NAME OF THE PROJECT

Car Price Prediction

Submitted By:

Arpan

Acknowledgement

This data is scrapped from the website named ‘OLX’. This scrapping is done according to the instructions given by the company under which my internship is going on. I just used a library named ‘Selenium’ in python notebook for scrapping the data from the website. No any other source I have used for scrapping the data.

Business Problem Framing:

According to demand in market, prices of car go up and down. Price also depends on the condition of vehicle and place from where we are purchasing. Keeping this in mind, we will build a machine learning model. With the help of model, anyone can be easily get, how the vehicle price is varying.

Conceptual Background of the Domain Problem:

We will do this project in python notebook using different libraries. We will use PANDAS, NUMPY, SEABORN, MATPLOTLIB, REGRESSION models, and regression related metrics, and many more libraries are there which are helpful for us during this project.

Review of Literature:

During this project, first of all we scrapped the data using 'selenium' in python notebook. Then, stored the scrapped data into 'csv' file. After that opened dataset into python notebook for model building process using different libraries.

Motivation for the Problem Undertaken:

We did this project in python notebook. Because python is very smooth and can easily understandable. Main purpose for building this model is that prices of cars go up and down according to the demand in market, if anyone wants to go in cars sell-purchase business then this project will help surely.

Analytical Problem Framing

Mathematical/Analytical Modelling of the Problem:

The data we have scrapped is a new data. All columns are having proper entries. Of course, there are some nulls present in columns, but are very little which can be easily manipulated. But whole dataset is in object datatype, so we encoded whole dataset. But we did not touch our target column only encoded it.

Data Sources and their Formats:

Data has scrapped from website named 'OLX'. Data has scrapped with the help of 'selenium' using python notebook. Data has stored in csv format. Four columns have scrapped because these four are most important for data processing.

Data Preprocessing Done:

Scrapped data is in the csv form. Data is present as object datatype. We encoded the data. Some nulls are also present. We treated with nulls. Checked, how data is distributed using distribution plot. We checked for outliers using boxplot and with the help of quantiles. Then plot heatmap to check multicollinearity problem.

Hardware and Software Requirements and Tools Used:

During this project we used our local machine. Anaconda Navigator in which python notebook is used to complete this project. Tools which are used during this project are pandas, numpy, ordinal encoder, matplotlib, seaborn, heatmap, standard scaler, train test split, regression models and metrics for regression model for checking error of models.

Model/s Development and Evaluation

Identification of possible problem-solving approaches:

Because it is a very simple dataset. This dataset has continuous data, so simply we apply regression models during the completion of this project.

Testing of Identified Approaches:

First – Linear regression

Second – Decision tree

Third – Random Forest

Fourth – SVM

Fifth – Ada Boost

After applying these algorithms we did hyperparameter tuning for two models i;e Random forest and Ada Boost.

Run and Evaluate selected models:

All models are giving low score on test data except Random Forest. When we apply hyperparameter tuning to Random Forest, its accuracy goes down, we tried with different parameters but goes down. So, we consider only Random forest for saving.

Visualizations:

First – we plot distribution plot to have a look how data is distributed.

Second – we plot boxplot to check outliers.

Third – we plot scatter plot to check how data is scattered.

Conclusion

We applied five algorithms to given dataset. But only one of five performed good. Training score was also good and test score was also nearabout fifty percent. So, we considered only it. If some more columns contribute in dataset, then it makes more interesting to complete this project.