

NAME OF THE PROJECT

***Car Price Prediction***

Submitted by:

Ashish Bomble

**ACKNOWLEDGMENT**

Data has been scrapped from Cars24.com .

With support from Mr. Keshav Bansal Sir’s guidance on some points I am able to solved some problems which I had never had been face earlier.

Thanks to them for their support on time.

**INTRODUCTION**

* **Business Problem Framing**

Due to covid -19 our whole world economy impacted very badly. And that’s also applicable for all business sectors except Pharma sector. Specifically manufacture sector suffered a lot from this pandemic Exmp. Vehicle, Enterprises goods, cloths.

In now days everything is available online weather it’s about selling or for purchasing anything from home to car ,or any clean up service to food ordering etc.

But this is also impacted for them those are running shops, hotels, malls. If everyone purchase online then no one will buy from local vendors and there business may shutdown completely.

Now for their survival in this online era they also can take advantage of this for running their shops and that is also with by earning good margin in business.

Online available data of vehicles is fair enough to give average market price of that vehicle.so buyer will get car in good price as well as seller will also not suffer.

* **Conceptual Background of the Domain Problem**

Car, Bike is type of vehicles that are primary important thing for transportation in now days. we also should understand change is the behaviour of nature .while any particular person using any type of item from long time, at certain point of stage weather he’ll try to replace or change it with next or better version of that item which is available in market .same thing apply in case of this project also.

We can observe there are many car makers in market like Tata, Mahindra, and Maruti etc. every car maker is trying to give good facilities with affordable prices to their customers. Not every customer effort new car so they will look for used cars which are available with online sellers as well as at local vendors shop.

Online sellers will easily get buyers as well as sellers because that is on stop solution for both buyers and sellers and within less time both buyers and sellers get connect with each other due to online site. If deal gets confirmed then seller will get satisfactory price which he/she expected and buyers requirement also get full filled.

In this case if deal got failed then no one will in loss .not online portal, not sellers also not buyer. But if there is local vendor instead of online portal the vendor will have to face loss, vendor may get many customers in future for that same car but he may or may not get good price for that car .it mean there is uncertainty in business.

* Review of Literature

Before purchasing car the buyer should know about all specification of car, so with help of that specification he can estimate what should be the ideal price of this used car.

Research done on all specifications that are buyers required to know and that specifications like transmission, engine type, model type, owner type about used car.

Then taking that all in consideration along with other things like company name, insurance ending date, Fuel type.

There are many more car sellers available on online sites like cars24.com, olx.com, cardekho.com. In different formats.

Now while surfing from these portals not every portal is that much of capable to give prosper details like OLX.com in which some of other things like car parts also there for selling which found on car result page. While fetching data this things can reduce model performance like there prices or specifications.

* Motivation for the Problem Undertaken

Motivation behind making this project is to good well trained model to local vendors so they can easily estimate car price which came at their local shop as per the car specification.

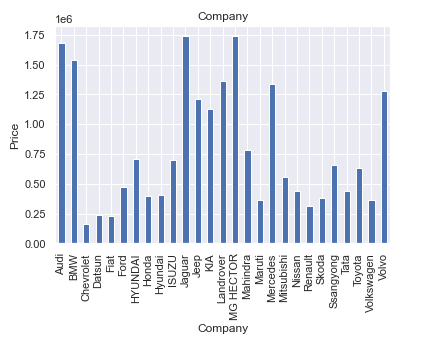
If car seller get good price of car and if he looking for other option then he definitely not refer any other source to search for car he will asked to that same vendor because he satisfied with amount which he got for he’s used car from vendor.

Our model is help to finalize the price of car on the basis of current trend or current price of that car. Seller have to just tell the specifications of car and buyers also just have to tell the specification which he/she looking for when vendor simply enter details he will get ideal price of car then with some modification he can sell same car to any buyers and can earn good profit.

**Analytical Problem Framing**

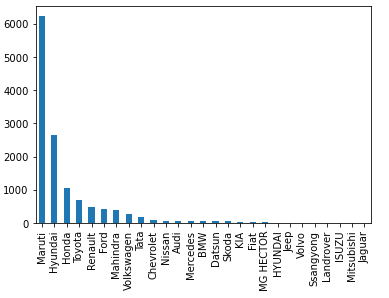
* Mathematical/ Analytical Modeling of the Problem

Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

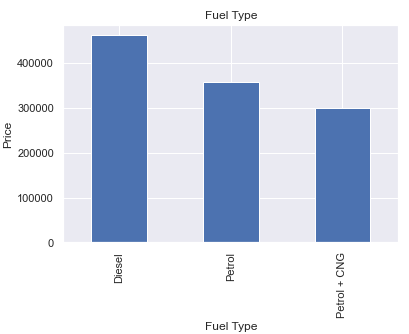


As per the available data from sellers:-

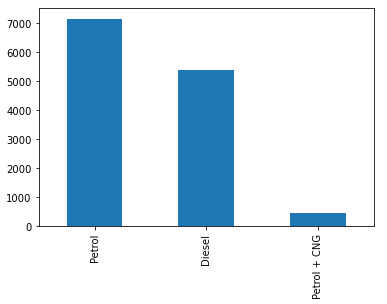
1. Chevrolet companies have fewer prices compare to other.
2. Jaguar & MG HECTOR have highest price than others.
3. Audi is also costing more after Jaguar & MG HECTOR.



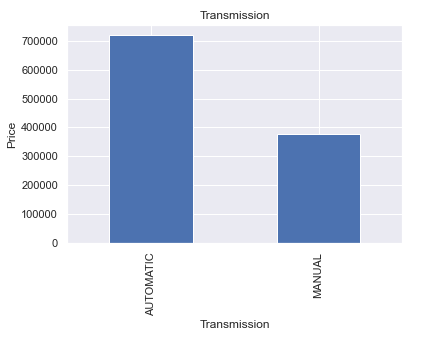
1. Maruti Suzuki Company’s has maximum cars for sell.
2. Where Jaguar‘s cars are very less in numbers for sell.



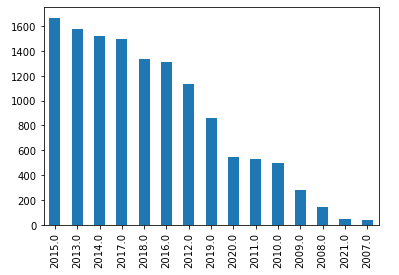
1. Diesel car price is more as compare to Petrol and Petrol+CNG.
2. Where Petrol + CNG have less price as compare to other 2 fuel types.



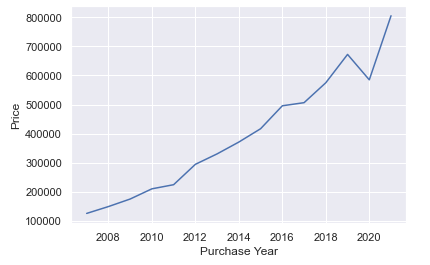
1. Those cars Fuel Type is Petrol that are available in large numbers for sell.
2. Where Petrol+CNG cars are available in very less numbers for sale .



1. Those cars have Automatic transmission that cars have high price than cars that have manual transmission.



1. From available cars maximum cars are manufactured in year 2015.
2. Where less cars are available which are manufactured in year 2007



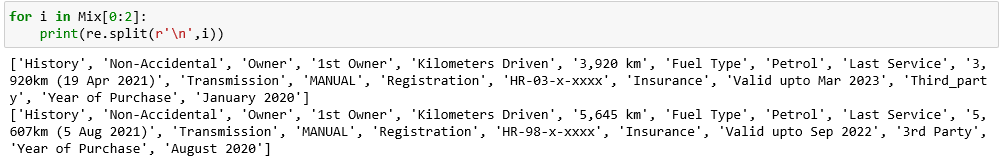
1. Recently purchased/manufactured cars selling price is higher.
2. Those cars are purchased years back are available in less price.

* Data Sources and their formats

Data has been scrapped from Cars24.com. Data available on this site is in a standard format, & it was reachable and manageable by Semantic Web tools. After scrapping it by using selenium web driver its was managed and manipulated (in case of formats) by using different python as well as Regular Expression techniques.

And only extracted that details which are required for analysis. First it was extracted in python list then it converted into data frame.

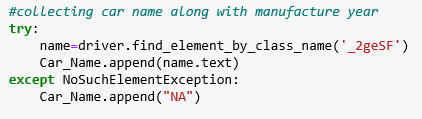
Example:

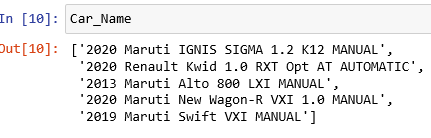


First 2 car specifications were scrapped from web site .and it was splatted on basis of space by using Regular Expression.

* Data Pre-processing Done

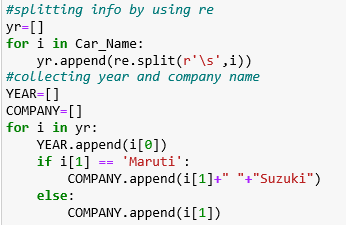
When data were scrapped from website it was in raw format. All text got scrapped from website some of non-necessary text also has been scrapped from site.

  
 As per above SS by using class name “\_2geSF” from cars24.com site we had scrapped all text which contain by this particular class.



In car name list we had scrapped all texts from class “\_2geSF”.

Manufactured year, company name, model name were scrapped together.



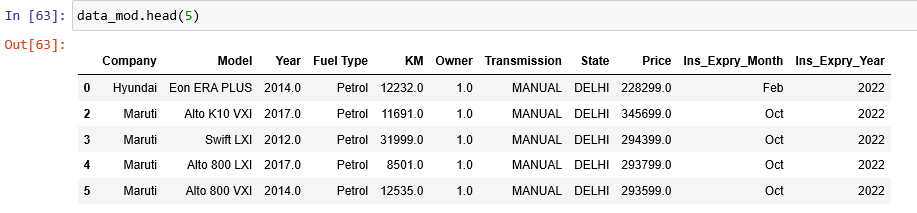
By using Regular Expression we splatted it and then append manufactured year in “yr” list and company name in “Company” list. Then it was converted in data frame format.

Same process has been followed to scrap other specifications of car.

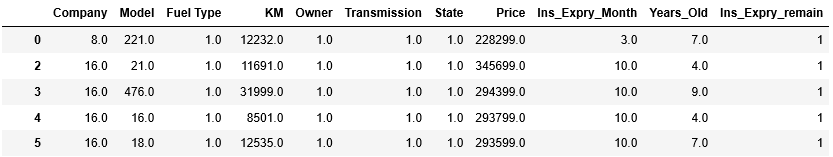
* Data Inputs- Logic- Output Relationships

For building the machine learning model text data needs to be converted into numeric format.it was converted with help of OrdinalEncoder () which convert text to numeric. Because machine can understand numbers not a text.

Before converting into numeric format.



After converting



* Hardware and Software Requirements and Tools Used

**Hardware & Software Requirement**

**RAM**: - Min 6 GB

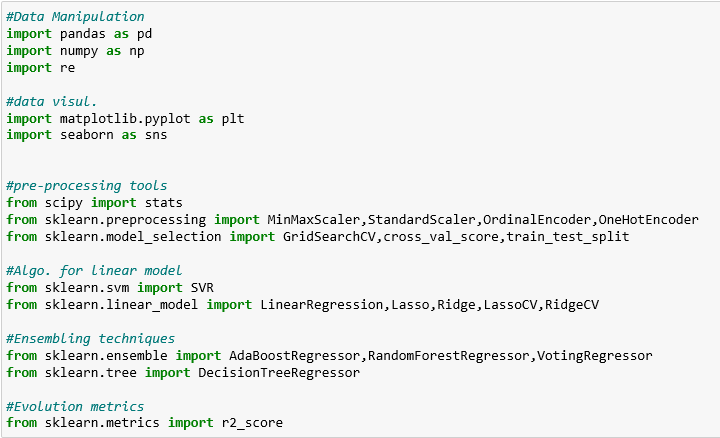
**Windows Edition**: -  Windows\* 7 or later, macOS, and Linux

**Processor**: - Intel Core i3

**Disk Space:-** 2 to 3 GB

**Python\* versions:-** 3.6.X

Libraries that used for Project :-

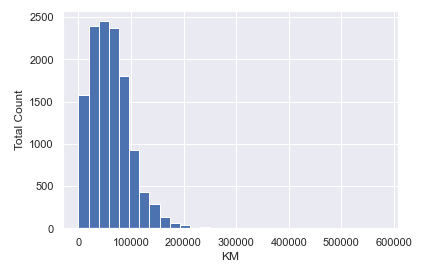


**Data Manipulation:-**

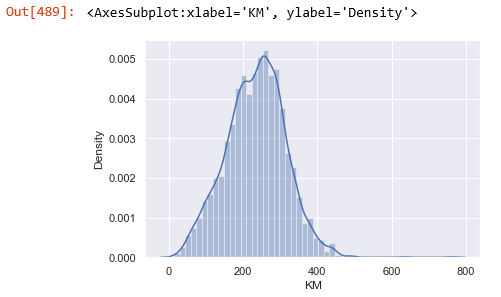
We used pandas library which help to view data in tabular format in python .with help of this we can perform operations on dataset.

Numpy is an abbreviation of Numeric Python .Numeric operation can perform on dataset by using it .we tried to convert “KM” columns in normal distribution format by using np.sqrt method.

Before Normal distribution.







**Data Visualization:-**

Used Seaborn and matplotlib for data visualization.

**Data Pre-processing:-**

**MinMaxScaler**:- converting all data points in single range(normalization of data points)

**OrdinalEncoder:-** Converting all text labels into numeric(float) data type.

**Ensembling Algorithms**:-

**RandomForestRegressor**:- Collection of decision tress that increase the accuracy score of overall model.it learn from its existing mistakes and improves the next one.

**Sklearn.Tree:-** From it we importDecisionTreeRegressor()

**Evolution – Metrics :-**

**r2\_score :-**  Used to check the score of model train as well as test model .

**cross\_val\_score**:- Cross check the model Score weather its over fitted or under fitted.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Initially the data had contain **13191 Rows X 11 Columns.**

1. Dropping the duplicate entries from dataset. But there is duplicate entry is not present.
2. Checking Dtype of columns by using Data.info()
3. Dropping rows which have Nan values in dataset.
4. After dropping available data is 12499 Rows X 11 Columns.
5. Dropping distinct value column. The available data is 12499 Rows X 10 Columns.
6. Checking **Categorical features & Numerical features** from dataset.
7. Comparing Car price with different features. How car price get affected by change in feature.
8. Converting text features into numeric(float) format
9. Data proceed for Building a model.

* Testing of Identified Approaches (Algorithms)

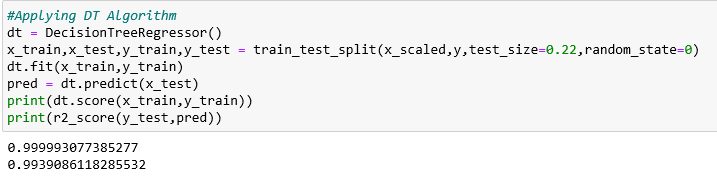
Training and Testing dataset :-

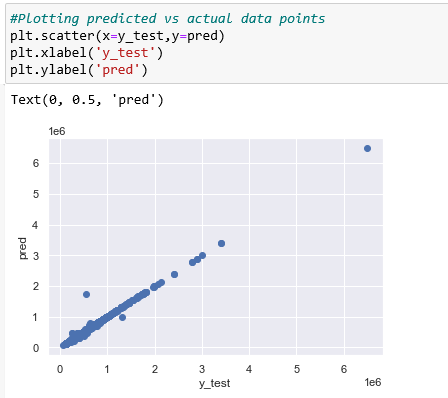
1. DecisionTreeRegressor() with its default parameters.
2. RandomForestRegressor() with its default parameters.

From which DecisionTreeRegressor() is finalized with its default parameter.and accuracy score has been cross verified by using cross Validation.

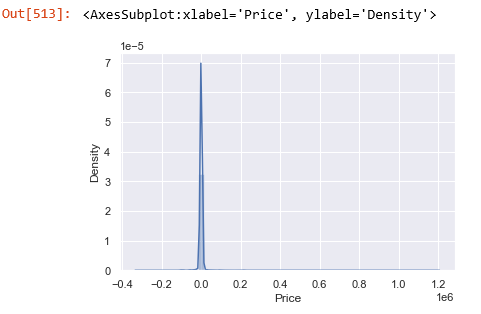
* Run and Evaluate selected models

For Evolution :-









* Interpretation of the Results

**Interpreted from Visualization**:

* In recent year manufactured cars are costlier.
* Brands like jaguar, MG HECTOR are having less numbers of cars for sell on online portals.
* Brands like Maruti Suzuki have more number of cars for sale.
* As per available data there are not electric cars available on portal for selling.

**Pre-Processing:**

* Try to fetch numeric data as much as possible from dataset.
* Converting continuous variable in normally distributed form. That helps our model to train on its best level without varying from target variable.

**CONCLUSION**

* Learning Outcomes of the Study in respect of Data Science

Visualization help to understand how dependent variable is acting or behaving as per each independent variable on initial stage before model building.

More categorical variable are present in dataset as compare to numerical variables. DecisionTreeRegressor() is find out a perfect fit algorithm for building model because of categorical data.

* Limitations of this work and Scope for Future Work

Currently present model is able to learn about 99.99% from data from which it can easily predict 99.39% of data.

By checking with RandomForestRegressor() with it default parameters or by applying proper hyper parameter tuning model accuracy may get increase .