CAR PRICE PREDICTION

Submitted by:

SURJEET SINGH

ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to **FlipRobo** who gave me the golden opportunity to do this wonderful project on the topic **Car Price Prediction** project, which also helped me in doing a lot of Research and I came to know about so many new things such that things to keep in mind while buying a used car. I am really thankful to them. The data used in this project was scraped from website of **cars24**.

INTRODUCTION

- ## Business Problem Framing

  During lockdown a lot of people faced problems whiles moving to their home because very less public transport was available. Now as world is slowly getting out of the pendamic, people are trying to get prepared for such a problems in future. Now people want their own vehicle. As in India, there are so many families which can't by a new car so they prefer a used car instead. So the structure of the used car market is changing. It is important for used car market dealer to accurately place the price of the car so that it can be sold at best price. So we need a machine learning model which can predict the price of used car by the condition of the car.

- ## Conceptual Background of the Domain Problem

  Before diving into the model building we should understand the structure of the used car market. Which type of car has more value in the market. For example, Maruti Suzuki Alto is one of the best selling because of it is small in size and less parking space is required. Places like Shimla, parking space is a big problem so people preferred compact cars.

- ## Review of Literature

  While working on this project, I have gone through several websites and articles to accurately evaluate the used car. I have learned several important parameters to evaluate the car. For example, fuel type of the car is important parameters. Usually, diesel cars has higher price because a diesel engine has a better mileage and will lower cost on regular use. Selling price of CNG cars is less because of availability of CNG. Here is an article which was very helpful for me to understand the problem better. https://indialends.com/car-loan/used-car-valuation-process

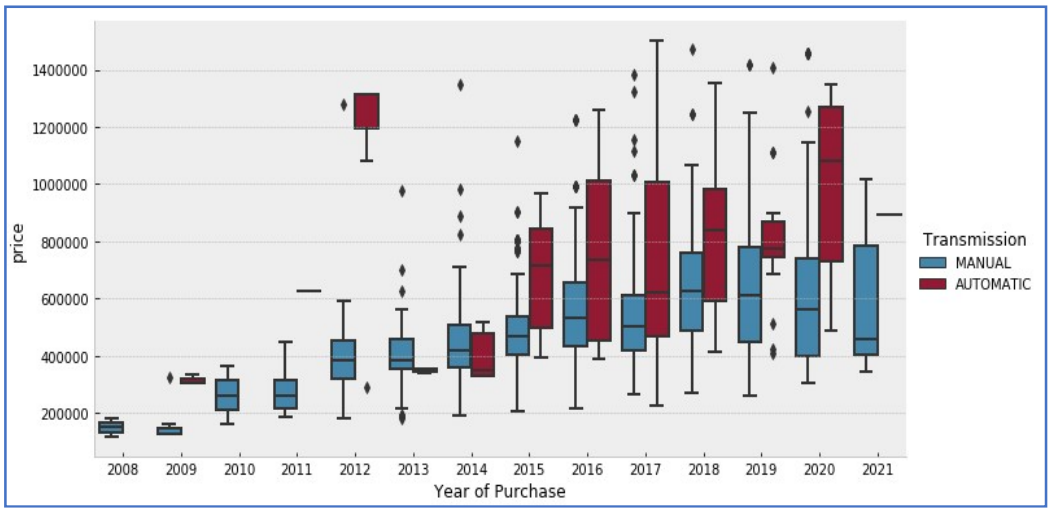
- **Motivation for the Problem Undertaken**

  The main objective of this project is to help the used car dealers to accurately place the price of the used so that they can get a deal quickly. Undervaluation of will hit the dealer with a losing deal while overvaluation of the car will delay in getting a deal or no deal at all. We will try several machine learning algorithms to figure out the best model working for the problem.

Analytical Problem Framing

- **Mathematical/ Analytical Modeling of the Problem**

  While analyzing the data for the given problem I came to know several outcomes which were helpful in creating the model. It is very obvious that an old car is depreciated and must have lower price than a relatively newer one. But it is not the only criteria to evaluate the price. A car having automatic transmission must have a higher price as shown in the graph below.



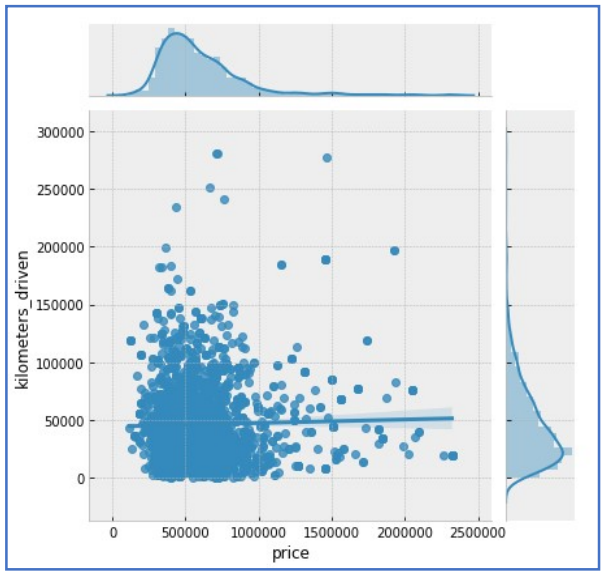
- **Data Sources and their formats**

  The data used in this problem is scraped from the website of cars24. I have included the parameters as much as I can by keeping in mind that these parameters will be helpful in evaluating the price

of the car. Most of the parameters scraped are categorical. A kilometre driven is a numerical parameter. The categorical features such as Engine Sound, 1 stands for perfect sound and 0 stands for defective sound. Same pattern followed in the other features also.

- **Data Preprocessing Done**

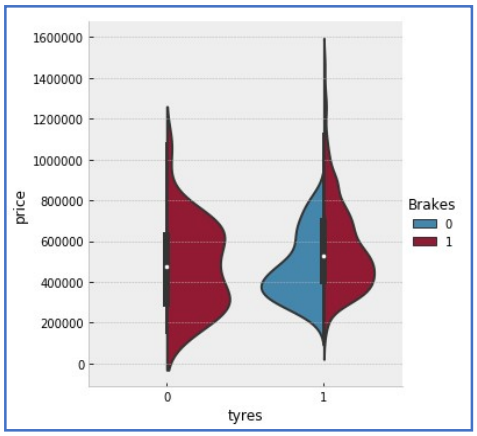
In the first go, I have filled the all the missing values which I have found only in two categorical variables. Since no missing values in the numerical features, no issue in the filling missing values because it won't alter the mean. Then I have merger the different columns which has similar data. For example, in some of the web pages, the music system was spelled as "Audio System" and "Audio system". Due to this separate columns have been created. Then I have merged more similar columns by doing some features engineering. For example, there were 4 columns for each door. I made a single column for doors in which 4 stands for all doors are perfect, 3 stands for 3 perfect door but 1 defected, 2 stands for 2 perfect doors but 2 defected, 1 stands for 1 perfect door but 3 defected doors and at last 0 stands for all the doors are defected. It is also found that, if all the doors of the car are defected, their ORVM are also defected. After plotting a jointplot of price and kilometre driven, it is found that most of the data lies below Rs. 1500000 price and most of the cars were driven less than 150000 km. The cars having Rs. 0 have been removed.

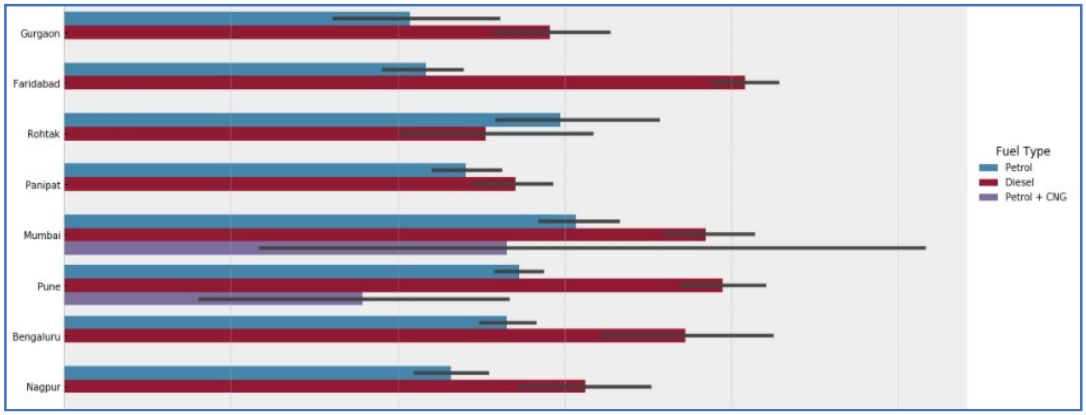


- **Data Inputs- Logic- Output Relationships**

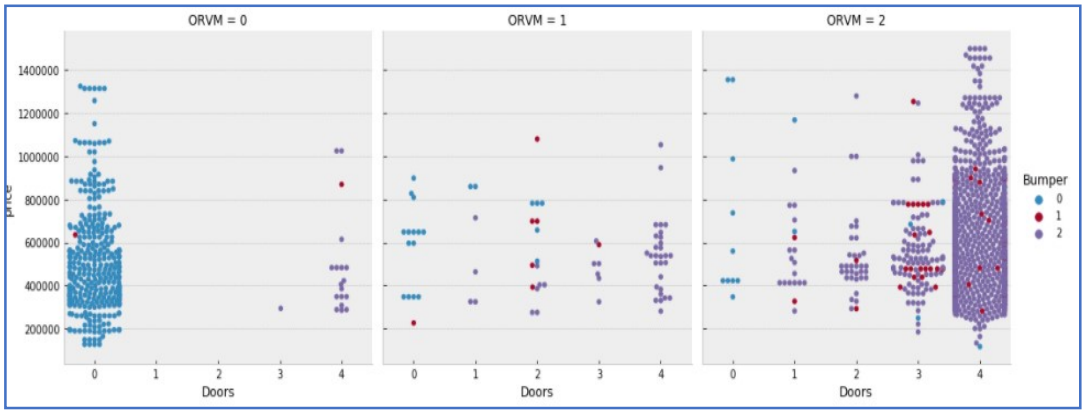
Since there are many features used as input so I will explain some them here. I have already discussed some of the input output logics

regarding the purchase year, transmission and fuel type. The car having good condition of tyres has relatively higher price. On the top of that if car has better condition of brakes along with the good tyres, it can be sold at higher price. Below I have attached a visualization of average price and fuel type for some of the cities. The analysis of this graph already discussed.



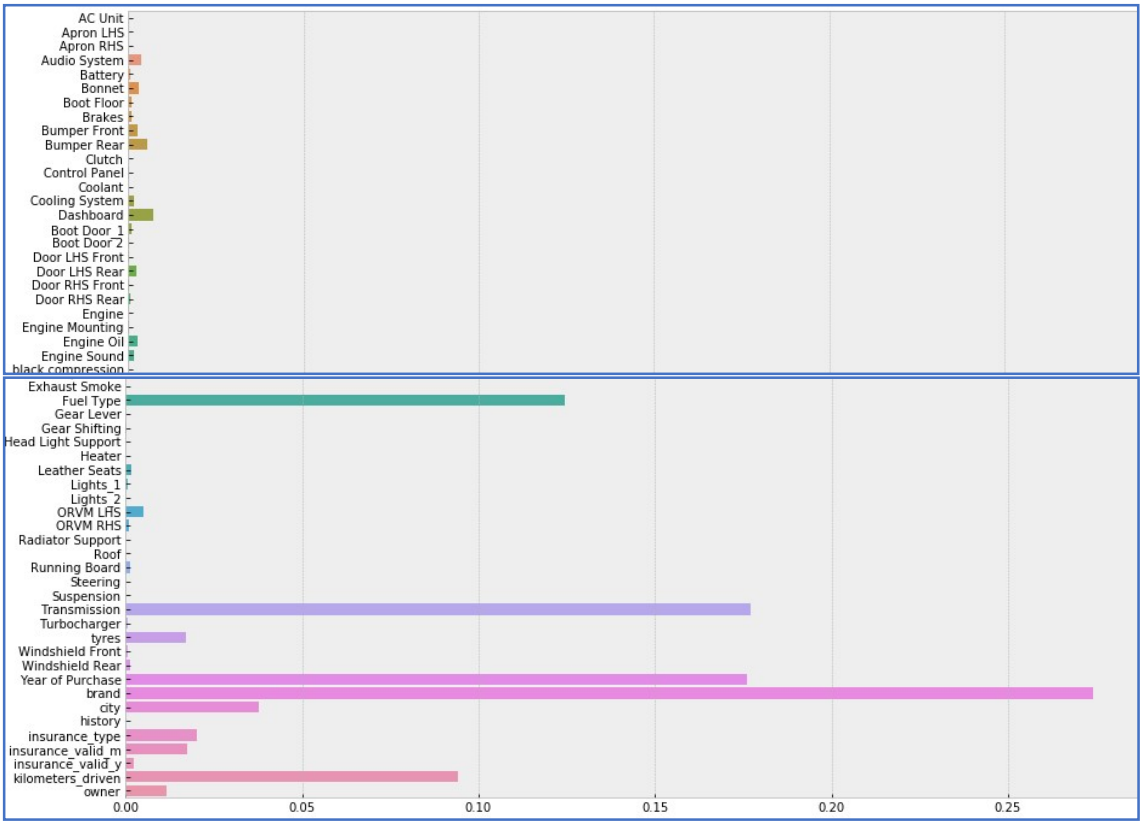


In the above section, I have explained the doors and ORVM, below is the visual representation of related data.



- **State the set of assumptions (if any) related to the problem under consideration**

  Since there are several parameters scraped, but all the parameters are not important to set the price of the car. I have considered 95% of importance of the features and discarded rest of the features. To figure out that, I have quickly trained a decision tree model and determined the feature importance. Then I choose the most important feature so that I can build a computationally simple model.



- **Hardware and Software Requirements and Tools Used**

  All the codes are written in python 3. Jupyter notebook used to create all visualizations, training, and testing of the machine learning models. Device used in this project : hp laptop with intel i5 process and 4 GB RAM.

Model/s Development and Evaluation

- ## Testing of Identified Approaches (Algorithms)

  Since it is regression problem, we can begin with simplest regression algorithm which is linear regression, and then we can try decision tree, random forest, gradient boost and extra tree algorithms.

- ## Run and Evaluate selected models

  1. **Linear Regression**

     Linear regression is an algorithm which fit a line to the data by assigning some weighs for each feature. It works well if there exists some linearity between the target and features.

```
from sklearn.linear_model import LinearRegression
LR = LinearRegression()
LR.fit(X_train, y_train)
y_pred_LR = LR.predict(X_test)
LR.score(X_test, y_test)
```

```
0.38710878116040104
```

     The accuracy of linear regression is extremely low which clearly shows that there is no linearity between the features and target.

  2. **Decision Tree Regression**

     Decision tree is an algorithms which distribute the data into several different bins by asking appropriate questions by itself. Then it give an output from the appropriate bin when we predict the value.

```
from sklearn.tree import DecisionTreeRegressor
DT = DecisionTreeRegressor()
DT.fit(X_train, y_train)
score = DT.score(X_test, y_test)
score
```

```
0.5897215129895266
```

3. **Random Forest Regression**

   Random forest used number of decision trees and combined the output of all the trees to give a better prediction.

```
from sklearn.ensemble import RandomForestRegressor
RF = RandomForestRegressor(n_estimators = 70, min_samples_split=3, max_features=15, max_leaf_nodes=800)
RF.fit(X_train, y_train)
accuracy = RF.score(X_test, y_test)
accuracy
```

```
0.77828750622543
```

4. **Gradient Boost Regression**

   Gradient boost is also a ensemble based approach but it is a bit different from the random forest.

```
from sklearn.ensemble import GradientBoostingRegressor
GB = GradientBoostingRegressor(n_estimators = 80)
GB.fit(X_train, y_train)
accuracy = GB.score(X_test, y_test)
accuracy
```

```
0.6384590708168367
```

5. **Extra Tree Regression**

   This algorithm implements a meta estimator that fits a number of randomized decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

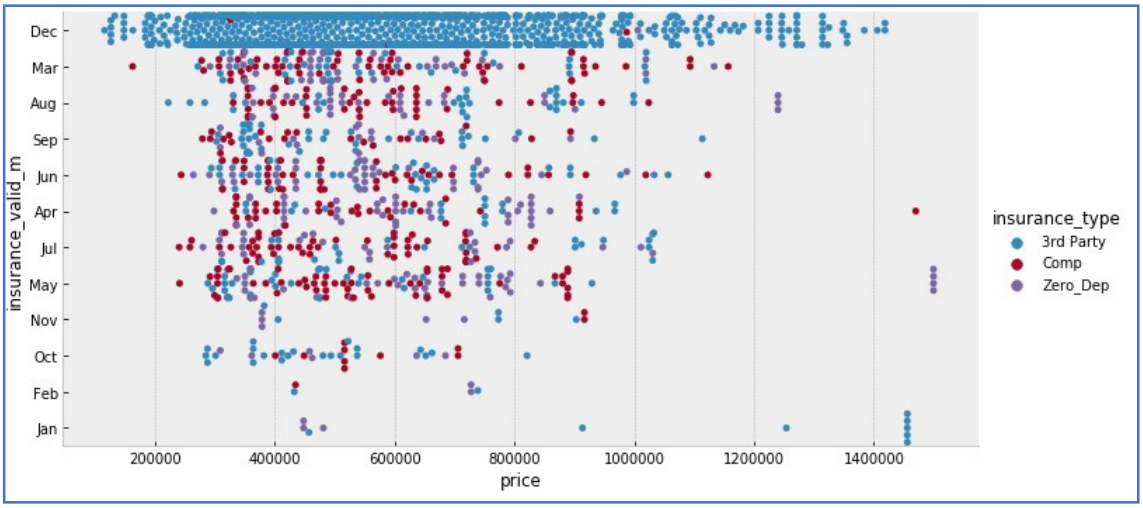
```
from sklearn.ensemble import ExtraTreesRegressor
ETR = ExtraTreesRegressor()
ETR.fit(X_train, y_train)
ETR.score(X_test, y_test)
```

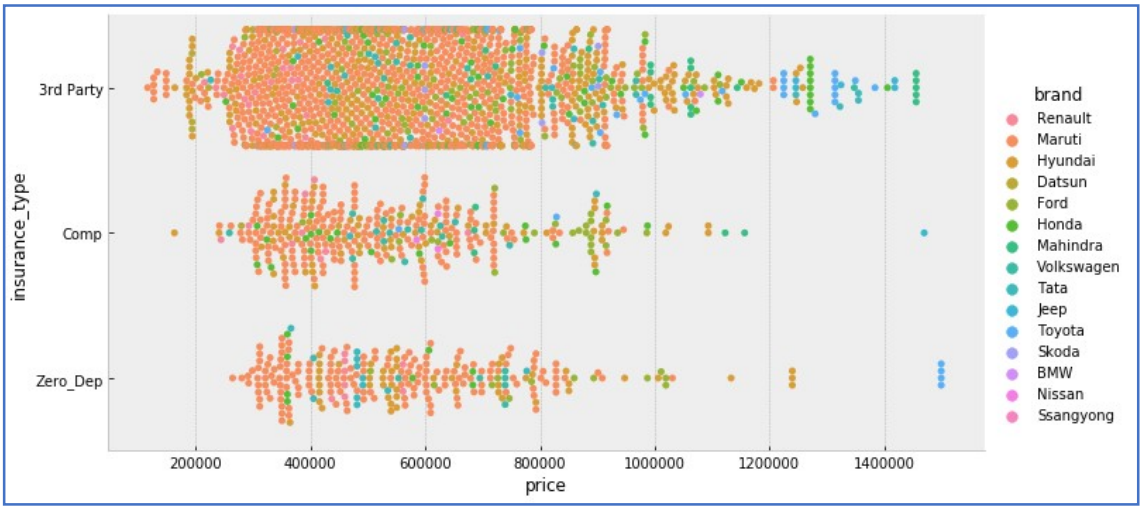
```
0.7484721308409915
```

- ## Visualizations

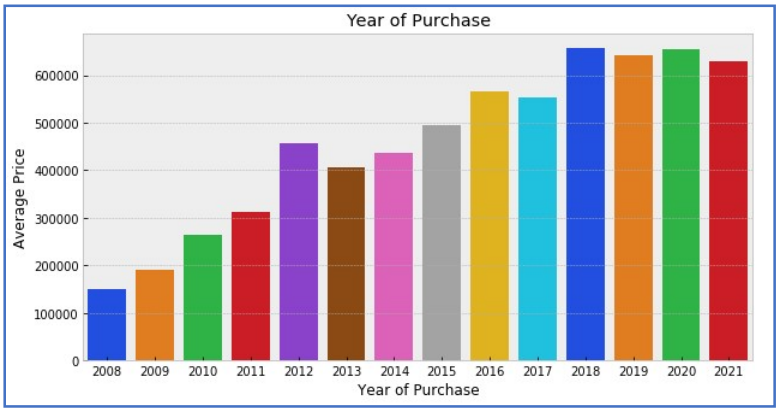
Some of the plots have been already explained above in some of the sections. Below is the plot for different brands scattered over the price axis along with the insurance type of the car. Most of the cars in india belongs to the brand Maruti Suzuki and Hyundai. Moreover, most of them secured their car with a third party car insurance which is going to end in the December month. After that many of the cars are secured with the comprehensive insurance but very rare of them are ending in the December month. It shows that people buy insurance in the month of January for one year. Endings

of comprehensive insurance has a spread over the year but very few have ending date in the month of January, February, October and November.





It is very obvious that the older vehicle will have less price so the plot between the year of purchases and average is shown below.



- **Interpretation of the Results**

  From the above analysis, it can be summarized that the most important parameters for setting a price of a used car are year of purchase, fuel type, transmission and overall look of the car. We will use only those important features for the training which will make 95% of importance.

CONCLUSION

- **Learning Outcomes of the Study in respect of Data Science**

  In this project, I have explored some new visualization methods which can give deep insight of the data. I have leaned the price placement of the used car. Some of the points I have learned in this product can also be used to analyze price placements for some other products.

- **Limitations of this work and Scope for Future Work**

  I have tried my best to complete this project but still I have noticed some points which can be used to further improve the accuracy of the model. The thing I missed in my web scraping is adding a column with the name of the model and body type of the car along which is also an important features. Moreover, we can also add the data whether the company is still producing the same model or not. Maruti Suzuki Alto has price in the range of Rs. 80000 to Rs. 300000. But on the other hand, as ford discontinued its production in india, its model become cheaper. Resale value of Ford Fiesta sedan car is around Rs. 100000 which is almost equivalent to the price of the Alto but a Ford Fiesta has better features than an Alto. So these features also will be helpful to get better accuracy.