard Business Review, 46*, 14-25.

Sheppard, S. (1999). Hedonic analysis of housing markets. *Handbook of Regional and Urban Economics, 3*, 1595-1635.

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., & Despouy, P. (2017). plotly: Create Interactive Web Graphics viaplotly. js. *R Package Version, 4*(1).

Sirisuriya, D. S., & others. (2015). A comparative study on web scraping. *Kotelewa Defence University International Conference*.

Stafford, J. E., & Enis, B. M. (1969). The price-quality relationship: An extension. *Journal of Marketing Research*, 456-458.

Tull, D. S., Boring, R. A., & Gonsior, M. H. (1964). A note on the relationship of price and imputed quality. *The Journal of Business, 37*, 186-191.

Walt, S. V., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering, 13*, 22.

Wang, L., Lei, Y., Zeng, Y., Tong, L., & Yan, B. (2013). Principal feature analysis: A multivariate feature selection method for fMRI data. *Computational and Mathematical Methods in Medicine, 2013*.

Wilson, C. (1980). The nature of equilibrium in markets with adverse selection. *The Bell Journal of Economics*, 108-130.

Wu, J.-D., Hsu, C.-C., & Chen, H.-C. (2009). An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. *Expert Systems with Applications, 36*, 7809-7817.

Yin, P.-L. (2006). Information Dispersion and Auction Prices. *Stanford Institute for Economic Policy Research Working Paper, 2*(24).

Zeithaml, V. A. (1988). Consumer perceptions of price, quality, and value: a means-end model and synthesis of evidence. *The Journal of Marketing*, 2-22.