National College of Ireland

Regression Analysis for Predicting Prices of Used Cars: A Study Utilizing Data from Car Trading Website

MSc Research Project
MSc Data Analytics

Selin ULUTURK
Student ID: x23160373

School of Computing
National College of Ireland

Supervisor: Dr. Christian Horn

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Selin Uluturk ……
Student ID:	x23160373 ……
Programme:	MSc Data Analytics …… **Year:** 2024……
Module:	Research Project ……
Supervisor:	Dr. Christian Horn … ……
Submission Due Date:	14.09.2024……
Project Title:	Regression Analysis for Predicting Prices of Used Cars: A Study Utilizing Data from Car Trading Website
Word Count:	7141…… **Page Count** 21 ……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Selin Uluturk……

Date: 11.08.2024……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Regression Analysis for Predicting Prices of Used Cars: A Study Utilizing Data from Car Trading Website

Selin Uluturk
x23160373

Abstract

The aim of this study is to predict used car prices using various regression models. Car data collected from a used car trading website in Türkiye in June 2024 was used. The study focused on the fifteen best-selling car brands in Türkiye in 2023. The dataset was divided into four groups based on price, and each group was analyzed separately. The regression models used included Multiple Linear Regression (MLR), Support Vector Regression (SVR), Random Forest, and XGBoost. The analysis results showed that the XGBoost model had the highest R-squared values, while the Random Forest model had the lowest Mean Absolute Percentage Error (MAPE) values. Additionally, it was determined that gear type, specific car brands, and models played an important role in price prediction. This study aims to provide more accurate price predictions in the used car market, enabling both individual consumers and businesses to make more informed decisions.

Key Words__ Regression, Used Car, Price Prediction, Random Forest, XGBoost

1 Introduction

The used car market holds significant importance for both buyers and sellers due to economic reasons, personal preferences, and investment purposes. Many individuals turn to the used car market to find cars that fit their budget and needs at more economical prices. Some engage in buying and selling used cars solely for trading purposes. Additionally, some sellers aim to sell their old cars in the used car market even when they are planning to purchase a new one. Thus, whether people are car owners or not, prefer used cars or new ones, they might find themselves dealing in the used car market at some point. Therefore, accurately determining used car prices impacts many individuals and necessitates continuous market monitoring. For these reasons, our study aims to provide accurate price predictions for parties involved or planning to enter the used car market, regardless of their role. Additionally, guidance is aimed to be provided to individuals or businesses seeking to streamline the sale of their used cars by identifying which features most significantly impact the price.

In our research, car prices in Turkiye were aimed to be predicted. The data of cars for June 2024 from a website that facilitates the buying and selling of used cars in Turkiye was used. Fifteen different car brands, chosen based on the best-selling and most preferred brands in Turkiye in 2023, were focused on. These car brands are illustrated in Figure 1. During the research, these car brands were split into four distinct groups and the analysis was continued accordingly. This grouping was necessary due to significant price disparities among the cars, which adversely affected the prediction accuracy. The details of this grouping, based on the average of all prices, and the brands included in each group are outlined under the data pre-

processing section. It is important to note that these car prices and groupings pertain to used cars, which may differ from new car prices, and that prices in the Turkish market may also vary compared to Europe.

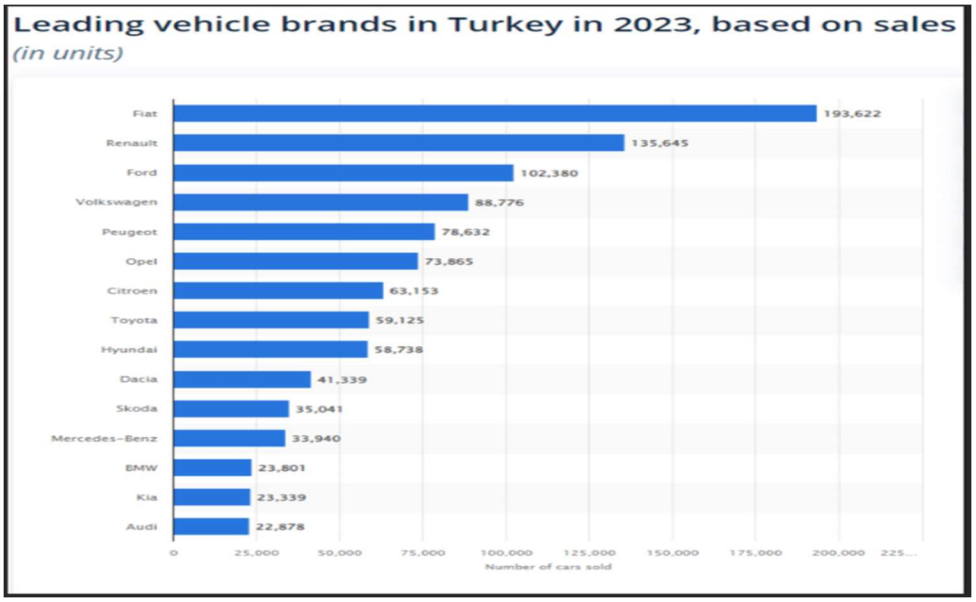


Figure 1: Leading car brands based on sales in Turkiye (Statista, 2023).

In addition to the contributions aimed to be made to the market by our study, the main purpose and the technical path it follows can be explained by stating the research question. “How to measure the estimated price and the variables affecting the price with regression analysis in the used car market?” From the perspective of the research question, the goal is not only to achieve the most accurate used car price prediction but also to compare different regression models to determine the best predictor. By identifying the most significant features in price prediction according to the final model, the curiosity of market participants regarding which attributes influence the price is aimed to be addressed, as well as providing accurate price predictions.

Prepared in line with these objectives, this report will follow a structured format including abstract, introduction, related works, methodology, exploratory data analysis, data pre-processing, design specification, implementation, evaluation steps and will end with the conclusion and future work section.

2 Related Work

In order to better understand the purpose of our study, it is important to examine related works. In this section, the results of studies conducted on used car price prediction and the methods used will be discussed, and the gaps in the existing literature and our innovative contributions will be emphasized.

To determine the most effective linear regression model for predicting used car prices, a literature review was initiated, beginning with a study that focuses on Multiple Linear Regression, Lasso Regression, and Ridge Regression models (Khan, 2022). The final selection was Lasso Regression, which achieved an R-squared value of 0.79. The study has an organized

flow and clear visualizations that aid in understanding the dataset. However, it only uses three independent variables for the prediction models. Including more variables would likely yield more credible results and provide greater insight into the features affecting the price. Additionally, the study did not check for outliers and only compared linear approaches. Including at least one non-linear model would have been beneficial to determine which approach better predicts the price. This is an approach taken in our study.

The aim of the study by Muti and Yıldız (2023) is to evaluate the effectiveness of the linear regression model in predicting used car prices. The study applied a linear regression model to a dataset containing the features and price information of cars in Turkiye for 2020, resulting in an R-squared value of 0.73. While the study is easy to understand due to its structured format and sufficient visuals, the lack of a literature review in the introduction and its absence under a separate heading can be confusing for readers. By applying only, the Multiple Linear Regression (MLR) model, the study missed the opportunity to compare with other linear and non-linear models that might offer higher accuracy.

