BRAC UNIVERSITY

Inspiring Excellence

Project Name: **Used Vehicle Price Prediction using Machine Learning Algorithms**

Submitted by

Shafin Ahad Siam
ID- 19301013

Nabil Hasan
ID- 19301222

Shihab Sharar
ID- 19241007

Group - 05

- ## Introduction

Vehicles or automobiles are now a crucial part of our everyday lives. It is afterall a important invention which allows us to move from one place to another quickly and effectively. It is hard to imagine the modern world without the existence of vehicles. As the world becomes more technologically advanced, people are becoming more reliant on automobiles. In 2022 it is estimated that there are almost 1.446 billion cars in the world. As people buy new vehicles there is also a need to sell old or used vehicles. And so there exists a huge market of buying and selling used cars and vehicles.

The reselling price of used vehicles depends on various criterias. The different features of the car are analysed and depending on their quality the price is set. So as all the different features of a vehicle need to be considered, it may be quite difficult and tiring to manually investigate each and every feature and set a price accordingly. So here comes the use of Machine Learning. By the use of machine learning we can train models that will help automatically predict the reselling price of the car quite accurately and with less hassle. It will be quite beneficial for the buyer and seller alike. So here we have such a dataset which contains various features of different models of used cars, and by the help of this dataset we will try and train some price prediction models using some machine learning algorithms.

- ## Methodology

➔ Dataset Description-

The dataset contains information containing the name, year, selling price, present price, distance driven, fuel type, seller type, transmission and owner of different models of used cars. There are 4341 rows and 9 columns in the dataset. A brief description of each of the columns is given below.

Car_Name - Name of the cars

Year - Year of the car when it was bought

Selling_Price - Price at which the car is being sold

Present_price - Current price of the car

Kms_Driven - number of kilometres the car is driven

Fuel_Type - Fuel type of car(petrol/diesel/CNG/LPG/electric)

Seller_Type - Tells if a seller is individual or Dealer

Transmission - Gear transmission of the car (Automatic/Manual)

Owner - Number of previous owners of the cars

➔ Pre-Processing Techniques applied -

In order to process the dataset at first we categorised the columns of the dataset into numerical and categorical data. This helps us to identify data of which columns need to be encoded numerically. We used .head() and .shape in order to get an overview of the dataset. Methods like .info and .shape allowed us to get a more in detail view of the data. We used the .isnull().sum() to look for any irrelevant null data present in any of the columns.

After that we found out the age of each car by subtracting the “Year” column data of each car from the current year which is 2022. So Age of each car = 2022 - “Year” of each car. After that we dropped the “Year” column from the dataset and replaced it with the newly found “Age” column. This helps us to understand the usage age of each car more accurately.

We furthermore used the value_counts() function to check the distribution of each categorical data. After that we used the .corr() function to check out the correlation between the datas in the columns. By importing the seaborn library we were able to plot a heatmap which allowed us to visualise the correlation between the datas even more clearly.

In order to make our dataset more easier to read we encoded the categorical datas into simple integer values so that they can be more easily distinguished. For example,for the fuel type column ‘Petrol’ was encoded as 0, ‘Diesel’ as 1 and ‘CNG’ as 2.

We also split the original dataset into training and testing datasets. We dropped the columns “Can Name” and “Selling Price” from the original dataset to create the input dataset and just used the “Selling Price” column to create the output dataset.

The input and output dataset were then further split into some ratio to create the test and training dataset.

So in summary the techniques applied were:

- .head( )
- .shape
- .info
- .describe
- .isnull( ).sum( )
- Calculating “Age” column and replacing the “Year” column
- .value_counts( )
- .corr( ) along with heatmap
- Encoding the categorical datas

➔ **Models applied -**

☐ **Decision Tree Regression -**

Decision tree is one of the most used practical approaches, which can be used both for regression and classification models as a form of tree structure. It is a tree-structured classifier with three types of nodes. The initial node is the root node which represents the whole sample and will get splitted into further nodes. Then, the Interior Nodes represent the features of a data set and the branches represent the decision rules. Finally, the leaf nodes represent the outcome or the result we are looking for. This algorithm is very useful for solving decision-related problems. For a particular data point, it runs completely through the entire tree by answering true or false till it reaches the leaf node. The final prediction which will be the result, is the average of the value of the dependent variable in that particular leaf node. It needs multiple iterations in order to reach a proper value for prediction. We are using the decision tree regression because the data we are using are continuous and not categorical. The problem that we are trying to solve, such as the price of a vehicle, is a regression problem and decision tree regression is good for this type of problem.

☐ **Polynomial Regression -**

Polynomial regression is a special multiple linear regression. It is a statistical method of determining the relationship between an independent variable (x) and a dependent variable (y) and model their relationship as the nth degree polynomial. Here, the independent variables do not depend on each other and the errors are independent, normally distributed with mean zero and a constant variance. With the help of a polynomial equation the relationship of the independent and dependent variable on the graph looks like a curvilinear relationship. Polynomial regression is used when there is no linear correlation fitting all the variables. So instead of looking like a line, it looks like a nonlinear function. Also, polynomial regression has the best approximation of the relationship between the dependent and independent variable. We have used polynomial regression rather than a linear regression because from the dataset when we plot the graph line it is not straight. In the polynomial regression model the scatterplots are scanned for a pattern and a line is drawn according to that pattern of the points. Also, polynomial regression does not make it mandatory for the data to have a linear relationship between them and because of that linear regression can not determine a linear relationship between variables but polynomial regression does. That is why we used polynomial regression that has a more accurate result than linear regression.

□ **Random Forest Regression -**

Random forest is a machine learning technique which builds multiple decision trees and merges their predictions together to get a more accurate and stable prediction than individual decision trees.It uses a statistical technique called bagging. Bagging process is the combination of two processes, bootstrapping and aggregation. Each tree in the random forest learns from a random sample of training observations. The samples are drawn with replacement which is known as bootstrapping. The predictions obtained from the trees generated from these different samples are then combined together which is known as aggregation. The main idea behind this is by training each individual tree in the forest with different samples, the entire forest overall will have a lower variance. Also even though the results of each tree may not be accurate, their combined average result will be more accurate. Averaging the results of individual trees also helps in reducing overfitting. Even though random forest can be used in both classification and regression, in our case we use the random forest regression as predicting selling price is a continuous data and not categorical and so we must use regression models.

- ## **Results**

We used some metrics to define our results. A brief description of them is given below:

Lastly we defined some metrics that we used in the algorithms so that new people who access the code can easily understand the metrics.The defined metrics were:

R2 score:- The r2 score is a very important metric that is used to evaluate the performance of a regression-based machine learning model. It is pronounced as R squared and is also known as the coefficient of determination. It works by measuring the amount of variance in the predictions explained by the dataset. It basically tells us how well our model fits in the dataset. A good value of r2 score varies from dataset to dataset.

Mean Absolute Percentage Error (MAPE):-

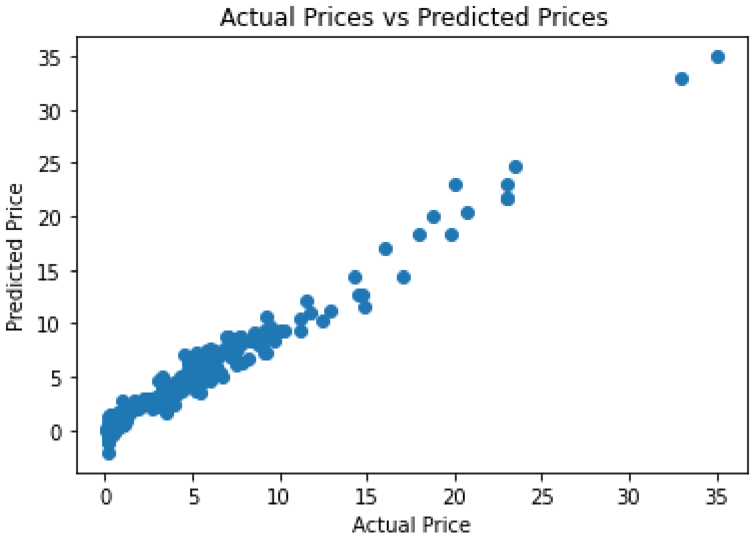
It is given by the formula:

mape = np.mean(np.abs((Y_actual - Y_Predicted)/Y_actual))*100

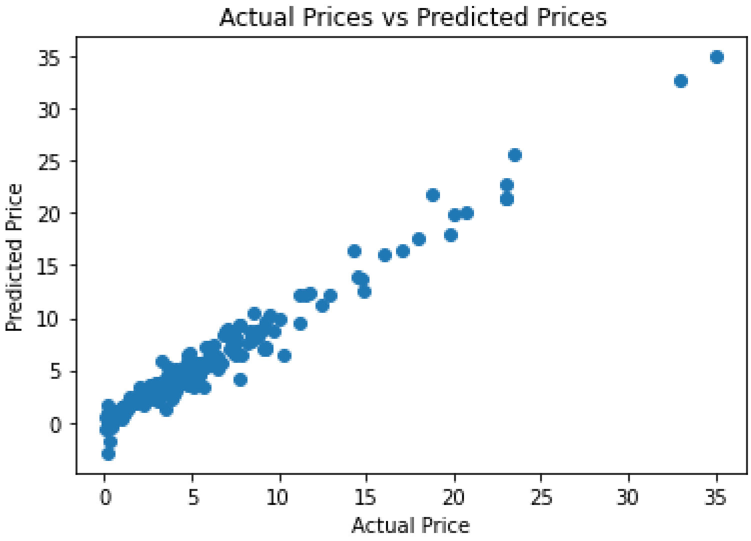
Results for Polynomial Regression:

For degree 2:

```
R squared error: 0.9716718364782485
 MAPE based accuracy = 35.70518406022421
```

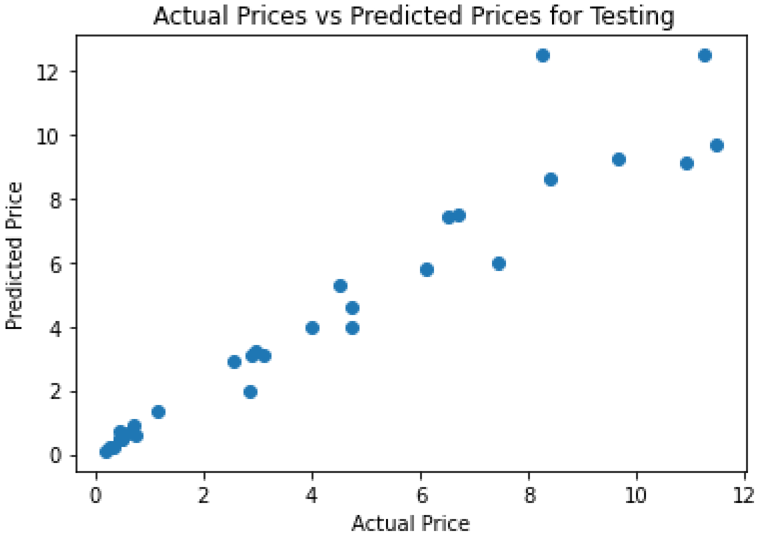


```
For degree 3:
     R squared error for degree 3: 0.9705915386501524
MAPE based accuracy for degree 3 = 37.258656840180095
```



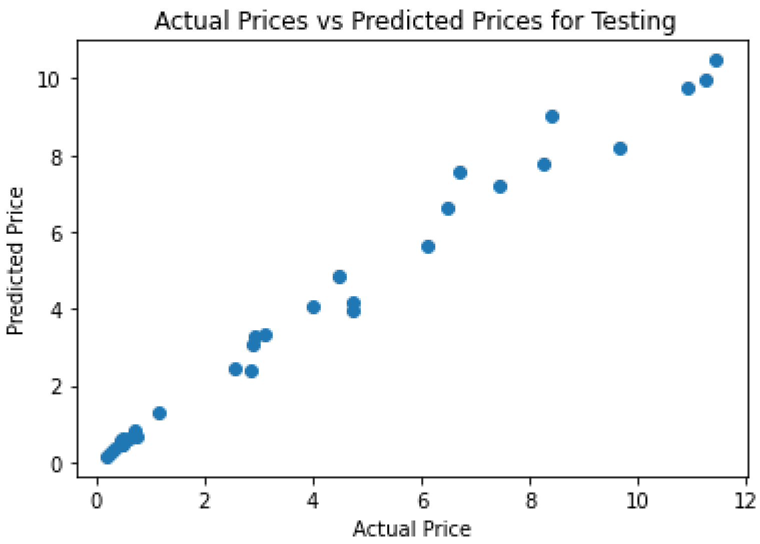
Results for Decision Tree Regression:

```
R squared error : 0.9212538394035072
Training Accuracy:  100.0 %
Testing Accuracy:  92.13 %
MAPE based accuracy = 42.80645161290322
```



Results for Random Forest Regression:

```
R squared error : 0.9828055582474547
Training Accuracy:  98.78 %
Testing Accuracy:  97.79 %
MAPE based accuracy = 63.25129032258066
```



- **Reference -**

1. https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho?select=car+data.csv
2. https://serokell.io/blog/polynomial-regression-analysis
3. https://www.voxco.com/blog/polynomial-regression-everything-you-need-to-know/
4. https://www.quora.com/How-does-random-forest-work-for-regression-1
5. https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda
6. https://www.youtube.com/watch?v=v6VJ2RO66Ag&t=307s
7. https://www.youtube.com/watch?v=ZVR2Way4nwQ&t=379s&ab_channel=NormalizedNerd
8. https://www.youtube.com/watch?v=Y90NTNG_yJg&t=201s&ab_channel=5MinutesEngineering