Demystifying second hand car market in Bangladesh using multimodal machine learning techniques

by

Roman Islam
19101596
Arman Rahman
19101318
Asifur Rahman
19101396
Wasiq Ferdous
19101316
Sk.MD.Golam Arman
18201054

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
January 2023

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material that has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Roman

Roman Islam
19101596

Arman Rahman

Arman Rahman
19101318

ASIFUR RAHMAN ASIF

Asifur Rahman
19101396

Wasiq Ferdous
19101316

S. Arman

Sk.MD.Golam Arman
18201054

Approval

The thesis/project titled "Demystifying second hand car market in Bangladesh using multimodal machine learning techniques" submitted by

1. Roman Islam (19101596)
2. Arman Rahman (19101318)
3. Asifur Rahman (19101396)
4. Wasiq Ferdous (19101316)
5. Sk.MD.Golam Arman (18201054)

Of Fall, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 19, 2023.

Examining Committee:

Supervisor:
(Member)

Farig Yousuf Sadeque

Dr. Farig Sadeque
Assistant Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

When a person plans to buy or sell a car in an overly-crowded second-hand market, it is often difficult to estimate accurate prices. People are frequently duped by scalpers into purchasing cars that are not worth the base price. People even sell cars for far less or far more than the market valuation. We aim to establish a system that allows people to see a car's most recent anticipated value. Apparently, using multimodal supervised machine learning approaches, this study illustrates a second hand car price prediction system. Here, we will be using Shallow machine-learning techniques and deep neural machine-learning techniques for such calculations. Then, the predictions will be compared and analyzed to find the precision of the best performance. We believe this research will provide us with the information we need to solve this challenge with high accuracy.

Keywords: Used cars, Multimodal, Shallow machine learning techniques, Deep neural machine learning techniques, Price prediction

Dedication

We dedicate this thesis to our parents for their constant guidance and support which has been pivotal towards the completion of our thesis. They motivated us, and we were able to finish our thesis efficiently.

Acknowledgement

We would like to express our sincere appreciation to Dr. Farig Yousuf Sadeque, our adviser, for providing us with direction, support, and encouragement all during the process of conducting research and preparing our thesis. His invaluable observations and years of experience have been a significant influence on the development of this work. In addition, I would like to extend my gratitude to the esteemed faculty members and individuals of BRAC University for the support and assistance they provided in all facets of this endeavor. In conclusion, we would like to express our gratitude to our family and friends for their unflinching support and unmatched level of understanding during the entirety of this project.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

MAE	Mean Absolute Error
MLP	Multi Layer Perceptron
MSE	Mean Squared Error
RF	Random Forest
$RMSE$	Root Mean Square Error
SVM	Support Vector Machine
$XGBoost$	eXtreme Gradient Boosting

Chapter 1

Introduction

1.1 Background Information

In the automotive industry, ”second hand” refers to a vehicle that has been sold to at least one other private buyer. They are quite common in Bangladesh and other developing countries. There are several factors that influence how much a secondhand car will cost. Customers often have a hard time finding an affordable used automobile that meets their needs. Therefore, if a customer knows what kind of vehicle they need, predicting the price still remains tricky. Within the scope of this research, we look into this problem and provide a method for making accurate price predictions for used automobiles in Bangladesh by using machine learning. An individual who wishes to buy a used car can use the model described in this article to make a better-informed choice.

Predicting the price of a car in Bangladesh would likely take into account a number of factors that are unique to the market and economy of that country. Some of these could be import tax, currency fluctuations, and the availability of car financing and loan options. Also, it would be important to think about the demand for different kinds of cars in Bangladesh, as well as the features and specifications that are most popular with Bangladeshi buyers. Compared to other countries, Bangladesh's car market is small, and only a few big brands like Toyota, Suzuki, and Honda make up the most of it. The used car market has a big impact on the market as well. The price of a used car depends on how old it is, how many miles it has been driven, its brand, model, and condition. Predicting car prices in Bangladesh is hard because there are many things that affect prices and demand. So, the prediction model would probably use advanced machine learning techniques and a lot of information about past car sales and other factors that are important.

1.2 Problem statement

Due to a lack of appropriate and pleasant public transportation in Bangladesh, many individuals have been obliged to own automobiles. Moreover, because of safety concerns and for a better life, a large number of people buy cars. In our country, most families are middle class. Because of the high cost of vehicles in Bangladesh, acquiring one has remained out of reach for middle-income households [15]. Part of the reason for the high cost of automobile ownership is the government's high import taxes. According to the government's national budget for the fiscal years 2020–2021, a buyer has to pay 128 percent import tax for cars between 0 cc and 1800 cc, 221 percent import tax for cars between 1800 cc and 2000 cc, 365 percent import tax for cars between 2000 cc and 3000 cc, and 628 percent import tax for cars between 3000 cc and 4000 cc [18]. That's why, middle class families cannot afford to buy a new car. So, they are more comfortable buying a used car, and for this reason, the used car market is very popular in our country. Purchasing a used car in Bangladesh is not as simple as you may imagine. Buyers encounter numerous challenges when purchasing a second-hand car. For starters, both buyer and the dealer have a lack of knowledge about the used car. The cost of a car is determined by the brand value, the year of manufacture, and other factors. It is difficult to obtain this information. So they can not perfectly determine the cost of the second-hand car by verifying the market condition. Therefore, sellers may occasionally deceive buyers. The seller deceives the buyer by providing false information. The second most important thing is to understand the mileage of a used car. Aside from the car's brand name and model, the mileage has the greatest impact on its worth. And it is quite difficult to detect dents and scratches in used car photos. The cost of the car varies depending on these factors. These are the issues that a buyer and a seller has while purchasing and selling a used car.

1.3 Research Motivation

In light of recent events, the automobile industry is thought to be one of the fastest-growing businesses in Bangladesh. The vehicle industry has substantially inflated throughout the last ten years, thanks to the socioeconomic development of the nation as well as the country's middle- and upper-middle classes' increasing purchasing ability. BRTA statistics show that, between the year 2011 and 2020, the CAGR(Compound Annual Growth Rate) for passenger automobile growth was 5.43 % [14]. Also, the transportation industry in Dhaka is undergoing a progressive change thanks to ride-sharing applications like Uber, Pathao, and others.Though the concept of ride-sharing services have been recently introduced in Bangladesh, mainly in Dhaka City; they are expanding throughout the country gradually. As the capital of a country, Dhaka really does have a significant client base, growing income as well as the internet consumer rate is increasing day by day. As a result, it is a perfect chance for the growth of profit making services. Within the last two years, ride-sharing services have become more popular in shared mobility [7]. Therefore, our major purpose is to entice buyers to purchase a used automobile. The previous owner's use of the vehicle may have resulted in difficulties that diminish the car's level of condition, such as a lower number of miles per gallon, scratches or dents on the outside, or problems with the vehicle's internal equipment, used automobiles are less expensive than new cars. As a result, used automobile pricing is at least largely determined by the number of miles driven, and as the number of kilometers traveled increases, the car's price decreases.

1.4 Research Objective

To solve the problem of price estimation, the followings are the research objectives:
1. To collect data about used cars, which will be cleaned and analyzed to uncover key elements that can be used to predict car prices, with no data being lost.
2. To make the data accessible for public.
3. Predicting the cost of secondhand automobiles using multi-modal supervised machine learning algorithms effectively.
4. Deploy a web application for end-users.

Chapter 2

Literature Review

The purpose of Gajera et al.[13] was creating a data set to predict second-hand car price. The model's accuracy is used to figure out the best technique. The scientists utilized a dataset of 92,386 events to develop the model .Distance traveled in kilometers, year since registration,car model, fuel type,gear type and automotive brand have all been proven to impact the cost of a car. Linear regressors, K Nearest Neighbors (KNN), XGBoost, Decision Tree and Random Forest were applied by the authors.The highest result was produced with Random Forest, with R-squared value and lowest root mean square error, according to observational data.

Gonggie et al.[4] presented an Artificial Neural Networks (ANN)-based system ,estimating the cost of second-hand vehicles. He looked at multiple aspects, including the amount of kilometers traveled, the estimated vehicle life, and the company. The suggested model was intended to handle nonlinear data interactions,It was different from earlier models that relied on basic linear regression techniques. The non-linear model outperformed standard linear models in predicting automobile prices.

Listiani et al.[1] calculated the value of a secondhand automobile by factoring in depreciation by using Support Vector Regression. Also for comparing the accuracy of the predictions, statistical regression models are being used. A completely automated approach to modify and execute Support Vector Regression (SVR) is designed utilizing revolutionary research ideas. The basic concept is powered by machine learning and is focused on actual statistics from a significant German carmaker.

Wu et al.[3] used an experience and understanding system to perform a vehicle price prediction study. These parameters were taken into consideration: year of production, engine type and company. Their model gave findings that were similar to the basic regression model. Moreover, they established an intelligent program known as ODAV (Optimal Distribution of Auction Vehicles) because there is a huge need for vehicle sellers to sell cars towards the conclusion of the rental year. Also, This approach offers details on the best vehicle price along with the location of the best bargain.

Gegic et al.[10] study numerous factors for consistently and precisely calculating used automobile costs in Bosnia Herzegovina. The authors constructed a statistical technique employing three important techniques based on machine learning which are: Artificial Neural Networks (ANN), Random Forests (RF) and Support Vector Machines (SVM). Rather than functioning independently, the machine learning techniques were merged to form a combination. The accuracy of the technique on the dataset was not more than 50 percent. Nevertheless, after combining the techniques, the accuracy improved to 92.38 percent which represents a significant improvement. Furthermore, in comparison to a single machine learning technique, the suggested process consumes considerably more computing resources.

Venkatasubbu et al.[11] used the following machine learning methods: multiple regression, regression tree and lasso regression to develop models that could predict the cost of a second-hand car depending on former customer information and a list of standards. All of these models had far lower prediction error rates than the allowed 5 percent, according to an analysis of their prediction accuracy. The analysis of variance (ANOVA) confirmed that the Regression Tree Model provided the highest result of mean error rate. Lasso regression and the multiple regression models produced lower mean error rate than Regression tree but at the same time they do not have a huge difference, according to the post hoc test.

Noor and Jan[6] developed a model using some machine model techniques to determine vehicle price prediction. The author first collects the data and then splits it for Pre-processing and testing. They use linear regression, so they have to use Minitab for proper use of regression. They used the Least square method for model estimation after getting the result from Minitab. The author also uses feature selection and at last, gets a 98 percent accuracy rate from their model.

Samruddhi et al.[12] suggested a machine learning model for evaluating second-hand automotive pricing based on the K-Nearest Neighbor (KNN) technique, considered to be best for smaller sets of data. Here, the classifier was put to the test using the data, and the ratio of train to test data was used to evaluate its effectiveness. To evaluate its performance, the same model is cross-validated using the K-Fold validation data approach. A dataset from the platform Kaggle was employed for training the model. In comparison to linear regression, which was shown to be just 71 percent accurate, the K-Nearest Neighbor method was 85 percent accurate. The experimental analysis demonstrates that the K-Fold validation data approach model is the best match.

Richardson et al.[2] used multiple regression analysis in his paper. He presented that hybrid automobiles have longtime value compared to normal vehicles. Richardson suggested that manufacturers create vehicles with greater durability. This originates from environmental fears about climate change, and also increases fuel efficiency.

Based on supervised machine learning, Monburinon and colleagues[8] considered the efficiency of regression models. A pricing strategy for secondhand automobiles was created using Random Forest Regression (RFR), Multiple Linear Regression (MLR) and Gradient Boosted Regression Trees, a pricing strategy for secondhand automobiles was developed. For training each model, the identical test data obtained from the German used car e-commerce websites was used. Comparing the error values, Gradient Boosted Regression Trees provided the highest outcome with the value of mean absolute error (MAE) of 0.28. RFR and MLR had the mean squared value (MSE) of 0.35 and 0.55 constitutively, which was not as good as Gradient Boosted Regression Trees.

