MACHINE LEARNING MODELING FOR CAR SELLING PRICE PREDICTION

BY

ABDULLAH AL MAMUN
ID: 191-15-12629

FATEMA ABDULLAH
ID: 191-15-12146

SANJIDA AFROZ
ID: 191-15-12150

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

DR. MD. TAREK HABIB
Assistant Professor
Department of CSE
DaffodilInternational University

Co-Supervised By

MR. MD. SADEKUR RAHMAN
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project titled "Machine Learning Modeling for Car Selling Price Prediction", submitted by Abdullah Al Mamun and Fatema Abdullah Famia and Sanjida Afroz to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on date.

BOARD OF EXAMINERS

Dr. Touhid Bhuiyan — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Subhenur Latif — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mohammad Monirul Islam — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

26.01-23
Dr. Dewan Md Farid — **External Examiner**
Professor
Department of Computer Science and Engineering
United International University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Dr. Md. Tarek Habib,** Assistant Professor, Department of CSE, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. Md. Tarek Habib
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Mr. Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Abdullah Al Mamun
ID: 191-15-12629
Department of CSE
Daffodil International University

Fatema Abdullah
ID: 191-15-12146
Department of CSE
Daffodil International University

Sanjida Afroz
ID: 191-15-12150
Department of CSE,
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to**Almighty Allah** for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Dr.Md. Tarek Habib**, Assistant Professor, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "**Machine Learning**" to carry out this project. His endless patience, scholarlyguidance, continualencouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Professor Dr. Touhid Bhuiyan,** Professor and Head**,**Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Forecasting car prices using machine learning (ML) refers to the use of ML algorithms and techniques to make assumptions of the future price of a car. This can be useful for a variety of purposes, such as helping car buyers and sellers make informed decisions, assisting car dealerships with inventory management, or providing insights for car manufacturers and other industry stakeholders. To predict car prices using ML, data is collected on a variety of factors that can affect ongoing costing of a car, such as its make and model, age, mileage, condition, and location. This data is then fed into XGBoost ML model, which uses statistical techniques to analyze the data and identify patterns and trends. The model performs 98% accurately in the tested portion of the dataset and ensures that the model can then be used to make predictions about the future cost of an automobile based on these patterns and trends.

TABLE OF CONTENTS

CONTENTS **PAGE**

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Introduction

A used car is a vehicle that has previously been owned by someone else and has been driven for a certain period of time. Used cars can be purchased from a variety of sources, such as dealerships, private sellers, or online marketplaces. The price of a used car is typically lower than the price of a new car, as the vehicle has already been driven and may have some wear and tear. A reconditioned car is a used car that has undergone repairs or renovations to bring it up to a higher quality standard. Reconditioning can include a variety of activities, such as fixing mechanical issues, cleaning and detailing the exterior and interior, and replacing worn or damaged parts. Reconditioned cars are typically more expensive than used cars that have not undergone any repairs, but they may also be in better condition and have a longer lifespan. Both used and reconditioned cars can be good options for people who want to save money on their vehicle purchase. However, it's important to carefully research and inspect a used or reconditioned car before buying to ensure that it is in good condition and a good value.

Forecasting car prices using machine learning (ML) refers to the use of ML algorithms and techniques to make assumptions of the future price of a car. This can be useful for a variety of purposes, such as helping car buyers and sellers make informed decisions, assisting car dealerships with inventory management, or providing insights for car manufacturers and other industry stakeholders. To predict car prices using ML, data is collected on a variety of factors that can affect ongoing costing of a car, such as its make and model, age, mileage, condition, and location. This data is then fed into a ML model, which uses statistical techniques to analyze the data and identify patterns and trends. The model can then be used to make predictions about the future cost of an automobile based on these patterns and trends.

There are numerous different ML algorithms and techniques to use and predict automobile prices for. The specific algorithm or technique used will depend on the characteristics of the data and the goals of the prediction. Overall, car price prediction using ML is a complex and multifaceted process that requires a deep understanding of ML algorithms and techniques, as well as knowledge of the specific characteristics of the car market.

A used car's pricing can be complicated to predict for a multitude of reasons. For individuals who are looking to buy or sell a used car, accurate price predictions can be helpful in making informed decisions. Buyers can use price predictions to determine whether a particular car is a good value, while sellers can use them to set a realistic price for their car. For car dealerships, accurate price predictions can be useful in managing inventory and making informed purchasing decisions. By understanding the demand for different types of used cars, dealerships can stock the types of cars that are most likely to sell and avoid having excess inventory that is difficult to move. Accurate price predictions can provide valuable insights into trends in the used car market and help car manufacturers and other industry stakeholders understand the factors that are driving demand for used cars. This can inform business strategies and help companies to make more informed decisions. By providing more accurate predictions about used car prices, ML models can help to refine the efficiency for the used car market by reducing uncertainty and helping buyers and sellers to make more informed decisions. This can lead to more efficient resource allocation and overall market performance. Overall, estimating the charge of a used car can be important for a variety of purposes, including assisting individuals and businesses in making informed decisions, providing insights into market trends, and improving the efficiency of the used car market.

As predicting the price of a quite used or reconditioned automobile is a quite troublesome fact, depending on various factors, prices can vary very often keeping pace with the ongoing economic situation. Using ML, although we cannot assure the predicted price would be the most accurate one or the fixed price it shall go on, we can assume that the range of the car that is going to be expected to be sold. In this study, several ML models as: Linear Regression (LR), decision tree regression (DT), K-Nearest neighbor (KNN), Lasso Regression (LR), Random Forest Regression (RF) and XGBoost algorithm has been tested and implemented to find out the best and most efficient one in terms of considering several performance metrices of these models.

The report is structured as followed: section 2 describes about the detailed discussion regarding the background studies of this topic, section 3 included methodological detailed explanations followed by the section 4 which visualizes some exemplary demonstrations of the implemented experiments. At the last, the conclusion of this topic report is stated along with the future studies and limitations are looked over.

1.2. Motivation

The motivation for selecting car price prediction as a project can vary depending on the individual or organization involved. Some possible motivations might include a personal interest in cars or the used car market, the relevance of the topic to a current or future career, the opportunity to make a contribution to the field, or the opportunity to learn and develop new skills. For example, a student interested in pursuing a career in the automotive industry might be motivated to choose this topic for a school project because it aligns with their interests and provides practical experience with ML algorithms and techniques. Alternatively, a car dealership might be motivated to develop a car price prediction model as a way to improve their inventory management and make more informed purchasing decisions. Ultimately, the motivation for selecting car price prediction as a project will depend on the specific goals and interests of the individual or organization involved. Our motivation towards selecting this topic includes:

1. Personal interest: We have a personal interest in cars or the used car market, we may be motivated to choose this topic because it aligns with our interests and passions.
2. Relevant to current or future career: We are pursuing a career in the automotive industry or are interested in using ML in a business context, car cost estimation may be a relevant and practical topic for your thesis.

3. Opportunity to make a contribution: Car price prediction using ML can have significant practical applications, such as assisting car buyers and sellers in making informed decisions, helping car dealerships manage their inventory, and providing insights for car manufacturers and other industry stakeholders. Selecting this topic for our thesis may allow us to make a meaningful contribution to the field.
4. Opportunity to learn and develop skills: Working on a thesis project related to car price prediction can provide an opportunity to learn and develop a range of skills, such as ML algorithms and techniques, data analysis, and research and writing skills.

Overall, there are many potential motivations for selecting car price prediction as a thesis topic, including personal interest, relevance to our career goals, the opportunity to make a contribution, and the opportunity to learn and develop new skills.

1.3. Problem Definition

To develop a ML model that accurately predicts the future price of a used car based on its characteristics and other relevant factors, such as its make and model, age, mileage, condition, and location. This problem definition identifies the main goal of the project, which is to develop a ML model that can predict the price of a used car. It also identifies the key inputs to the model, which are the characteristics and other relevant factors of the car, and specifies that the model should be able to predict the future price of the car.

Other potential elements of a problem definition for this project might include the target audience or stakeholders for the model (e.g., car buyers and sellers, car dealerships, car manufacturers), the specific business or research question that the model is intended to address, and any constraints or limitations that the model will need to consider (e.g., the availability of data, computational resources, time constraints). This problem definition identifies the primary goal of the study, which is to evolve a ML model that can estimate the cost of a used car. It also identifies the key inputs to the model, which are the characteristics and other relevant factors of the car, and specifies that the model should be enabled to distinguish the future price of the car.

1.4. Research Questions

In order to find the right track for the most feasible solution of this research, at first, we need to point out the accurate research questions that we are attempting to solve through our study. Some of our research questions depending on the circumstances are:

1. What factors most significantly impact the price of a used car?
2. How does the age of a car affect its price?
3. How does the mileage of a car impact its price?
4. How does the condition of a car impact its price?
5. What is the relationship between the make and model of a car and its price?
6. How does the location of a car impact its price?
7. What ML algorithms and techniques are most effective for predicting used car prices?
8. How does the size of the training dataset impact the accuracy of the algorithm?
9. How does the inclusion of additional features (e.g., fuel type, horsepower) affect the efficiency of the model?
10. Can the model be used to predict the price of a used car with a high level of accuracy?

These research questions can help us to guide the development and evaluation of a ML model for predicting used car prices, and can be refined and revised based on the specific goals and needs of the project.

1.5. Research Methodology

A research methodology for car price prediction using ML might involve the following steps: First, the research problem and objectives should be clearly defined, including the target audience or stakeholders for the model and the specific business or research question that the model is intended to address. Next, relevant data should be identified and collected, including data on the characteristics of used cars and data on the prices of used cars. The data should then be preprocessed and cleaned to ensure that it is in a suitable format for use in the model. The data should be split into a training set and a test set, with the training set used to fit the model and the test set used to evaluate its performance. Finally, ML algorithms and techniques should be used to train and evaluate the model, with the results compared to determine the most accurate model. This process should be iterative, with the model refined and improved as needed based on the results of the evaluation. Overall, using this methodological approach can aid in ensuring that its model is both precise and dependable that it can be often used anticipate the cost of used automobiles with confidence.

A research methodology for car price prediction using ML might include the following steps:

1. Defining the research problem and objectives: Clearly defining the problem that the model is intended to solve and the specific goals and objectives of the project. This may involve identifying the target audience or stakeholders for the model, the specific business or research question that the model is intended to address, and any constraints or limitations that the model will need to consider.
2. Identify and collecting relevant data: Determine the types of data that will be needed to train and evaluate the model, and collect this data from appropriate sources. This may include data on the characteristics of used cars (e.g., make and model, age, mileage, condition) as well as data on the prices of used cars.
3. Preprocess and cleaning the data: Preprocess and cleaning the data to ensure that it is in a suitable format for use in the model. This may or may not involve removing any missing or irrelevant data, scaling or normalizing the data, and handling any outliers or anomalies.
4. Splitting the data into training and test sets: Dividing the data into a training set and a test set, with the training set used to fit the model and the test set used to evaluate its performance. This can help to reduce the risk of overfitting and improve the generalizability of the model.
5. Training and evaluating the model: Use ML algorithms and techniques to train the model on the training set and evaluate its performance on the test set. This may

involve using a variety of algorithms and techniques and comparing the results to determine the most accurate model.

6. Fine-tuning the model: Once the most accurate model has been identified, fine-tune it by adjusting the model parameters and the training dataset.

1.6. Research Objectives

A research objective for a car price prediction ML model might be to develop a model that is accurate and reliable in predicting the future price of a used car based on its characteristics and other relevant factors. The model should be able to predict the price of a used car with a high level of accuracy, and it should be generalizable to a wide range of used cars. Additionally, the model should be efficient and easy to use, and it should be able to handle a variety of different input data and make predictions in real-time. Other potential research objectives for a car price prediction model might include identifying the factors that most significantly impact the price of a used car, understanding the relationship between the age, mileage, and condition of a car and its price, and developing a model that is relevant and useful for a specific target audience or stakeholder group (e.g., car buyers and sellers, car dealerships, car manufacturers). Ultimately, the specific research objectives of a car price prediction model will depend on the specific goals and needs of the project. Some of the most notable ones are as followed:

1. We will be predicting the prices of used cars.
2. We will be building various ML models and Deep Learning models with different architectures.
3. We will see how ML models perform in comparison to deep learning models.
4. To create a model that accurately forecasts the cost of a used car based on user inputs.
5. To attain high accuracy. to provide a user-friendly User Interface (UI) that accepts input from the user and forecasts the pricing
6. Because it needs a lot of work and specialized knowledge, predicting the price of cars has attracted a lot of study interest.