The aim of Sun et al.'s study (2017) is to develop a model that accurately evaluates used car prices using the optimized Backpropagation (BP) Neural Network Algorithm. The goal is to analyse price data for various cars and create an appropriate price evaluation model, thereby contributing to the industry by enabling more efficient trade for buyers and sellers through better used car price prediction. Although the study is well-structured and includes useful visualizations, it lacks numerical accuracy values despite frequent mentions of accuracy. The study could be improved by applying multiple models to achieve more reliable results.

The aim of another study by Yadav et al. (2021) is to identify the best machine learning model to accurately predict the price of used cars based on their features. By creating an estimation system or application for predicting used car prices from both buyer and seller perspectives, the study aims to facilitate online car purchases. A dataset of various car models' sales prices in different Indian cities, sourced from Kaggle, was analysed. The study found Multiple Linear Regression and Random Forest models yielded the best results. However, the study contains unnecessary Jupyter notebook cells for visualizations, and some visualizations are placed under references, indicating poor structure. Additionally, no R-squared or accuracy values are provided under the results section, and the conclusion mentions "good R2" without specifying numbers. The study requires improvements in both structure and content.

The contribution of Satapathy et al.'s study (2022) to the market is to solve the dilemma of accurately estimating the price of a used car for sellers and to enable users to make more informed purchases. The study also aims to determine the best model for predicting used car prices based on their current features through comparative analysis. For this, Satapathy et al. used Multiple Linear Regression, Lasso Regression, Ridge Regression, XGBoost, and Random Forest machine learning algorithms. XGBoost was determined to be the most optimal model, achieving an R-squared value of 0.92. The study has a clear and understandable structure with effective visualizations to help recognize the dataset. However, when examining the data preprocessing steps, it is noted that there is no mention of handling outliers or applying transformation steps. The authors only applied dummy variables, omitting other crucial data cleaning steps.

The aim of Zhang's study (2022) is to develop a price evaluation model using the LightGBM algorithm to predict used car prices. The model evaluates car price data through big data analysis, aiming to determine the most appropriate price for each car type. The pre-

processing steps are detailed, and the study is well-structured. However, while the title suggests a focus solely on LightGBM, the study also compares different models and reports better results. To align the title with the content, adding a phrase indicating comparative analysis would benefit researchers.

The aim of the study by Satioglu and Tugrul (2021) is to estimate the price of a used car and to thoroughly examine the positive and negative effects, as well as the impact rates, of the car's features on its price. The data was collected from a website in Turkiye using the web scraping method. By providing detailed visual representations of the dataset, the authors made the study easily understandable for a wide range of readers. However, the fact that the study is based solely on multiple linear regression analysis for price prediction does not allow comparison of different prediction methods. Additionally, the absence of a "Related Works" or "Literature Review" section is a notable shortcoming. Incorporating different algorithms into the study can provide a more comprehensive analysis and increase the robustness of price predictions.

A study conducted by Monburinone et al. (2018) tried using car price estimation using data from a German e-commerce site. This study is similar to ours in terms of data collection and price estimation. The primary aim was to predict prices using regression methods and compare these methods. The methods used were Multiple Linear Regression, Random Forest Regression, and Gradient Boosted Regression Trees, with the latter providing the best performance. However, the study's quality is affected by comparing only MSE values and not including R-squared values, which are crucial for explaining the dependent variable's variance. Additionally, while the study presents a car price evaluation model, it primarily focuses on statistical results, lacking adequate discussion on market contributions. Detailing the areas of contribution, explaining the application of methods, and including R-squared values would enhance the study's utility and comprehensiveness for industries and individuals.

Another contribution to the used car trading market is done by Hankar et al. (2022) while applying K-nearest Neighbour Regression (KNN), Random Forest regressor (RFR), Gradient Boosting Regressor (GBR), and Artificial Neural Network (ANN) models to predict used car prices using data obtained from a local e-commerce site in Morocco. The Gradient Boosting Regressor yielded the highest R-squared value (0.80) and was selected as the best predictor. Additionally, the assumptions of "Multicollinearity, Linearity, Normality, and Homoscedasticity" given in the article as assumptions of all regression models belong only to MLR. Although other regression models share some of these assumptions, they often have their own assumptions tailored to their methodology. The study utilized three numerical and three categorical independent variables. Higher prediction rates could be achieved by increasing the number of variables, thereby enhancing the contribution to and expansion of the study.

In a study by Shanti et al. (2021), the aim was to automate the used car buying and selling process by developing a mobile application, employing advanced machine learning techniques. Another objective was to identify the best price prediction model among Random Forest, Gradient Boosting, Neural Network, and Support Vector Regression using data from a commercial website in Palestine. The Random Forest model, with an R-squared value of 0.90, was deemed the most suitable and used to develop the mobile application. The study is noted for its clarity, quality structure, effective pre-processing steps, visualizations, and comparison tables. Future plans include automating data collection, preprocessing, and model training for periodic updates, potentially increasing the application's usability and impact.

The dataset used in Samruddhi and Kumar's study (2020) is provided by Kaggle and pertains to the Indian car market. The authors applied the K-nearest neighbour (KNN) algorithm and experimented with different test-train splits (15%, 20%, and 25%), finding that a 15% test split yielded the best accuracy at 0.85. However, they did not check for outliers during the preprocessing steps. The study lacks visualizations for the dataset in the EDA section, presenting only two bar plots for accuracies at the end. Additionally, the conclusion mentions linear regression and its accuracy, even though MLR was not applied in the study. The inconsistent use of terms like "second-hand" and "used" and the limited application of algorithms are notable weaknesses. Incorporating more than one algorithm and including proper data visualizations could have strengthened the study further.

The study by Wang et al. (2021) compares supervised learning algorithms, including Extra Trees (ET), Random Forest, Ridge Regression, and Decision Tree, to enhance the efficiency and competitiveness of the used car market by directly displaying prediction results from internet-entered data. The Extra Trees model was identified as the best performing model with an R-squared value of 0.98, but this value is too high to be reliable. All necessary pre-processing steps were clearly explained and implemented. However, the literature review includes only two related works, which is insufficient for a study of this nature. Additionally, the lack of visualizations to show correlations and distributions within the dataset is a notable weakness. Incorporating more visualizations would enhance the study's comprehensibility and the dataset's interpretability.

The study by Venkatasubbu and Ganesh (2019) aimed to predict used car prices by comparing Regression Tree, Multiple Linear Regression (MLR), and Lasso Regression algorithms. Although better results are claimed for MLR and Lasso, the study does not present R-squared values and skips pre-processing steps, offering only a table of Mean Squared Error (MSE) values for model evaluation. The structure is overly complex due to the abundance of tables and figures (12 tables and 15 figures in 8 pages). Additionally, the abstract does not provide a comprehensive overview, and using a dataset from 2005 in a 2019 study further reduces paper's clarity and relevance.

The aim of Jin's study (2021) is to create a model that predicts used car prices by considering various car features. Various regression methods, including linear regression, polynomial regression, support vector regression, decision tree regression, and random forest regression, were applied. Random Forest Regression was found to be the most successful model, the highest R-squared value of 0.90. The article is well-structured, easy to understand, and includes images and tables. However, the study only used data from Mercedes brand cars. Including other brands to compare different brand effects on prices could improve the study.