For car price prediction in Mauritius, Pudaruth [5] employed a range of methods which are: decision trees, naive bayes, k-nearest neighbors and multiple linear regression analysis. The prediction model's data was manually collected from local newspapers. He researched multiple characteristics which are: cubic capacity, brand, model, mileage in kilometers (KM), manufacture year, transmission type, price and exterior color. Furthermore, the author observed that Decision Tree and Naive Bayes were incapable of categorizing as well as predicting a set of numbers. Furthermore, due to the minimal quantity of dataset examples, high classification performance, i.e. less than 70 percent accuracy, was not attainable.

Researcher Nabarun Pal [9] used Random Forest to predict second hand car prices. Prices for used cars were predicted using data from Kaggle. By training 500 decision trees using Random Forest, a thorough data exploration was conducted to investigate the pricing consequence of the variables. It is often used for classification, but by transforming the issue to a similar regression problem, they were able to convert it into a regression model. Experimental results showed training accuracy, 95.82 percent and testing accuracy, 83.63 percent. Choosing the highest related characteristics, the machine learning figure can estimate used car cost with high accuracy.

Chapter 3

Methodology

The research methodology for this thesis paper on predicting used car prices will involve collecting data on various attributes of used cars, such as brand, model, year, mileage, and condition, from reputable online sources such as automobile classified websites. This data will be cleaned and pre-processed to ensure its quality and consistency. Next, statistical techniques such as regression analysis will be applied to the data to develop a predictive model. The model will be evaluated using metrics such as mean absolute error and coefficient of determination. Finally, the model will be tested on a separate dataset to verify its performance and make any necessary adjustments. In addition, this research will also make use of Machine Learning algorithms such as Random Forest, XGBoost.

3.1 Data Interpretation

In the description of the data research methodology has been presented, indicating the data acquisition where the data gathering via scraping is explained. Alongside pre-processing, missing values are handled, plus correlation of features is mentioned in the data analysis in addition to the process of feature selection. The research methodology is detailed in the following data description information. In this part the acquisition of data is described in the methodology, where scraping is conducted for collecting data. In addition to the process of feature selection, the data analysis also addresses missing values and mentions the correlation between features.

3.1.1 Data Acquisition

To get this information, we scraped the website https://bikroy.com/en, a renowned Bangladeshi internet marketplace. We selected this site since it was one of the first in Bangladesh to purchase and sell used vehicles, and it holds a large amount of data. We prioritized html requests for data extraction above other scraping techniques. Other libraries like BeautifulSoup and Selenium experienced some issues with data scraping. Since there is a top down option that hides the complete description, it was difficult to obtain the entire description for Beautiful the soup from the Bikroy website. When attempting to use Selenium's pagination functionality, users frequently encountered errors such as "failed to create a new connection: [winerror 10061] no connection could be made since the target computer actively refused it. A portion of the message gives the impression that the server to which you are attempting to connect does not allow connections on the port that was selected. It's possible that the server is disabled, that it's not set up to accept connections on that port, or that a firewall is preventing the connection from being made which is fixed or withdrawn. The aforementioned difficulties may be resolved with the use of an HTML request in this site. HTML requests and Beautiful Soup are both libraries that may be used to interact with web pages, but their aims are distinct. The first dataset contains over seven thousand sets of data and 20 features. HTML requests is a library for sending HTTP requests to a web server and retrieving the response. It can get a web page's HTML source code, which can subsequently be parsed using a library such as Beautiful Soup. The Beautiful Soup library parses and navigates HTML and XML texts. It is used to extract specific information from a web page's HTML source code. It offers a straightforward and intuitive method for navigating, searching, and modifying the parse tree of an HTML or XML document. Therefore, HTML Queries is used to retrieve the HTML source code of a web page, whereas Beautiful Soup is utilized to extract specific information from the HTML source code. Together, these libraries can be used to extract the desired information from a web page. First, you send a request to the web server using the requests library, then you use Beautiful Soup to parse the HTML source code and extract the required information. By iterating and altering the string value of the url, pagination was handled using html requests. This webpage has a car list. The individual vehicle url was then retrieved from the corresponding url by obtaining the href values from the relevant car module using its html class. All of the automobiles belong to the same class, which provided us with unique car urls. Finally, after obtaining the

Features	**Illustration**
CarName	Includes the car name
brand	Brand of cars
model	Model of cars
manufacture_year	Manufacturing year of the cars
trim	Feature of the particular version of a car model
condition	Includes whether it's used or reconditioned
transmission	Includes whether it's automatic or manual
body_type	Different types of car body
fuel_type	Types of fuel
engine_capacity	Includes the engine capacity in cc
mileage	Car's traveled distance in kilometers
price	Selling price of cars
description	Includes the description given by the seller
seller_info	Includes whether the vehicle is being sold via a shop or an individual
color	Car color
owner_info	Whether the seller is a member or not
AC_system	Includes if the car has Air Conditioning system or no
Accident_history	Car's accident history
camera_facility	Includes if the car has camera facility or no
Light_facility	Includes if the car has light system or no

Table 3.1: Description of selected attributes

automobile url, we extract each feature using the html.find() method, in which the class name of the feature is obtained by analyzing the htm page.We extracted a lot of features from the website, where we also did feature engineering in case of seller info and description given on the website.

3.1.2 Data Pre-Processing

Before the data can be analyzed, they must first go through a procedure that is referred to as ”data preparation,” which comprises organizing, transforming, and cleaning the data. This must be done before the data can be evaluated. This includes removing any values that are invalid, missing, or out of range, as well as reformatting the data, scaling the variables, and creating new variables. Formatting the data, reformatting the data, and adding new variables are some more procedures that need to be completed. Following the completion of our preliminary analysis of the dataset, we were able to identify a few problems. There is a risk that the collected dataset will not always be in a form that can be exploited by the algorithms that are involved in machine learning. It is essential to clean up data that has become damaged over time in order to ensure that machine learning algorithms can be employed effectively. This will ensure that the desired results can be achieved. During the process of extracting data from the Bikroy.com website, we found that the information referring to the various types of automobiles was laid up in rows that were kept entirely to themselves. This option is not available for all cars, despite the fact that a significant number of users on the website have provided information

Features Name	**Unique Values**
CarName	5110
brand	39
model	238
manufacture_year	NaN
trim	2986
condition	2
transmission	3
body_type	8
fuel_type	50
engine_capacity	NaN
mileage	NaN
price	NaN
description	7140
seller_info	1507
color	16
owner_info	2
AC_system	2
Accident_history	2
camera_facility	2
Light_facility	4

Table 3.2: Number of unique values for all the features

regarding the year of registration for their own cars. As a result of this, when we attempted to scrape the data, we found that our dataset contained features with null values for the Condition and Manufacture year columns. This was a discovery that came about as a direct consequence of the previous point. For handling missing values, we dequeued rows where the values are missing. Then, we discovered that our dataset had a significant amount of duplicate values. Python code was used in order to get rid of these occurrences.
Information for each fuel type's feature counts:

Fuel_type	**Value Counts**
Octane	1991
CNG, Octane	1596
Petrol, Octane	1008
Hybrid	766
Hybrid, Octane	687
Petrol, Hybrid, Octane	582
Octane, LPG	209
Petrol, CNG, Octane	204
Diesel	183
Petrol, CNG	144

Table 3.3: Value counts of fuel type feature

The following are the value counts for each body type:

Body_type	**Value Counts**
Saloon	3662
MPV	1487
SUV / 4x4	1380
Hatchback	785
Estate	367
Coupé/Sports	36
Convertible	8

Table 3.4: Value counts of body type feature

As we have a lot of Toyota brand cars, Axio car model has the most value as shown in the table, then comes Premio model. We select the top ten models from the dataset, as this a Bangladeshi website. We have these second hand cars a lot. There are some other names inlcuded as for the top ten. Hiace, Noah, Corolla, Fielder, Allion, X-Trail, Esquire, Aqua car models have some major counts. Even though, from the website we tried to extract the feature is a dynamic website and the website uploads car selling information every minute, it's hard to keep up the pace where could high brand models even more. There are a few expensive brand models were found in the dataset.
Also, when it comes to transmission feature, we have the most value in Automatic. There are some information of cars where transmission wasn't given. Also, a few cars have manual transmission. These values have good impacts in our paper. The following are the value counts for models as well as the transmission count:

Model	**Value Counts**
Axio	965
Premio	712
Hiace	656
Noah	570
Corolla	502
Fielder	448
Allion	441
X-Trail	327
Esquire	215
Aqua	181

Table 3.5: Value counts of body type feature

Transmission	**Value Counts**
Automatic	7490
Manual	235

Table 3.6: Value counts of transmission feature

The following are the value counts for each manufacture year:

Manufacture Year	**Value Counts**
2017	2232
2018	594
2014	377
2011	377
2012	352
2004	341
2010	323
2015	310
2013	294
2005	271

Table 3.7: Value counts of Manufacture year feature

3.1.3 Feature Engineering

We mainly used natural language processing in order to extract key features from the column that was already there in the dataset. Before utilizing NLP, the dataset that we were working with contained 15 features. 'description' have so much external information. We thought that we would be able to make use of that information to generate new features for our project. To begin, we did some basic cleaning, which included things like getting rid of duplicate values and converting bold text to regular text, etc. Following that, we began the NLP part. We made use of NLTK, which is a Python library for working with data relating to human language. In the NLP part, firstly, we tokenized the values of the columns 'CarName', 'trim' and 'description' into words. Tokenization is an important part of Natural Language

Processing (NLP) that can be used to find useful keywords in the description of a dataset of used cars. For that, we used “word_tokenize” from NLTK. Then we downloaded ‘stopwords’ and ‘punkt’ from NLTK. These were put to use in order to remove any stopwords and punctuation from the values. Stop words are words that are used a lot but don’t add much to what is being said. They can be taken out of the analysis to make it work better. ”The” ”and” and ”is” are all examples of stop words. Getting rid of punctuation marks in a text can make it easier to read and understand. So, it is a very important step in the process. After that, we used stemming. Stemming is a very interesting part of NLP. Stemming is the process of reducing a word to its root form. In most cases, it gets rid of the suffixes or endings that are attached to a word, such as -ing, -ed, -ly, etc. The process of stemming aims to reduce a word to its most fundamental form in order to make it simpler to examine and evaluate in relation to other words. For stemming, we used ”PorterStemmer” from ’nltk.stem’ Then, we used Lemmatization. The process of reducing a word to its most basic form while guaranteeing that the resulting word is a valid dictionary word is referred to as lemmatization. In order to determine fundamental structure of a word, it examines the surrounding information, such as the context and the part of speech of the word, as well as its vocabulary and grammar.For lemmatization, we used ’WordNetLemmatizer’ from ’nltk.stem’. We also downloaded ’wordnet’ from NLTK.After that we searched the most used word from all the values of ‘description’ and got a list of how many times a word was used.

