

CAR PRICE PREDICTION

Submitted by:

REBANTA DUTTA

**ACKNOWLEDGMENT**

I would like to acknowledge some of the websites from where I have taken help:

* <https://www.analyticsvidhya.com/blog/2022/02/a-quick-guide-to-bivariate-analysis-in-python/>
* https://www.geeksforgeeks.org/seaborn-barplot-method-in-python/

**INTRODUCTION**

* Business Problem Framing

The impact of COVID-19 has heavily affected the car market. Some cars are in demand and therefore the sellers have increased their price. As a result, the traders are facing problem with valuation of the car and thus, a new model needs to be designed with new data.

* Conceptual Background of the Domain Problem

The price of the cars depends on a number of variables or factors like the brand, year of manfacture, kilometres driven, accidental history, fuel and the type of transmission

* Review of Literature

The COVID-19 has affected all industries and second hand car market is one of the industries. The traders in the industry are facing challenges with valuations of the cars due to rise in demand for some cars. Therefore, with data and analytics they are trying to properly make valuation of a car.

* Motivation for the Problem Undertaken

Selling of second hand cars is an important part of revenue generation, which contributes to the economy. Further it associates multiple other industries and businesses like car garages, mechanics, and spare parts.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

The data was scrapped from cars24 using selenium tool and it was imported in Jupyter notebook. Further, the data was stored as a dataframe and using plots and charts like bar plot, distribution plot, each column was analysed. Further, bar chart helped to under the relationship with target variable price. After this, log1p() was used for removal of skewness and through use of zscore, the outliers were removed.

* Data Sources and their formats

The dataset contains 10 rows and columns.

Brand : object

model : object

year : int64

history : object

ownership : object

km\_driven : int64

fuel : object

transmission : object

location : object

price : int64

* Data Preprocessing Done

After loading the dataset, the km\_driven was scrapped as string from the website cars24. I have used the pd.to\_numeric function to convert it into integer type. Following this, zscore technique was used for removal of outliers, and skewness was removed.

* Data Inputs- Logic- Output Relationships

The heatmap was used to visualise the relationship between the independent variables and the target variable. It was seen the price variable was positively and negatively correlated with some of the independent features.

* Hardware and Software Requirements and Tools Used

The Jupyter notebook was used for analysis and building of the machine learning model. In addition to this different libraries were used like the matplotlib, seaborn, pandas and numpy.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

After identification of object type data, I have used Label encoding technique to convert it into numeric type

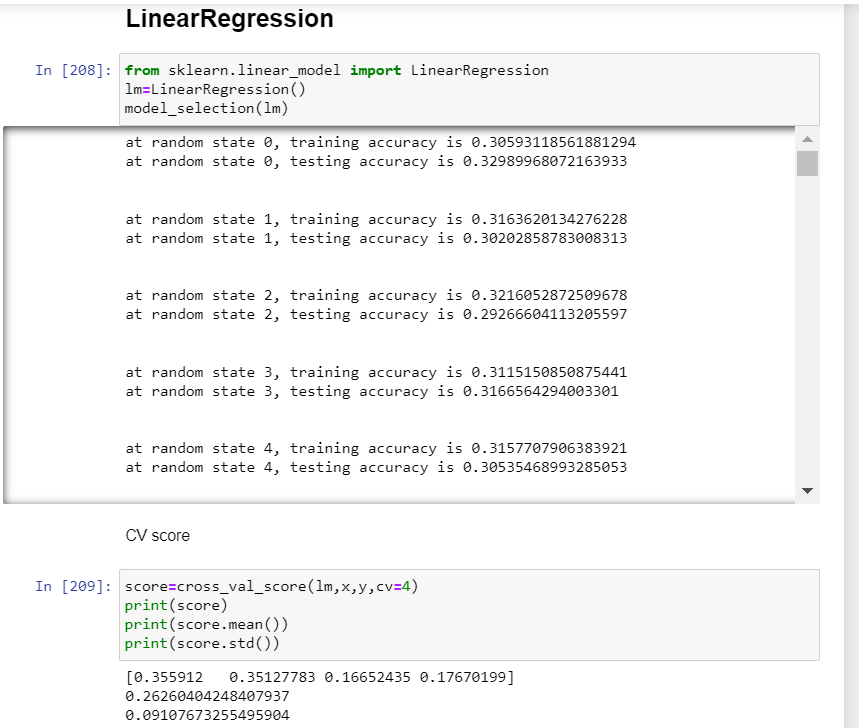
* Simple imputation was used for removal of null values
* Label encoding technique was used to convert object type data into numeric type.
* Skewness was controlled using log1()
* Testing of Identified Approaches (Algorithms)

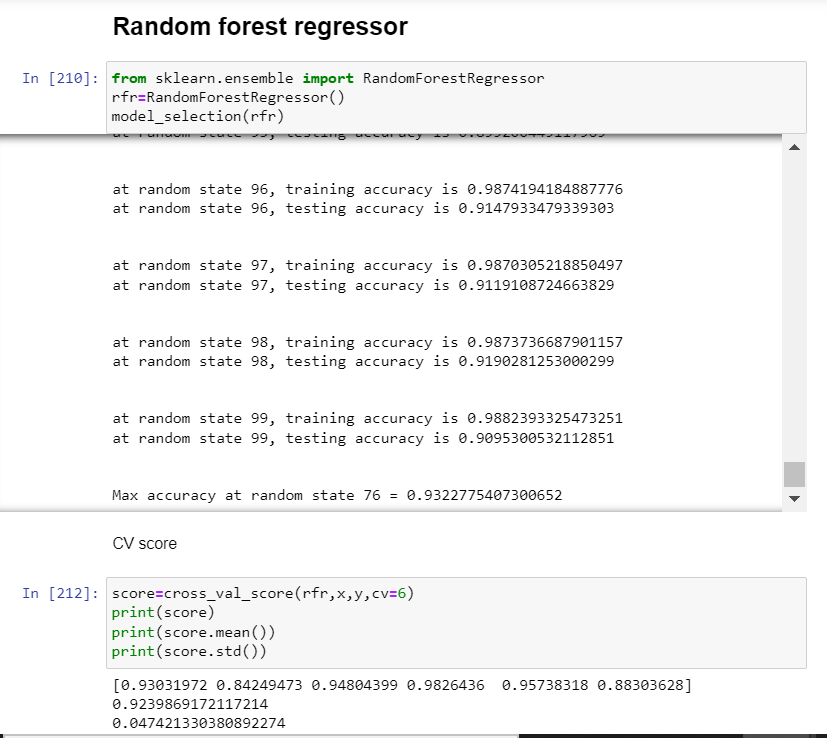
Linear regression

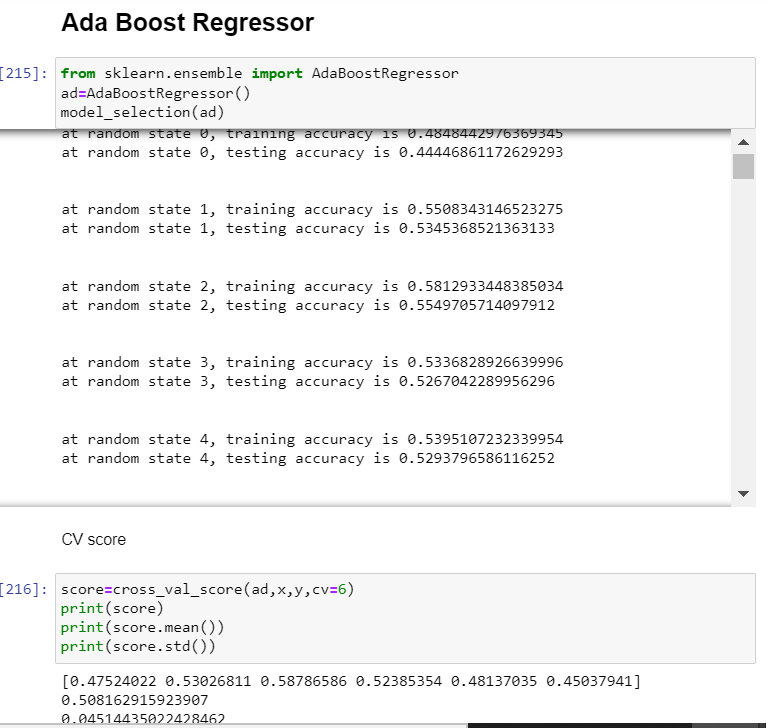
Random forest regressor

Ada Boost Regressor

* Run and Evaluate selected models





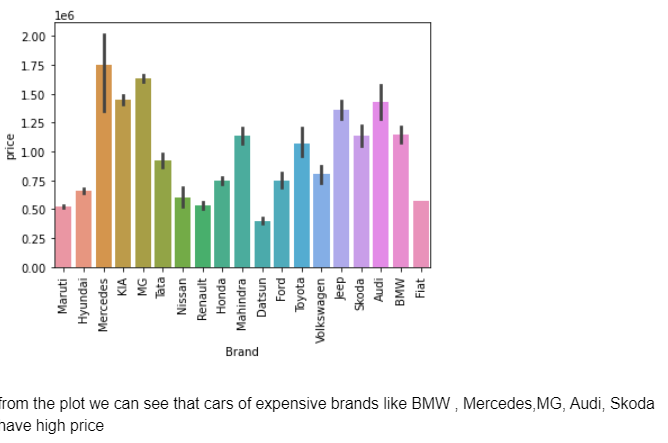


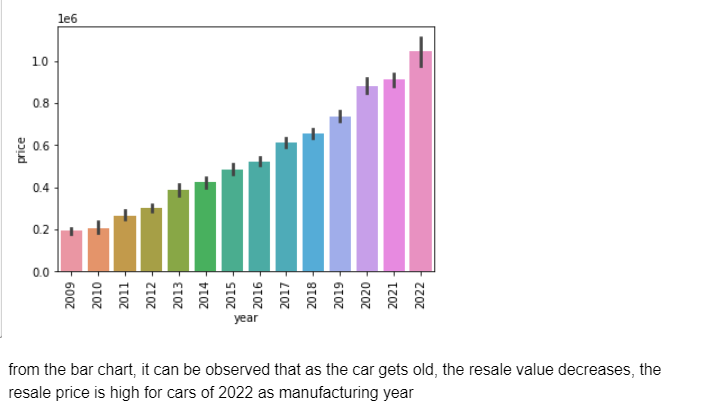
* Key Metrics for success in solving problem under consideration

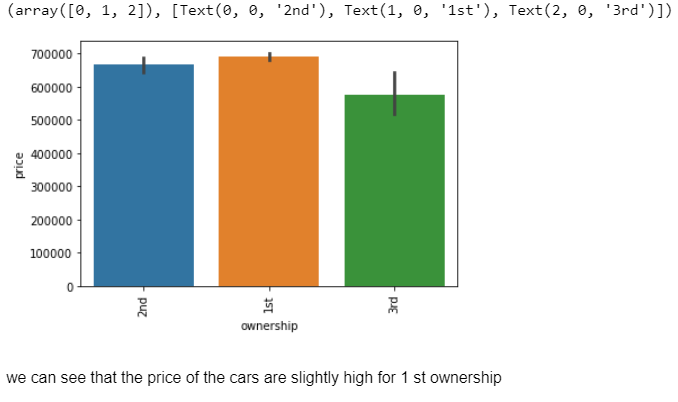
The key metrics that have been used are r2score() and mean absolute error.

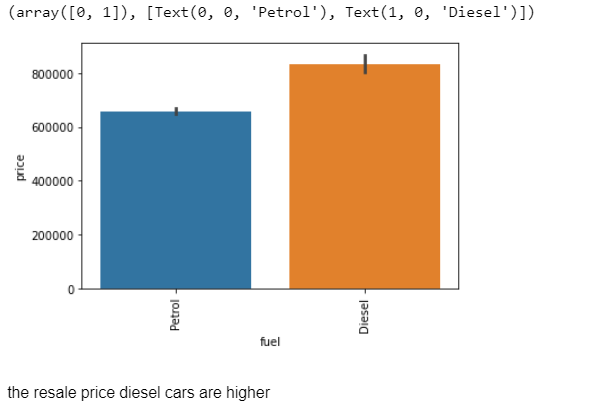
* Visualizations

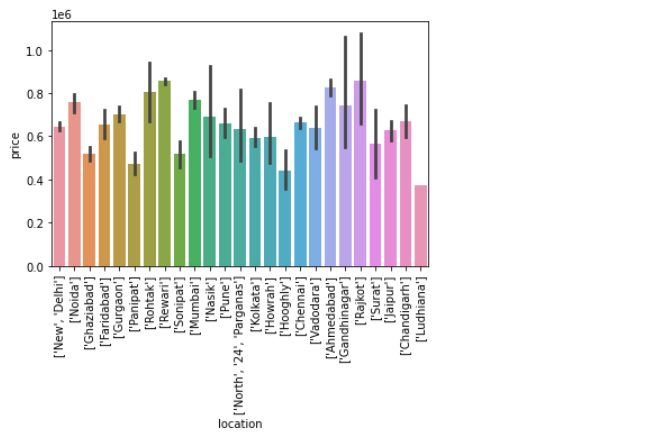
***Barplots***

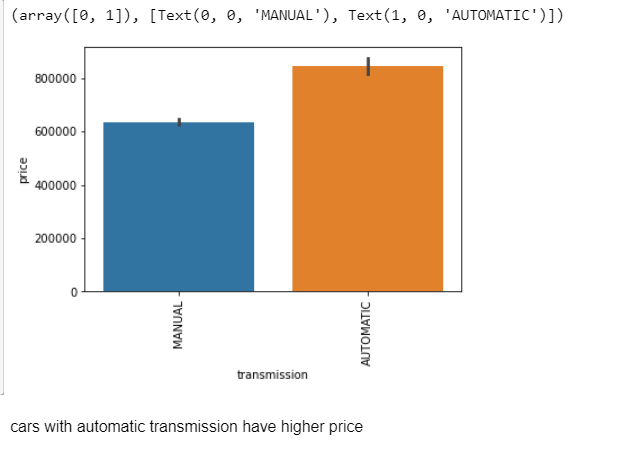


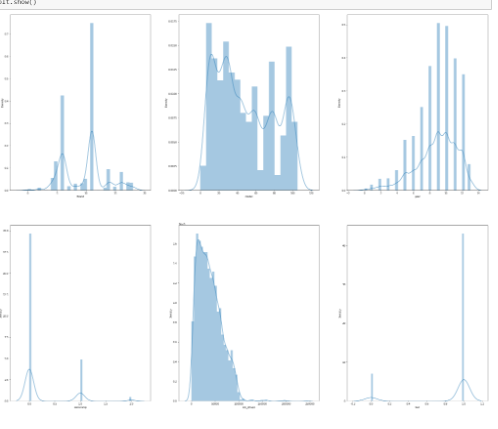


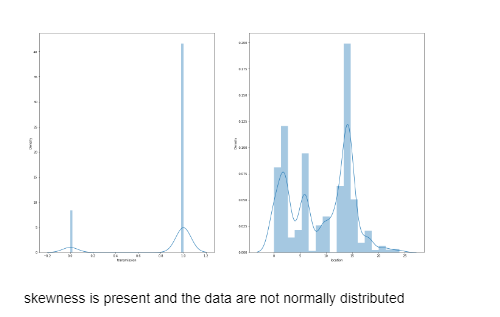




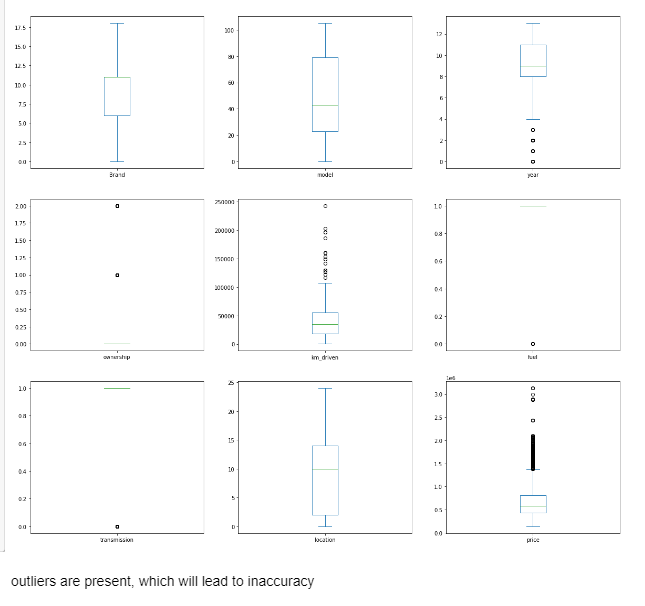








***Boxplot***



* Interpretation of the Results

From the visualisation, it could be interpreted that the price of cars is dependent on the type of fuel, transmission, location, km driven and brand of the car.

**CONCLUSION**

* Key Findings and Conclusions of the Study

For building a car price prediction model, it has very essential to scrape large sets of data since in the process of data cleansing, large amount of data could be lost due to presence of outliers. Moreover, with change of external business environment factors new model needs to be developed for accurate prediction of price.

* Learning Outcomes of the Study in respect of Data Science

For prediction of car price, the K Neighbors Regressor was the best model that gave results with high levels of accuracy.

* Limitations of this work and Scope for Future Work

The limitations was in aspect of presence of outliers because removing all outliers is leading to large scale data loss. So we removed outliers to some extent.