Cars Prices Prediction Using Machine Learning Algorithms

Prepared By: Ali Ali

Introduction

- In this project we will attempt to predict cars prices using different machine learning algorithms on a properly structured cars prices data-set.
- First we will Do Exploratory Data Analysis (EDA) and Data Cleaning on the data-set.
- After that we will perform data processing and modeling on the data-set.
- Finally discussing the results of the models.

EDA And Data Cleaning

- After reading the data we look to the data generally in porpoise to know about the size, the data types of columns and the columns names also which is our features.
- Checking where the nulls are and the types of data of the columns to see if it is an appropriate data type for those values.
- Also editing the content of some columns to be more flexible to use the data in it.

EDA And Data Cleaning

- I dropped the ID column from the data-set because an id of a car is just a label of this car and doesn’t affect its price.
- I got more insight of the data by knowing each value how many times is repeated as a number and how many different values are there.
- Then I got the essential statistical information about the data which gives a high importance vision about the data.

EDA And Data Cleaning

- Visualizing the data with Uni-variate Analysis - By Box-plot which help detecting the outliers.
- The comparing between median and mean gives us an intuition about the distribution of the data.
- When the mean is bigger than the median that tells us that the values to the right of the median is bigger which is actually makeing the mean bigger than median also, otherwise when median is bigger.

EDA And Data Cleaning

- Visualizing the data with Bi-Variate Analysis by pair-plot which gives us a relationships between the features.
- Visualizing the correlation coefficient between features which clarify how and how much one feature affects another one and describe the relation between them.
- Visualizing the relationships between features using scatter matrix plot.

Data Processing

- There are some values doesn’t make sense usually be too high or too low which we call outliers.
- dealing with outliers and eliminate it.
- Also sorted the cars from the most frequent to the least frequent manufacturer and the mean its prices which gives us what is affordable and which is the expensive through all the manufacturer.

Data Processing

- Calculating essential statistical values of the some features with the price like the mean of the prices of cars that have turbo or not and so on.
- Visualizing histograms of the features to see the distribution of the data and visualizing scatter matrix after eliminating the outliers.
- Also I recalculate the correlation coefficient matrix after eliminating the outliers.

Data Processing

- Erase the price column from the data preparing it for training.
- Then use simple Imputer to fill in the nulls with median also use the one hot encoder to symbolize the categorical features.
- Use feature scaling to make the algorithms smoother and faster.
- Splitting the data as train data and test data.

Model Setup, Tuning and Evaluation

- Applying Linear Regression on the data displaying root mean square error (rmse) and mean absolute error (mae) to see the algorithm performance.
- It was a not that good with an error about 10K so decision tree have been applied
- It seems that it has over-fitting with a low error so we have created a cross-validation sets.
- Also (rmse) and (mae) have been calculated.

Model Setup, Tuning and Evaluation

- The model have done better but also the error is a bit high and we will try to decrease it by applying Random Forest.
- A final error was acceptable with 5.8K and a confidence interval with confidence of 0.95 as the error is between [5.5K – 6K] which means that 95% the error is in the above interval.
- We calculated features importance and displayed it.

Model Setup, Tuning and Evaluation

- In order to decrease over-fitting and training time we can eliminate the features that don’t add more information.
- Applying Random forest after feature selection.

Findings And Result

- Random forest is the most efficient algorithm to this problem because it have the best error also it avoids over-fitting unlike the decision tree algorithm.
- Linear Regression had a considered error and it is not the best algorithm for this problem.
- Using features which is the most important is better.
- Outliers should be eliminated to do better.