According to the used word list, we were able to determine that ”color” was one of the keywords that were used in the ”description” column the most frequently. Therefore, we updated our dataset by including a new feature called ”Color”. Firstly, we replaced colour value with color. After that, we added the value of the color to those rows where we had received the color. For instance, we might have received the name of the color black in the 22nd row of the ’description’ column. Therefore, we put a value for black to the 22nd row of the column ”Color”. We also saw that there were some color values in the “CarName” and “trim” column. We also updated the row values of “Color” column where we got color value in those ‘CarName’ and ‘trim’ column. Based on the used word list, we were able to determine that the word ”ac” was the second most common word that was used in the ’description’ column. Therefore, we decided to make ”AC_system” the new feature of our dataset. First of all, we changed all instances of the words ”air” ”super” and ”cool” to ”ac” After that, we did the same thing for the “AC_system” column that we did for the “Color” column, adding the value ”Air Conditioned” to those rows where we had the word ”ac” in the column’s value. After that, we modified the value of ”AC_system” as ”Air Conditioned” for all of the vehicles, with the exception of the Toyota brand. We also updated the value of “AC_system” as “Air Conditioned” for all the hybrid trim and for those cars that were manufactured after 2014.Then we filled all the null row values in the ”AC_system” column with “Non_Air Conditioned”. According to the results of the used word list, ”accident” was the third most common word in the column named ”description.” As a result, we added a new column to the database and gave it the name ”Accident_history”. To begin, we changed accidental to accident throughout the ’description’ column. Then we searched for ‘accident’ in the description column. If we got any rows, we just updated the value of the “Accident_history” column with “No Accident”. Because the term ”accident” was originally used to denote the absence of an accident. However, word tokenization separated the ‘no’ from the word ‘accident’. After that, we filled the null value of “Accident_history” with “Unknown”. After that, we discovered that the term ”camera” was the fourth most common word in the ”description” column. As the new feature for our dataset, we decided to create a new column called ”Camera_facility”. In the first place, we changed the words ”rare” and ”back” to ”camera”. Then, if the word ”camera” appeared in any of the rows of the description column, we just updated the value of “Camera_facility” for those rows with ‘Rare Camera’. After that, all of the null values in the ’Camera facility’ are replaced with the value ”No Info.” The word ”light” was the last word that was used the most frequently in the ’description’ column, as evaluated by the used word list. Therefore, we expanded our dataset by including the ’Light_facility’ feature as a new column. Firstly, we looked over each row of the table and looked for the word ”light” in the column called ”description” If we found that, all that needed to be done was to replace the value of the ”Light facility” column in those rows with ”LED.” In addition to this, we got the words ”led” ”headlight” and ”backlight” which all refer to different types of light. Therefore, all we need to do is change the value of ”Light_facility” to either ”LED” ”LED Headlight” or ”LED Backlight” depending on whether or not we got the word ”led” ”headlight” or ”backlight” appropriately. Lastly, we simply replaced the value ’No LED’ for any null values that were present in the “Light_facility” After NLP, we had 20 features. However, it is important to note that this process is not a one-step solution, it is an iterative process that needs to be improved and fine-tuned as per the data availability, research question, and specific requirements of the research. The results obtained from NLP techniques may not be perfect, but they can serve as a good starting point for further analysis and can provide valuable insights that can be used to improve the efficiency and accuracy of the used car buying process for both buyers and sellers.

3.1.4 Data Encoding

In our dataset, there were several categorical features such as car name, model, manufacture year, version, condition, transmission, body type, fuel type, engine capacity, mileage, price, description, seller info, brand, color, owner membership, air conditioning system, accident history, camera facility and light facility. The data type of these features was object. Because the model cannot train object type values, we converted categorical feature values into numerical representations. We used the pandas get dummies() function to convert the category features object values to numerical values. It transforms a category variable into an indicator or dummy variable. It generates separate columns for all of the unique values of categorical features. Then drop the main categorical feature’s column. In our dataset, each brand, model, condition, transmission, body type, and fuel type consists of unique values. So, we used the pandas to get the dummies function to get the numerical representations. Pandas is a data processing and analysis software package created for the Python programming language. The get_dummies function creates separate columns for each unique value.

3.1.5 Feature Selection

There are only a few features in the dataset that are useful in creating the model for machine learning; the rest of the features could be more efficient. If we include all of these variables that are redundant and have no specific impact on the dataset, it will cause noise. It may have a negative impact on the overall performance and accuracy of the model. Therefore, before we can train the model on the data, we need to compute the influence of each of the features. According to the computation, we need to determine the attributes that influence the output or dependent feature we are interested in. Our dataset contains four continuous variables, three of which are independent and one dependent. To begin, we looked into whether or not there was a connection between the independent and dependent features. If there is a high level of correlation between two different features, it suggests that those features provide information similar to one another. As a result, having both is unnecessary and unneeded. On the other hand, a high correlation between independent and dependent features suggests that the independent feature has the most impact on the dependent feature. To visualize, we used a correlation matrix with a heatmap.

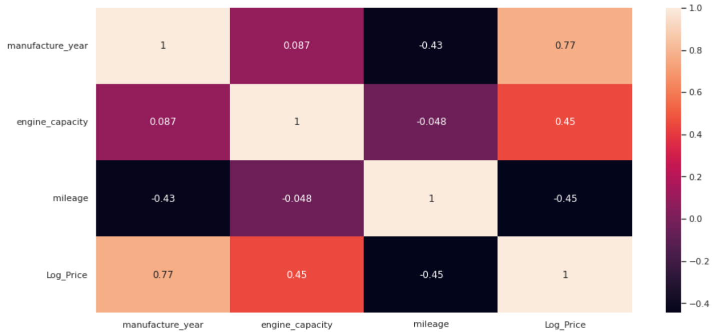


Figure 3.1: Feature-by-feature heat map

According to the heatmap, the manufacture_year feature has the highest correlation of 0.77. so manufacture_year has the highest impact on price. However, the correlation between the two independent features could be stronger. Mileage and engine_capacity have a -0.048 correlation, engine_capacity and manufacture_year have a 0.087 correlation, and mileage and manufacture_year have a -0.43 correlation. The correlation between independent features didn't exceed 80-90 percent in terms of correlation. As a result, we were unable to remove any features. Secondly, we can determine how important each feature in the dataset is using the model's feature importance property. Feature importance gives each piece of data a score. The higher the score, the more important or relevant the amount of data is to our dependent feature. Feature importance is a built-in class that comes with Tree-Based regression. We obtained the graph below after applying the Tree-based regression to our data.

Based on the graph, we discovered that manufacture_year was the most important feature in our dataset, with an impact greater than 0.35. Following that, engine capacity, condition_used and body_type_SUV have a good influence. The least important feature was the "Not MEMBER" value from "owner_info", among the top 6 important features that impact Machine learning models.

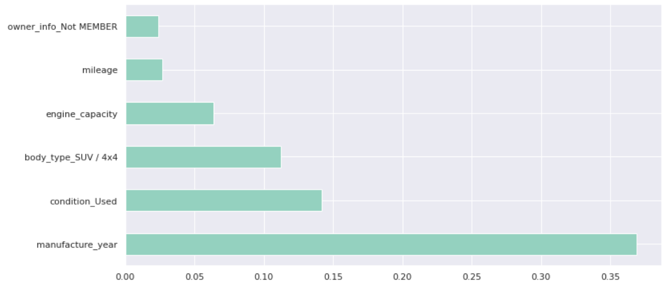


Figure 3.2: Feature importance

3.1.6 Data Splitting

In the field of machine learning, data splitting is a prominent procedure that helps prevent overfitting. At the outset of the machine learning process, the dataset must be split into a train set and a test set. By classifying the current data, we may assess how effectively a model responds to new information. Because train data and test data are kept in two separate locations, the model won't have access to the test data until after it has been trained. Because of this, we are able to conduct an accurate evaluation of the performance as we are testing them on the test set. Our data set was split at 80 percent to 20 percent. To further guarantee that our findings can be reliably replicated, we divided the dataset based on a fixed parameter value for the random state.

3.1.7 Feature Scaling

In our dataset, we used the Standard Scaler function to scale our dataset. This function was imported from the Scikit-learn library [21]. Many characteristics in our dataset that span a wide variety of possible values. Because of these variances, some attributes may emerge as more prominent than others throughout the training process for the data. It is crucial that our dataset be scaled up so that we can prevent being in this scenario. Here, we used the mean of feature values and variability in attribute values as measured by their standard deviation. Basically, in the process of standardization, each input variable is rescaled on its own by dividing by the standard deviation after subtracting the mean. This leads to a change in the distribution so that the mean eventually becomes zero and the standard deviation becomes one. The standardization formula is as follows.

$$x' = (x - \bar{x})/\sigma$$

The original feature vector is $\bar{x}$ is the mean of feature values and σ stands for standard deviation. The procedures of data scaling is given below:
On the training data, we used the fit procedure to get optimal values for the variables. Utilizing these newly acquired characteristics, we scaled up the original training set. The test set was also transformed into a scaled test set using the same learned parameters.

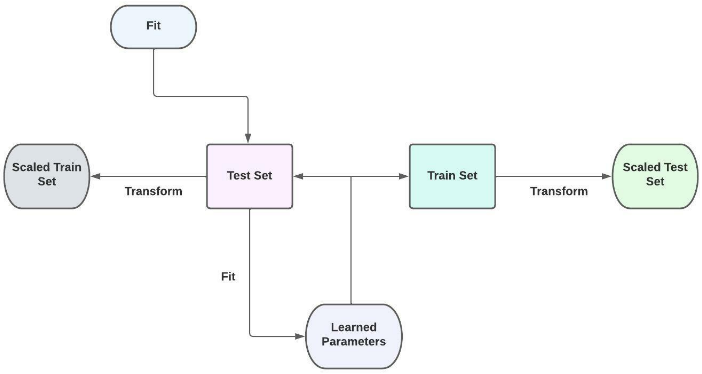


Figure 3.3: Data scaling procedures

3.2 Exploratory Data Analysis

When taking body type into consideration, the majority of people in Bangladesh drive saloons, followed by SUVs and then MPVs, with MPVs being used more frequently than SUVs. This is illustrated in the figure. On the other hand, hatchbacks and estates are not as prevalent as they once were, while sports cars and convertibles are practically extinct. The numbers that represent the values of their variables are rather low. (as may be seen in Figure 3.4).

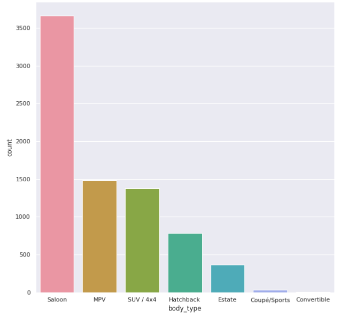


Figure 3.4: The amount of cars that exhibit a variety of body types

Based on the data analysis, it can be concluded that consumers purchase more automatic cars than manual transmission cars on the used automobile market. Consequently, automatic cars will be more prevalent than manual ones.

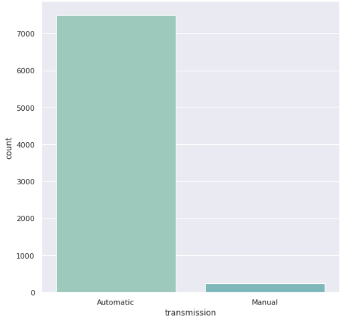


Figure 3.5: Cars transmission count

The characteristics of a box plot are provided below to facilitate an understanding of the relationship between average price and attributes: In this instance, the brand vs price box plot illustrates that luxury vehicles, such as Range Rovers, come in at an average price that is significantly higher than that of other cars. Aside from that, the typical prices of automobiles manufactured by Mercedes-Benz, Land Rover, and BMW are more than those of the more common branded vehicles. However, another significant finding emerges when we see that a brand like Toyota, which has the highest number of units sold, also has the greatest amount of variation.

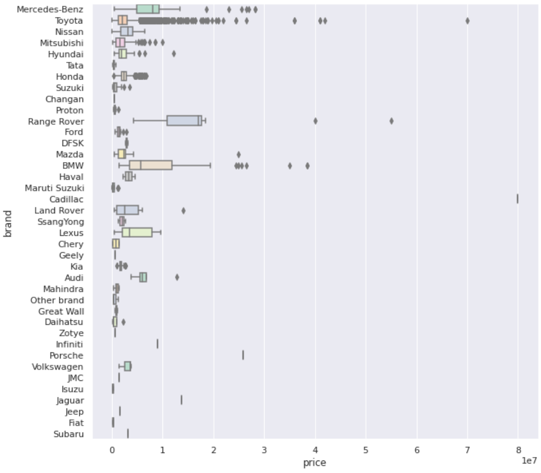


Figure 3.6: Box plot for brand vs price

When it comes to the various types of transmissions, the automatic transmission often has the highest average price. To put it simply, an automatic transmission is preferred over a manual one. When compared to the cost of a car with a manual transmission, the price of an automobile equipped with an automatic transmission is significantly more expensive.