The study by Narayana et al. (2022) aims to create a fairer market environment for buyers and sellers by providing an objective estimation of used car prices, thereby reducing uncertainties in current pricing methods. To achieve this, the researchers applied a Weighted Mixed Regression (WMR) model, using Random Forest and XGBoost regression models to train the WMR datasets. The WMR model, with an R-squared value of 0.79, was selected as the best model compared to other regression models like Random Forest, XGBoost, Multiple Linear Regression, Gradient Boosting Decision Trees, and LightGBM. Including models with R-squared values below 0.10 was unnecessary, and using more acceptable results would increase the study's quality.

The study by Agrahari et al. (2021) aims to assist the used car market, consumers, and dealers in making more accurate price predictions. By doing so, it helps users learn the market value of cars more quickly and reliably without browsing through websites, thus facilitating decision-making. Another objective is to compare the accuracy of linear regression and lasso regression models in predicting prices and determining the best prediction model. The dataset is sourced from Kaggle, but there is no information about which market, country, or years it pertains to, nor any accuracy or error values for each model. The conclusion part does not address the study's findings. There is no preprocessing step mentioned. While the study includes a sufficient literature review, other aspects are inadequate for a comparative study of this nature.

In another study, Varshitha et al. (2022) predicted used car prices using artificial neural networks and machine learning techniques, selecting the Random Forest model as the best estimator. The study could benefit from more visualizations. Pudaruth (2014) and Ponmalar and Christinal (2022) examined price determination using machine learning approaches comprehensively. Since Pudaruth's (2014) study data is less than a thousand, more reliable results could be achieved with larger datasets. Jain and Punia (2024) and Çelik and Osmanoğlu (2019) used linear regression for price prediction. Including some non-linear techniques in these studies could enhance the results. Additionally, Jain and Punia (2024) did not present evaluation criteria such as accuracy and error rates, which could be provided numerically for clarity. Rane et al. (2023) obtained effective results using the Random Forest Regressor approach, but the study could benefit from including different models like Bhatt et al. (2023) provided a comprehensive empirical analysis of different algorithms. In another study, Yılmaz and Selvi (2023) used the combination of web scraping and machine learning for price prediction. Using real-time data increases the reliability of study results. Lastly, Lessmann and Voß (2017) examined the effect of regression methods on car price prediction accuracy, focusing on a single brand and splitting it into six subgroups based on brand models, each modeled as a separate dataset. This approach is similar to our work, but they offer a more specific version by working with a single car brand.

Author(s) and Year	**Description**	**Models**	**R-Squared**
Khan (2022)	Only linear approaches were used	Lasso Regression	0.79
Muti and Yıldız (2023)	Only one model was applied	Multiple Linear Regression	0.73
Jin (2021)	Just one brand was used	Random Forest Regression	0.90
Hankar et al. (2022)	A small number of variables were used	The Gradient Boosting Regression	0.80

Table 1: Comparison Table of Related Works

In Table 1, there are four different studies that evaluate the study with R-squared values, similar to our study. (Jin, 2021) obtained an R-squared value closest to our study but examined a single car brand, while our study offers a broader framework with fifteen different car brands.

(Khan, 2022) tried only a linear approach, Muti and Yıldız (2023) applied a single linear model, Hankar et al. (2022) built a model with a small number of independent variables, and all these studies obtained a lower R-squared value compared to our study. In our study, we created a more comprehensive study by working with four different linear and non-linear models. Additionally, unlike all other papers in which we reviewed the literature, we continued our study by splitting car brands into four different groups according to the average price we calculated, and this reveals another difference of our study.

3 Research Methodology

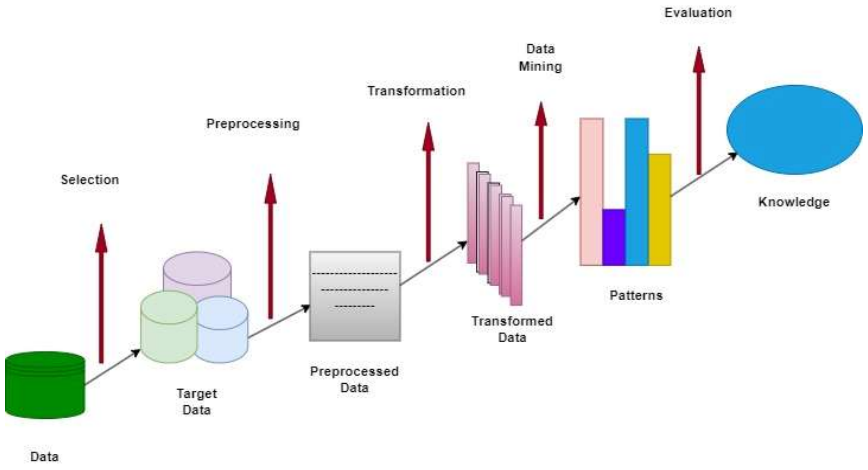


Figure 2: KDD (Knowledge Discovery in Databases)

The KDD (Knowledge Discovery in Databases) methodology was followed throughout the study, as illustrated in Figure 2. This section provides details on how the data collection process was conducted, the steps taken for preprocessing and transformation, and the subsequent procedures that were followed to derive knowledge by evaluating the models created, in accordance with the selected methodology.

3.1 Data Collection

Selenium and BeautifulSoup were used to collect data from an online used car trading platform in Turkiye (arabam.com). The listing page for the cars was scraped in one loop, and the results were saved. Then, the results were enriched by browsing the details page for each listing. A data set of approximately twelve thousand rows and 14 columns was obtained. The contents were collected into a JSON file for further use.

3.2 Data Description

A total of fourteen variables are explained below in the study. “Price” is identified as the dependent variable, while thirteen other independent variables are used to predict the Price using the regression models. The variables and their types, which are considered to affect the price prediction, are listed below for a clearer understanding of the study and dataset.

Variables	Data Type	Description
Price	numeric	Sales price of the used car (Turkish Lira)
Product_year	numeric	Year the car was produced
Mileage	numeric	Total kilometres travelled by the car
Manufacturer/Brand	categorical	Manufacturer of the car (Brand)
Category	categorical	The category of the car (e.g. sedan, SUV, hatchback)
Engine-Volume	numeric	Engine volume of the car (cc)
Engine_Power	numeric	Engine power of the car (in horsepower)
Avr-fuel-consumption	numeric	Average fuel consumption of the car (ℓ/100 km)
Fuel_Tank	numeric	Fuel tank capacity of the car (ℓ)
Model	categorical	Model of the car
Colour	categorical	Colour of the car
Drive-type	categorical	The car's drive type (e.g. FWD, RWD, AWD)
Gear_type	categorical	Car's gear type (e.g. manual, automatic)
Fuel-type	categorical	The fuel type of the car (e.g. gasoline, diesel, electric)

Table 2: Variables of Dataset

3.3 Exploratory Data Analysis

In this section, various visualizations of our datasets will be presented to facilitate a comprehensive understanding of their distributions, average values, outliers, and correlations. These visual representations will enhance familiarity with the datasets and enable visual analyses based on the statistical information encompassed. This approach will provide valuable insights into the underlying patterns and relationships within the data.

