ANALYSING THE SELLING PRICE OF USED CARS USING PYTHON

Author: DEBANJAN DUTTA

This project is about Analysing the selling price of used cars using Python. In this project, we have gone through the process step by step, including data collection, data preprocessing, exploratory data analysis (EDA), feature engineering, model selection, training, and evaluation.

Data Collection:

For this project, we have used the dataset containing information about used cars and their selling prices. The dataset under study, is a collection of information about used cars that aims to facilitate analysis and insights into factors influencing the pricing of second-hand vehicles. It contains a variety of features related to the cars' specifications, such as brand, model, year of manufacture, kilometres driven, fuel type, transmission type, engine specifications, power output, and the selling price.

This dataset consists of 24,199 used vehicles listed for sale in 2023. These vehicles were listed for sale within 25KM proximity of downtown Toronto, Ontario, Canada. The vehicles were found for sale on Autotrader.ca. This dataset consists of 24,199 used vehicles listed for sale in 2023. These vehicles were listed for sale within 25KM proximity of downtown Toronto, Ontario, Canada. The vehicles were found for sale on Autotrader.ca.

The columns consist of:
-Year
-Make
-Model
-Kilometres
-Body Type
-Engine
-Transmission
-Drivetrain
-Exterior Colour
-Interior Colour
-Passengers
-Doors
-Fuel Type
-City

-Highway
-Price

The dataset is aimed at providing insights into the factors that affect the pricing of used cars. This can include understanding how features like the car's age, brand, fuel type, engine specifications, and more impact its resale value. By analysing this dataset, we can uncover relationships between these features and the selling price, and potentially build predictive models that estimate the price of a used car based on its attributes.

The dataset's dimensions and diversity of features make it suitable for various types of analysis, including exploratory data analysis, regression modelling, and feature importance assessment. It also offers opportunities to experiment with advanced machine learning techniques to create more accurate predictive models for used car prices.

Data Preprocessing:

In this step, we have loaded the dataset. We observed that there are various missing values and hence we handled those missing values in various ways for various feature variables. For example, in the 'Kilometres' column, the missing values are filled with the mean of that column for each corresponding 'Year' group. Another example is for the 'Body Type' variable, where the missing values are filled with the assumption that vehicles of the same 'Model' are likely to have the same 'Body Type'. Similarly for different variable, based on the data, various methods have been employed to fill out the missing values.

After cleaning the data, we have dropped the unnecessary columns.

Exploratory Data Analysis (EDA):

We have explored the dataset to understand its structure, summary statistics, and relationships between features.

We have performed 'Correlation Analysis' and have obtained from the heatmap that '*City', 'Highway',* and '*Cylinder'* have moderate correlation with price, while kilometres have moderate negative correlation with priceWe have performed Bivariate Analysis as well as checked the normality of the variables which have moderately high correlation with our target variable, 'Price'.

It is worth noting that the 'City' variable almost follows a Normal Distribution, as is evident from the plotted Q-Q Plot.

Observing the presence of many outliers, we have removed the existing outliers from the data to gain a better understanding of the data.

Feature Engineering:

We have encoded the categorical variables by the method of one-hot encoding and converted our categorical variables into numerical format so that our Machine Learning algorithms can use them.

Modelling:

We have chosen various regression models model our data. They are:

1. Lasso Regression
2. Decision Tree
3. Random Forest Regression
4. Support Vector Regression

For the better fitting and evaluation of the model, we have divided the entire data into *Training Set* and *Test Set* with training set as *70%* of the total data. After training the model based on the training set, we used the model to make the required predictions and evaluated how good the model is performing based on the test set.

Model Evaluation:

For the purpose of evaluating how well the model is performing, we have used 3 measures. Those measures are:

- MSE (Mean Squared Error)
- RMSE (Root Mean Squared Error)
- R2 (Co-efficient Of Determination)

Here is table for the comparative study of the various regression models used in this project.

Measure Regression Model	MSE (Mean Squared Error)	RMSE (Root Mean Squared Error)	R2 (Co-efficient Of Determination)
Lasso Regression	0.245	0.494	0.535
Decision Tree	0.056	0.235	0.895
Random Forest Regression	0.031	0.176	0.941
Support Vector regression	0.067	0.257	0.874

Interpretation and Conclusion:

From the analysis done in this project, it can said that as far as predicting the price of sale price of used cars are concerned, the **performance of *Lasso regression*** is ***not satisfactory***, whereas, ***Random Forest Regression*** is ***performing the best*** among the 4 regression models. In regards to MSE value, Random Forest Regression has the lowest value, similar is the case for RMSE, where Random Forest Regression has the lowest value and when comparing with help of the value of co-efficient of determination, it has the highest value (= 0.94 approximately), which shows that in the light of the given data, our fitted Random Forest Regression Model performed quite good.