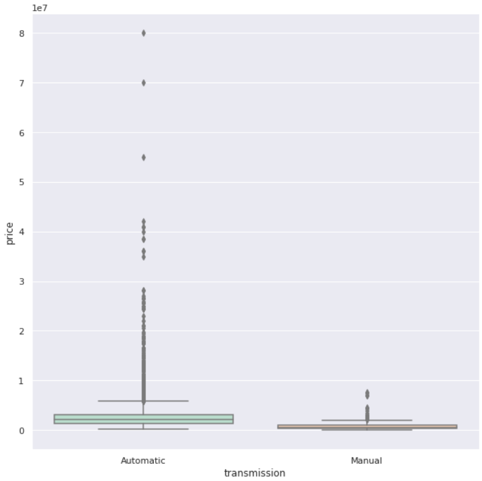


Figure 3.7: Transmission vs Price box plot

When the many types of bodies are considered, the average cost of an SUV is the greatest. The number of instances of convertible body types is significantly lower in this dataset in comparison to SUV, Saloon, MPV, and Estate body types. According to the preceding description of the box plot in terms of price, it is possible to conclude that exclusive automobiles in terms of body style and brand are less likely to be skewed. This conclusion is based on the fact that exclusive vehicles are more expensive. Saloon quantity price fluctuation is smaller compared to suv. The reason why body types like convertibles and coupes/sports cars have minimal variance is because there are so few of them.

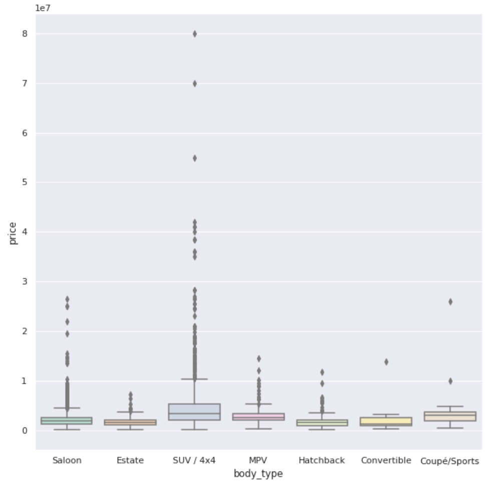


Figure 3.8: Body type vs Price box plot

We can also see that being a member of the bikroy.com website results in a variation in the price that is being asked for a car. Here, the sellers with no member badges have a somewhat lower average price than sellers with member tags, it can be said that sellers with member tags have a greater range of asking prices for vehicles. Aside from that, the median line is located in the box plot of non-members at a somewhat lower position.

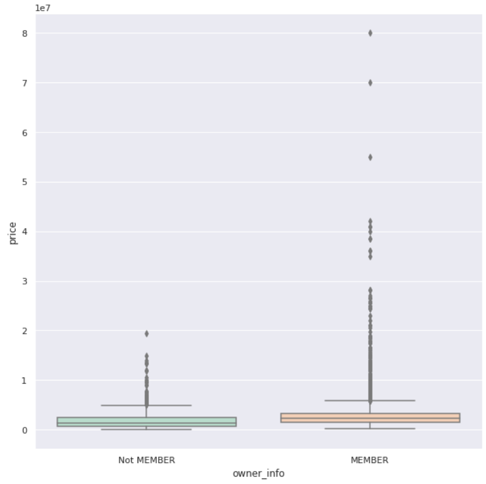


Figure 3.9: Owner info vs price box plot

It has been shown that the average price box plot for a vehicle that does not have air conditioning is somewhat cheaper than that of a vehicle that does have air conditioning . Based on the information we have, we can deduce that the number of cars with air conditioning is about three times more than the number of cars that do not have air conditioning. As a result, the average price of used cars without air conditioning is lower than the cost of cars that are installed with air conditioning.

Given this, it comes as a surprise that cars that do not have air conditioning sell for far more than the typical price of non-air-conditioned cars.

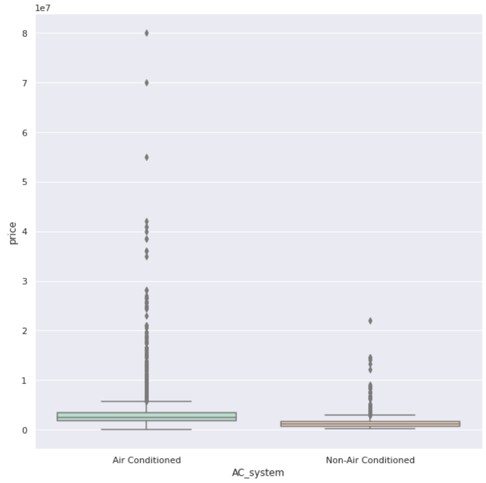


Figure 3.10: Ac system vs Price box plot

In light of the information presented thus far, it can be deduced that the density of AC cars increased between the years 2010 and 2020. On the other hand, automobiles that are in any condition have a value that is on average between the years 2000 and about 2010, which places them just below the air condition plot in terms of the year they were manufactured. Previous research has shown that there is a far smaller population of people driving vehicles without air conditioning. Because of this, the height of the visual depiction of the non-air condition vehicle box plot is greater than the height of the air condition car box plot. This explains why there is a significantly higher concentration of air condition autos in that particular box plot location. Surprisingly, there are still automobiles with air conditioning that were produced between the years 1980 and the early 2000s.

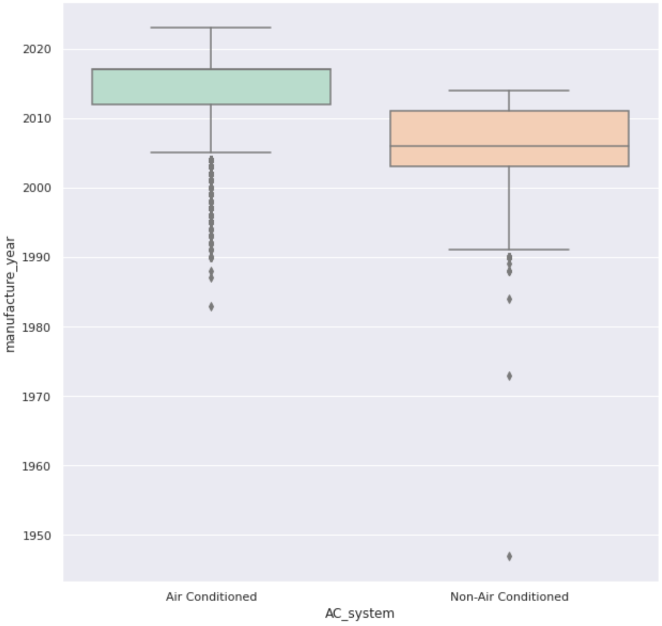


Figure 3.11: AC system vs Manufacture year box plot

Also, there are features like Camera facility and Light facility which is used to predict the cost of a second hand car, but these features have less correlation to the predictive models as for some data rows of the dataset have less information, LED headlights and LED backlights information was given as LED as the most frequency for the dataset. Also, for Camera facility, only Rare camera information was extracted for a few dataset and other includes no information as given in the Bikroy.com website. From the preprocessed dataset, we had a few encounters with outliers. Here, the mileage outlier doesn't have much of an effect on how the data is processed. However, mileage is dictated on the number of kilometers a car has traveled since it was made. Thereby mileage is a key factor in figuring the amount a vehicle will cost. Based on the mileage distribution diagram from the processed data, the mileage exceeds 400,000 km, which is highly implausible. As a result, we included outliers in the respective distribution for the elimination of outliers. The redundant numbers in this case are quantiles of 0.01 percent. The values that were incorrectly entered by users. We remove the outliers and get a new distribution after deleting the redundant values. From the new milage distribution graph, it can be observed that the upper bound comes down to 250000 km.

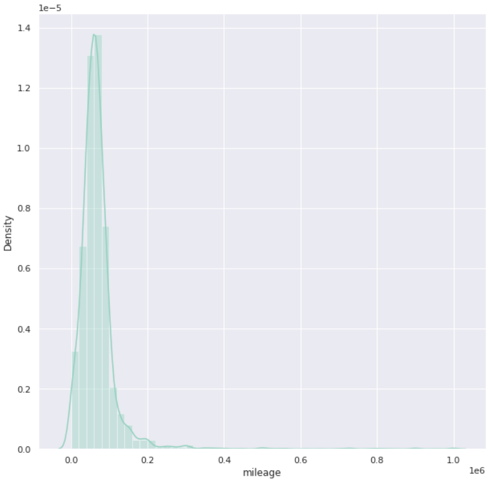


Figure 3.12: Mileage distribution plot before removing outlier

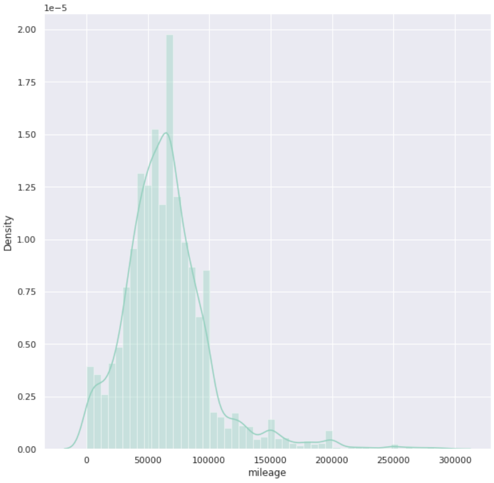


Figure 3.13: Mileage distribution plot after removing outlier

During the price scaling process, we dropped the actual car price and put in the log price instead. This is performed when the differences between the variables are on different scales. This implies that there is a large variance in terms of vehicle price in our dataset consisting of vehicle price with large and small magnitude.
In the figure below, the y-axis shows the price of the scattered diagram based on Engine, manufactur_year, and mileage. The top three plots have a price as the y-axis, while the bottom three plots have a log price as the y-axis. Here, we can observe that the price goes up as the year of manufacture goes up. In the same way, when the size of an engine goes up, so does the price. But when it comes to mileage, the price goes down as mileage goes up. For this justification, it can be stated that the value of a car depreciates as it travels farther or covers more area in its life cycle. In the lower half of the graph, where the log price is on the y-axis, the similar characteristics can be found. But doing that, it is also estimated that the bottom half of the figure is less skewed than the top half which dictates the lower portion of the distribution or scattered diagrams to be asymmetrical resulting in the plot being oriented towards a certain direction.

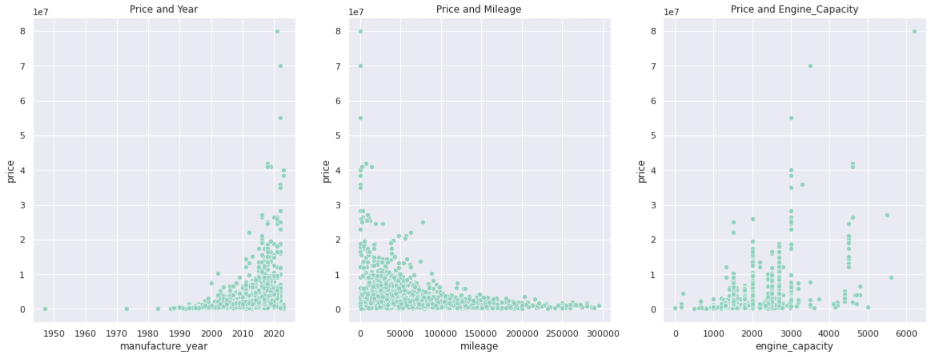


Figure 3.14: Figure of price vs manufacture year, mileage and engine capacity

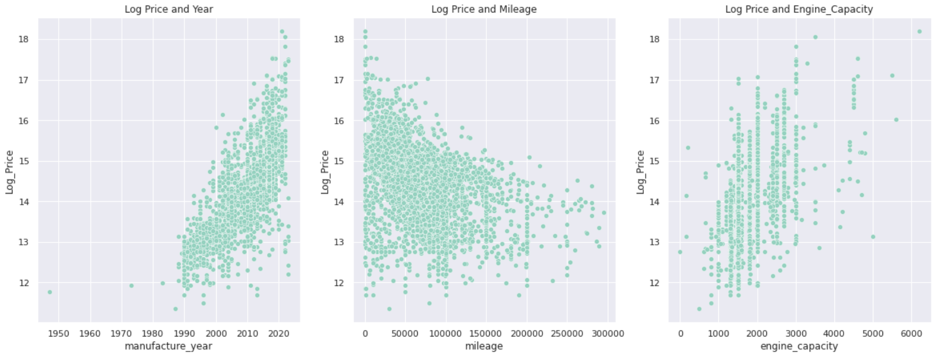


Figure 3.15: Figure of log price vs manufacture year, mileage and engine capacity

3.3 Models used

The models we used in our dataset are Random Forest Regression, Support vector Machine, and Neural Network model (Multi-layer Perceptron Classifier), XGBoost (Extreme Gradient Boosting). Here is the description of the models that we used in our dataset.

Random forest: Random Forest Regression is a form of ensemble learning method that makes predictions by using several decision trees. It is a change to the Random Forest algorithm, which is often used to sort things into different groups. By combining the results of several different decision trees, the Random Forest Regression model aims to boost overall accuracy and decrease prediction variation. In a Random Forest Regression model, the decision trees are made by taking a random subset of the data and a random subset of the features and putting them together. This means that each tree in the forest grows on its own and has its own splits and points where it has to make a choice. This variety of trees in the forest helps to prevent overfitting and improve the model's overall accuracy. The Random Forest Regression model averages the predictions made by all the decision trees in the forest to make a prediction. This method is called averaging or bagging, and it helps to make the predictions less different from each other. The final prediction is the average of all the predictions made by the decision trees in the forest. The Random Forest Regression model's ability to handle large datasets and high-dimensional data is one of its best features. The decision trees in the forest can automatically learn how to handle missing values and categorical data, eliminating the need for preprocessing. Additionally, Random Forest Regression is less sensitive to overfitting than traditional decision tree models. Another advantage of Random Forest Regression is its interpretability. The decision trees in the forest are easy to understand and interpret, as they are made up of simple if-then statements. This makes it easy to understand how the model is making its predictions, which is useful for understanding the relationships between the features and the target variable.

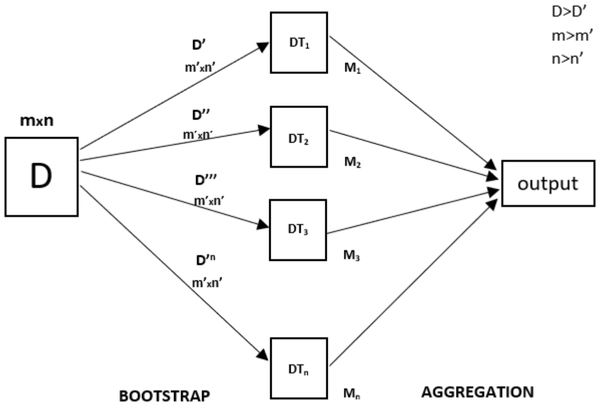


Figure 3.16: Random Forest regression model

Learning Algorithm	R2	MAE	MSE	RMSE
Random Forest	0.9202	0.1116	0.0412	0.2030

Table 3.8: Random forest predection details

Support Vector Machine: Support vector machines are a type of supervised learning technique that can be used to do a number of things, such as classifying data or finding analytical deviations. All of these are the standard ways to do things in the field of machine learning. We used Support vector machine to figure out how much used cars would cost (SVM). Support Vector Machine is a popular supervised learning algorithm that is used to solve classification and regression problems. This is a type of machine learning that builds on what we already know. Perceptions can't learn certain patterns on their own because they are set up in a vector space. Based on the nearest extreme point or the vectors on the n-axis plot, SVM makes the best n-dimensional line, called a hyperplane. This places the new data point in the appropriate section. SVM consists of The positive and negative hyperplane, representing the line nearest to the vectors on the top and bottom of the hyperplane. It also includes the Maximum margin, which is a generated straight line, and the maximum margin, the sum of the distance between the maximum margin hyperplane with the positive hyperplane and negative hyperplane SVM ensures that the data is separated by the largest margin possible. SVM can handle a very large dataset, works better with high-dimensional data, and does not encounter issues with under-fitting or over-fitting.

Learning Algorithm	R2	MAE	MSE	RMSE
Support Vector machine	0.8381	0.5305	0.4194	0.6476

Table 3.9: SVM predection details

MLP Neural Network: A multilayer perceptron (MLP) is an artificial neural network used to solve challenging complex computations such as stock analysis, image identification, spam detection, and predictive modeling in machine learning. This model's purpose is not to produce realistic brain models but rather to develop robust algorithms and data structures that can be used to visualize complex issues. Like the human brain, this model is made up of interconnected neurons that communicate information to one another. Each neuron has a value given to it. The model may be roughly broken down into three distinct levels. To begin, we have the input layer. It is at this layer that information is taken in order to generate the desired result. Then the second layer is the hidden layers. Basically, layers that are not directly connected to the environment are called hidden layers. The model must include at least one hidden layer that executes calculations and operations on the input information to generate something meaningful. An output layer is the last one. In this layer, neurons evaluate the data and form an opinion or prediction based on their findings. When using MLP, information flows from the input layer to the output layer. That's why, MLP is also known as FeedForward Neural Network. There are weights applied to the link between the layers .The neurons in the MLP are trained with the back propagation learning algorithm.This is basically a technique for optimizing an MLP's weights by using the outputs as inputs. Here, random weights are assigned to all the connections. These random weights distribute data throughout the network to generate the actual result. Obviously, this result would differ from what was anticipated. The discrepancy between the two numbers is called the mistake. Backpropagation sends this mistake back through the network, automatically changing the weights such that the discrepancy between the actual and predicted output is now less than it was before. This causes the output of the current iteration to become the input for the subsequent iteration and vice versa. This is repeated until the desired outcome is attained. The procedure is repeated until the error has reached the smallest possible value [22]. After implementing MLP on our dataset, we get this values as given on the table below. Here, accuracy score is best suited to show the prediction value for a classification mode.

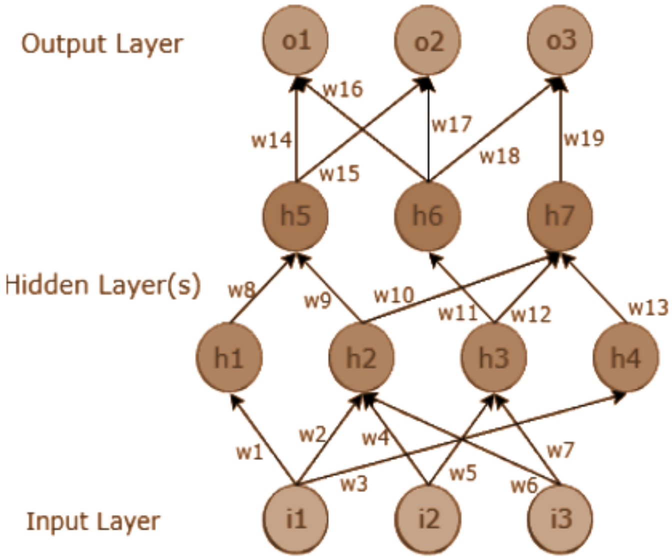


Figure 3.17: Multilayer perceptron model

Learning Algorithm	Accuracy _score	MAE	MSE	RMSE	R2
MLP Neural Network	0.8476	0.1667	0.2020	0.4495	0.6693

Table 3.10: MLP predection details

XGBoost: Extreme Gradient Boosting, or XGBoost, is a powerful machine learning library for gradient boosting that is used by a lot of people. It is an improved version of gradient boosting that is made to work with large datasets and achieve state-of-the-art performance on a wide range of machine learning tasks. In this paper, we will look at the main benefits and features of XGBoost, as well as its underlying algorithm and how it can be used in different areas. Gradient boosting is a powerful ensemble learning method that uses weak models, like decision trees, to make a strong model. Gradient boosting works by adding new models to the ensemble one at a time, with each new model being trained to fix the mistakes of the old models. The final ensemble model is a combination of all the weak models, and it is expected to do better than any of the individual models. XGBoost is an improved version of gradient boosting that is designed to work with large datasets and achieve state-of-the-art performance on a wide range of machine-learning tasks. One of the best things about XGBoost is that it can deal with missing values and variables that have more than one type. It can handle missing values by filling them in with the mean or median of the column, and it can handle categorical variables by making a binary indicator variable for each category. This lets XGBoost work with many different kinds of data and deal with datasets with missing or categorical variables. Another important thing about XGBoost is that it can deal with overfitting. Overfitting is a common problem in machine learning. This is when a model is trained to fit the training data too closely and does poorly with new data. XGBoost has a number of ways to stop overfitting, such as regularization and stopping early. Regularization is a technique that adds a penalty term to the loss function to discourage complex models, and early stopping is a technique that stops the training process when

the model's performance starts to get worse on the validation set. XGBoost also supports parallel processing and distributed computing, which makes it possible to train big models on big datasets. Because of this, XGBoost is a great tool for working with big data. It is used a lot in natural language processing, computer vision, and recommendation systems, among other things.
In conclusion, XGBoost is a powerful machine learning library for gradient boosting that is used by many people. It is an improved version of gradient boosting that is designed to work with large datasets and get state-of-the-art results on a wide range of machine learning tasks. It stands out from other libraries because it can deal with things like missing values, categorical variables, overfitting, parallel processing, and distributed computing. Because of how well-known it is and how many ways it can be used, many professionals and researchers use it as their go-to library [23].

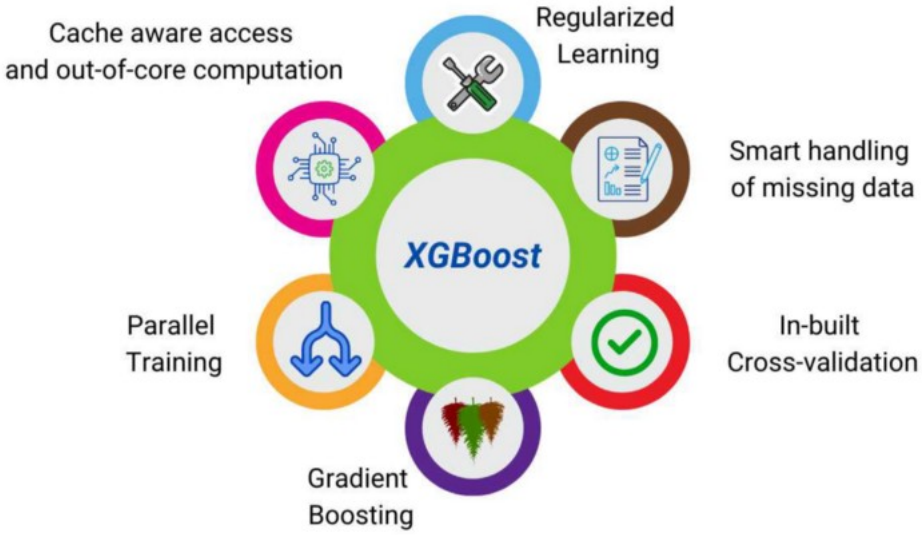


Figure 3.18: XGBoost model

Learning Algorithm	R2	MAE	MSE	RMSE
XGBoost	0.9136	0.1458	0.0497	0.2230

Table 3.11: XGBoost predection details

KeyBERT (NLP): KeyBert is a pre-trained transformer model used for natural language processing (NLP) tasks such as keyphrase extraction, text summarization, and text classification. It is based on the BERT (Bidirectional Encoder Representations from Transformers) architecture and has been trained on a large corpus of text data. KeyBert is designed to identify keyphrases in text, which are phrases that summarize the main topics or ideas in a piece of text. It uses a combination of unsupervised and supervised learning techniques to extract keyphrases from text, and can be fine-tuned on specific datasets to improve performance for specific tasks or domains. The KeyBert model can be fine-tuned for a wide range of NLP tasks, such as text classification, text summarization, and sentiment analysis. KeyBert is trained with a large corpus of text data, which allows it to understand the context of the input text and make predictions based on that context [19]. This allows it to perform well on a wide range of NLP tasks, and it can be fine-tuned on specific datasets to improve performance for specific tasks or domains. Here we use key bert to extract the most important keywords from description.

3.4 Model Evaluation Parameters

Accuracy Score: Accuracy is a frequently used evaluation parameter for classification-based models. It shows how many of the model's predictions turned out to be right out of the overall amount of predictions. We can get a percentage value for the accuracy score by dividing the number of correct predictions by the total number of predictions and then multiplying by 100. But when it comes to models based on regression, accuracy is not always the best way to judge them. This is because regression models are made to predict continuous values, not categorical labels. Most of the time, other metrics like mean absolute error (MAE), mean squared error (MSE), and R-squared are used to decide how well regression models work.
R-squared: R-squared, also called the coefficient of determination, is a mathematical metric that indicates how much of the variation in the dependent variable can be clarified by the independent variable(s) inside a regression model. It can be between 0 and 1, where 1 means that the model explains all of the differences in the dependent variable. R-squared is found by dividing the total amount of change in the dependent variable by the amount of change that can be explained by the independent variables in the model. In other words, it is the ratio of the variation that can be explained to the variation that can be explained. A value of 1 for R-squared means that the independent variables in the model explain all of the changes in the dependent variable. A value of 0 means that none of the changes in the dependent variable can be explained by the independent variable(s). R-squared is a common way to measure how well a regression model works, but it also has some problems. A high R-squared value doesn't always mean that the model fits the data well. It can also mean that the model is too good at fitting the data, or that the independent variable doesn't help explain the dependent variable. In summary, R-squared is a measure of how well the independent variable(s) in a regression model explain the variation in the dependent variable. It ranges from 0 to 1, and a high value of 1 means that the model fits the data well, but it should be used with other metrics and visualizations to evaluate the model's performance.

$$r^2 = 1 - RSS/TSS$$

Here, RSS is the sum of the squared differences between what was expected and what occurred. TSS is the sum of the squared differences between the actual values and the average of the actual values.

MSE: MSE stands for Mean Square Error, which is a measurement of the average square deviation among values that were predicted and the values that were actually observed. In the fields of regression analysis and machine learning, it is frequently employed as a loss function or a performance indicator for the purpose of determining how well a model represents the data. To determine the MSE, just take the average of the squared difference that exists among the values that were to

be predicted and those that were actually observed for each data point. The MSE value should be as low as possible for the model to have the best performance. The MSE is determined by the following formula:
The mean square error, or MSE, may be calculated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where the observed values are represented by Yi, Yi is the predicted values, and number of observations represented by n.

RMSE: Root Mean Square Error (RMSE) is a measurement that may be used to determine the average disparity between the values that were predicted and the values that were actually observed. It is commonly used in the fields of regression analysis and machine learning as a loss function or a performance metric for the purpose of determining how accurate a model is. It is computed by taking the square root of the mean square error (MSE), which dictates the accuracy of . The RMSE number should be as low as possible because it indicates how well the model is doing. The following is the formula for calculating RMSE: RMSE is calculated by taking the square root of MSE, which is equal to the square root of (1/n) times the difference between the actual value and the predicted value. The number of observations is denoted by n. Note that RMSE has the same unit as the original data, therefore it enables us to grasp the amount of the mistake in the same unit as the original data, but MSE does not have any dimensions associated with it.

MAE: Mean Absolute Error, also known as MAE, is a metric that is used to assess the average difference that exists between the values that were predicted and the values that were actually obtained. To determine it, just take the average of the absolute difference that exists between the values that were predicted and those that were actually observed for each data point. It is a performance statistic that is frequently used in the field of regression analysis and machine learning for the purpose of measuring the degree to which a model is accurate. The MAE number should be as low as possible for the model to have the best performance. The MAE is determined by the following formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|$$

MAE is calculated as (1/n) times the difference between the actual value, Xi, and the predicted value, Yi and n is the number of observations.

3.5 Data fair usage policy

Research is the process of looking into a fact or a problem to learn more about it or come up with a good solution. Legal research is the process of finding the law about a specific issue in an organized way. It thoroughly searches all the available legal resources, such as laws, new laws, and court decisions. If we want to learn more about law, we need to look into the reasons or ideas behind the laws already

in place. The way these things are done should be planned ahead of time. Before the data can be licensed, the person who owns it and wants to use it must agree on something in a legal setting. The license says how the data can be used and under what conditions. It also says if there are any limits or restrictions on how the data can be used. This can include information like how long the data can be used, what applications it can be used for, and if there are any fees. There are a few different ways to get permission to use data, such as:
1. Public domain: The term "public domain" describes the freely available information to the public and can be used by anyone, for any reason, without permission or payment.
2. Creative Commons: Creative Commons is a license that lets you use data as long as you follow certain rules. For example, you might have to give credit to the source or use the data for something other than making money.
3. Open Data: Open data is made available for public use with minimal or no restriction for any purpose, commercial or non-commercial.
4. Open Source: Under an open-source license, anyone who wants to use the software can get access to and change its source code.
5. Proprietary: Privately owned data can only be used with the owner's permission and often requires payment.
6. Royalty-free: Royalty-free licenses are a type of license where the user pays a one-time fee to use the data, but they don't have to pay extra royalties when they use the data again in the future.
7. Rights-managed: Rights-managed licenses are a type of license in which the user must pay a fee for each use of the data, and the data owner is responsible for carefully defining and controlling the license terms.
The Creative Commons (CC) organization is a non-profit that gives authors different licenses they can use to make their work available to everyone. Here are the different kinds of licenses that Creative Commons offers:
1. CC-BY: To follow the rules of the CC-BY license, credit must be given. People can use and share the work, but they must give credit to the original author.
2. CC-BY-SA: "Attribution Share-Alike" is what "CC-BY-SA" is short for. People can use and share the work, but they must give credit to the original creator. Any changes or new works made using the original work must also be released under the same license as the original work.
3. CC-BY-ND: The CC-BY prefix is used for the license type "Attribution-NoDerivatives." People can use and share the work, but the original file cannot be changed, and the person who made it must be acknowledged for their contribution.
4. CC-BY-NC: The license type is Attribution-NonCommercial, often written as CC-BY-NC. Allows other people to use and share the work for non-commercial purposes, but the original author must be credited.
5. CC-BY-NC-SA: "Attribution-NonCommercial-ShareAlike" is what "CC-BY-NC-SA" stands for. People can use and share the work for non-commercial reasons, but they have to give credit to the original creator and share any changes or new works they make based on the original work under the same license.
6. CC-BY-NC-ND: Under the CC-BY-NC-ND license, the type of license for shared works is Attribution-NonCommercial-NoDerivatives. It lets other people use and share the work as long as they don't do it for profit. However, they can't change the original work and must give credit to the original creator.

Everyone knows that the used car market in Bangladesh is less well-known than markets in other countries. Because of this, we needed help and assistance gathering information about the cars. After looking into the topic, we found that the website bikroy.com is a great place to get the information because it has a lot of information about cars. So, we looked at the website bikroy.com for information. For our investigation to go further, we asked the company if we could use their data. After that, we got their permission, and they permitted us to use that data only for research. Also, they approved our data as legalized data , which means that it is accurate information that can be used for our research according to the rules of the Creative Commons Attribution-NonCommercial-NoDerivatives license (CC BY-NC-ND). The Creative Commons Attribution-NonCommercial-NoDerivatives (CC BY-NC-ND) license is a type of copyright license that lets the user reuse and distribute the copyrighted work of someone else as long as they give credit to the original creator and do not use the work for commercial purposes. The Creative Commons group developed the Creative Commons Attribution-NonCommercial-NoDerivatives (CC BY-NC-ND) license. The license also says that the user can't change the work in any way or do new work based on it. The ND clause of the license says that the user is not allowed to change or modify the original work in any way. This includes changing the original work by taking out or adding parts[16]. This license type is usually used for works whose creators want people to reuse and utilize their work but doesn't want it to be changed or sold in any way. Let's say we want to use someone else's work in our thesis paper, and we want to do it under a Creative Commons license that says we have to give credit, we can't use it for profit, and we can't change it in any way. If this is the case, we will need to give credit to the original author and not use the work for commercial purposes. We will also have to ensure that we don't use the work in a way that could be considered plagiarism. This will be something we'll have to take care of. In addition, we will have to make sure that the work has stayed the same. Even though the CC BY-NC-ND license lets us use and share someone else's work, we need to know that this license does not give us ownership of the work. This fact must always be kept in mind. Even though we have permission to use the work, the original creator still owns it and can take back permission at any time. On the Creative Commons website, we can fully explain the CC BY-NC-ND license and how it works. To get the license, we must meet the following preliminary requirements:

1. Attribution: We have to give credit to the person who did the work first, include a link to the license, and say if the work has been changed. Attribution.

2. NonCommercial: We can not use the work in any way that makes us money or helps your business.

3. NoDerivatives: We can not change the work in any way, and we can not use it to do any other work.

Anyone or any group that uses this dataset must give credit where credit is due, link to the license, and say whether or not changes were made. The data that bikroy.com permits us to use can only be used for research. We do not want the other user to change or alter these data in any way. We also have no plans to sell it or use it in any way that would make money. We would like it to be illegal to use this information for anything other than research. That's why our thesis will be licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives license, written as CC BY-NC-ND [20] [17].

Chapter 4

Results and Discussion

4.1 Comparison and analysis of the results

As we used Random Forest, Support vector regression, Neural Network (MLP Classifier), and XGBoost models for our dataset. For the evaluation process, Random forest predicted 92.02% accuracy, Neural network (MLPClassifier) prediction was 84.76% Support vector machine predicted the accuracy of 83.81% and for XGBoost, the prediction was 90.36%. After completing the training, the accuracy of the models was predicted on the test dataset. The predicted values can differ slightly within that range. Here, the Random forest model gives the best-predicted outcome in comparison. We also measured the RMSE, and MSE of those models.
The RMSE, or root mean squared error, is found by taking the square root of the mean of the square of error. Also, the mean absolute error (MSE) of the model in relation to the test set can be found by taking the average absolute value of all the individual prediction errors in the test set. Taking the average of the absolute error values gives you the MAE score. When doing regression and not wanting outliers to have a big effect, we usually use MAE. R squared is a noise measure that shows how well the line fits the data. When the value is low, it means that the values are far from the regression line or prediction. A high R-squared value means that all of the data points are close to the regression line.