Since the prices of cars from different manufacturers in the data set are very different and some brands have high price averages and some have low or medium price averages, divided the data set into 4. These groups and manufacturers are as follows: High price (BMW, Mercedes, Audi), Medium-high price (Renault, Volkswagen, Toyota, Skoda), Medium-low price (Ford, Opel, Kia), Low price (Fiat, Citroen, Peugeot, Hyundai, Dacia). You can see the number of car brands in the dataset in table 3.

Low Priced		Medium-Low Priced		Medium-High Priced		High Priced	
Dacia	979	**Ford**	980	**Renault**	282	**Mercedes**	746
Citroen	943	**Kia**	363	**Toyota**	707	**BMW**	966
Hyundai	455	**Opel**	980	**Skoda**	979	**Audi**	613
Fiat	979			**Volkswagen**	980		
Peugeot	970						

Table 3: Number of Car Brands in the Dataset

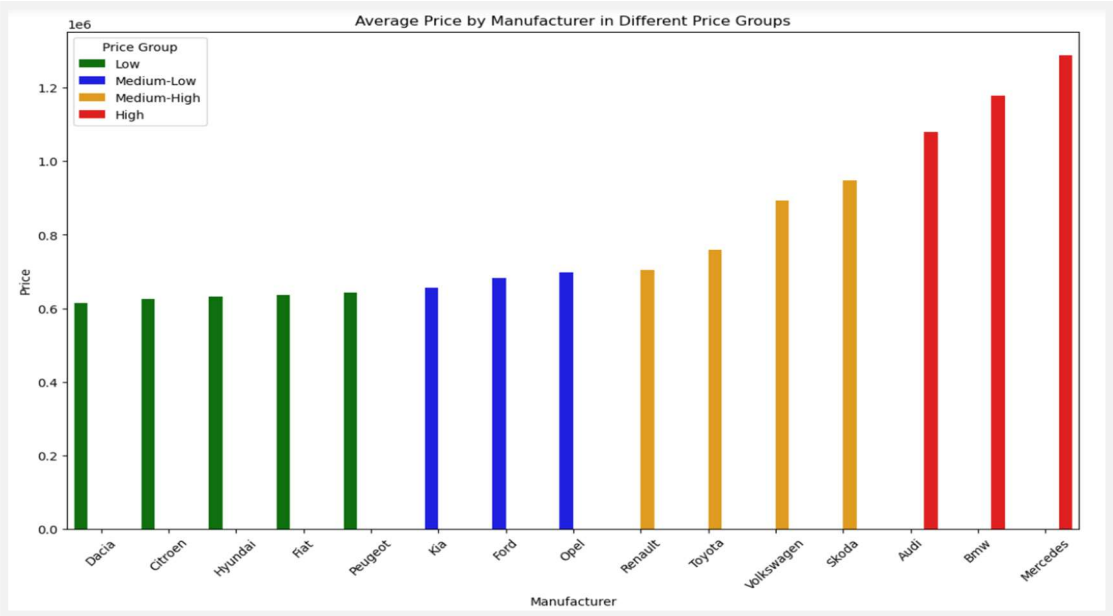


Figure 3: Average Price by Manufacturer in Different Price Groups

Figures 3 shows the car manufacturers and the average prices of the brands of these manufacturers for the low-priced, medium-low priced, medium-high priced and high price data sets, respectively. The data set, which included brands from 15 different manufacturers, was divided into four separate data sets and studied separately due to the very different price averages, and this enabled the accuracy rates in modeling to increase. If grouping was not done, using the model as a categorical variable and then converting it to numerical form would be quite complicated for regression analyzes due to the large number of models each brand has, and the number of dummy variables would be hundreds Therefore, after splitting the data set into four groups, it was more meaningful and easier to include the "Model" variable in the regression model.

When examining car price averages and the groups separated accordingly, it should be noted that this data is obtained from a data set consisting of used cars, which may not always carry the same information as new cars. Another thing to consider is that not every car brand or model may be evaluated the same way in every country. This study covers the Turkish market and real automobile information. Since all the cars included in the research are imported to Turkiye from different countries, their arrival prices vary. Moreover, the prices increase significantly with the added taxes, making it difficult to compare with any European country.

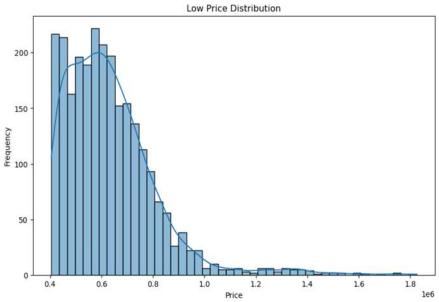


Figure 4: Distribution of Price - Low

Figure 5: Distribution of Price – Med-Low

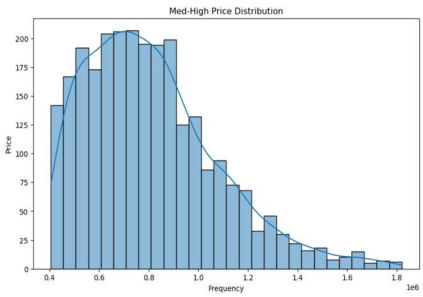


Figure 6: Distribution of Price - Med-High

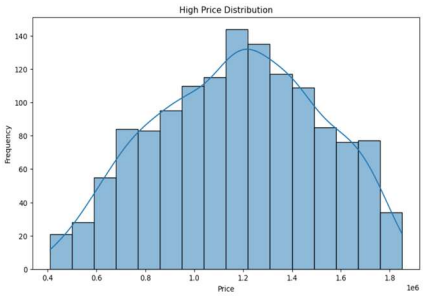


Figure 7: Distribution of Price - High

Figures 4, 5, 6 and 7 show histogram plots and density curves of the Price variable for each dataset. Given that the distributions deviated significantly from the normal distribution before splitting the datasets into four subsets, these visualizations show that partitioning the dataset by price was an appropriate decision. Current distributions now fit closer to the normal distribution, confirming the effectiveness of this approach.

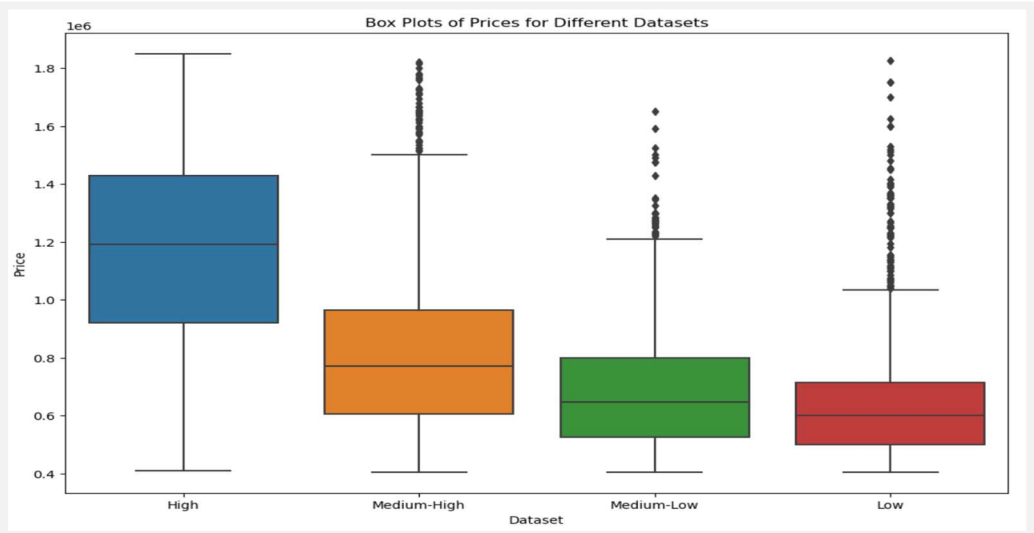


Figure 8: Box Plots of Prices of Different Datasets

Figure 8 contains visualizations showing the box plots of the Price variable of each data set. A box plot displays a five-number summary of a data set. The five-number summary includes minimum, first quartile, median, third quartile, and maximum values.In a box plot, we draw a box from the first quarter to the third quarter. The vertical line inside the box indicates the median. The lines extend from each quadrant to the minimum or maximum values. Apart from high-priced cars, we see the existence of outliers exceeding the quarterly values of the Price variable in other groups. This shows the importance of checking the box plot before data preprocessing steps.

