CAR PRICE PREDICTION PROJECT

Submitted by:

VAISHNAVI

ACKNOWLEDGMENT

[1] Sameerchand Pudaruth, “Predicting the Price of Used Cars using Machine Learning Techniques”;(IJICT 2014)

[2] Enis gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric,” Car Price Prediction Using Machine Learning”; (TEM Journal 2019)

[3] Ning sun, Hongxi Bai, Yuxia Geng, Huizhu Shi, “Price Evaluation Model In Second Hand Car System Based On BP Neural Network Theory”; (Hohai University Changzhou, China)

INTRODUCTION

- Business Problem Framing

The impact of COVID 19 on the market has resulted in numerous changes in the automobile industry. Now, certain cars are expensive because they are in high demand, while others are less expensive because they are not. One of the clients deals with small merchants who market second-hand automobiles. They are having issues with their earlier machine learning models for valuing automotive prices due to the changes in the market brought on by the influence of COVID19. They are therefore searching for fresh machine learning models using fresh data.

- Conceptual Background of the Domain Problem

In 2020, the Indian used automobile industry brought in $18.3 billion, and between 2021 and 2030, it is anticipated to increase at a CAGR of 14.8%. The growing consumer focus on high-quality, reasonably priced vehicles and rising preference for imported cars are the key factors driving the market for used cars in the nation. Additionally, the increased quality of old cars following thorough servicing by the dealers is changing consumers' perceptions of how long they will survive.

The used automobile industry in India has been impacted by the COVID-19 outbreak, notably the numerous lockdowns that were implemented in its wake. Due to the pandemic's severe effects on market participants' businesses, workers no longer receive remuneration, and there has been a decline in global trade. In particular, consumer demand, communication, operations, and the supply chain have all been affected negatively by the outbreak. The economic downturn brought on by the coronavirus has also significantly decreased consumer

purchasing power, which has a negative impact on used automobile sales.

- Review of Literature

Predicting the Price of Used Car Using Machine Learning Techniques is the first study [1]. In this research, they look into how supervised machine learning techniques can be used to estimate the cost of used vehicles in Mauritius. The forecasts are supported by historical information gathered from daily publications. The predictions were made using a variety of methodologies, including multiple linear regression analysis, k-nearest neighbours, naive bayes, and decision trees.

Car Price Prediction Using Machine Learning Techniques is the second paper [2]. For the dependable and accurate prediction, a large number of unique attributes are considered. They have employed three machine learning approaches (Artificial Neural Network, Support Vector Machine, and Random Forest) to create a model for forecasting the cost of second-hand cars in Bosnia and Herzegovina.

The third paper [3] presents a second-hand car price evaluation model using BP neural networks. The price evaluation model based on big data analysis is put forth in this research. It makes use of widely disseminated vehicle data as well as a sizable amount of vehicle transaction data to evaluate the pricing data for each type of car using the BP neural network method that has been tuned. In order to determine the price that best fits the car, it attempts to build a model for evaluating used car prices.

Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

This project consists of two phases- Data Collection Phase and Model Building Phase. In Data Collection phase, the used cars data is scraped using Web Driver and the data is exported in CSV format. The data is again imported for data analysis and model development and testing which gives an efficient model to predict the price of used cars.

- Data Sources and their formats

For the data collection, data of 7980 cars are scraped from "car dekho" portal using Selenium Web Driver. The three parameters based on which the used cars data is collected are- Fuel, Kilometres-driven, Transmission and Price. After analysing the data, the following conclusions are made-

- The database dimensions are 7980 rows and 4 columns. Fuel, Transmission and Kilometers-driven are of object type hence Kilometers-driven column has to be converted to numeric type.
- There are no null values in the dataset.
- Kilometers_Driven denotes the number of kilometers the used car has driven before the sale.
- Fuel provides the type of fuel in the car, values for this feature are-"Petrol","Diesel","CNG".
- Transmission denotes the type of transmission of the car-Automatic or Manual.
- Price column gives the selling price of the used cars in lakhs.

When graphical analysis was performed on the features of the data, the following points were observed:

- The feature-Kilometer-driven is ranging from 5000 to 125000 and 1300-1400 cars have 30000 to 50000 kilometres-driven.

- 4500 cars have Petrol fuel and 3500 cars have Diesel fuel and very few cars have CNG type fuel.
- 6200 cars have Manual Transmission and 1600 cars have Automatic Transmission.
- 5500 cars have price from 2-5 lakhs followed by 1600 cars have price from 5-15 lakhs.

