FLIP ROBO

Used Car Price Prediction

Submitted by:

Abhishek Jain

ACKNOWLEDGMENT

I take great pleasure to thank and acknowledge the help provided by **Flip Robo Technologies**. I extend whole hearted thanks to Mrs. Khushboo Garg who become my Mentor and with whom I worked and learned a lot and for enlightening me with her knowledge and experience to grow with the corporate working. Her guidance at every stage of the Project enabled me to successfully complete this Project which otherwise would not have been possible without her consent encouragement and motivation. Without the support it was not possible for me to complete the report with fullest endeavour.

Abstract

Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle’s price on the market. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models across cities in the India. Our results show that Random Forest model and K-Means clustering with linear regression yield the best results, but are compute heavy. Conventional linear regression also yielded satisfactory results, with the advantage of a significantly lower training time in comparison to the aforementioned methods.

INTRODUCTION

- Business Problem Framing

  Today, one of the foundations of the economy is thought to be the transportation sector. In wealthy countries, the automobile industry is referred to as the "Industry of Industries." Industry experts claim that India's automobile sector has experienced impressive growth. It represents its global prominence in addition to being the country with the automobile industry's quickest growth. Similar to most other nations, India is seeing a significant increase in the popularity of cars among both the native populace and the ex-pat community who work there. In India, used automobiles of every make are available for purchase.

  Nowadays, almost everyone wants their own automobile, but many people choose to buy used cars due to issues with price or the state of the economy. Because used car prices depend on so many different features and conditions, it takes expertise to anticipate them accurately. Prices for used cars fluctuate on the market, therefore both buyers and sellers require an intelligence system to accurately anticipate the price. The collection of the dataset, which includes all crucial details like the car's manufacturing year, gas type, condition, miles driven, horsepower, doors, the number of times a car has been painted, customer reviews, the car's weight, etc., is the most challenging task facing this intelligent system.

  Before feeding the data straight into the data mining model, it is important to pre-process and transform the obtained data into the appropriate format. The dataset was first statistically examined and plotted. The presence of duplicate, null, and missing values was found and corrected. Correlation matrices were used to select and extract features. The most correlated features were kept, while others were removed in order to create an effective model. Given that it falls inside the supervised learning domain, this prediction problem can be categorised as a regression problem. Numerous regression models, including bagging regression, linear regression, and random forest, were trained and contrasted. In this project, a random forest Regressor performed better than all others, hence it was selected as the primary algorithm model.

- Statement of problem

The goal of this study is to forecast the pricing of used automobiles in India using data mining techniques, scraping information from websites that market used cars, and examining the various features and factors that influence the real valuation of used car prices. By just presenting the computer with a set of desired car's qualities, buyers will be able to determine the actual value of their current or wanted vehicle.

The goal of this research is to comprehend, assess, and design a method for forecasting used automobile prices using data mining techniques in India.

- Project goals

In order to assist those wishing to purchase or sell cars and to provide them with a better understanding of the automotive industry, this project intends to supply price prediction models to the general public. Because some dealers are infamous for using dishonest sales techniques to complete a deal, purchasing a used automobile from a dealer can be a tedious and unsatisfactory experience. Therefore, this study aims to arm customers with the necessary tools to aid them in their purchasing experience and help them avoid falling prey to such strategies.

The project's exploration of novel approaches to assessing used automobile pricing and comparing their accuracy is another objective.