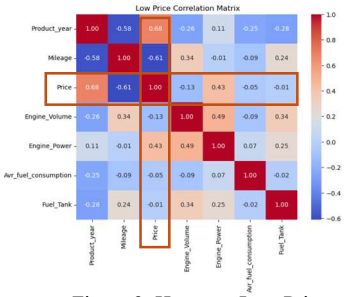


Figure 9: Heatmap-Low Price

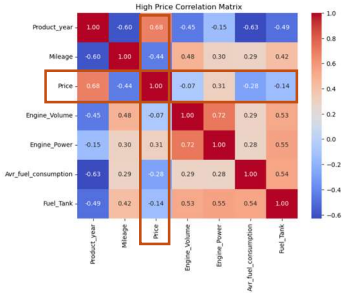


Figure 10: Heatmap-High Price

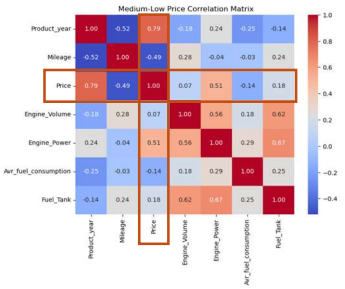


Figure 11: Heatmap-Med-Low Price

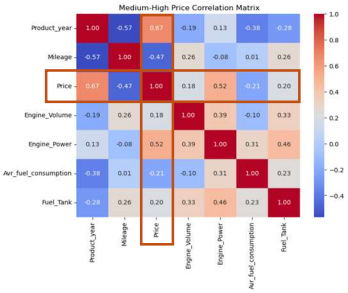


Figure 12: Heatmap-Med-High Price

In figures 9, 10, 11, and 12, we see heat maps showing the linear relationship and strength of correlation between two or more numeric variables. There can be a positive or negative relationship between two variables, positive correlation means that two variables change together in the same direction, while negative correlation means they change in the opposite direction. When we examine the figures, we see that the variables with the highest linear correlation in prices are the product year, engine power and mileage. Although there is a relatively high correlation between engine power and engine volume, which are independent variables, we see that the linear relationship between engine volume and the dependent variable price is quite weak. However, this does not mean that there is no non-linear relationship.

3.4 Data Pre-Processing

Since the data was collected through web scraping, it was raw and required extensive data cleaning and processing steps. In this section, all the steps taken to improve the quality of the data and prepare it for the modeling phase, as well as visuals of the initial and final versions of the data, will be shared.

The process began by taking the necessary steps to combine 15 different JSON files from 15 different manufacturers, which were scraped separately. While reading the files, the manufacturer name was added to the dataset as a new variable. Although they all actually represent the same information, some columns were dropped because they were registered twice with different names in some files. After the number of columns, types and names were equalized, all these data sets were combined to create a single data set and saved as a single JSON file. All variables that were supposed to be numeric but had some punctuation marks and text in them (all except Product year) were cleared and converted from categorical to numeric. All categorical variables and their classes in the Data Set were checked and they were translated from Turkish to English. Additionally, all variable names have been renamed to more commonly used and English names. Thus, the data set became completely English.

When we look at the production years of the cars, the fact that there were old cars that were almost scrap made the data set poor quality. For this reason, data for cars older than 20 years old were deleted and study continued with cars from the last 20 years of production. Excessively high (billion) or very low-price data that did not comply with logic due to incorrect data entry were deleted. Null values in the dataset were checked and since some variables had a relatively larger number of null variables, these were cleared from the dataset. The reason for this is to prevent the sample from losing accurate representation by replacing many null values of the variable with mean values. The "mean" value mentioned here is the "mean" values of the variables of each data set separated according to price groups. However, variables with relatively fewer null values were replaced with mean values to further prevent data loss. After the data cleaning process, you can examine Tables 4 and 5 respectively to understand the before and after situation of the dataset more clearly.

Some categorical variables have many classes, and some classes have only a few data. When I applied one-hot coding before creating the regression model, it would be unnecessary to turn classes that did not contribute to the model into dummy variables. I checked all classes against four datasets and removed some classes with less than 10 data in them. For example, the number of classes of the categorical variable "Model" such as Coupe, Sonata, Xsara was only 1 and these were removed from the dataset.

After completing data cleaning steps and grouping the data under four main categories, it is time to proceed with other necessary preprocessing steps before modelling. After grouping, the means of price and other numerical variables for each dataset were checked, and new average values and distributions were naturally observed. Outliers were identified using the box-plot method, and based on these box plots, values outside 2 or 3 standard deviations were replaced with mean values for each variable. (The "mean" values mentioned here are the values calculated separately for seven different numerical variables in four different data sets, that is, they were calculated separately for 28 variables.) Because some outliers were concentrated at the right end of the box plot while others were concentrated at the left end, different standard deviations were used to clean up the left and right ends to compensate for this imbalance. This method ensures that outliers do not disproportionately affect the results, thereby leading to more robust and generalizable predictive models. The distributions of all numerical variables (except Price) were checked using histograms, and logarithmic transformation was applied where appropriate. The Price variable was not log-transformed to make it easier to interpret the errors and model outputs at each stage of the modelling process. Additionally, applying a log transformation to the Price variable negatively impacted model results. Subsequently, dummy variables were created for categorical variables using one-hot encoding. The dataset is now ready for the application of regression models.