- Data Pre-processing Done

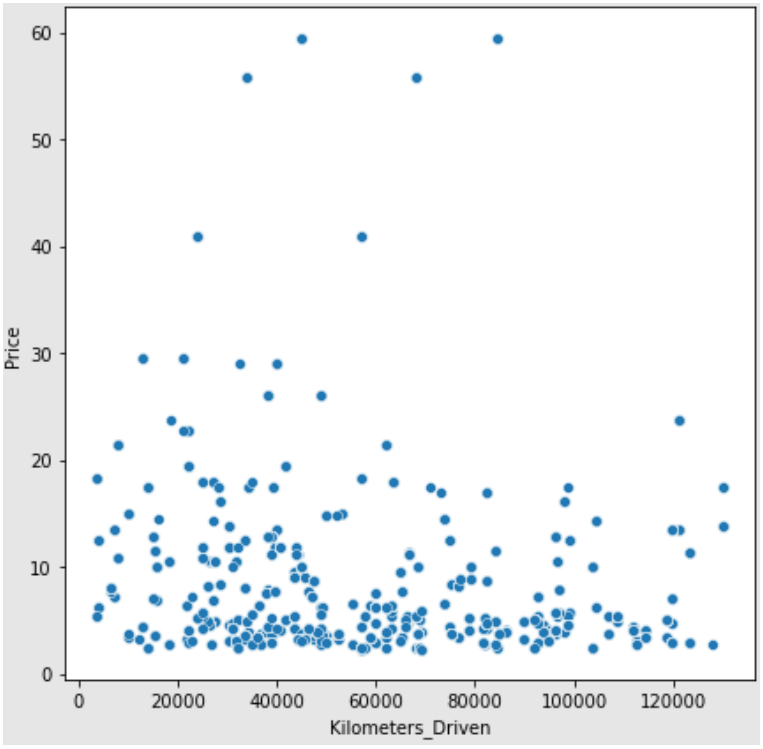
The column “Unnamed: 0” is removed since it replicates the index. “Kilometers_Driven” column consists of the distance covered by the car which is of numeric type but the type is of object which is converted to integer type.

“Fuel” and “Transmission” are of object type which are hence encoded using Label Encoder. There are few data points beyond the range as observed in the box plot. After determining the z-score for the dataset the data points of z-score beyond the value 3 are eliminated. The elimination of outliers gave rise to a data loss of 2.14%.

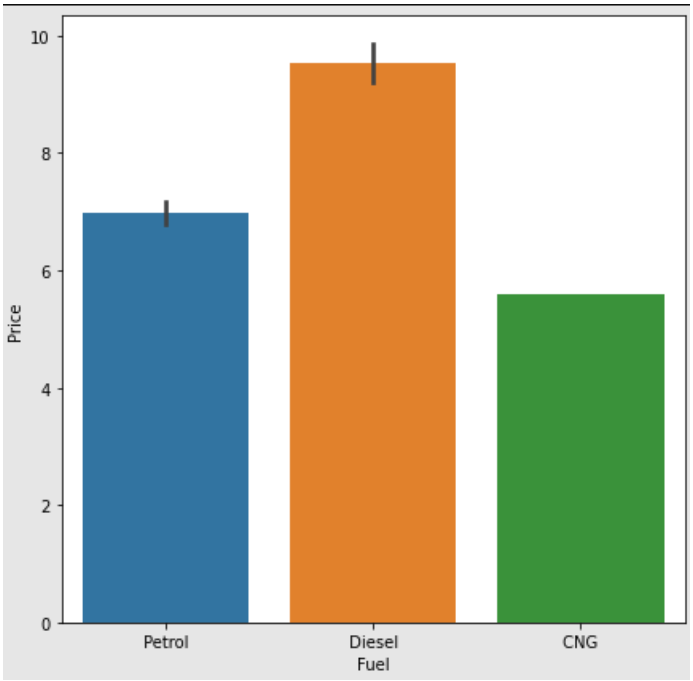
The features – Kilometers_Driven, Fuel and Transmission and target – Price are split from the dataset. The features are scaled using Min Max Scaler algorithm, this estimator scales and translates each feature separately so that it falls within the training set's defined range.

To give features a more Gaussian appearance, a class of parametric, monotonic adjustments called power transforms is used.