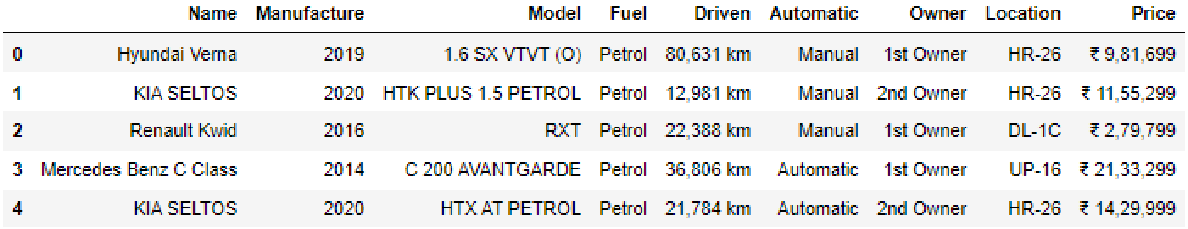
Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Data Preliminary data analysis must be performed to gain a deeper understanding of the quality of the data, in terms of outliers and the skewedness of the figures, descriptive statistics, and other factors. Understanding and preparation are essential parts of building a model because they provide insight into the data and what corrections or modifications shall be made before designing and executing the model. To do that, category and numerical variables were statistically analysed. Additionally, it helps to be aware of the key factors that influence how prices are determined. This was accomplished by creating a correlation matrix for each attribute to comprehend the relationships between the various components.

- Data Sources and their formats

The project deals with India used cars. Using Selenium, the benchmark dataset from cardekho.com and car24.com was scraped in order to build the effective intelligent model.



	Name	Manufacture	Model	Fuel	Driven	Automatic	Owner	Location	Price
0	Hyundai Verna	2019	1.6 SX VTVT (O)	Petrol	80,631 km	Manual	1st Owner	HR-26	₹ 9,81,699
1	KIA SELTOS	2020	HTK PLUS 1.5 PETROL	Petrol	12,981 km	Manual	2nd Owner	HR-26	₹ 11,55,299
2	Renault Kwid	2016	RXT	Petrol	22,388 km	Manual	1st Owner	DL-1C	₹ 2,79,799
3	Mercedes Benz C Class	2014	C 200 AVANTGARDE	Petrol	36,806 km	Automatic	1st Owner	UP-16	₹ 21,33,299
4	KIA SELTOS	2020	HTX AT PETROL	Petrol	21,784 km	Automatic	2nd Owner	HR-26	₹ 14,29,999

9 Features have been scrapped.

1. Name: Car complete name
2. Manufacture: Car manufacturer company name
3. Model: Car Model/ variant
4. Fuel: Which fuel is being used in the car (Petrol/Diesel/CNG etc)
5. Driven: Total driven km by car
6. Automatic: if the car is Manual or Automatic
7. Owner: How many owners have been changed of the car
8. Location: which state of India
9. Price: Selling price of the car

- Data Preprocessing Done

  In order to reduce the complexity of the model, the dataset was pre-processed after data collection to remove samples with missing values, remove non-numerical parts from numerical attributes, convert categorical values into numerical (if necessary), fix any discrepancies in the units, and remove attributes that don't affect price evaluations.

- Data Preliminary data analysis must be performed to gain a deeper understanding of the quality of the data, in terms of outliers and the skewedness of the figures, descriptive statistics, and other factors. Understanding and preparation are essential parts of building a model because they provide insight into the data and what corrections or modifications shall be made before designing and executing the model. To do that, category and numerical variables were statistically analysed. Additionally, it helps to be aware of the key factors that influence how prices are determined. To understand the relationships between the various components, a correlation matrix for each attribute was used.

- Data Inputs- Logic- Output Relationships

  The data is then organised and translated into a format that the data mining technology can process. Various data mining techniques have been developed to forecast used automobile prices and values. The construction of three models employing the Logistic Regression model approach, Random Forest Regressor, and Bagging Regressor is suggested in this work. First, the data was divided into sections for testing and training. Portioning percentages can be tried with various ratios to analyse various results. The four evaluation matrices known as model score, mean square error (MSE), mean absolute error (MAE), and root mean square error were used to evaluate all three models (RMSE). The Random Forest Regressor outperformed them all.

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

  With the pandemic-related shortages of semiconductors throughout the past year, the secondhand car market has undergone a significant transformation. As a result, there was a rapid change in automobile prices during this study, which will have an impact on future predictions of actual car pricing. The autos on the market will be undervalued by the present dataset. Therefore, the ideal solution would be a model that is based on real-time data and can be easily integrated into a mobile app for general use.