	0	1	2	3	4	5	6	7	8	İlan Tarihi	...	Motor Hacmi	Motor Gücü	Çekiş	Araç Durumu	Ort. Yakıt Tüketimi	Yakıt Deposu	Boya-değişen	Takasa Uygun	Kimden	Brand
0	/ilan/sahibinden-satilik-peugeot-508-1-6-e-hdi-access/sahibinden-peugeot-508-1-6-e-hdi-access-2012-model/25571062	Peugeot 508 1.6 e-HDi Access	Sahibinden Peugeot 508 1.6 e-HDi Access 2012 M...	2012	190.000	Beyaz	650.000 TL	02 Haziran 2024	İstanbul\nEyüpsultan\n\n\n\n\n\nKarşılaştır\nK...	02 Haziran 2024	...	1560 cc	114 hp	Önden Çekiş	İkinci El	4,4 lt	72 lt	1 değişen, 2 boyalı	Takasa Uygun	Sahibinden	Peugeot
1	/ilan/sahibinden-satilik-peugeot-301-1-6-hdi-active/sahibinden-peugeot-301-1-6-hdi-active-2014-model/25467323	Peugeot 301 1.6 HDi Active	Sahibinden Peugeot 301 1.6 HDi Active 2014 Model	2014	295.000	Beyaz	385.000 TL	02 Haziran 2024	İstanbul\nÜmraniye\n\n\n\n\n\nKarşılaştır\nKar...	02 Haziran 2024	...	1560 cc	93 hp	Önden Çekiş	İkinci El	4,3 lt	50 lt	Belirtilmemiş	NaN	Sahibinden	Peugeot
2	/ilan/sahibinden-satilik-peugeot-206-1-6-executive/sahibinden-peugeot-206-1-6-executive-2007-model/25569857	Peugeot 206 1.6 Executive	Sahibinden Peugeot 206 1.6 Executive 2007 Model	2007	78.825	Siyah	490.000 TL	02 Haziran 2024	İstanbul\nSarıyer\n\n\n\n\n\nKarşılaştır\nKarş...	02 Haziran 2024	...	1587 cc	112 hp	Önden Çekiş	İkinci El	7,1 lt	50 lt	5 boyalı	Takasa Uygun Değil	Sahibinden	Peugeot

Table 4 : First Version of Dataset

	Product_year	Mileage	Price	Model	Category	Engine_Volume	Engine_Power	Avr_fuel_consumption	Fuel_Tank	Manufacturer	color	Drive_type	Gear_type	Fuel_type
0	2015	113000	1849900	E	Coupe	1991.0	211.0	5.6	66.0	Mercedes	Navy Blue	Rear-Wheel Drive (RWD)	Automatic	Gasoline
1	2012	72000	1330000	C	Sedan	1796.0	156.0	6.7	67.0	Mercedes	Black	Rear-Wheel Drive (RWD)	Automatic	Gasoline
2	2016	147000	1165000	A	Hatchback/5	1461.0	109.0	4.4	50.0	Mercedes	Gray	Front-Wheel Drive (FWD)	Semi-Automatic	Diesel
3	2010	220000	830000	C	Sedan	1597.0	156.0	6.5	66.0	Mercedes	White	Rear-Wheel Drive (RWD)	Automatic	Gasoline
4	2018	80250	1325000	A	Hatchback/5	1461.0	116.0	4.5	43.0	Mercedes	White	Front-Wheel Drive (FWD)	Semi-Automatic	Diesel

Table 5: Last Version of Dataset

4 Design Specification

In the study, Random Forest, XGBoost, SVR and MLR models were used in used car price prediction. The reason for choosing these models is that each of them has the ability to make high accuracy predictions using different machine learning techniques. In addition, MLR and SVR are linear, while RFR and XGBoost are non-linear approaches, and these two approaches were also compared. The combination of these four models was preferred in order to compare model performances and determine the model that gives the best result. In this section, the algorithms, functions, and definitions applied in the study will be explained in detail. The techniques, architecture, and framework underlying the application will be described and supported with appropriate visuals, diagrams, or equations.

4. 1 Random Forest

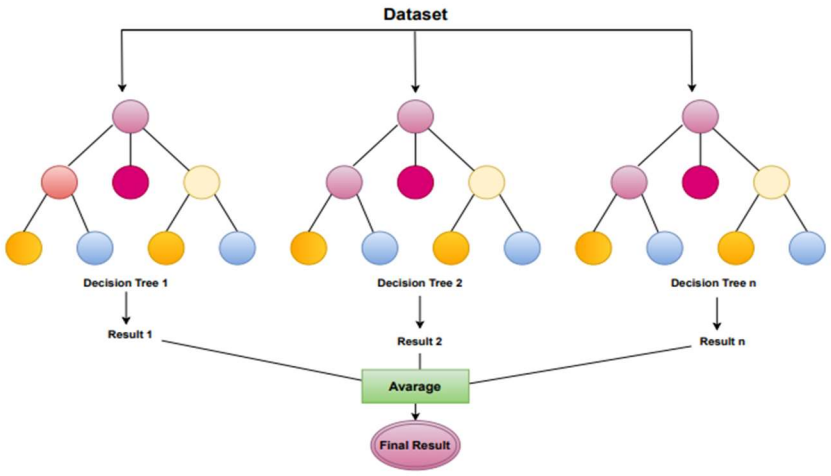


Figure 13: Random Forest Architecture

Random Forest is a widely used method in both regression and classification problems. We can use RFA for medical diagnosis and biomedical applications, marketing and customer analysis, retail and supply chain management, price predictions, financial analysis and risk management, etc. Random forest algorithm is the process of combining many decision trees that work independently of each other and selecting the value with the highest score among them. We can think of a decision tree as a game where, at each level, yes-no questions are asked. Each question narrows down the possibilities and brings us closer to the final outcome. The ensemble nature of random forest makes it an effective estimator.

4.2 Support Vector Regression

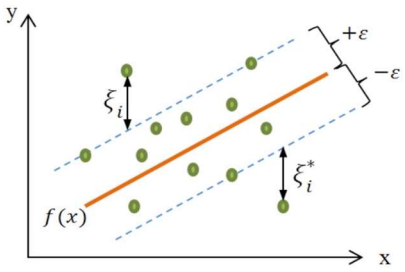


Figure 14: SVR (Yavuz, 2024)

SVR aims to find a function that predicts a continuous target variable and maximizes the margin between predicted values and actual data points. Support vectors are the data points closest to the regression line and determine the model. SVR can model nonlinear relationships using kernel functions. It is robust to outliers because it focuses on data points close to the margin. In the figure 14, the orange line denoted f(x) represents the estimated regression line. The dashed blue lines around this line indicate the margins called ε (epsilon). These margins indicate the allowable range of error between predicted values and actual data points. Green dots represent data points, while symbols ξ_i and ξ^*_i denote errors that fall outside the epsilon margins. SVR aims to provide the best prediction accuracy by trying to stay within these margins.

4.3 Multiple Linear Regression

Multiple linear regression is a statistical method that aims to predict a dependent variable using multiple independent variables and analyzes the linear effects of the independent variables on the dependent variable. The structure of the regression equation is as follows:

$$\Upsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

Υ = Dependent variable
$X_1/X_2...X_n$ = Independent variables
Slope ($\beta_1/\beta_2 \ldots \beta_n$): The amount of change in Υ caused by a one-unit change in $X_1/X_2...X_n$ The sign gives information about the direction of the relationship between the dependent and independent variables.
B_0= constant (point where it intersects the y-axis)
ε= Error term or residual

4.4 XGBoost

XGBoost (Extreme Gradient Boosting) regression is a powerful and efficient machine learning algorithm based on decision trees. In regression problems, it is used to predict the target variable (for example car price). XGBoost implements a boosting method that corrects errors by adding consecutive trees, and each new tree aims to reduce the errors of previous trees. This process increases the model's accuracy and generalization capability. Due to its speed, performance, and regularization capabilities, XGBoost is ideal for large datasets and complex prediction problems.

5 Implementation

The "Pandas" library was used to load, process, and analyse the data. Efficient computation was achieved by using “Numpy” for numerical operations and data processing tasks. “Matplotlib” and “Seaborn” were used for data visualization, allowing effective visualization of data distributions, correlations, and model performance. "Statsmodels" was implemented for statistical modelling and hypothesis testing, proving very useful, especially when conducting regression analysis.