Models	**Predicted Price**	**Accuracy_score (%)**
Random Forest	.9202	92.02
Neural Network (MLP Classifier)	.8476	84.76
Support Vector Machine	.8381	83.81
XGBoost	.9036	90.36

Table 4.1: Accuracy percentage of the models

The table above shows the accuracy percentage of the models, where we used accuracy_score that was imported from "sklearn.metrics", which is a reliable criterion for evaluating classification and regression model fit. However, R2 cannot be relied upon as a reliable metric for assessing the fit of a classification model. Because of this, the R-squared score for determining accuracy in SVM and neural network (MLP) models is lower. It works well with the Random Forest model. When dealing with continuous variables, the R2 coefficient is the best predictor to use. It is believed to be the percentage of the variance of the dependent variables that is effectively reproduced by the model, and when dealing with continuous dependent variables, it may range from 0 to 1 and has a possible range of values. If the value of R2 in the model is 1, then it is able to reconstruct the dependent variable accurately, but if it is 0, then the model is completely incapable of doing so. From the table below we can see, the predicted value of SVM and Neural Network (MLP) are lower than of the values that are given in Figure 4.1

Learning Algorithm	**R2**	**MAE**	**MSE**	**RMSE (%)**
Random Forest	0.9202	0.1116	0.0412	0.2030
Neural Network (MLP Classifier)	0.6693	0.1667	0.2020	0.4495
Support Vector Machine	0.8381	0.5305	0.4194	0.6476
XGBoost	0.9036	0.1458	0.0497	0.2230

Table 4.2: Check car price prediction with RF model

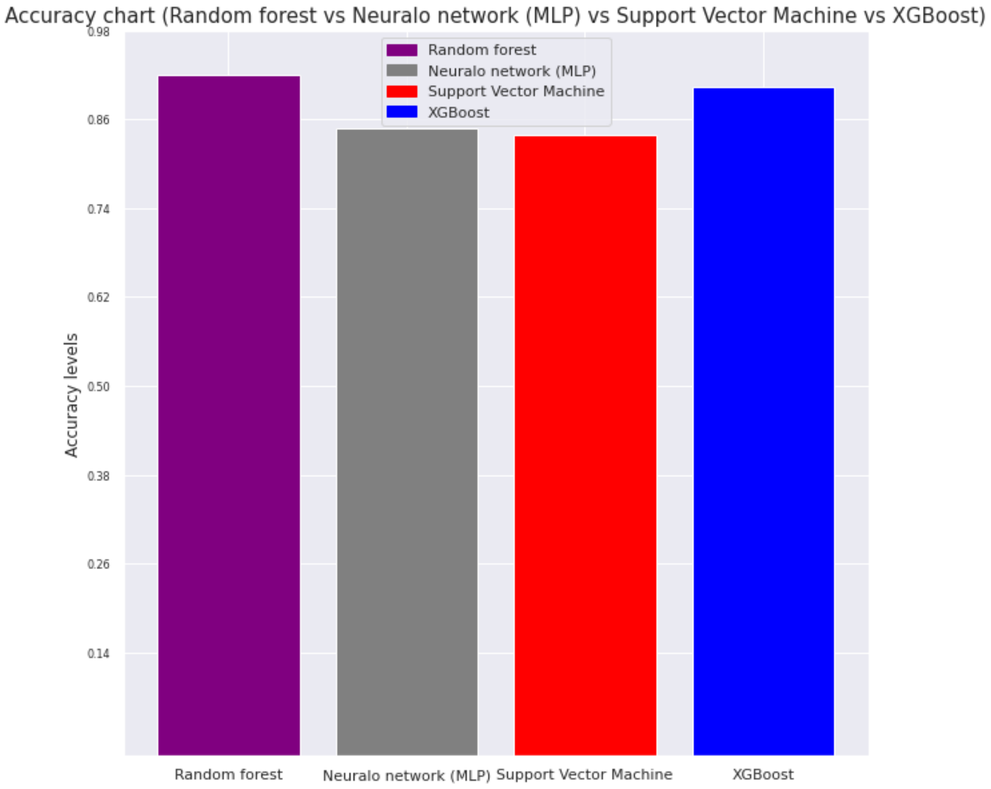


Figure 4.1: Comparison between the models

Then additionally, we went for Voting regression, which is an ensemble method that combines the predictions of multiple individual models in order to improve the overall accuracy and robustness of the final prediction. In the context of used car price prediction, voting regression could be used to combine the predictions of three different models: Random Forest, Support Vector Regression (SVR), and XGBoost. As it's illustrated, Random Forest regression is a popular ensemble learning method that combines multiple decision trees. It is known for its ability to handle high-dimensional data and is often used for regression tasks. Support Vector Regression (SVR) is a type of support vector machine that is commonly used for regression tasks. It finds the best hyperplane that maximizes the margin between the target

and the predictions. XGBoost (eXtreme Gradient Boosting) is a powerful and efficient gradient boosting library that is known for its ability to handle large datasets and improve model performance. The predictions of these three models would be combined using a voting system, where the final prediction is the average of the predictions of the individual models. And the predicted score of the voting regression is 91.35%. But we get better prediction value in case of Random Forest regression model.
The manual check for these predictions is given below in the table. Here, we calculate the difference between the targets and the predictions. We choose the Random forest model for that as it gives better accuracy. Since the Random forest model provides better accuracy, we choose to use it.

Car Index	Predicted Price	Actual Price	Residual	Difference (%)
1519	1323111.41	1320000.00	-3111.41	0.24
1521	4625676.02	4700000.00	74323.98	1.58
1522	2180567.25	2330000.00	149432.75	6.41
1524	5169987.11	5000000.00	-169987.11	3.40
1525	2224016.64	2160000.00	-64016.64	2.96
1526	1269708.78	1280000.00	10291.22	0.80
1527	1457022.38	1560000.00	102977.62	6.60
1528	1797090.61	1950000.00	152909.39	7.84

Table 4.3: Check car price prediction with RF model

A scatter plot graph of Random Forest regression based model, the actual price vs the predicted price given below:

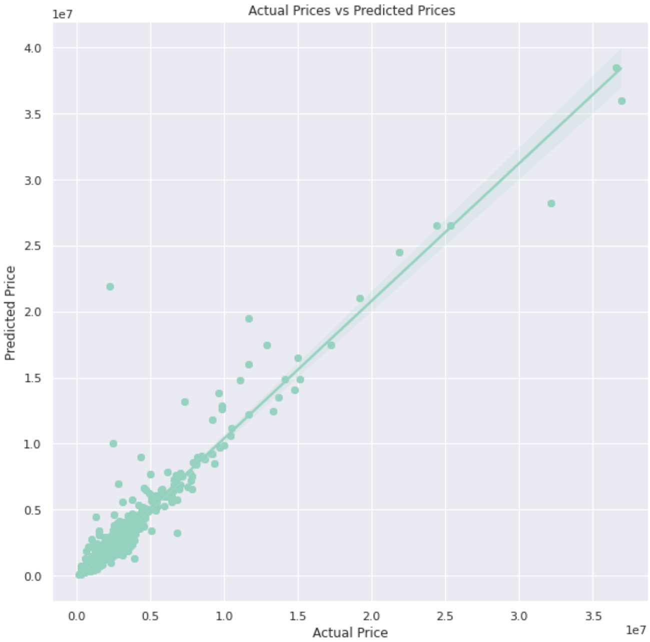


Figure 4.2: Actual price vs Predicted price

In a thesis paper on car price prediction, the real price of a vehicle would be the genuine market worth of the vehicle, as established by some characteristics. The predicted price, on the other hand, is the value that the model employed in the thesis paper judges the automobile to be worth based on input data and the technique used for prediction. The purpose of the thesis would most likely be to evaluate the model's accuracy by comparing anticipated prices to actual prices.

Chapter 5

Real-Time Implementation

In order to put this research into action in the real world, we have developed a website. On the website, a seller can find out what the current price of his or her vehicle is, and a buyer can discover the price of the second-hand car that they are interested in purchasing. At this very moment, the beta version of our website is live. In the future, people will be able to obtain information about the name of the shop from which they can purchase the item. For the front end of the website, we used ReactJS, and for the back end we used Flask. In the back end, we also used our model to make predictions about the pricing of cars. Firstly, we dumped our model into a file so that it could be used in the backend of the system. The random forest model was eliminated from the evaluation. We made use of joblib's ”dump” command. Simply stated, we made use of joblib in order to store the model in a file on the local disk. Then, when we got to the backend, the first thing that we did was load our model file into the backend. In order to accomplish this, we relied on the load function found in joblib. After that, we developed the Flask application. Additionally, we used the ”CORS” protocol for cross-origin. Within the Flask application, we created a route referred to as ”/predict.” Both ”GET” and ”POST” requests can be sent through this route. By using the ”POST” request, the front end can send information to the back end. After that, we modified the input data. simply because our model requires a very certain kind of input. Basically, all that needs to be done is convert some values into dummy types. Our model predicts the price of the car based on the input. The price response is then sent to the front end. JSON is the primary format for the output, which is sent. After that, we hosted the backend on the host server. In the end, we built our front end using the ReactJS library. For the styling, we used Tailwind CSS, and Framer Motion was used for the animation. This is the home page of our website.

If we click on the button called ”used car price,” it will take us to the page where we can enter information about the car. This is the page where you enter your information.

If we enter the necessary information and then click the ”go” button, the entire set of data will be transmitted to the backend API. Then, after a price has been predicted, our model will transmit that price as a response to the front end. The price will be displayed on a separate page that will open shortly after the prediction. The page that was predicted appears like this:

We are confident that visitors to our website will be able to obtain reliable information regarding the cost of a vehicle.

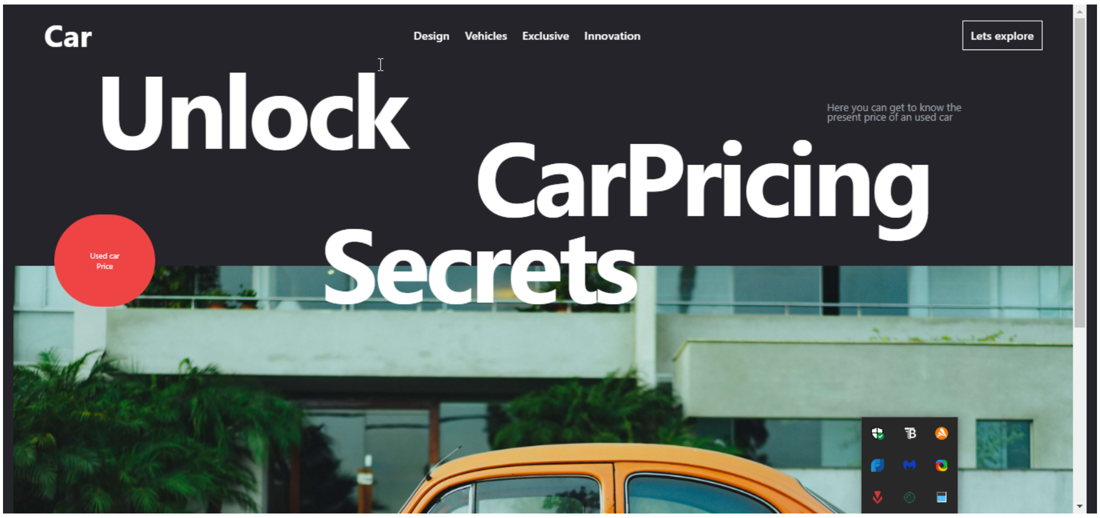
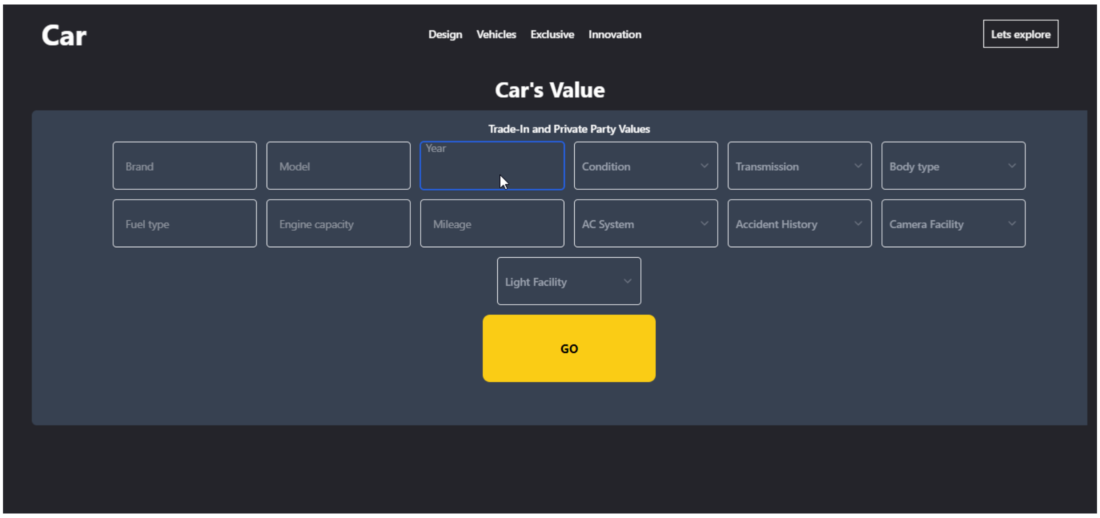


Figure 5.1: Home Page



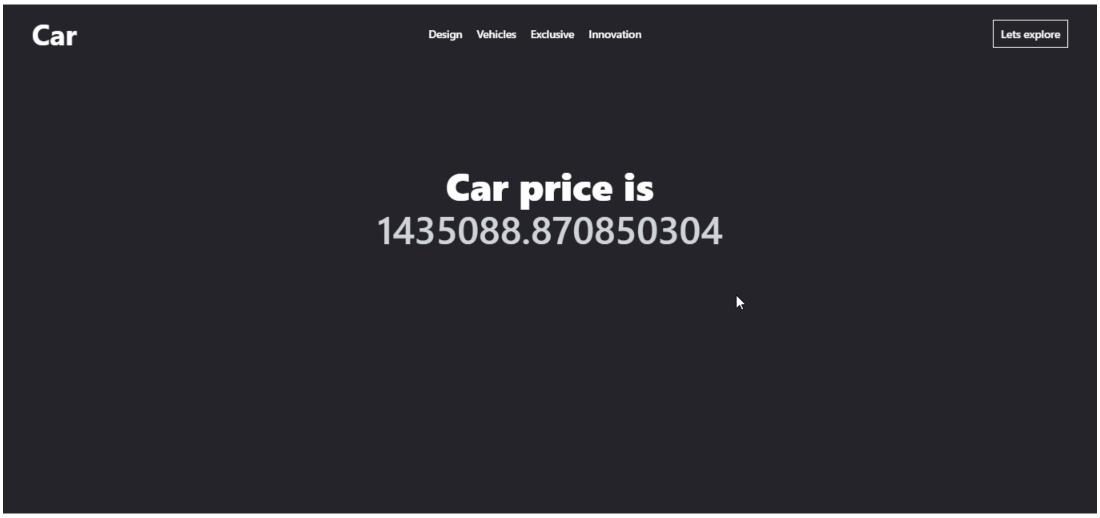
Car's Value

Trade-In and Private Party Values

Brand
Model
Year
Condition
Transmission
Body type
Fuel type
Engine capacity
Mileage
AC System
Accident History
Camera Facility
Light Facility

GO

Figure 5.2: Car Input Page



Car price is
1435088.870850304

Figure 5.3: Predict Car price page

Chapter 6

Conclusion

6.1 Summary of the research

Demystifying the price of used cars, which is applicable to Bangladesh, can be pretty challenging as there are multiple attributes and characteristics of a used car, and each of them should be considered for a more accurate price. So, for research reasons, we gathered data on the following: car name, model, manufacture year, trim, condition, transmission, body type, fuel type, engine capacity, mileage, price, description, seller info, brand, color, owner membership, air conditioning system, accident history, camera facility and light facility. The necessary data which will be used in our paper are collected from Bikroy.com. To analyze our obtained data for determining the price, we used a variety of machine learning algorithms. We wish to come up with a solution that is appropriate for our research. We believe that by integrating machine learning approaches, we may combat deception and efficiently deal with relevant problems.

6.2 Limitations and challenges

During the research, there were a lot of problems that we faced in order to conduct the research. Some of the problems that were seen were caused by the fact that the bikroy.com websites didn't have the required information for the dataset. Not having an insufficient registration year was one of the important concerns that were raised. We had to get rid of some of the rows containing unknown registration years. When that was done, another problem arose, and it caused the dequeued rows to lose the description that was in the dataset. During the investigation, unique features were taken from the description of the research. In terms of data extraction from the description, we prioritized the use of the key-bert natural language processing model for tokenization and obtaining the most important keywords from the process. However, by doing so, we were able to obtain useful keywords, many of which lacked the essential keyword that was required for our research, which led to an increased number of missing values. Moving forward, it was determined that dropping rows with missing values allowed for the detection of more than 2500 colors, which was insufficient. and because of this reason, our next approach was to identify car colors through the application of image processing. The objective was to identify the vehicle while simultaneously retrieving the color name of the pixel with the highest degree of involvement. Another difficulty was that some of the automobile photographs had been taken in direct sunlight, which caused color shifts that ultimately led to inaccurate color representations. Because the color that we fetched from image processing did not correspond with the color that was described by the seller, we decided to pause the portion of the research that dealt with image processing. Aside from that, there were inconsistencies of keywords all over the data frame, which led to more unique values.

6.3 Implications and future work

The implications of writing a thesis paper on the topic of car price prediction could include the ability to predict the pricing of cars in the future with a higher degree of accuracy. This would be advantageous for car manufacturers, dealerships, and individual buyers. This might result in pricing methods that are more effective, which might in turn lead to better offers for consumers. In addition, investors in the automotive industry could benefit from using a price prediction model for automobiles in order to make more educated judgments regarding the purchase or sale of stock in automobile manufacturing businesses. On the other hand, it is essential to keep in mind that the precision and dependability of any prediction model will be reliant on the quality of the data that was utilized to train the model as well as other methodologies that were utilized in our study. In the course of our future work, one of our goals is to create software that can automatically extract features from photographs of vehicles and estimate their prices on behalf of the model. Based on the findings of the research, one can draw the conclusion that the year of manufacture carries a significant amount of weight. However, information such as the manufacturing year cannot be determined solely from an image. Because of this, one of our goals is to determine whether or not there is a correlation between the manufacturing year and the number plate of the car in order to derive a price estimate for the car. Apart from that, our software will detect dent, scratch, crack in the windshield and any missing components of the car via image processing, which will provide users with the predicted car price. Therefore, it is necessary to develop a more complex model in order to consider the impact changes occurring over the time. These changes may include changes in market conditions, rise in the cost of fuel, and technological advancements in vehicles.

Bibliography

[1] M. Listiani *et al.*, “Support vector regression analysis for price prediction in a car leasing application,” *Unpublished. https://www. ifis. uni-luebeck. de/˜ moeller/publist-sts-pw-andm/source/papers/2009/list09. pdf*, 2009.

[2] M. S. Richardson, “Determinants of used car resale value,” Ph.D. dissertation, Colorado College., 2009.

[3] J.-D. Wu, C.-C. Hsu, and H.-C. Chen, “An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference,” *Expert Systems with Applications*, vol. 36, no. 4, pp. 7809–7817, 2009.

[4] S. Gongqi, W. Yansong, and Z. Qiang, “New model for residual value prediction of the used car based on bp neural network and nonlinear curve fit,” vol. 2, pp. 682–685, 2011.

[5] S. Pudaruth, “Predicting the price of used cars using machine learning techniques,” *Int. J. Inf. Comput. Technol*, vol. 4, no. 7, pp. 753–764, 2014.

[6] K. Noor and S. Jan, “Vehicle price prediction system using machine learning techniques,” *International Journal of Computer Applications*, vol. 167, no. 9, pp. 27–31, 2017.

[7] S. M. Kamal and N. A. Ahsan, “Uber-pathao’ride-share’s impact on dhaka,” *The Financial Express.[online] Available at: https://thefinancialexpress. com. bd/views/uber-pathao-ride-sharesimpact-on-dhaka-1524842540 [Accessed 26 Oct. 2019]*, 2018.

[8] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya, and P. Boonpou, “Prediction of prices for used car by using regression models,” in *2018 5th International Conference on Business and Industrial Research (ICBIR)*, IEEE, 2018, pp. 115–119.

[9] N. Pal, P. Arora, P. Kohli, D. Sundararaman, and S. S. Palakurthy, “How much is my car worth? a methodology for predicting used cars’ prices using random forest,” in *Future of Information and Communication Conference*, Springer, 2018, pp. 413–422.

[10] E. Gegic, B. Isakovic, D. Keco, Z. Masetic, and J. Kevric, “Car price prediction using machine learning techniques,” *TEM Journal*, vol. 8, no. 1, p. 113, 2019.

[11] P. Venkatasubbu and M. Ganesh, “Used cars price prediction using supervised learning techniques,” *Int. J. Eng. Adv. Technol.(IJEAT)*, vol. 9, no. 1S3, 2019.

[12] K. Samruddhi and R. A. Kumar, “Used car price prediction using k-nearest neighbor based model,” *Int. J. Innov. Res. Appl. Sci. Eng.(IJIRASE)*, vol. 4, pp. 629–632, 2020.

[13] P. Gajera, A. Gondaliya, and J. Kavathiya, "Old car price prediction with machine learning," *Int. Res. J. Mod. Eng. Technol. Sci*, vol. 3, pp. 284–290, 2021.

[14] *Bangladesh automobile industry: Current trends and future - business inspection bd*, https://businessinspection.com.bd/automobile-industry-of-bangladesh/, (Accessed on 09/16/2022).

[15] J. Chakma, *Realising the potential of bangladesh automotive industry — the daily star*, https://www.thedailystar.net/supplements/four-wheeler-special/news/realising-the-potential-bangladesh-automotive-industry-2904441, (Accessed on 09/16/2022).

[16] *Creative commons — attribution-noncommercial-noderivs 2.0 generic — cc by-nc-nd 2.0*, https://creativecommons.org/licenses/by-nc-nd/2.0/, (Accessed on 01/17/2023).

[17] *Creative commons licenses*, https://www.openaccess.nl/en/creative-commons-licenses, (Accessed on 01/17/2023).

[18] *How much tax do you pay on a car in bangladesh? - cardokan.com*, https://cardokan.com/how-much-tax-do-you-pay-on-a-car-in-bangladesh/, (Accessed on 09/16/2022).

[19] *Keyphrase extraction with bert transformers and noun phrases — towards data science*, https://towardsdatascience.com/enhancing-keybert-keyword-extraction-results-with-keyphrasevectorizers-3796fa93f4db, (Accessed on 01/17/2023).

[20] *Open access licenses - elsevier*, https://www.elsevier.com/about/policies/open-access-licenses, (Accessed on 01/17/2023).

[21] *Sklearn.preprocessing.standardscaler — scikit-learn 1.2.0 documentation*, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html, (Accessed on 01/17/2023).

[22] *What is a multi-layered perceptron?* https://www.educative.io/answers/what-is-a-multi-layered-perceptron, (Accessed on 01/17/2023).

[23] *Xgboost documentation — xgboost 1.7.3 documentation*, https://xgboost.readthedocs.io/en/stable/, (Accessed on 01/17/2023).