- Hardware and Software Requirements and Tools Used

  Hardware:

  Software: Latest Anaconda for Jupyter

  Python Libraries:

  Pandas , Numpy, seaborn, matplotlib, scikit-learn,

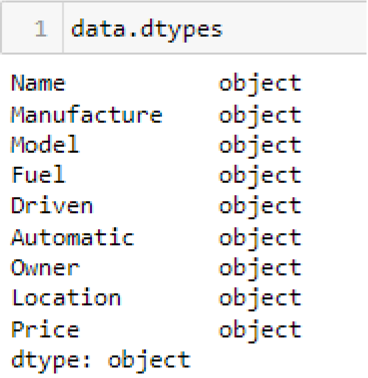
Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)
  1. Pre-processing is a Data Mining approach that entails putting unstructured data into a format that can be understood. Real-world data frequently lacks particular information on activity or trends and also contains erroneous information. As a result, this can lead to bad data gathering, which could then lead to poor models that are built using the data. The data can be pre-processed to address such issues.
  2. Input modification or encoding is the process of pre-processing in machine learning, which makes data easier for the computer to parse. The algorithm can now correctly comprehend the data as a result.

3. The dataset in this project is pre-processed using the techniques below.

4. 1. Dataset collection: We obtained the dataset from three of the top websites that deal with the sell-buy of old cars.Pre-Processing: scrapped data was very messy. We have to perform many pre-processing on that dataset.

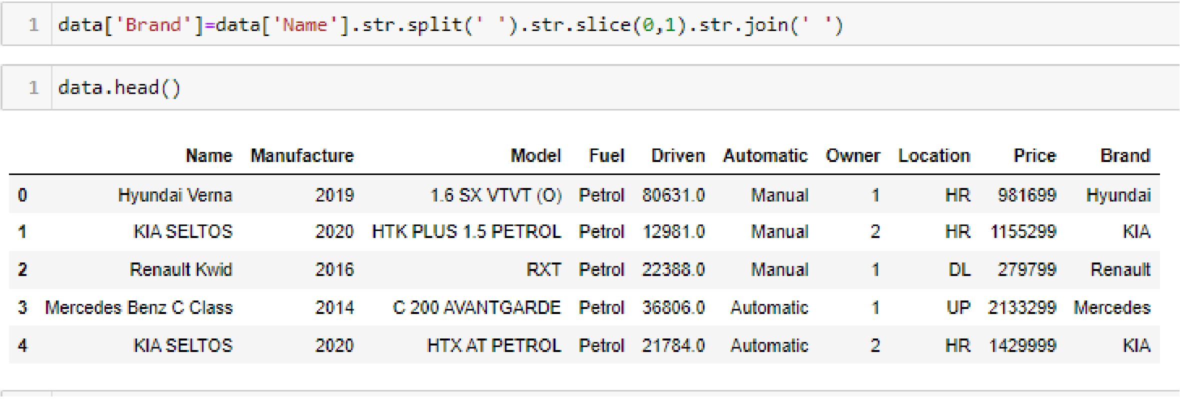
   As the data is scrapped from the websites, all features were in Object form, even the integer features also were in object form.



```
data.dtypes
```

```
Name           object
Manufacture    object
Model          object
Fuel           object
Driven         object
Automatic      object
Owner          object
Location       object
Price          object
dtype: object
```

   Name feature was having the complete name of the car, while we required the car manufacture company name so we extract this from the car name.