7. A large number of unique characteristics are analyzed for the precise and dependable prediction.
8. In this project, the customer will be able to take as much as he needs, it will save a lot of time, he will be able to buy as much as he needs.

One potential economic objective of a car price prediction model might be to improve the efficiency of the used car market by providing more accurate and reliable price predictions. This could help to reduce uncertainty for buyers and sellers and facilitate more informed decision-making, leading to more efficient resource allocation and overall market performance. Accurate price predictions could also help to reduce transaction costs by reducing the time and effort involved in negotiating prices, and could help to reduce the risk of fraud or other unethical practices. Additionally, a car price prediction model could be used to help car dealerships and other industry stakeholders make more informed decisions about inventory management, pricing strategies, and other business practices. Overall, the economic objective of a car price prediction model could be to improve the efficiency and performance of the used car market and provide value to a range of stakeholders.

A technical objective of a car price prediction model might be to develop a model that is accurate and reliable in predicting the future cost of a used automobile based on its characteristics and other relevant factors. This might involve identifying the most appropriate ML algorithms and techniques to use, optimizing the model parameters and hyperparameters to achieve the best performance, and developing a model that is able to handle a wide range of input data and make predictions in real-time. Other potential technical objectives for a car price prediction model might include developing a model that is efficient and scalable, able to handle large and complex datasets, and able to handle missing or incomplete data. Ultimately, the specific technical objectives of a car price prediction model will depend on the specific goals and needs of the project.

1.7. Research Layout

Chapter 1: The introduction, purpose, problem definition, research question, research methodology, and anticipated results of our project will all be presented in Chapter 1.

Chapter 2: The framework of this study and its related activities, as well as the project's present status from Bangladesh's perspective and information on government priorities and rules, will all be made clear in Chapter 2.

Chapter 3: Chapter 3 will outline how the suggested models are put into practice to estimate the price of used car in Bangladesh.

Chapter 4: The effectiveness of the algorithms in price forecasting will be covered in Chapter 4.

Chapter 5: Comparing the outcomes and choosing the top-performing algorithm are the main topics of Chapter 5.

Chapter 6: The research's result is found in Chapter 6.

Chapter 7: All of the references consulted for this research are displayed in Chapter 7.

CHAPTER 2
BACKGROUND

2.1. Introduction

Predicting used car prices correctly can have a variety of effects on people, businesses, and society. One of the key advantages of used car price prediction is that by lowering uncertainty and enabling better-informed decision-making, it can assist to increase the efficiency of the used automobile market. This may result in more effective resource management and market performance overall, making it simpler for buyers and sellers to locate the ideal vehicle at the ideal cost. Predicting used car prices may also save time and effort when purchasing or selling a car by giving people and businesses a quick and simple way to figure out the value of a used car. Accurate price forecasts can also help to lower transaction costs by cutting down on the time and effort required for price negotiations, which can facilitate the buying and selling of used cars. In order to enhance fairness and lower the possibility of fraud or other unethical actions, used vehicle price prediction can also help to make sure that buyers and sellers have access to accurate and reliable information about the value of used cars. Last but not least, used car price forecasting can promote the purchase of more environmentally friendly and fuel-efficient used cars, hence lowering greenhouse gas emissions and other negative environmental effects.

2.2. Related Works

Jeon et al. (2018) offer a novel methodology that combines artificial neural networks, stepwise regression analysis, and dynamic time warping. To validate the model, we employ JaroWinkler distance with Synthetic Aggregate approximation (SAX) as a prediction accuracy metric. Using big data open sources, we build a big data processing framework to manage the whole processes. They offered a novel methodology that combines artificial neural networks, stepwise regression analysis, and dynamic time warping. To validate the model, we employ JaroWinkler distance with SAX as a prediction accuracy metric. Using big data open sources, we build a big data processing framework to handle the overall procedures [1]. Yadav et al., 2021 are aware of the idea of object detection, such as the car detection, and use automatic ML techniques to investigate the cost of a used car.

Chandar, 2019 [2] seeks to identify an effective soft computing method for stock prediction. For target projection, the car's price is regarded as a dependent variable (S, 2020)[10]. Chandar, 2019 [2] seeks to identify an effective soft computing method for stock prediction. For target prediction, the car's price is regarded as a dependent variable (S, 2020). Partial dependence plots (PDPs) are used to look at nonlinear correlations between shared-car use and various predictors, and a gradient boosting regression model (GBRT) is used to forecast shared-car use at the station level. 2020 Wang et al. [3]. Yadav et. al., 2021 [4] using the idea of object detection, such as car detection, to investigate the cost of a used car utilizing automatic ML techniques.

In Shalini et al., 2021 [5], ML approaches are used to optimize prediction models, and two techniques are compared: one that will be implemented through ML techniques like LR, and one just through optimization algorithms like gradient descent and stochastic gradient descent. Making poor decisions can result in significant losses or possibly the closure of a corporation. A gradient boosting regression model (GBRT) is used to forecast shared-car use at the station level, and partial dependence plots (PDPs) are utilized to examine nonlinear connections between shared-car use and other predictors. 2020 Including Wang et al.

to provide a creative response to this problem The study by Narayana et al., 2021 [6] focuses on the used automobile sales industry, one of the retail industries. A CNN-BiLSTM-AM approach to forecast the stock closing price of the following day is proposed by Lu et al. in 2021 [7]. The development of a system to anticipate the price of used Tesla vehicles using machine learning is described Arefin, 2021 [8]. Samruddhi et al., 2020 [9] is yet another important piece of literature.

2.3. Bangladesh Perspective

The reusing of reconditioned cars can be a cost-effective and sustainable alternative to buying a new car, particularly in countries where the cost of new cars may be prohibitively high for many people. Reconditioned cars are often used cars that have been restored to a good condition and are sold at a lower price than a new car. They can be a good option for people who want to own a car but are on a budget, or for people who are looking for a more environmentally-friendly transportation option. In some cases, reconditioned cars may also be eligible for government incentives or subsidies, which can further reduce the cost of ownership. Overall, the reusing of reconditioned cars can be a viable and attractive option for many people in Bangladesh and other countries.

There are a number of economic aspects that Bangladesh as a country might consider when evaluating the use of reconditioned cars. Some potential considerations might include:

- Cost: Reconditioned cars are typically cheaper than new cars, which can make them a more affordable option for many people. This can be particularly important in countries where the cost of new cars is high or where many people have limited financial resources.
- Employment: The reconditioning of used cars can create jobs in the automotive repair and maintenance sector, which can contribute to economic growth and development.
- Trade: The import and export of reconditioned cars can contribute to international trade and help to balance a country's trade deficit or surplus.
- Environmental sustainability: Reconditioned cars can be a more environmentally-friendly transportation option, as they reuse existing resources and emit fewer greenhouse gases than new cars. This can be beneficial for a country's environmental and energy policies.

- Consumer protection: Reconditioned cars may be subject to regulations or standards to ensure that they meet certain quality and safety standards, which can help to protect consumers and promote fair and transparent trade practices.

Considering all these, it is a good and safe choice for both of the user and the government to encourage more usage of reconditioned cars indicating a greatly positive aspect towards our future economic progress.

CHAPTER 3
RESEARCH METHODOLOGY

3.1. Introduction

In this experiment, we developed a ML model to predict the future price of a used car based on its characteristics and other relevant factors. Our goal was to develop a model that was accurate and reliable, and that could be used to assist car buyers and sellers, car dealerships, and other stakeholders in making informed decisions. To achieve this goal, we followed a systematic approach that involved defining the research problem and objectives, identifying and collecting relevant data, preprocessing and cleaning the data, splitting the data into training and test sets, and training and evaluating the model. In the following sections, we will describe each of these steps in detail and discuss the methods and techniques that we used to develop and evaluate the model.

3.2. Dataset Properties

The Car Price dataset from custom collecteda showroom named Car View in Bangladesh is a collection of data on used car prices and characteristics. Some properties of this dataset might include:

- Number of observations: The dataset includes more than 25,000 observations, each representing a different used car.
- Variables: The dataset includes a variety of variables, including the make and model of the car, chasis_no, registration_year, transmission of the car, the mileage

of the car, the condition of the car engine, the fuel of the car, door, dimensions and the price of the car in Bangladesh.

- Data types: The variables in the dataset are a mix of categorical (e.g., make and model, condition) and numerical (e.g., age, mileage, price) data.
- Missing values: There are a small number of missing values in the dataset, which will need to be addressed before the data can be used in a ML model.
- Data sources: The data in the Car Price dataset was collected from a variety of sources, including online classified websites, car dealerships, and private sellers.

Overall, the Car Price dataset from custom collecteda showroom named Car View in Bangladesh is a large and comprehensive collection of data on used car prices and characteristics that could be useful for developing a ML model for predicting used car prices.

The table 1 stated below shows a visual description about the dataset we have collected:

Table 3.1: Dataset Example

Car_Name	Model	Chasis_No	Registration_Year	Mileage	Engine_Size	Steering	Transmission	Fuel	Passangers	Dimension	Price_In_BD
2012 Toyota 86 G	DBA-ZN6	ZN6-026**	2013	188753	2000	Right	Manual	Gasoline/Petrol	4	9.96	2049018.72
2006 Toyota Allex 4Grade	DBA-NZE121	NZE121-5108***	2006	25000	1490	Right	Automatic	Gasoline/Petrol	5	10.36	1767651.32

2004 Toyota Allex XS 150 TOYOTA ALLEX	CBA-NZE121	NZE121-0316***	2004	91000	1500	Right	Automatic	Gasoline/Petrol	5	10.45	1574574.7
2001 Toyota Allex ALLEX XS150	TA-NZE121	NZE121-0063***	2001	65000	1490	Right	Automatic	Gasoline/Petrol	5	10.36	1504349.12
2001 Toyota Allex	NZE121	0032***	2001	92420	1500	Right	Automatic	Gasoline/Petrol	4	10.36	1442247.92

Figure 1 shows the density of engine size curve:

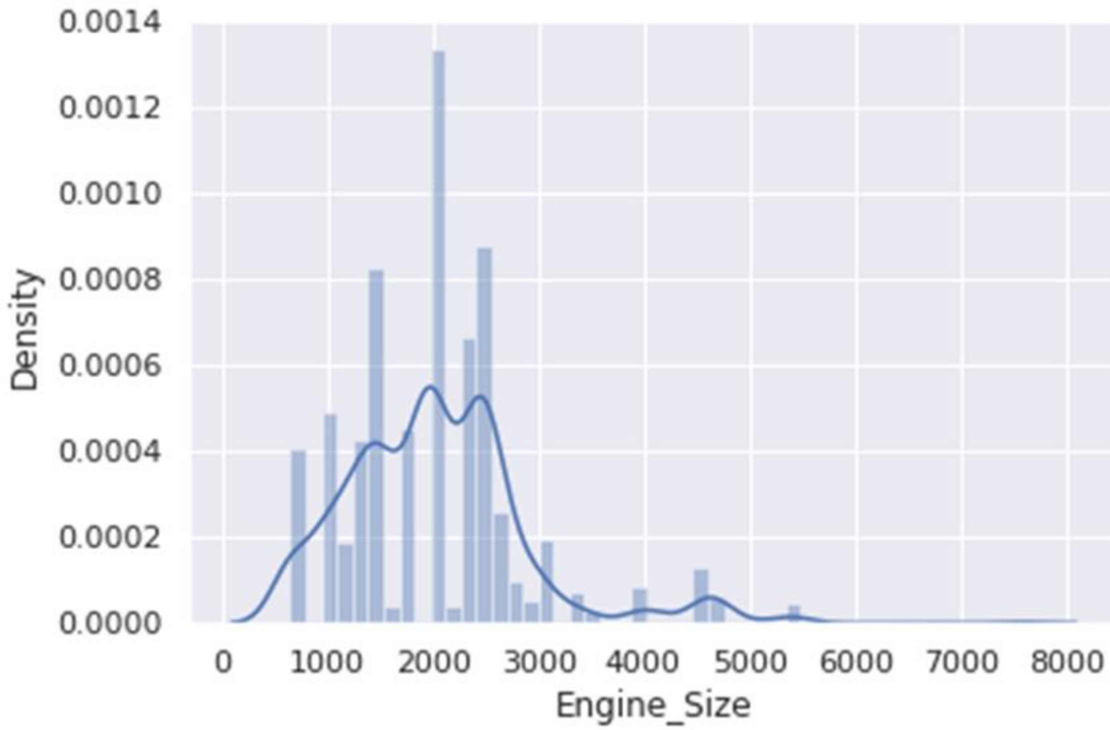


Figure 3.1: Distance between Engine Sizes

A figure showing a visualization curve for engine sizes might be a graph that displays the distribution of engine sizes among a group of used cars. The x-axis of the graph might represent the different engine sizes, while the y-axis might represent the number of cars with each engine size. The graph might show a curve or a series of bars, with each bar representing the number of cars with a particular engine size. The curve or bars might be color-coded to indicate different characteristics of the cars, such as their age or price.

Figure 2 visualizes mileage comparisons of the dataset:

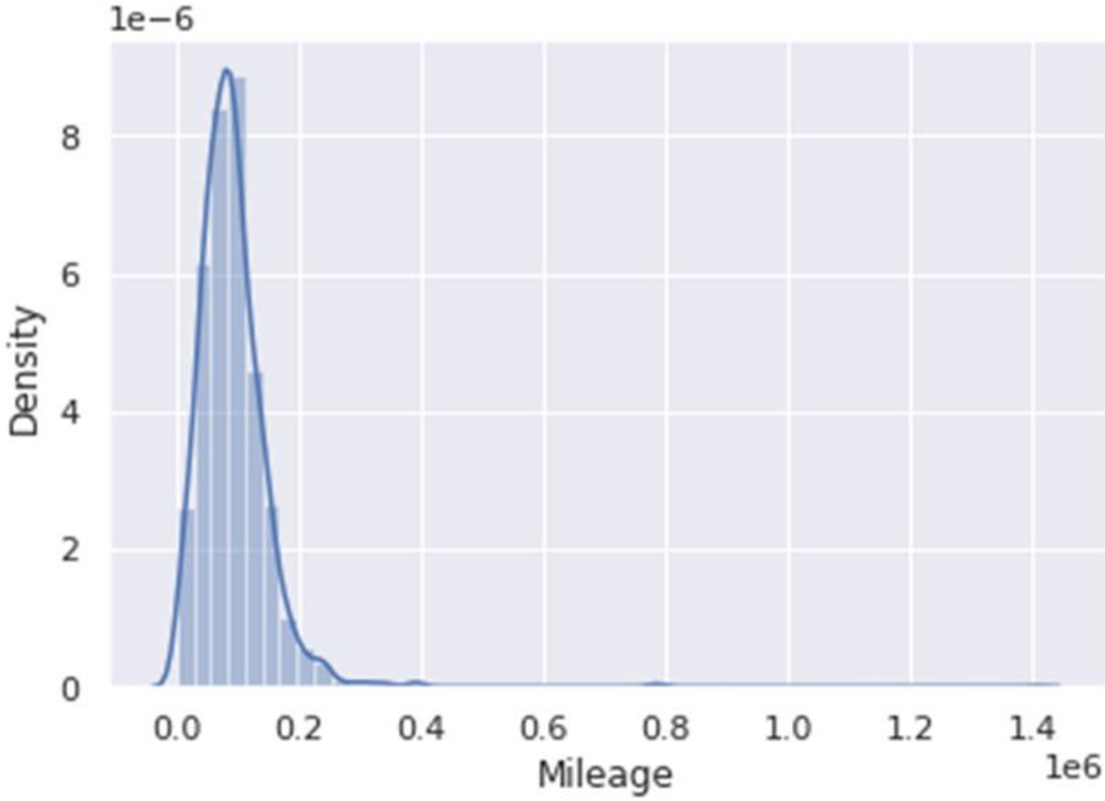


Fig 3.2: Mileage of dataset cars

A figure showing mileage per liter might be a graph that displays the fuel efficiency of a group of used cars. The x-axis of the graph might represent the different cars, while the y-axis might represent the mileage per liter (MPL) of each car. The MPL of a car is a measure of its fuel efficiency, calculated by dividing the number of miles that the car can travel on a single liter of fuel by the number of liters of fuel that the car consumes.

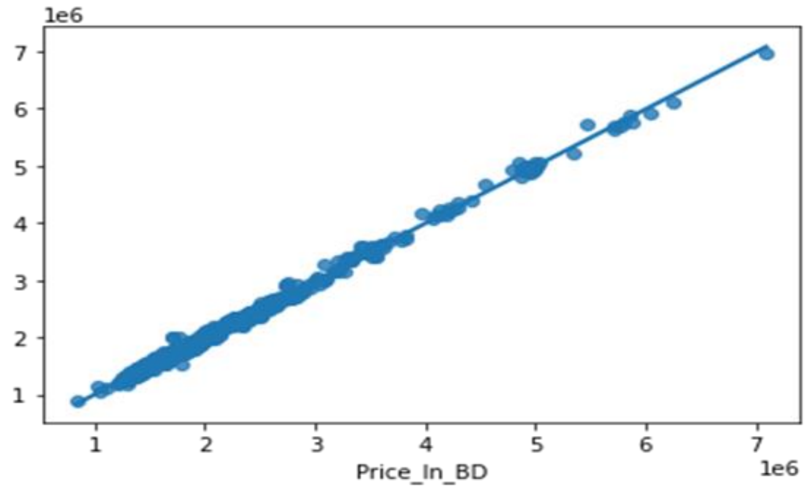


Fig 3.3: Price in Bangladesh

Figure 4 shows dimensions of the dataset:

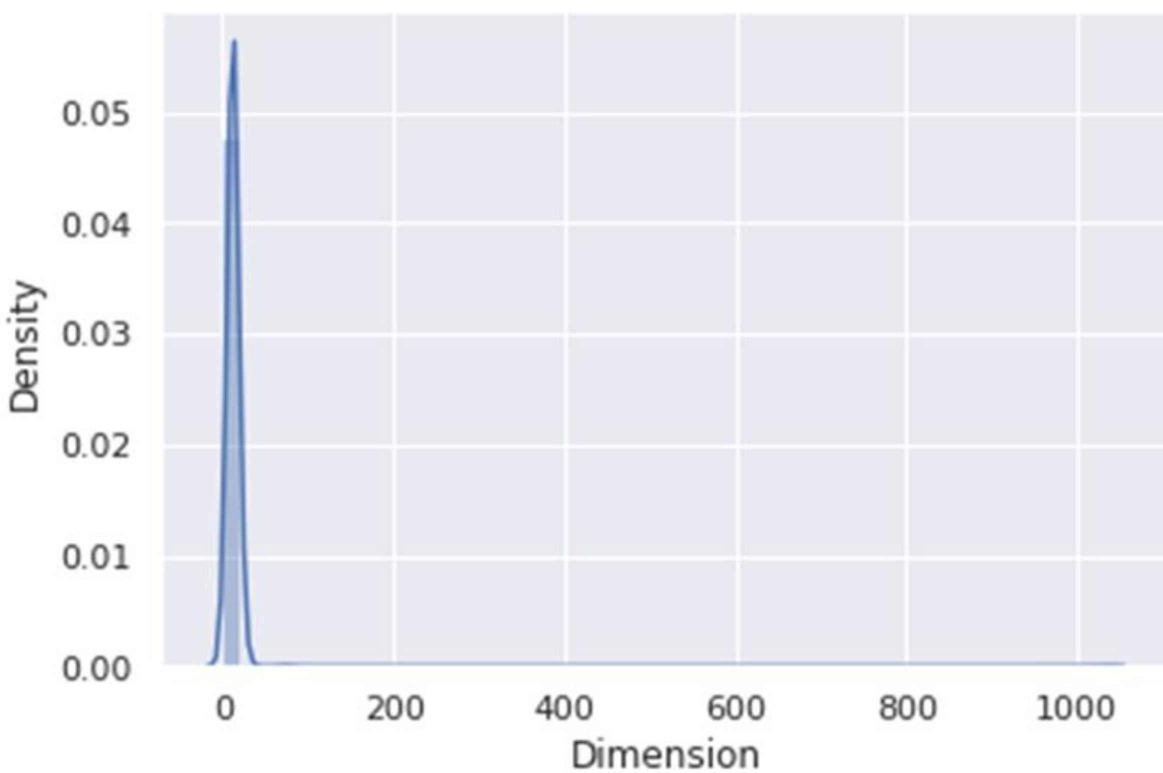


Fig 3.4: Dimension of Dataset

Figure 5 shows registration year verities:

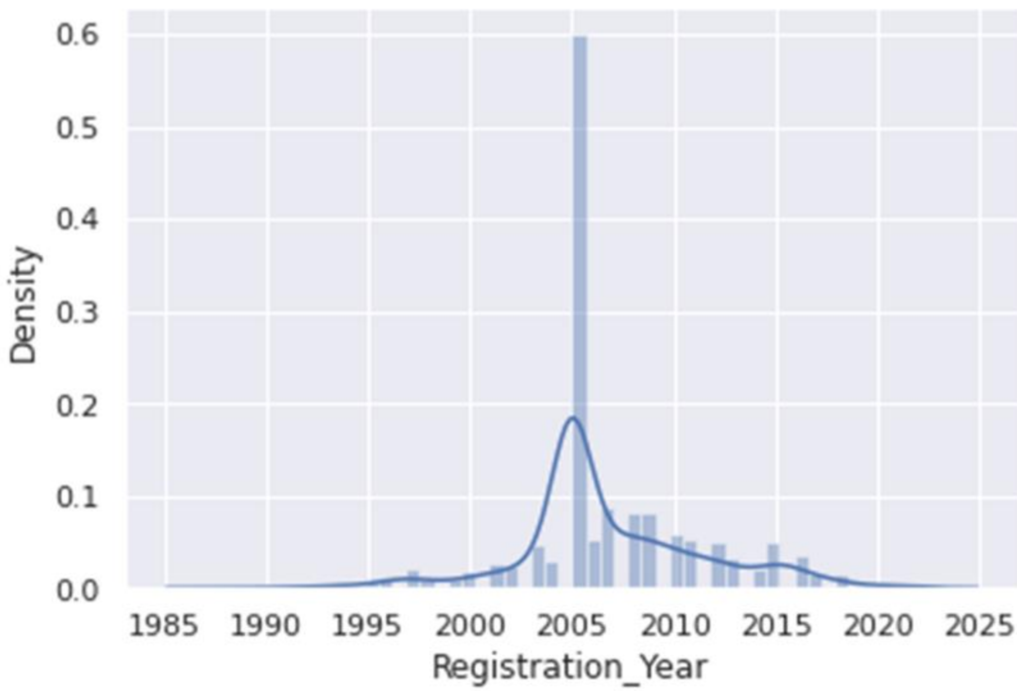


Fig 3.5: Registration Year of dataset

3.3. Dataset Preprocessing

Preparing data for an ML model is known as data preparation. In the context of car price prediction, data preprocessing might involve a variety of tasks, including:

1. Data cleaning: Removing any missing, irrelevant, or inconsistent data from the dataset.
2. Data scaling: Normalizing the data to ensure that all variables are on the same scale, which can help to improve the efficiency of the model.
3. Data transformation: Transforming the data into a suitable format for use in the model, such as converting categorical variables into numerical variables.
4. Data splitting: Separating the data into two sets a training set and a test set, and using the training set to modify the model and the test set to assess how well it performed.
5. Data augmentation: Generating additional data points to increase the size and diversity of the dataset.

The preprocessed data can be visualized as followed:

Table 3.2: Preprocessed Dataset

sl	Registration_Year	Mileage	Engine_Size	Passangers	Door	Dimension	Price_In_BD
0	2013	188753	2000	4	5	9.96	2049019
1	2006	25000	1490	5	5	10.36	1767651
2	2004	91000	1500	5	5	10.45	1574575
3	2001	65000	1490	5	5	10.36	1504349
4	2001	92420	1500	4	5	10.36	1442248

Overall, data preprocessing is an important step in the process of developing a ML model for car price prediction, as it helps to ensure that the data is in a suitable format for use in the model and that the model is able to learn from the data as effectively as possible.

3.4. Applied Regression Algorithms

K-Nearest Neighbor

KNN (k-NN) is a ML algorithm that is used for classification and regression tasks. It is a type of instance-based learning, which means that it does not build a model to make predictions, but instead stores the training data and uses it to make predictions based on the similarity of new data points to the stored training data. KNN (KNN) is a ML algorithm that can be used for predicting used car prices by finding the K nearest neighbors in the dataset and using their prices to make a prediction. This algorithm works well when there is a strong relationship between the independent variables and the used car price.

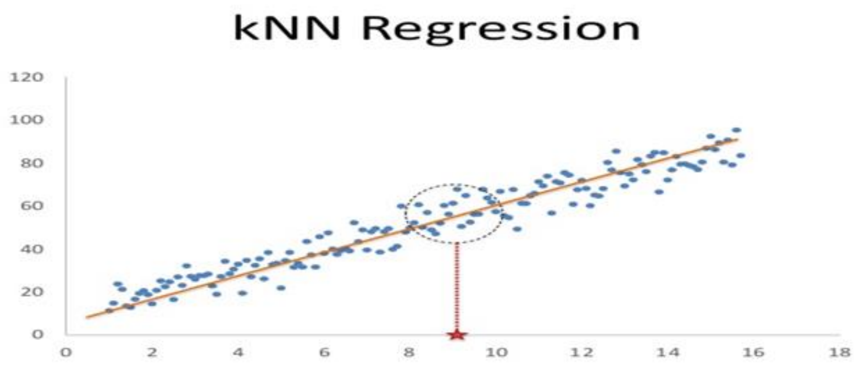


Fig 3.6: K-NN Regression

To make a prediction using k-NN, the algorithm calculates the distance between a new data point and the k training data points that are most similar to it (the "nearest neighbors"). The prediction is then based on the values of the nearest neighbors. For example, in a classification task, the prediction might be the most common class among the nearest neighbors, or in a regression task, the prediction might be the average value of the nearest neighbors.

k-NN is a simple and effective algorithm that is widely used in a variety of applications, including pattern recognition, image classification, and anomaly detection. It is particularly useful when the relationships between the variables in the data are complex and non-linear, as it is able to capture these relationships without the need for explicit feature engineering. However, it can be sensitive to the choice of the value of k and can be computationally intensive for large datasets.

Linear Regression

LR is a statistical method that is commonly used to model the relationship between a dependent variable and one or more independent variables. It is based on the idea that there is a linear relationship between the variables, meaning that the change in the dependent variable can be predicted by a linear function of the independent variables.

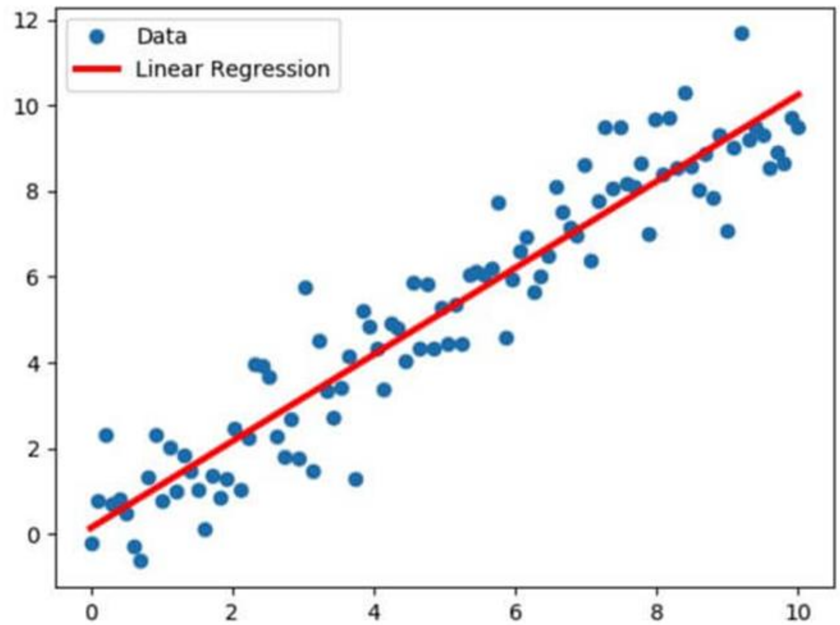


Fig 3.7: Linear Regression

In the context of car price prediction, LR could be used to model the relationship between the price of a used car and its characteristics, such as its make, model, age, and mileage. The model would be trained on a dataset of used car prices and characteristics, and the goal would be to learn the coefficients of the linear function that best predict the price of a used car based on its characteristics.

To use LR for car price prediction, we would first need to gather a dataset of used car prices and characteristics. This dataset would be used to train the LR model. Once the model is trained, we can use it to make predictions on new data points by inputting the characteristics of the car into the linear function and solving for the predicted price.
LR is a simple and widely-used method for modeling the relationship between variables, and it is particularly useful when the relationship is expected to be linear. However, it can be less effective when the relationship is more complex or non-linear.

LassoRegression

LR is a type of LR that is used to model the relationship between a dependent variable and one or more independent variables. It is similar to standard LR, but it includes a regularization term that helps to shrink the coefficients of the model towards zero, which can help to reduce overfitting and improve the interpretability of the model.

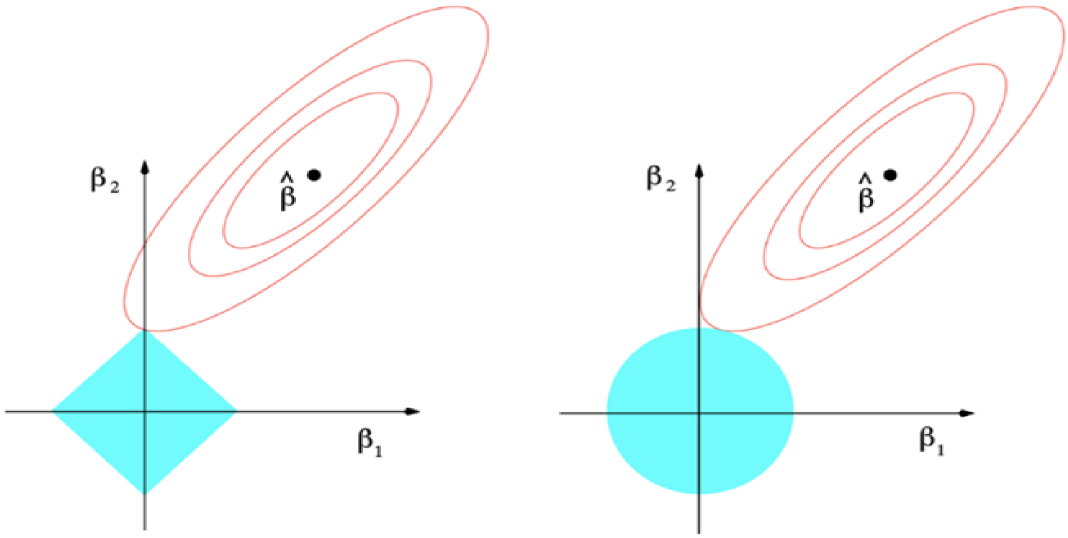


Fig 3.8: LassoRegression

In the context of car price prediction, LR could be used to model the relationship between the price of a used car and its characteristics, such as its make, model, age, and mileage. The model would be trained on a dataset of used car prices and characteristics, and the goal would be to learn the coefficients of the linear function that best predict the price of a used car based on its characteristics, while also reducing overfitting and improving interpretability.

To use LR for car price prediction, we would first need to gather a dataset of used car prices and characteristics. This dataset would be used to train the LR model. Once the model is trained, we can use it to make predictions on new data points by inputting the characteristics of the car into the linear function and solving for the predicted price.

LR is a useful tool for car price prediction because it can help to reduce overfitting and improve the interpretability of the model. It is particularly useful when there are many independent variables and it is important to identify the most important ones, as the regularization term helps to shrink the coefficients of the less important variables towards zero. However, LR can be less effective when the relationship between the variables is more complex or non-linear.

Decision Tree Regression

DTis a ML algorithm that is used to model the relationship between a dependent variable and one or more independent variables. It works by constructing a tree-like model that makes predictions based on a series of decision rules that are learned from the training data.

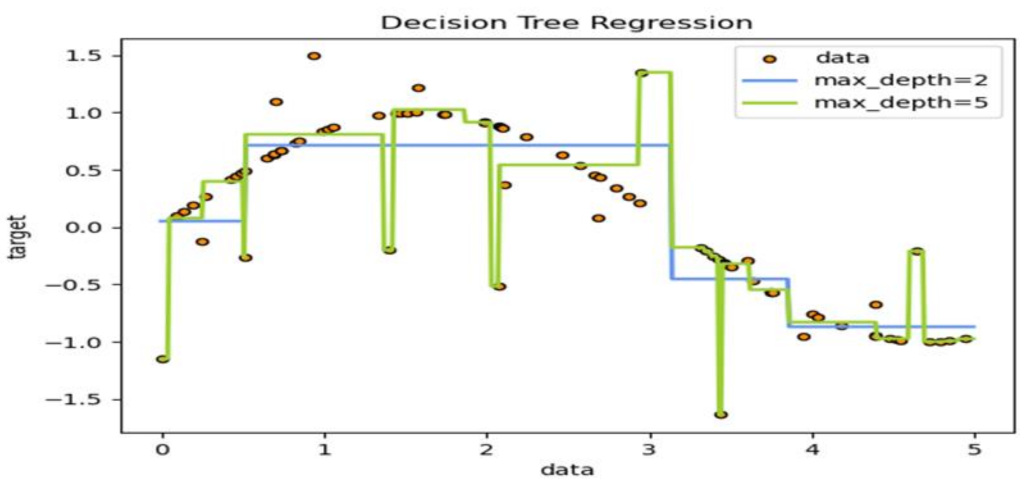


Fig 3.9: Decision Tree Regression

In the context of car price prediction, DTcould be used to model the relationship between the price of a used car and its characteristics, such as its make, model, age, and mileage. The model would be trained on a dataset of used car prices and characteristics, and the goal would be to learn a set of decision rules that can be used to predict the price of a used car based on its characteristics.

To use DTfor car price prediction, you would first need to gather a dataset of used car prices and characteristics. This dataset would be used to train the DTmodel. Once the model is trained, you can use it to make predictions on new data points by following the decision rules of the tree to arrive at a prediction for the price of the car.

DTis a flexible and interpretable algorithm that can handle both numerical and categorical data. It is particularly useful when the relationships between the variables are complex and non-linear, as it is able to capture these relationships using a series of decision rules. However, decision tree models can be prone to overfitting, particularly when the tree becomes too deep or complex, and they may not generalize well to new data.

Random Forest

RF is a ML algorithm that is used to model the relationship between a dependent variable and one or more independent variables. It is an ensemble method that combines the predictions of multiple decision tree models to produce a more robust and accurate prediction.