Data preprocessing, modelling, and evaluation steps were carried out using these libraries. The code structure is organized starting from data loading and preprocessing steps to model training and evaluation. The tools and methods used at each stage are aimed at making the data analysis process more effective and understandable. After completing the model building steps, R-squared values and Mean Absolute Errors were calculated to evaluate and compare the results.

5.1 Results

Models	High Price	Medium-High Price	Medium-Low Price	Low Price
MLR	0.83	0.84	0.84	0.83
SVR	0.53	0.66	0.72	0.70
RFR	0.94	0.95	0.96	0.95
XGBoost	0.95	0.95	0.96	0.95

Table 6: Comparison Table - R-Squared Values of Machine Learning Models

Table 6 shows the R-squared values of four different regression models used for used car price prediction in different price groups. Likewise, Table 7 shows the MAPE values of four different regression models in different price groups. Analysis was conducted for four separate data sets split into four separate groups according to the price variable (high price, medium-high price, medium-low price and low price).

Models	High Price	Medium-High Price	Medium-Low Price	Low Price
MLR	11.7%	8.5%	8.8%	7.9%
SVR	16.3%	12.4%	10.8%	9.3%
RFR	10.7%	8.5%	8.2%	7.7%
XGBoost	11.2%	8.3%	8.4%	7.7%

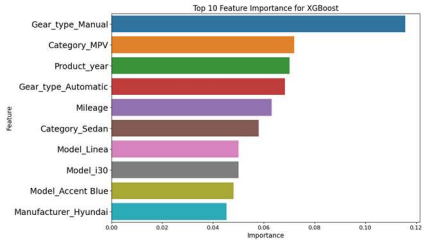
Table 7: Comparison Table-Mean Absolute Percentage Error (MAPE) Values of Machine Learning Models

6 Evaluation

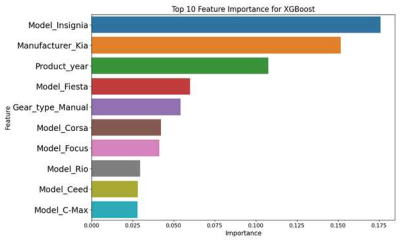
When examining Table 6, the XGBoost model generally exhibits the highest R-squared values, demonstrating superior performance, particularly in the High Price and Medium-Low Price categories. The Random Forest Regressor (RFR) model also performs closely to XGBoost, with similarly high R-squared values. Conversely, the Multiple Linear Regression (MLR) model shows lower performance compared to the RFR and XGBoost models in terms of R-squared values, whereas the Support Vector Regression (SVR) model exhibits the lowest R-squared values across all categories. This indicates that SVR's linear structure is inadequate for capturing the complex relationships within the dataset.

Table 7 evaluates the models based on their MAPE values. Here, the RFR model generally has the lowest MAPE values, yielding the best results in the High Price, Low Price, and Medium-Low Price categories. The XGBoost model, however, achieves the lowest MAPE value in some categories, especially in the Medium-High Price category, and overall displays a competitive performance. The MLR and SVR models, on the other hand, present higher MAPE values and consequently make larger estimation errors compared to RFR and XGBoost.

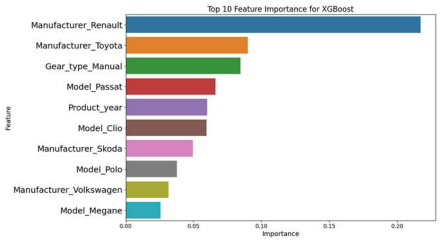
In conclusion, the XGBoost and RFR models emerge as the most effective for predicting used car prices. The XGBoost model is capable of better capturing complex relationships in the dataset, as evidenced by its high R-squared values and competitive MAPE values. The RFR model consistently delivers the lowest MAPE values, providing the most accurate predictions with smaller prediction errors. This analysis underscores the importance of considering both R-squared and MAPE metrics when selecting the optimal model for price prediction.



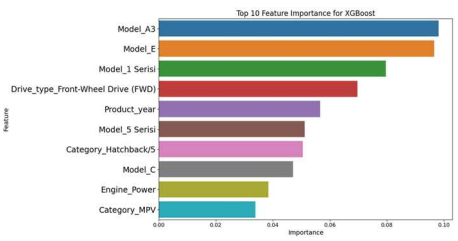
a) Most important features by low price



b) Most important features by med-low price



c) Most important features by med-high price



d) Most important features by high price

Figure 15: Most Important Features by Price Groups

For the purpose of the study, different regression models were compared, and the most important features affecting the price were investigated as the second aim. Bar plots were created in Figures 15 a, b, c, and d, showing the top 10 features according to the XGBoost model, one of the best prediction models selected. These plots highlight the features with the highest importance scores, providing insight into the key variables influencing the model's predictions.

In the high price group, Model_A3(Audi) was determined as the most important feature. This is followed by Model_E (Mercedes). In the low-price group, Gear_Type_Manual and Category_MPV stand out as the two most important features. Model_insignia(Opel) has been determined as the most important feature in medium-low priced cars. In the medium-high price group, Manufacturer_Renault and Manufacturer_Toyota stand out as the most important features.

Actual	Predicted	Percentage Error
757,900	793,235	4.6%
669,000	672,931	0.6%
515,000	493,321	-4.2%
522,000	567,322	8.7%
425,000	431,126	1.4%
850,000	813,520	-4.3%
497,500	511,719	2.9%
756,500	729,598	-3.6%
525,000	520,256	-0.9%
770,900	785,196	1.8%

Table 8: Actual-Predicted Prices(Turkish Lira) and PE-Low Priced

Actual	Predicted	Percentage Error
1010000	986368	-2.3%
999750	1146902	14.7%
1490000	1519327	1.9%
805000	774898	3.7%
1499000	1295090	-13.6%
1660000	1766291	6.4%
795000	848834	6.8%
1250000	1388543	11%
1470000	1587742	8%
950000	1065041	12.1%

Table 9: Actual-Predicted Prices(Turkish Lira) and PE-High Priced

Tables 8 and 9 above are generated from 10 randomly selected data and show the predicted results using the Random Forest model. Both tables contain the "Actual" and "Predicted" prices along with the MAPE (Mean Absolute Percentage Error) values for two datasets grouped as low-priced and high-priced.

The tables reflect the overall performance of the model positively. Generally, the predicted values are quite close to the actual values. For example, low MAPE values (e.g. 0.59, 1.44) reinforce the accuracy of the model. However, although the model generally has a high accuracy rate, there are deviations in some cases and indicate that the model needs to be improved in some specific cases.

6.1 Discussion

Different features' varying importance in different price groups shows that each price segment has its own dynamics. For example, the importance of manual gear in low-priced cars shows that buyers in this segment prefer manual gear, and this feature has a significant impact on price. Similarly, the fact that specific models and manufacturers come to the fore in high-priced cars shows that cars in this segment are expected to have desirable brands and models. For mid-priced cars, we can say that gear type and specific manufacturers have a significant impact on the price.

These differences indicate that each price group has its own unique customer demands and market dynamics, thus the characteristics of each group vary in importance. Additionally, it is crucial to remember that this analysis pertains specifically to the used car market in Turkey. The preferences and market dynamics could be entirely different for new cars or in other countries.