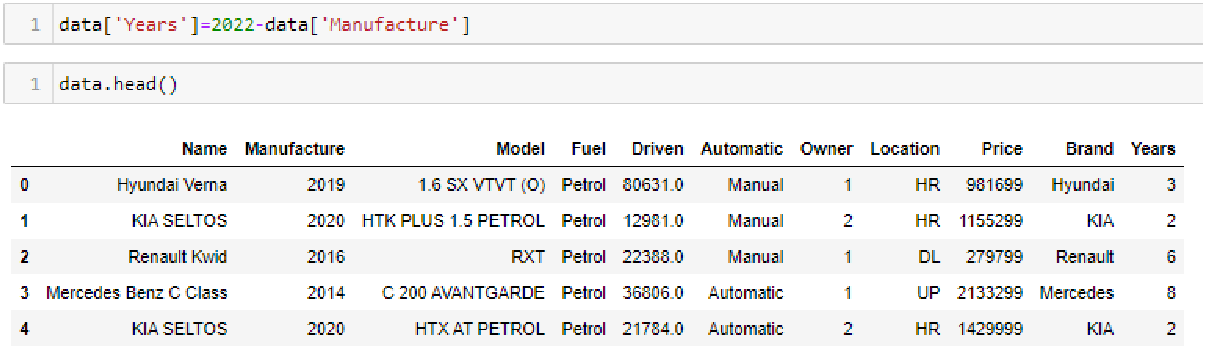


```
data['Brand']=data['Name'].str.split(' ').str.slice(0,1).str.join(' ')
```

```
data.head()
```

	Name	Manufacture	Model	Fuel	Driven	Automatic	Owner	Location	Price	Brand
0	Hyundai Verna	2019	1.6 SX VTVT (O)	Petrol	80631.0	Manual	1	HR	981699	Hyundai
1	KIA SELTOS	2020	HTK PLUS 1.5 PETROL	Petrol	12981.0	Manual	2	HR	1155299	KIA
2	Renault Kwid	2016	RXT	Petrol	22388.0	Manual	1	DL	279799	Renault
3	Mercedes Benz C Class	2014	C 200 AVANTGARDE	Petrol	36806.0	Automatic	1	UP	2133299	Mercedes
4	KIA SELTOS	2020	HTX AT PETROL	Petrol	21784.0	Automatic	2	HR	1429999	KIA

   Manufacture feature provided the year of car manufacture which is required basically to know the age of that particular car at instance. So we calculated the Age of car at this moment



```
data['Years']=2022-data['Manufacture']
```

```
data.head()
```

	Name	Manufacture	Model	Fuel	Driven	Automatic	Owner	Location	Price	Brand	Years
0	Hyundai Verna	2019	1.6 SX VTVT (O)	Petrol	80631.0	Manual	1	HR	981699	Hyundai	3
1	KIA SELTOS	2020	HTK PLUS 1.5 PETROL	Petrol	12981.0	Manual	2	HR	1155299	KIA	2
2	Renault Kwid	2016	RXT	Petrol	22388.0	Manual	1	DL	279799	Renault	6
3	Mercedes Benz C Class	2014	C 200 AVANTGARDE	Petrol	36806.0	Automatic	1	UP	2133299	Mercedes	8
4	KIA SELTOS	2020	HTX AT PETROL	Petrol	21784.0	Automatic	2	HR	1429999	KIA	2

Fuel feature were having many garbage inputs so we have replaced the garbage inputs with the median of feature

Driven feature were having small null values, we have decided to fill those null values with median of Driven feature

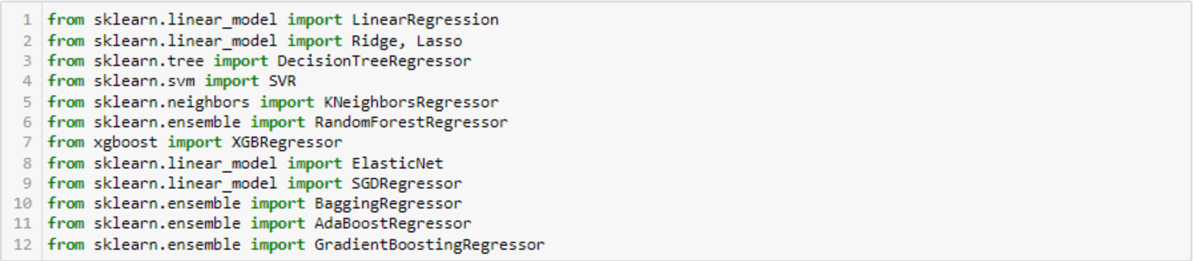
Location features were having the registration code for the car, so we have to scrap the state of the India. Because due to taxation, every state have different on-road price of car.