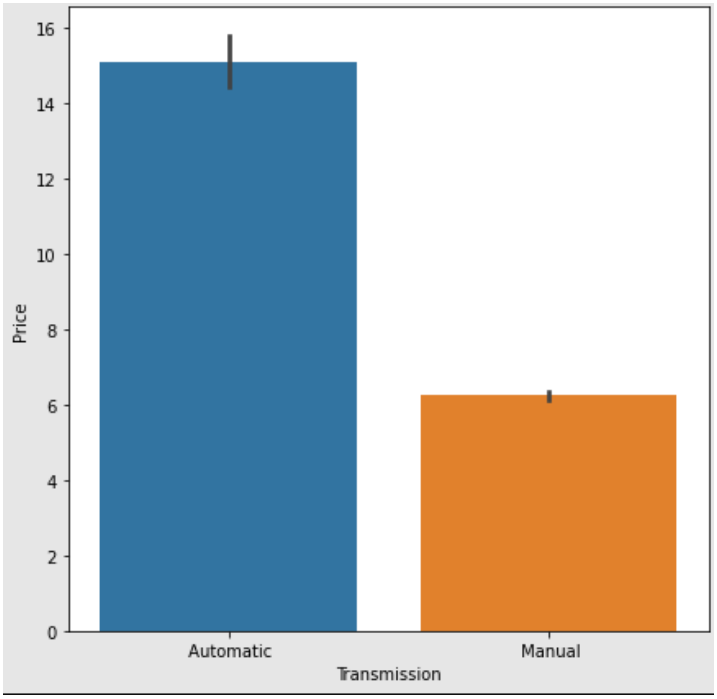
- Data Inputs- Logic- Output Relationships
  - When the car has less kilometres driven the price is higher.



  - With respect to fuel, used cars with Diesel fuel have higher price followed by Petrol and CNG.



- Automatic Transmission cars have higher price compared to Manual Transmission cars. Automatic cars have price of 15 lakhs and Manual cars have price of 6 lakhs.



- Hardware and Software Requirements and Tools Used

Hardware requirements:

- Operating system- Windows 7,8,10
- Processor- dual core 2.4 GHz (i5 or i7 series Intel processor or equivalent AMD)
- RAM-4GB

Software Requirements

- Python
- Pycharm
- Jupyter Notebook
- Chrome

1) Pandas- In pandas library the functions used are-
    - Pd.read function is used to load the data.

- The DataFrame's information is printed via the info() method. The data includes the total number of columns, their labels, data kinds, memory use, range index, and the number of cells in each column (non-null values)
- The DataFrame's data is described via the describe() method. If the DataFrame includes numerical data, each column's description will provide the following details, count is the total amount of non-empty values, mean – a measure of average value, the standard deviation, or std, min is the lowest value, 25 percent – the 25 percentile*, the fifty percentile is 50%, the 75% percentile* is at 75%, max is the highest value.
- The isnull() method returns a DataFrame object where all the values are replaced with a Boolean value True for NULL values, and otherwise False.
- Pandas DataFrame.dtypes attribute return the dtypes in the DataFrame. It returns a Series with the data type of each column.
- DataFrame from Pandas. The DataFrame's dtypes are returned via the dtypes attribute. A Series with the data types of each column is what is returned.

2) Python scripts can be used to create 2D graphs and plots using the Matplotlib module. With features to control line styles, font attributes, formatting axes, and other features, it offers a module called pyplot that makes things simple for plotting.
3) A package called Seaborn uses Matplotlib as its foundation to plot graphs. In order to see random distributions, it will be used.
4) A free machine learning library for Python is called Scikit-learn. It supports Python's NumPy and SciPy libraries as well as a number of methods, including support vector machines, random forests, and k-neighbors.

Model/s Development and Evaluation

- Testing of Identified Approaches (Algorithms)

  The features and target are split using the train_test_split function in sklearn, and a loop is defined to obtain the random state value for the function which yields best results.

  The algorithms which were used in training and testing the model are:

  1. Linear Regression
  2. Random Forest Regression
  3. Decision Tree Regression
  4. K Neighbours Regression
  5. Support Vector Regressor
  6. Gradient Boosting Regressor
  7. XGB Regressor
  8. Ada Boost Regressor

- Run and evaluate selected models

  When the features and target were split using the train_test_split function and results were analysed for different random state values, the following observations were made:

  - Linear Regression models had minor difference in r2 score of training and testing data. The r2 score for training data obtained is 27 and for testing data 29.
  - Decision Tree Regressor had very less difference of r2 score between training and test data. R2 score for training data is 70 and test data is 71.
  - Random Forest Regressor has same results as Decision Tree Regressor.
  - The difference of r2 score for K Neighbours Regressor is 4 between training and test data.

- Key Metrics for success in solving problem under consideration

The two metrics used in this project were R2 score and Mean Absolute Percentage Error. The R-Squared statistic in a regression model is a measure of how much variance in a dependent variable is explained by one or more independent variables.

One of the most used measures of model prediction accuracy is mean absolute percent error (MAE), which is also known as its percentage equivalent. A model's average magnitude of error (MAPE), or the average deviation from predictions, is measured.

- Interpretation of the Results

The Random Forest Model and Decision Tree Model are efficient models since Linear Regression and K Neighbours Models are having low r2 score for both training and test data.

CONCLUSION

- Key Findings and Conclusions of the Study

The Random Forest Model was hyper tuned using Random Search algorithm and an improvement of 0.17 was found in r2 score of train data.

- Limitations of this work and Scope for Future Work

The model is developed for the data consisting of 3 features, if more features are implemented to the dataset while data collection using web scraping the model development will be more efficient.