When choosing the best predictive model, it is challenging to select a single model since four different datasets and two different evaluation metrics are used. When examining the R-squared and MAPE evaluation metrics, these metrics indicate different models in some datasets. In this case, there is no harm in choosing both models as the best predictors. However, it would be more meaningful to choose the MAPE metric in evaluating the success of a model because its main purpose is to make predictions with the least error. Therefore, it would be appropriate to conclude that XGBoost is the best predictor for medium-high priced cars, while the Random Forest model is the best predictor for other price groups.

7 Conclusion and Future Work

The study used various regression models to predict used car prices and determine the most accurate prediction model. When completed, both Random Forest and XGBoost models were identified as highly effective and achieved R-squared values in the range of 0.94-0.96. The Random Forest model achieved the lowest Mean Absolute Percentage Error (MAPE) values for 3 price groups, while the XGBoost model achieved the lowest MAPE value in the medium-high price group. Considering their performance, XGBoost can be used for the medium-high price group, and the Random Forest model can be used for the other price groups. However, since the Random Forest model minimizes the MAPE values in the majority of the groups, it would be a good decision to choose it as the best estimator.This preference is based on the statistical principle that the primary goal in prediction is to minimize the prediction error. In addition, the findings revealed that gear type, specific car brands, and models play an important role in determining prices, but these features vary according to price groups.

Future expansion of our study will be possible by working with more data or by including different and more diverse car brands. Thus, the performance and general validity of the models can be increased with a larger and more diverse data set. In addition, the regression models we applied in our study can be diversified and potentially higher accuracy rates can be obtained by adding new models. Advanced modelling techniques, especially deep learning methods, can provide more successful results in price predictions.

Acknowledgements

I would like to sincerely thank my supervisor Dr Christian Horn for his invaluable guidance and support throughout the entirety of this project. His meticulous supervision and dedication at every stage, down to the finest details, were instrumental in ensuring the quality and success of this research. I am deeply fortunate to have had the opportunity to work under such a highly professional and disciplined mentor, whose expertise and insights have greatly enriched my learning experience.

I would also like to thank my dear husband, Anıl Ulutürk, who encouraged me to start my master's program and supported me throughout the year. Without him, I would not have had the courage and strength to complete my education in a foreign country.

References

Agrahari, K., Chaubey, A., Khan, M. & Srivastava, M., 2021. Car price prediction using machine learning. *International Journal of Innovative Research in Technology*, 8(1), pp.572-575.

Bhatt, N.S. et al., 2023. An Empirical Analysis of Machine Learning Algorithms for Used Car Price Prediction System. In *2023 Global Conference on Information Technologies and Communications (GCITC)*, pp.1-5. IEEE.

Çelik, Ö. & Osmanoğlu, U.Ö., 2019. Prediction of the prices of second-hand cars. *Avrupa Bilim ve Teknoloji Dergisi*, (16), pp.77-83.

Hankar, M., Birjali, M. & Beni-Hssane, A., 2022. Used car price prediction using machine learning: a case study. In *2022 11th International Symposium on Signal, Image, Video and Communications (ISIVC)*, pp.1-4. IEEE.

Jain, S. & Punia, S.K., 2024. Accurate Estimation Technique Development for Second Hand Car Through Linear Regression Algorithm. In 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT), Vol. 5, pp.723-727. IEEE.

Jin, C., 2021. Price prediction of used cars using machine learning. In *2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT)*, pp.223-230. IEEE.

Khan, Z., 2022. Used car price evaluation using three different variants of linear regression. *International Journal of Computational and Innovative Sciences*, 1(1), pp.12-20.

Lessmann, S. & Voß, S., 2017. Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy. *International Journal of Forecasting*, 33(4), pp.864-877.

Monburinone, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S. & Boonpou, P., 2018. Prediction of prices for used car by using regression models. In *2018 5th International Conference on Business and Industrial Research (ICBIR)*, pp.115-119. IEEE.

Narayana, C. V., Madhuri, N. O. G., NagaSindhu, A., Aksha, M. & Naveen, C., 2022. Second Sale Car Price Prediction using Machine Learning Algorithm. In *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, pp.1171-1177. IEEE.

Ponmalar, P.P. & Christinal, A.C., 2022. Review on the pre-owned car price determination using machine learning approaches. In *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, pp.274-278. IEEE.

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

Rane, M. et al., 2023. Random Forest Regressor Approach for Predicting Resale Value of Used Vehicles (RFRVP). In *2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pp.1-6. IEEE.

Samruddhi, K. & Kumar, R.A., 2020. Used car price prediction using K-nearest neighbor based model. *International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)*, 4(3), pp. 2020-686.

Satapathy, S.K., Vala, R. & Virpariya, S., 2022. An automated car price prediction system using effective machine learning techniques. In *2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, pp.402-408. IEEE.

Satioglu, M. C., Ar, Y. & Tugrul, B., 2021. Automobile Price Prediction in Turkey Marketplace with Linear Regression. In *2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp.329-333. IEEE.

Shanti, N., Assi, A., Shakhshir, H. & Salman, A., 2021. Machine Learning-Powered Mobile App for Predicting Used Car Prices. In Proceedings of the 2021 3rd International Conference on Big-data Service and Intelligent Computation, pp. 52-60.

Statista. (2023). Leading car brands based on sales in Turkey. Available at: https://www.statista.com/statistics/473806/turkey-leading-car-brands-based-on-sales/ (Accessed 07 Jul. 2024).

Muti, S. & Yıldız, K., 2023. Using linear regression for used car price prediction. *International Journal of Computational and Experimental Science and Engineering*, 9(1), pp.11-16.

Sun, N., Bai, H., Geng, Y. & Shi, H., 2017. Price evaluation model in second-hand car system based on BP neural network theory. In *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pp.431-436. IEEE.

Varshitha, J., Jahnavi, K. & Lakshmi, C., 2022. Prediction of used car prices using artificial neural networks and machine learning. In *2022 International Conference on Computer Communication and Informatics (ICCCI)*, pp.1-4. IEEE.

Venkatasubbu, P. & Ganesh, M., 2019. Used cars price prediction using supervised learning techniques. *International Journal of Engineering and Advanced Technology (IJEAT)*, 9(1S3).

Wang, F., Zhang, X. & Wang, Q., 2021. Prediction of used car price based on supervised learning algorithm. In 2021 International Conference on Networking, Communications and Information Technology (NetCIT), pp.143-147. IEEE.

Yadav, A., Kumar, E. & Yadav, P.K., 2021. Object detection and used car price predicting analysis system (UCPAS) using machine learning technique. *Linguistics and Culture Review*, 5(S2), pp.1131-1147.

Yavuz, A. (2024). Destek Vektör Regresyonu ve Makineleri. Available at: https://yavuz.github.io/destek-vektor-regresyonu-ve-makineleri/ (Accessed: 8 July 2024).

Yılmaz, S. & Selvi, İ.H., 2023. Price Prediction Using Web Scraping and Machine Learning Algorithms in the Used Car Market. *Sakarya University Journal of Computer and Information Sciences*, 6(2), pp.140-148.

Zhang, H., 2022. Prediction of Used Car Price Based on LightGBM. In 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), pp. 327-332. IEEE.