Price feature were in object form, we removed unwanted rupee sign and convert into integer datatype. Because this is our target feature here.

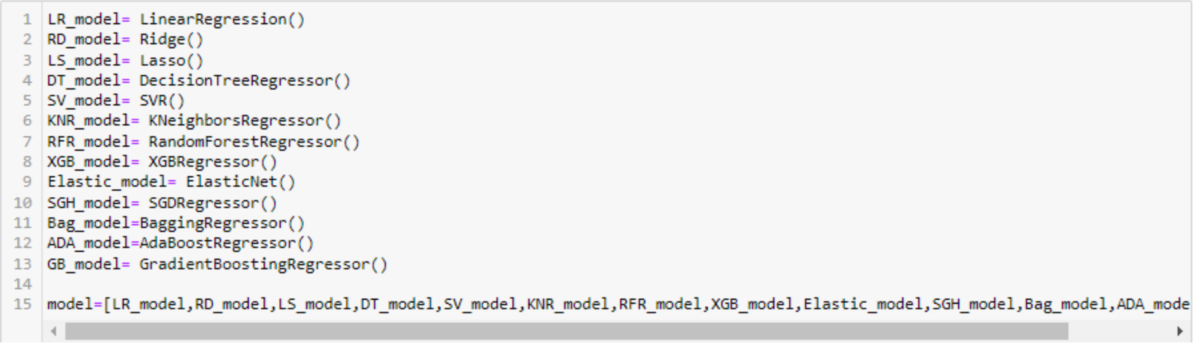
We have used StandardScaler to standardize the continuous features and OneHotEncoding to encode categorical features into integer feature.

- Testing of Identified Approaches (Algorithms)

We have used several available Regressor Algorithms



```python
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Ridge, Lasso
from sklearn.tree import DecisionTreeRegressor
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.ensemble import RandomForestRegressor
from xgboost import XGBRegressor
from sklearn.linear_model import ElasticNet
from sklearn.linear_model import SGDRegressor
from sklearn.ensemble import BaggingRegressor
from sklearn.ensemble import AdaBoostRegressor
from sklearn.ensemble import GradientBoostingRegressor
```

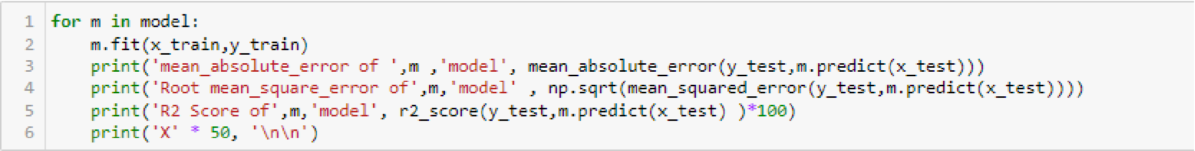


```python
LR_model= LinearRegression()
RD_model= Ridge()
LS_model= Lasso()
DT_model= DecisionTreeRegressor()
SV_model= SVR()
KNR_model= KNeighborsRegressor()
RFR_model= RandomForestRegressor()
XGB_model= XGBRegressor()
Elastic_model= ElasticNet()
SGH_model= SGDRegressor()
Bag_model=BaggingRegressor()
ADA_model=AdaBoostRegressor()
GB_model= GradientBoostingRegressor()

model=[LR_model,RD_model,LS_model,DT_model,SV_model,KNR_model,RFR_model,XGB_model,Elastic_model,SGH_model,Bag_model,ADA_mode
```

We have calculated RMSA for all the available Algos then decide to which one to proceed for model building

- