

PREDICTING USED CARS PRICE

&

SCRAPPED DATA

**SUBMITTED BY**:

Pankaj Pagare

**ACKNOWLEDGMENT**

I would like to express my gratitude towards FlipRobo Technologies for their kind co-operation and encouragement which help me in completion of this project.

I would like to express my special gratitude and thanks to industry persons and my mentor (Ms. Shristi Maan) for giving me such attention and time as and whenever required.

**INTRODUCTION**

* **Business Problem Statement**

Build a model which can be used to predict the price of the used cars for the company, so that they can further used sell the cars to customer and can earn profit with that. The demand of used cars nowadays is incasing year by year. So, this model helpful to predict the price.

* **Conceptual Background of the Domain Problem**
* Determining whether the listed price of a car is a challenging task, due to many factors that drive a used vehicles’ price on the market.
* The focus of this project is developing machine learning models that can accurately predicted the price of a used car based on its features, in order to make informed purchases.
* We implement and evaluate various learning methods on the dataset consisting of the sale prices of different makes and models across cities from cars24 site.
* Our results show that Random Forest model yield the best results.
* **Review of Literature**

The used car market in India has been the center of attention in the slow growing automotive industry in India. In the last year, demand for used cars has increased.

A few years ago, the ratio of new cars to used cars was 1:1.2 which is now at 1:2.2. Basically, when 10 new cars are sold, 22 used cars are available for sale in the market.

In 2008-09, the estimated sale of pre-owned cars was at 37 lakhs. For the 2018-19 period, projected sales are at 62 lakhs estimated to be worth Rs.1.62 lakh crores. The average holding time of a new car has come down to just 3 years which was 5-6 years earlier. For the first time aspiring buyer, the used car category has opened up many options.

* **Motivation for the Problem Undertaken**

The prices of new cars in the industry are fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But due to the increased price of new cars and the incapability of customers to buy new cars due to the lack of funds, used cars sales are on a global. There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features. Even though there are websites that offers this service, their prediction method may not be the best. Besides, different models and systems may contribute on predicting power for a used car’s actual market value. It is important to know their actual market value while both buying and selling.

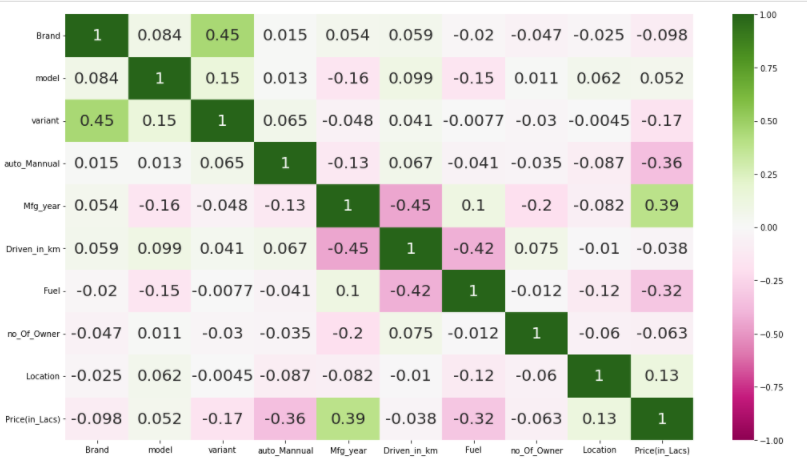
**Analytical Problem Framing**

* **Mathematical/ Analytical Modelling Of The Problem**
  + There were some values in Brand that are same but are different because of their lower and upper cases.
  + There were less records for the No. of owner for 4 and 5 and also for Fuel used in cars.
  + The maximum values for Distance covered while travelling by the pre-owner are not acceptable.
  + There were variables like Car is automatic or manual, Manufacturing year of car, Car driven, Fuel, No. of owner, Location such features were skewed. There we need to reduce the skewness. So, that features so more normally distributed.
* **Data Sources And Their Formats**
  + In this section we scrapped data from Cars24.com for different location with different Brand of car used Selenium to scrap the data.
  + The data was raw we transformed variable like:
    1. Split the car driven, fuel used, no. of owner from one list that we scrapped.
    2. Change price data type from object to integer and converted into lacs.
    3. Split brand, model, year of manufacturing car from one list.
    4. Split the car is automatic or manual and its variant from another list we scrapped.
  + Once the scrapping done saved the data to excel sheet.
* **Data Pre-Processing Done**

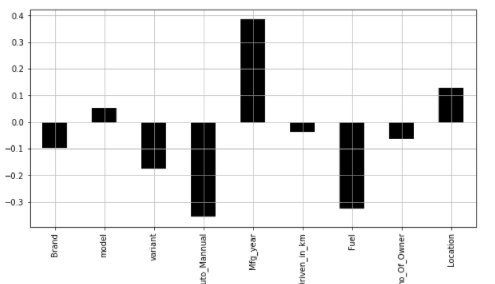
In this section we have seen that:

* + There were 5484 no. of records with 10 variables including target.
  + The no. of unique values of all variables.
  + The count of categories in Brand variables we have seen that there were some unique values. That we removed so that model should not be overfitted/ underfitted. Similarly, we did for No. of owners & Fuel variables.
  + The unique value for brand there we see the Hyundai & Renault were also in records with upper cases. So, we replace such values.
* **Data Inputs- Logic- Output Relationships**
  + In this dataset the Inputs variables are:
    1. **Brand**: Brand of Cars.
    2. **Model**: Model of the car.
    3. **Auto**\_**Manual**: Car is Automatic or manual.
    4. **Variant**: Type of car.
    5. **Mfg\_year**: Manufacturing year of Car.
    6. **Driven in\_km**: Car derived in kilo meters.
    7. **Fuel**: which type of fuel does the car used.
    8. **no**\_**Of**\_**Owner**: No. of owners of the car.
    9. **Location**: Location of the car from which city that car is used.
  + In this dataset the Output variables are: Price: Price i.e., the target variable in lacs in Indian rupee...

**Features Co-relation:**



* **Relation of features with the Target Variables.**



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

**Software Used:**

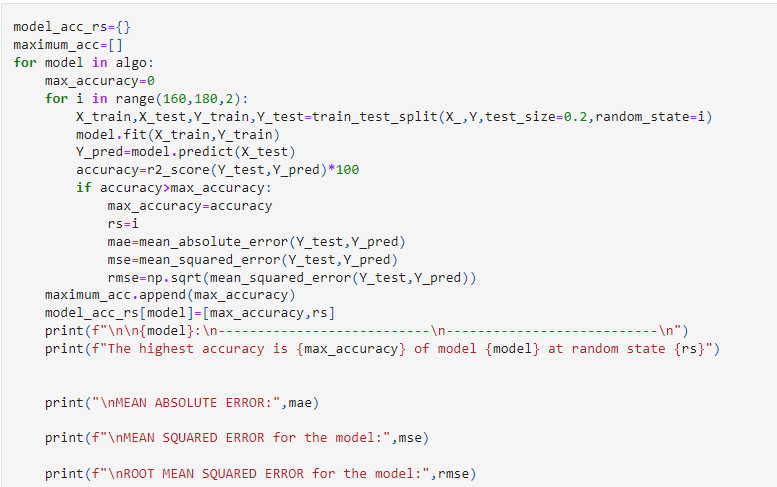
* + Jupyter Notebook
  + Ms-Paint
  + MS-PowerPoint
  + MS-Word

**Hardware used:**

* + Laptop
  + Good internet connectivity

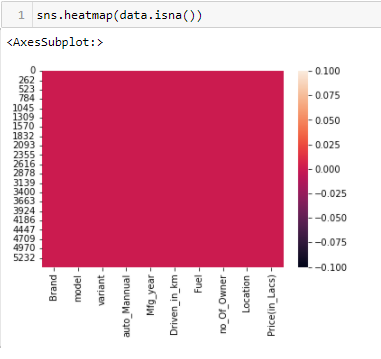
**Model/s Development and Evaluation**

* **Identification of possible problem-solving approaches** (methods)
  + Used Z-score techniques to remove outliers for Car Driven.
  + Used Power Transformer with yeo-johnson method to reduce skewness in features Car is automatic or manual, Manufacturing year of car, Car driven, Fuel, No. of owner, Location.
  + Standardized the features using StandardScalar().
* **Testing of Identified Approaches (Algorithms) :**
* Linear Regression
* Decision Tree
* Random Forest
* AdaBoost
* Bagging
* Support Vector Machine
* **Run And Evaluate Selected Models**

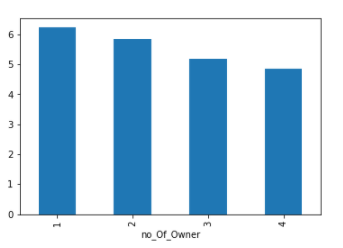


The above code shows the highest accuracy of each model at their best random state between 160 to 180 with the evaluation metrics like Mean Absolute Error and Mean Squared Error and Root mean Squared Error for each model.

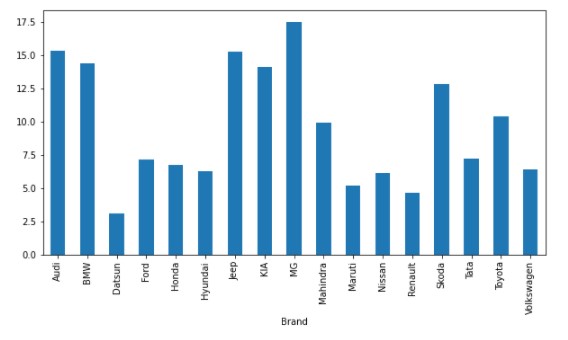
**Output**

* **Key Metrics for success in solving problem under consideration**
* What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.
* **Visualizations**
* 
* **Observations:**

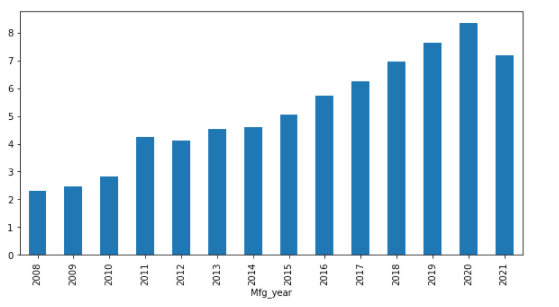
No null values present in our dataset.

* 
* **Observations:**

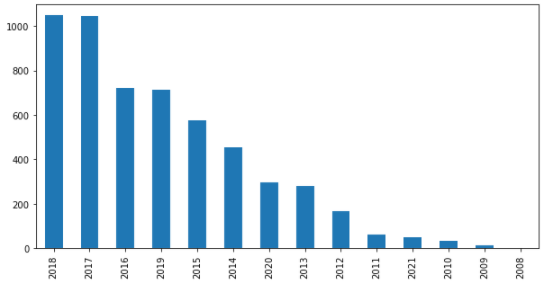
Less no of Owner their Car price is high.



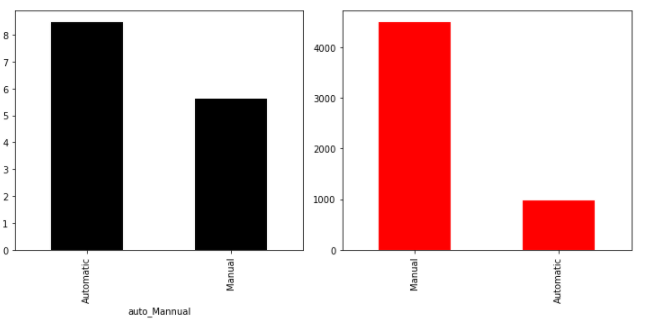
* **Observations:**
  + Mercedes type Brand's car has the highest price.
  + Datsun type Brand's car has the lowest price.



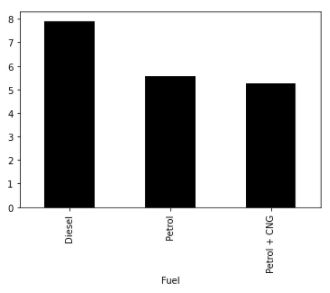
* **Observation**
  + As the year increases the price of car increasing.



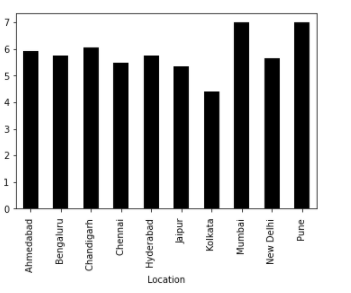
* **Observations:**
  + The Demand of cars is highly increased in the year of 2017 and 2018.



* **Observations:**
  + As we can see the no. for manual car type are used more than the automatic but the price are high for automatic cars.

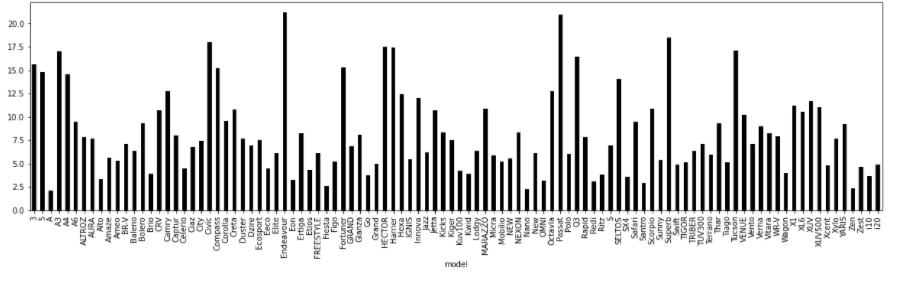


* **Observations:**
  + Diesel fuel container type of cars has high prices.

****

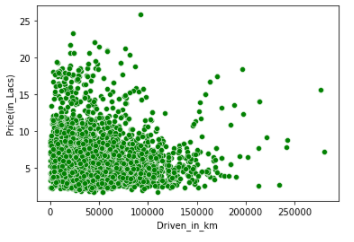
**Observation:**

* + Mumbai and Pune are having high selling of cars.



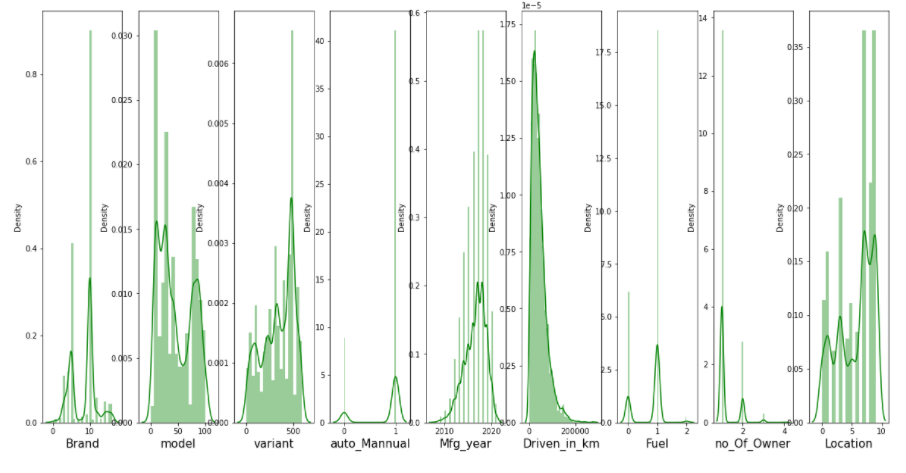
**Observations:**

* + Benz type of model has highest Price.
  + A’s type of model has the lowest Price.



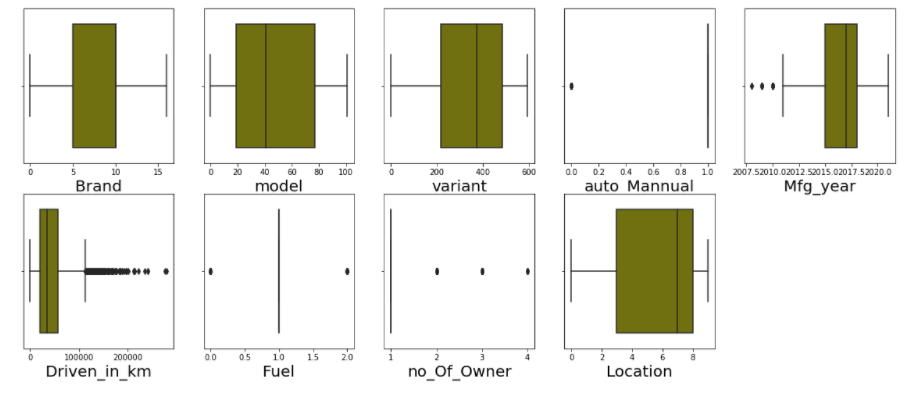
**Observations:**

* + There is slight relations between car driven and prices of cars that car is driven more prices are less.



**Observations:**

* Auto\_Manual, Mfg\_year, Fuel, no\_Of\_Owner these columns doesn't seems to normally distributed.

****

**Observations:**

* In Column Auto\_Manual boxplot shows outliers but this may be because of less no. of records for automatic cars
* similarly in columns Mfg\_year, Fuel, no. of Owner
* **Interpretation of the Results**

1. **Pre-processing:**
   * + There were 5484 no. of records with 10 variables including target.
     + The no. of unique values of all variables.
     + The count of categories in Brand variables we have seen that there were some unique values. That we removed so that model should not be over fitted/ under fitted. Similarly, we did for No. of owners & Fuel variables.
     + The unique value for brand there we see the Hyundai & Renault were also in records with upper cases. So, we replace such values.
2. **Visualizations:**

* There were no null values in the dataset.
* The car prices are increasing with respect to decreasing no. of owner.
* Mercedes type of Brand's car has high selling prices while Datsun Brand’s car are having lowest selling prices.
* As the demand of car is increasing with the increase in selling price of car.
* Mostly Cars are sold in the year of 2018 & 2017.
* Selling price of Automatic Cars are high while having mostly cars manual.
* Mostly cars are sold of Diesel type of fuel containing.

**3 Modelling**

* + Before modelling we removed Outliers from Car driven using Z-Score technique with standard deviations.
  + We reduced skewness of the skewed columns using Power Transformer with yeo-Johnson method.
  + We used Standard scalar to standardized the features.
  + Used 6 algorithms: Linear Regression, Decision Tree, Random Forest, Adobos, Bagging and Support vector machine.
  + For every model we get highest accuracy at their best random State in between 160 to 180. With their evaluation metrics like Mean absolute error, mean squared error, Root mean squared error, r2 score.
  + We also trained and test our model with CV in range of 2 to 7. And get best CV Score of all model.
  + We finalized our model with the least between the best CV score of each model and highest r2 score of each model.
  + Among all the algorithm we get Random Forest with least difference that’s our final model. That gives highest accuracy of 94.17%
  + We Hyper tuned the Random Forest model the Best Parameters: {'criterion': 'squared error', 'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'estimators': 116, 'n\_jobs': -1}

**CONCLUSION**

* **Key Findings and Conclusions of the Study**
* When we scrapped data then data that were combined, we split and also change data types for such attributes like manufacturing year, price, no of owner.
* Variant & model attributes are in real world helpful to know the prices but it’s gives very less contribute to predict the price.
* Reduce Outliers and skewness of the data that helps for model complexity.
* Chosen **Random Forest Classifier Algorithm** as the finalized model with the best r2 score i.e., 94.17 and after hyper tuning the best accuracy i.e., 93.33%.
* We finalized our model with the least between the best CV score of each model and highest r2 score of each model.
* **Learning Outcomes of the Study in respect of Data Science**
* As we can see that after hyper tuning the model with Random Forest Regress or, it is getting over fitted.
* Hence, we can go with normal Random Forest Regress or model which is also giving good accuracy.
* Finally, we have saved the Random Forest Regressor Model.
* **Limitations of this work and Scope for Future Work**
* In future this machine learning model may bind with various website which can provide real time data for price prediction.
* Also, we may add large historical data of car price which can help to improve accuracy of the machine learning model.
* We can build an android app as user interface for interacting with user. For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset.
* This allows owners of used cars to be able to sell them in a single visit at any branch of the platform.