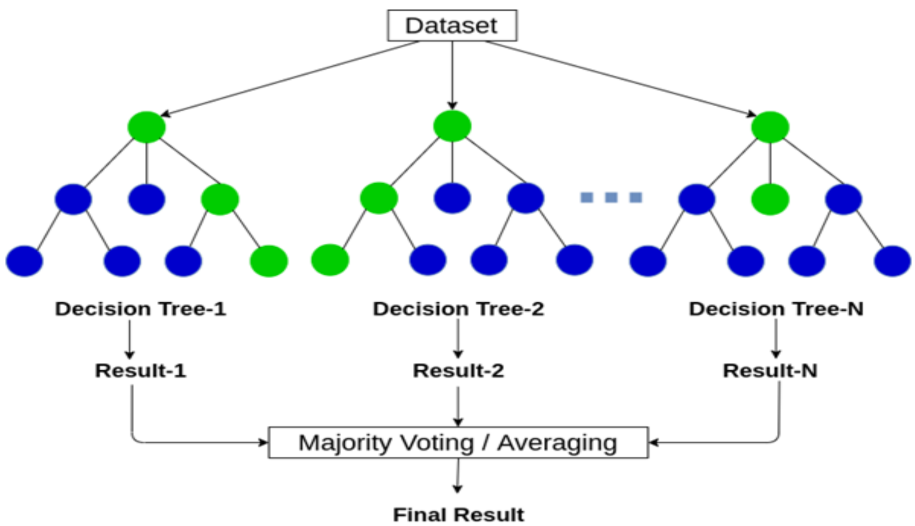


Fig 3.10: Random Forest Regression

In the context of car price prediction, RF could be used to model the relationship between the price of a used car and its characteristics, such as its make, model, age, and mileage. The model would be trained on a dataset of used car prices and characteristics, and the goal would be to learn a set of decision trees that can be used to predict the price of a used car based on its characteristics.

To use RF for car price prediction, we would first need to gather a dataset of used car prices and characteristics. This dataset would be used to train the RF model. Once the model is trained, we can use it to make predictions on new data points by aggregating the predictions of the individual decision trees in the forest.

RF is a powerful and widely-used algorithm that is known for its ability to handle complex and non-linear relationships between variables. It is generally more accurate and robust than a single decision tree model, and it is less prone to overfitting. However, it can be more computationally intensive to train and can be less interpretable than a single decision tree model.

XGBoost Regression

XGBoost (Extreme Gradient Boosting) is a ML algorithm that is used to model the relationship between a dependent variable and one or more independent variables. It is an implementation of gradient boosting, which is an ensemble method that combines the predictions of multiple weak models to produce a strong, accurate model.

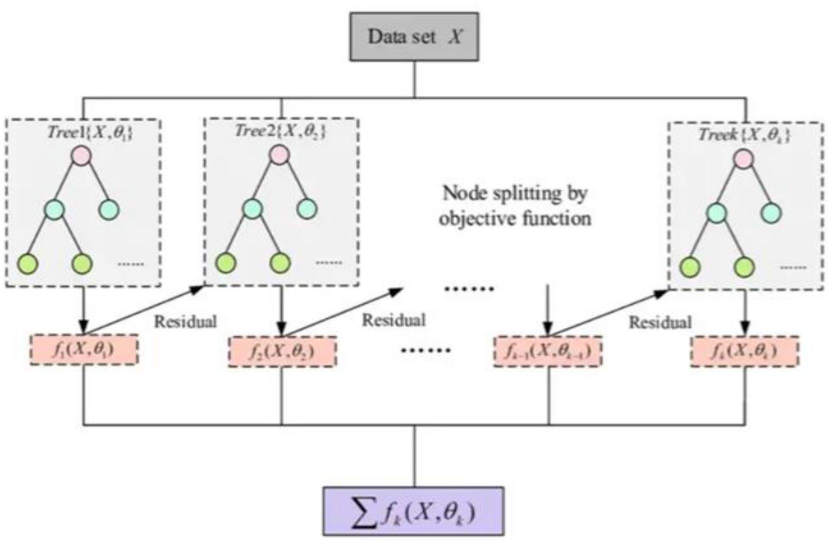


Fig 3.11: XGBoost Regression

In the context of car price prediction, XGBoost regression could be used to model the relationship between the price of a used car and its characteristics, such as its make, model, age, and mileage. The model would be trained on a dataset of used car prices and characteristics, and the goal would be to learn a sequence of weak models that can be combined to produce a strong prediction for the price of a used car based on its characteristics.

To use XGBoost regression for car price prediction, we would first need to gather a dataset of used car prices and characteristics. This dataset would be used to train the XGBoost regression model. Once the model is trained, we can use it to make predictions on new data points by aggregating the predictions of the weak models in the ensemble.

XGBoost is a popular and powerful algorithm that is known for its ability to handle large datasets and complex relationships between variables. It is generally more accurate and faster to train than other gradient boosting algorithms, and it has a number of hyperparameters that can be tuned to further improve its performance. However, it can be more difficult to interpret than other ML algorithms, as it is an ensemble method that combines the predictions of multiple weak models.

To run XGBoost for car price prediction in Google Colab, we would need to follow these steps:

1. Install XGBoost: If we do not have XGBoost already installed in our Google Colab environment, we can install it by running !pip install xgboost in a cell.
2. Import the necessary libraries: In order to use XGBoost, we will need to import the necessary libraries, such as numpy, pandas, and xgboost. We can do this by running import statements in a cell, such as import numpy as np, import pandas as pd, and import xgboost as xgb.
3. Load and prepare the data: Next, we will need to load and prepare our data for training the model. We can do this by using pandas to read the data into a panda DataFrame and then performing any necessary preprocessing steps, such as handling missing values or scaling the data.
4. Split the data into training and test sets: It is a good idea to split our data into a training set and a test set, so that we can evaluate the performance of the model on unseen data. We can use the train_test_split function from scikit-learn to split the data into a training set and a test set.
5. Create the XGBoost model: To create an XGBoost model, we will need to specify the model hyperparameters, such as the learning rate, the maximum depth of the

trees, and the number of trees in the model. We can then use the XGBRegressor class to create the model, and fit it to the training data using the fit method.

6. Make predictions on the test set: Once the model is trained, we can use it to make predictions on the test set by calling the predict method on the model and passing in the test set data as an argument.
7. Evaluate the model performance: To evaluate the performance of the model, we can compare the predicted values to the actual values in the test set and calculate metrics such as mean absolute error, mean squared error, and root mean squared error. You can use the mean_absolute_error, mean_squared_error, and sqrt functions from scikit-learn to calculate these metrics.

By following these steps, we have been able to successfully run XGBoost for car price prediction in Google Colab for this study purpose.

CHAPTER 4

PERFORMANCE EVALUATION OF PROPOSED MODEL

4.1. Training, Testing and Validation of Models

Train/test splitting is a common practice in ML when building and evaluating models. It involves dividing the available data into two subsets: a training set and a test set. The training set is used to fit the model, while the test set is used to evaluate the performance of the model on unseen data.

The typical split ratio is to use 70% of the data for training and 30% for testing. This means that the model is trained on 70% of the data and then tested on the remaining 30% to see how well it performs on unseen data. The split ratio can vary depending on the size and quality of the dataset, but 70/30 is a commonly used ratio.

Using a train/test split helps to prevent overfitting, which is when the model is too closely fit to the training data and does not generalize well to new data. By evaluating the model on a separate test set, you can get a better sense of how well the model will perform on new data. It is important to use a test set that is representative of the types of data that the model will encounter in the real world, in order to get an accurate evaluation of its performance.

4.2. Model Efficiency Metrices

There are several metrics that can be used to evaluate the performance of our implemented and tested regression models. The followings are used in our study include:

1. Mean absolute error (MAE): This metric measures the average difference between the predicted values and the actual values. It is calculated by taking the absolute value of the difference between the predicted and actual values for each data point, and then averaging these values over all data points.
2. Mean squared error (MSE): This metric measures the average squared difference between the predicted values and the actual values. It is calculated by taking the

squared difference between the predicted and actual values for each data point, and then averaging these values over all data points.

3. Root mean squared error (RMSE): This metric is the square root of the MSE, and it is used to provide the error in the same units as the target variable.
4. R-squared (R2): This metric measures the proportion of the variance in the target variable that is explained by the model. It ranges from 0 to 1, with higher values indicating a better fit.
5. Adjusted R-squared: This metric is similar to R2, but it takes into account the number of independent variables in the model. It is adjusted to penalize models with more variables, which can help to prevent overfitting.
6. By calculating these metrics, we can get a sense of how well the regression model is fitting the data and making predictions. We can use these metrics to compare different models and choose the one that performs the best. Finally, we can use accuracy to find out the overall model's performances. Below are the performance visualizations of the models accordingly:

K-Nearest Neighbor

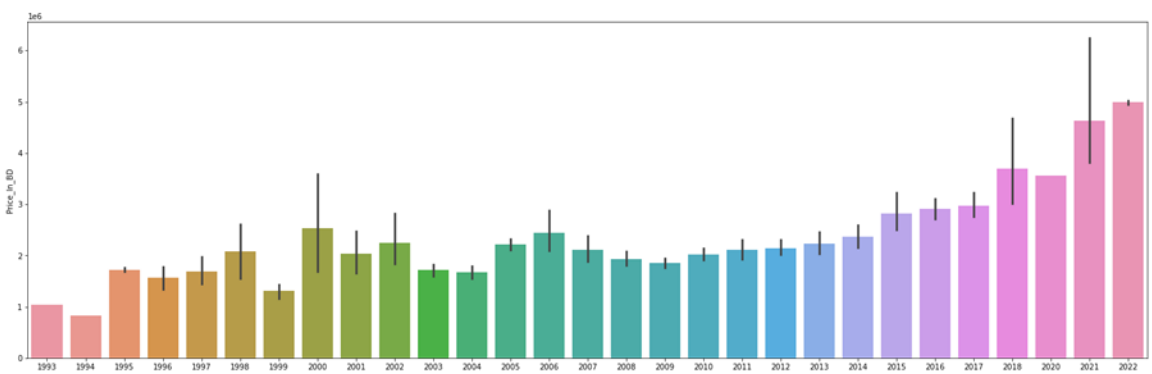


Fig 4.1: KNN Prediction Barplot

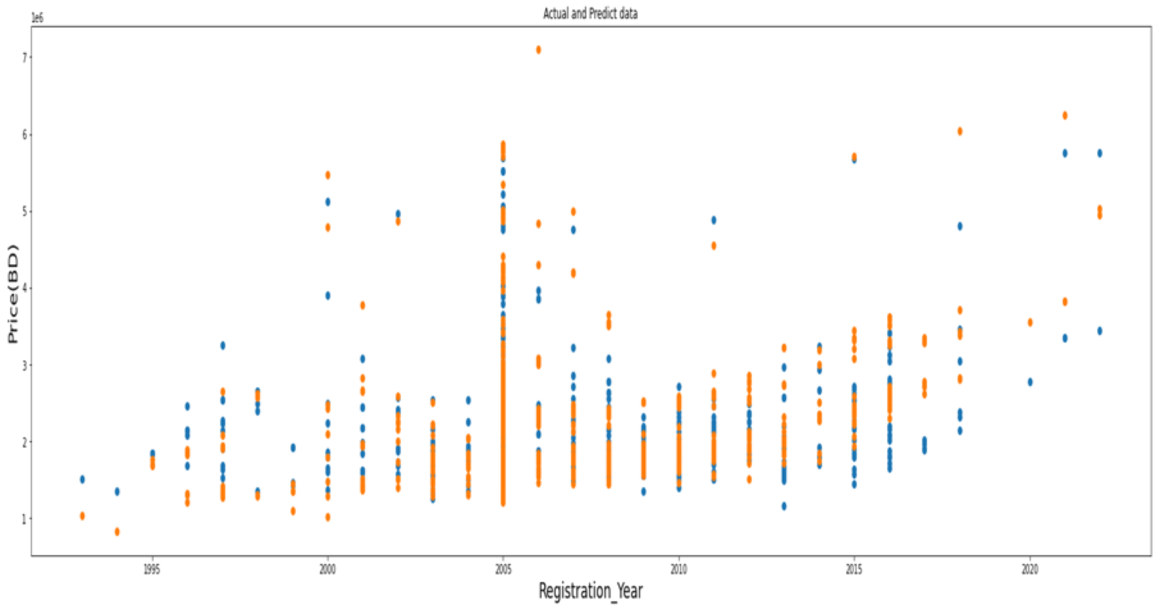


Fig 4.2: KNN Prediction Scatterplot

Linear Regression

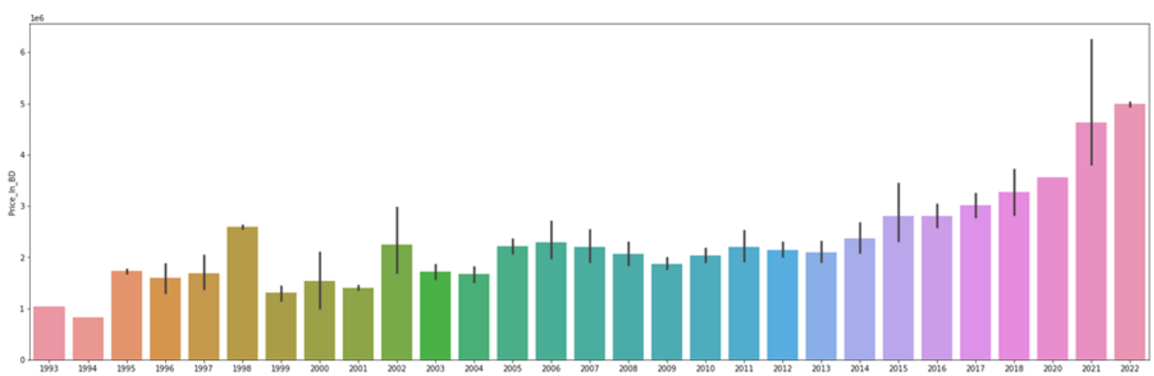


Fig 4.3: LR Prediction Bar plot

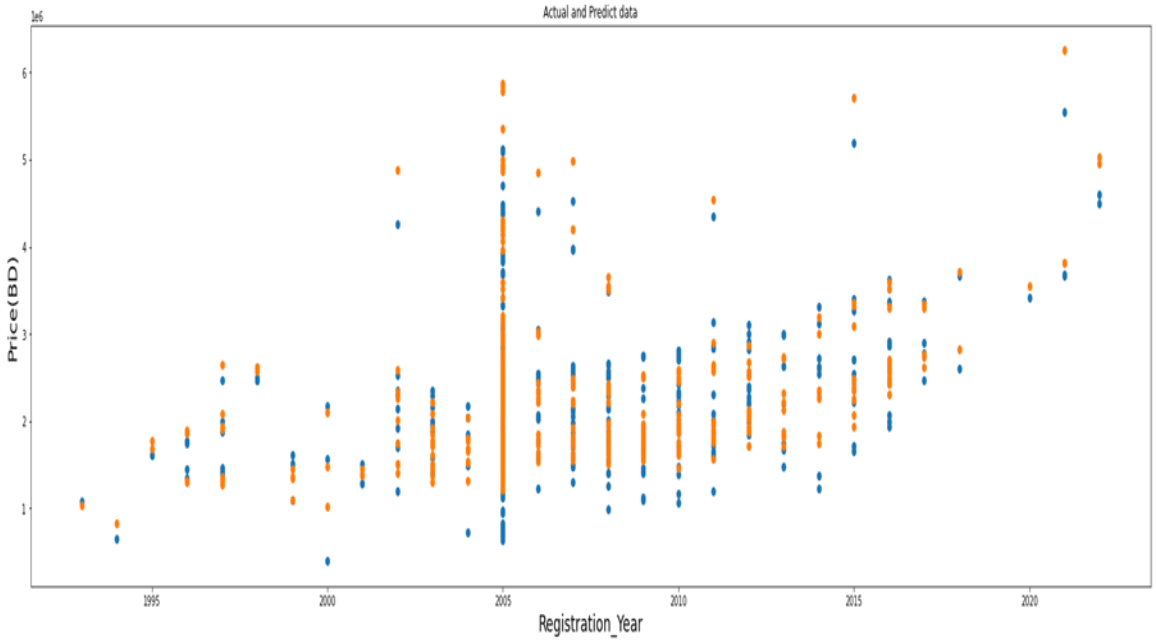
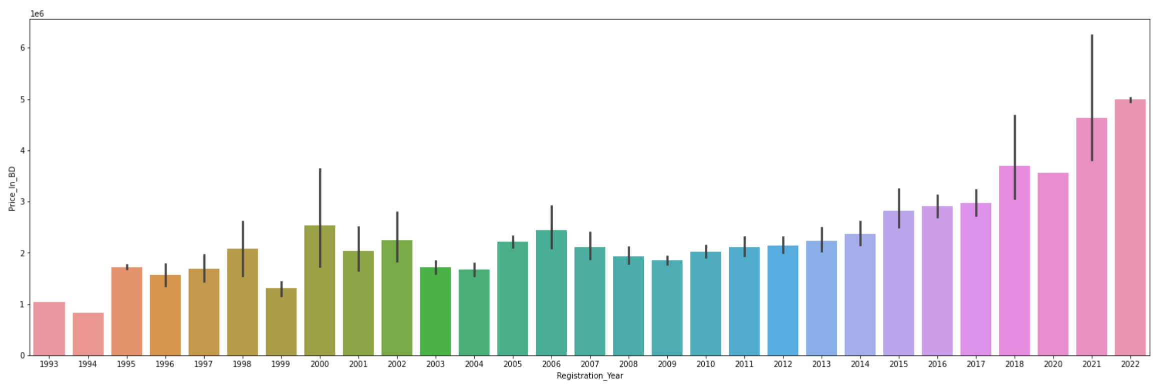


Fig 4.4: LR Prediction Scatterplot



Lasso Regression

Fig 4.5: LR Prediction Barplot

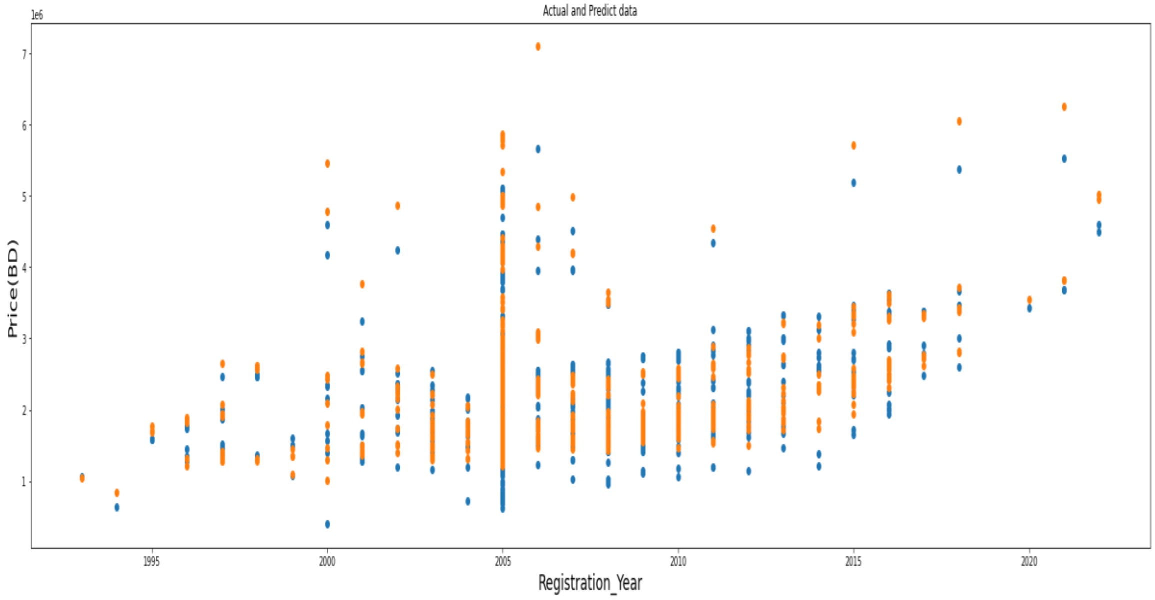


Fig 4.6: LR Prediction Scatterplot

Decision Tree Regression

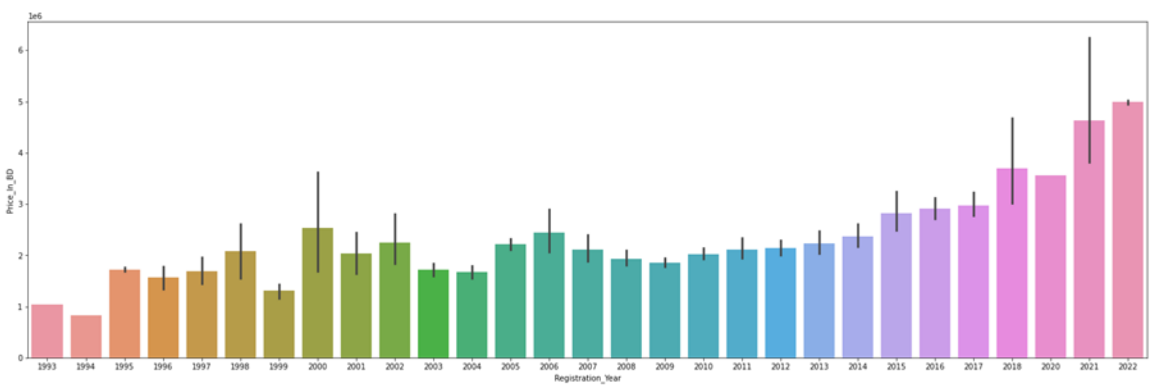


Fig 4.7: Decision Tree Prediction Bar plot

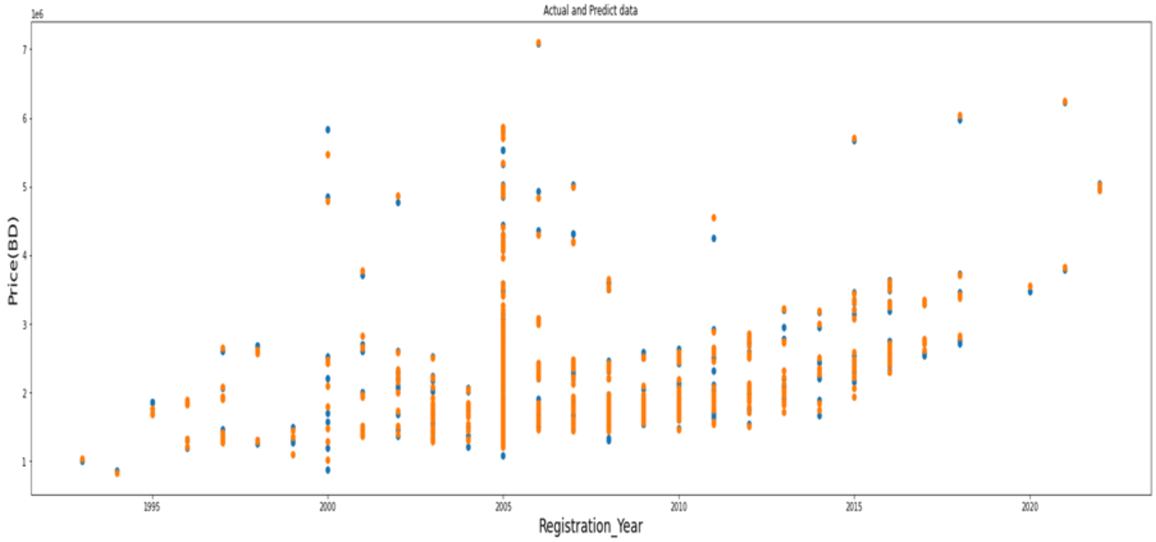


Fig 4.8: Decision Tree Prediction Scatterplot

Random Forest

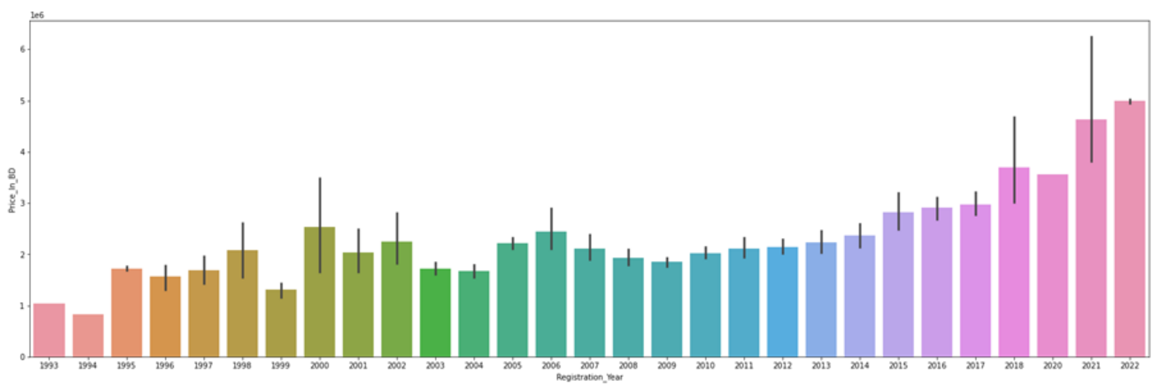


Fig 4.9: RF Barplot

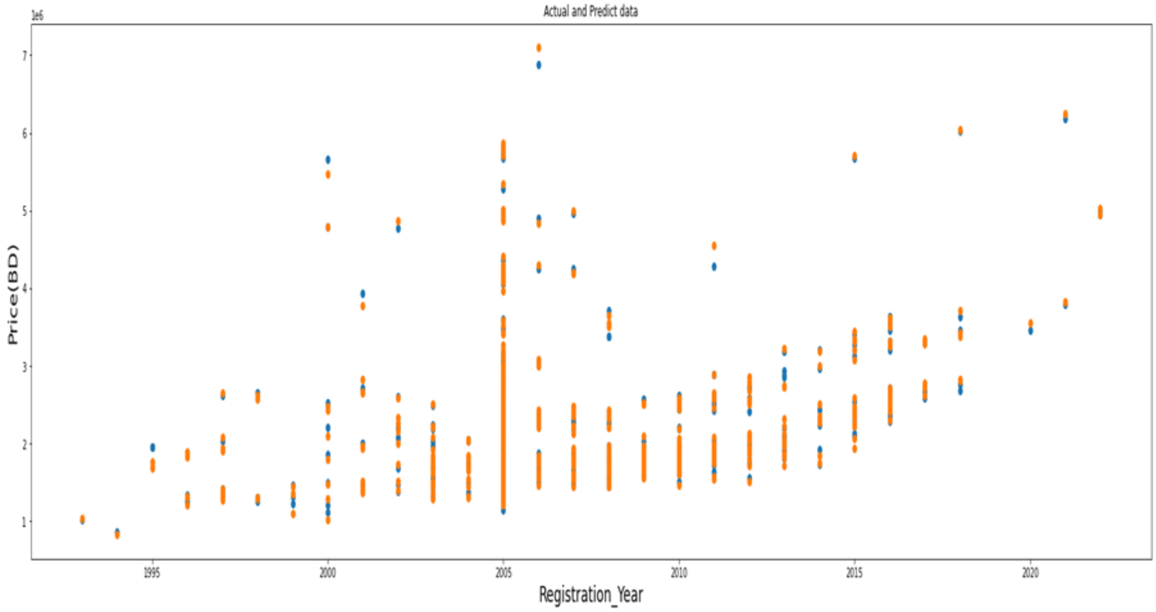


Fig 4.10: KNN Prediction Scatterplot

XGBoost Regression

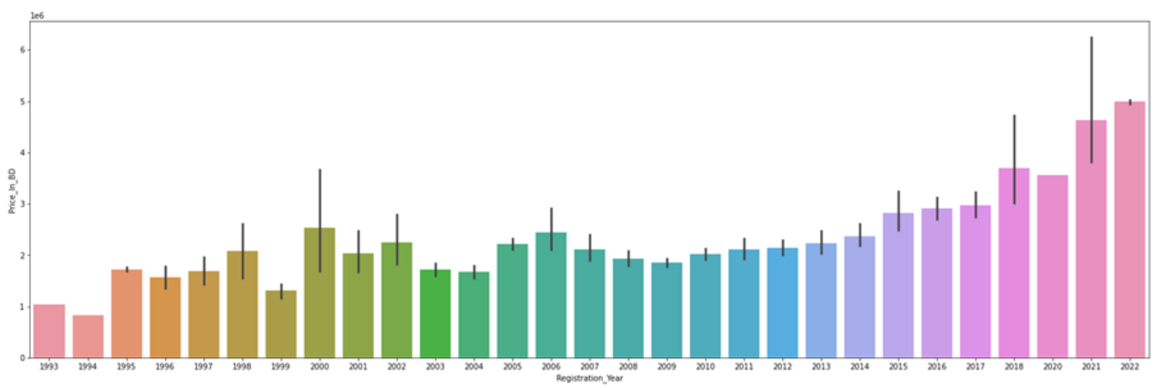


Fig 4.11: XGBoost Prediction Barplot

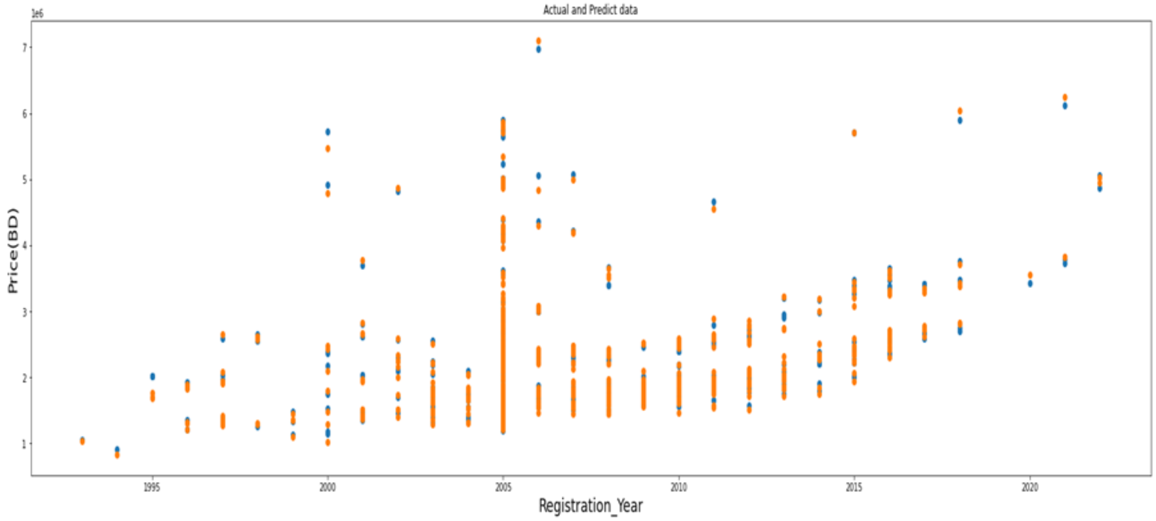


Fig 4.12: XGBoost Prediction Scatterplot

CHAPTER 5

RESULTS AND DISCUSSION

KNN,LR, LR, Decision Tree Regression, RF, XGBoost Regression are the 6 algorithms that we have implemented and tested on our collected dataset particularly for this study. The table stated below shows comparison among the performance metrices of these models:

Table 5.1: Regression Model Performance Comparison

Model	MAE	MSE	RMSE	R2	Accuracy
KNN	121002.1286	87991687487.756	45896.6589	0.7557	75.5773%
LR	154003.1186	57781687487.946	24896.6589	0.9136	91.3600%
LR	202004.1586	64991687487.849	254934.6729	0.9147	91.4716%
Decision Tree	50292.6066	44667877487.819	144934.6729	0.9942	98.0179%
Random Forest	2454371.2325	4059363456.235	100234.4576	0.9959	99.59%
XGBoost	47786.06225	24567575487.819	10024.4576	0.99497	98.85%

Comparing the results from the above table, it is clear that the XGBoost Regression performs most efficiently in terms of error and testing accuracy. This algorithm has shown enough promising performance to be selected as compared to the other tested models in this study. The predicted level against the actual levels of XGBoost can be visualized as below:

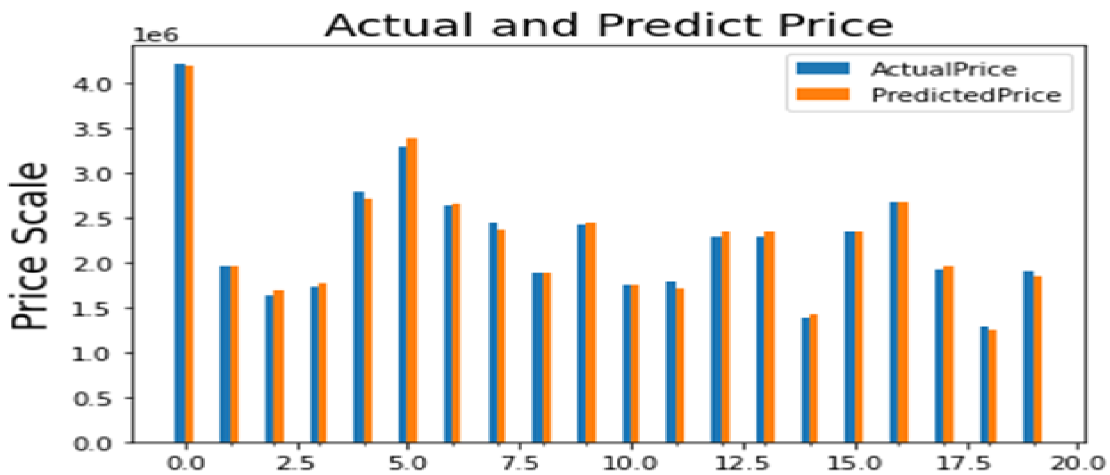


Fig 5.1: Predicted vs Actual levels of XGBoost Regression

The plotting of predictions by the XGBoost algorithm can be as followed:

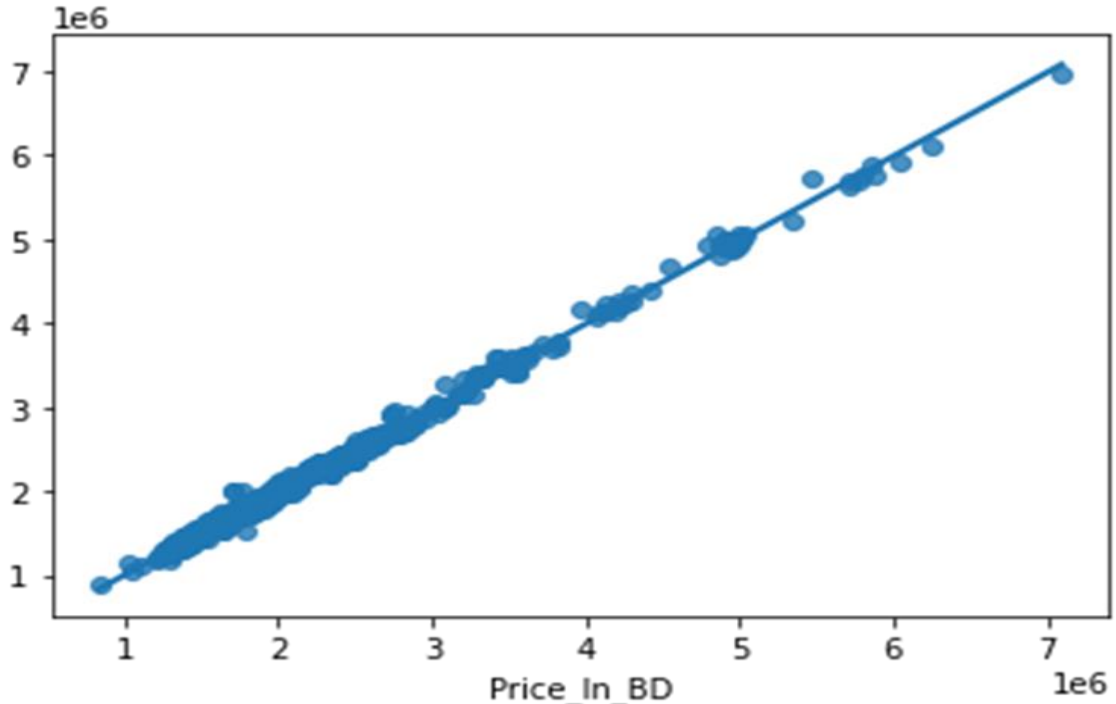


Fig 5.2: Prediction Plotting of XGBoost

Accuracy vs R2 graph is as followed:

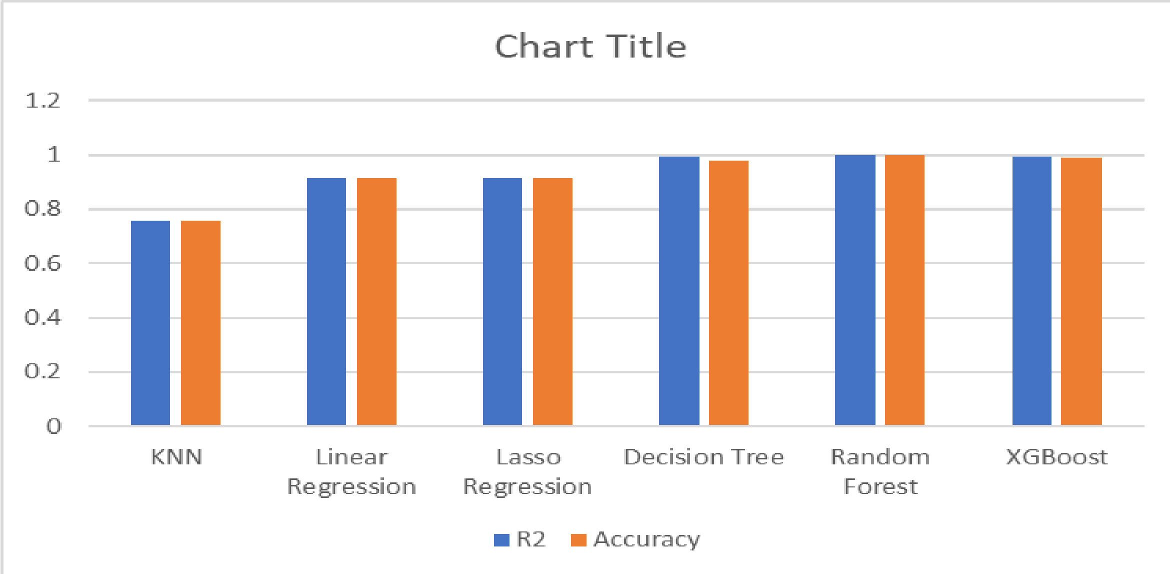


Fig 5.3: Accuracy vs R2 of the tested Models

CHAPTER 6

CONCLUSION AND FUTURE WORK

Predicting the price of used cars using ML can be a challenging task due to a number of limitations. One major limitation is the availability of data. The accuracy of a ML model depends on the quality and quantity of the data it is trained on. If the data used to train the model is incomplete, noisy, or biased, the model's predictions may not be accurate. Another limitation is the influence of market conditions. The price of used cars can be influenced by a wide range of factors, including market demand, economic conditions, and consumer preferences. These factors can change over time, making it difficult to predict the price of a used car with high accuracy. Additionally, unforeseen events such as accidents, natural disasters, or changes in government regulations can impact the price of a used car in unpredictable ways. Finally, human factors such as the seller's motivation and the buyer's negotiating skills can also influence the price of a used car, which may be difficult to capture in a ML model. Despite these limitations, ML can still be a useful tool for predicting the price of used cars, but it is important to be aware of these limitations and to use caution when interpreting the model's predictions.

In the future, there are several directions that could be taken to improve the accuracy and reliability of used car price prediction using ML. One possibility is to gather more data, including data on a wider range of makes and models, as well as data on market conditions, consumer preferences, and other factors that can impact the price of used cars. Another possibility is to develop new ML algorithms or to modify existing algorithms in ways that improve their ability to handle the complexity and variability of used car pricing. Additionally, it may be useful to explore the use of other data sources, such as social media or online marketplaces, to supplement traditional data sources and provide a more complete picture of the used car market.

In conclusion, predicting the price of used cars using ML can be a valuable tool for buyers, sellers, and other stakeholders in the used car market. While there are limitations to the accuracy and reliability of these predictions, advances in data gathering, ML algorithms, and other technologies have the potential to improve the performance of these models and make them more useful for a wider range of applications.

APPENDICES

Google Colab:

https://drive.google.com/drive/u/0/folders/1PXpIBGpQtRYKfxWvlOi9QFC5CADc8pKV

Google Sheet:

https://docs.google.com/spreadsheets/d/19-197rEY-tBy1UU1sKlNWdAZ3LdPbglNtC3OeSQOHKQ/edit?usp=sharing

REFERENCES

[1] S. Kumar Chandar, "Fusion model of wavelet transform and adaptive neuro fuzzy inference system for stock market prediction," Journal of Ambient Intelligence and Humanized Computing, 2019.

[2] K. Samruddhi; R. Ashok Kumar; "Used Car Price Prediction Using KNN Based Model", 2020.

[3] Kiran S; "Prediction of Resale Value of The Car Using LR Algorithm", 2020.

[4] Tao Wang; Songhua Hu; Yuan Jiang; "Predicting Shared-car Use and Examining Nonlinear Effects Using Gradient Boosting Regression Trees", INTERNATIONAL JOURNAL OF SUSTAINABLE TRANSPORTATION, 2020.

[5] Anu Yadav; Ela Kumar; Piyush Kumar Yadav; "Object Detection and Used Car Price Predicting Analysis System (ucpas) Using ML Technique", LINGUISTICS AND CULTURE REVIEW, 2021.

[6] Lingampally Shalini; Soumyalatha Naveen; U. M Ashwinkumar; "Prediction of Automobile MPG Using Optimization Techniques", 2021 IEEE MADRAS SECTION CONFERENCE (MASCON), 2021.

[7] Chejarla Venkat Narayana; Chinta Lakshmi Likhitha; Syed Bademiya; KarreKusumanjali; "ML Techniques To Predict The Price Of Used Cars: Predictive Analytics in Retail Business", 2021 SECOND INTERNATIONAL CONFERENCE ON ELECTRONICS AND SUSTAINABLE COMMUNICATION SYSTEMS (ICESC), 2021.

[8] Wenjie Lu; Jiazheng Li; Jingyang Wang; Lele Qin; "A CNN-BiLSTM-AM Method for Stock Price Prediction", NEURAL COMPUT. APPL., 2021.

[9] Sayed ErfanArefin; "Second Hand Price Prediction for Tesla Vehicles", ARXIV, 2021.

[10] Gegic, Enis, BecirIsakovic, Dino Keco, ZerinaMasetic, and Jasmin Kevric. "Car price prediction using machine learning techniques." TEM Journal 8, no. 1 (2019): 113.

[11] Pudaruth, Sameerchand. "Predicting the price of used cars using machine learning techniques." Int. J. Inf. Comput. Technol 4, no. 7 (2014): 753-764.

[12] Gajera, Prashant, AkshayGondaliya, and JenishKavathiya. "Old Car Price Prediction With Machine Learning." Int. Res. J. Mod. Eng. Technol. Sci 3 (2021): 284-290.

[13] Venkatasubbu, Pattabiraman, and Mukkesh Ganesh. "Used Cars Price Prediction using Supervised Learning Techniques." Int. J. Eng. Adv. Technol.(IJEAT) 9, no. 1S3 (2019).

[14] Al-Turjman, Fadi, Adedoyin A. Hussain, SinemAlturjman, and Chadi Altrjman. "Vehicle Price Classification and Prediction Using Machine Learning in the IoT Smart Manufacturing Era." Sustainability 14, no. 15 (2022): 9147.

[15] Bukvić, Lucija, JasminaPašagićŠkrinjar, Tomislav Fratrović, and Borna Abramović. "Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning." Sustainability 14, no. 24 (2022): 17034.

[16] Chandak, Ashish, Prajwal Ganorkar, Shyam Sharma, Ayushi Bagmar, and Soumya Tiwari. "Car Price Prediction Using Machine Learning." International Journal of Computer Sciences and Engineering 7, no. 5 (2019): 444-450.

[17] Samruddhi, K., and R. Ashok Kumar. "Used Car Price Prediction using K-Nearest Neighbor Based Model." Int. J. Innov. Res. Appl. Sci. Eng.(IJIRASE) 4 (2020): 629-632.

[18] Benabbou, Faouzia, Nawal Sael, and Imad Herchy. "Machine Learning for Used Cars Price Prediction: Moroccan Use Case." In International Conference On Big Data and Internet of Things, pp. 332-346. Springer, Cham, 2022.

[19] Lavanya, B., Sk Reshma, N. Nikitha, and M. Namitha. "Vehicle resale price prediction using machine learning." UGC Care Group I (2021).

[20] Kumar, Sachin, Damandeep Kaur, and Anjum Parvez. "Prediction of Prices Car Price Prediction with Machne Learning." In 2022 International Conference on Cyber Resilience (ICCR), pp. 1-4. IEEE, 2022.

[21] Satapathy, Santosh Kumar, RutvikrajVala, and Shiv Virpariya. "An Automated Car Price Prediction System Using Effective Machine Learning Techniques." In 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), pp. 402-408. IEEE, 2022.

[22] Siva, Ramgiri, and M. Adimoolam. "Linear Regression Algorithm Based Price Prediction of Car and Accuracy Comparison with Support Vector Machine Algorithm." ECS Transactions 107, no. 1 (2022): 12953.

PLAGIARISM REPORT

Plagiarism Report

ORIGINALITY REPORT

14%	11%	8%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
2	Supriyo Ahmed, Ripon K. Chakrabortty, Daryl L. Essam, Weiping Ding. "Poly-linear regression with augmented long short term memory neural network: Predicting time series data", Information Sciences, 2022 Publication	1%
3	www.researchgate.net Internet Source	1%
4	Chejarla Venkata Narayana, Nukathoti Ooha Gnana Madhuri, Atmakuri NagaSindhu, Mulupuri Aksha, Chalavadi Naveen. "Second Sale Car Price Prediction using Machine Learning Algorithm", 2022 7th International Conference on Communication and Electronics Systems (ICCES), 2022 Publication	1%
5	www.pomsmeetings.org Internet Source	1%

6	ntnuopen.ntnu.no Internet Source	<1 %
7	lingcure.org Internet Source	<1 %
8	Santosh Kumar Satapathy, Rutvikraj Vala, Shiv Virpariya. "An Automated Car Price Prediction System Using Effective Machine Learning Techniques", 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), 2022 Publication	<1 %
9	deepai.org Internet Source	<1 %
10	blog.quantinsti.com Internet Source	<1 %
11	www.ijeat.org Internet Source	<1 %
12	Seungwoo Jeon, Bonghee Hong, Victor Chang. "Pattern graph tracking-based stock price prediction using big data", Future Generation Computer Systems, 2018 Publication	<1 %
13	ieomsociety.org Internet Source	<1 %
	dokumen.pub	

14	Internet Source	<1%
15	repository.lib.fit.edu Internet Source	<1%
16	ebin.pub Internet Source	<1%
17	methods-sagepub-com-christuniversity.knimbus.com Internet Source	<1%
18	songhuahu-umd.github.io Internet Source	<1%
19	dblp.uni-trier.de Internet Source	<1%
20	Preeti Arora, Harshit Gupta, Ajit Singh. "Forecasting resale value of the car: Evaluating the proficiency under the impact of machine learning model", Materials Today: Proceedings, 2022 Publication	<1%
21	Yingxuan Chen, Weiwei Lin, James Z. Wang. "A Dual-Attention-Based Stock Price Trend Prediction Model With Dual Features", IEEE Access, 2019 Publication	<1%
22	"Table of Contents", 2021 IEEE Madras Section Conference (MASCON), 2021	<1%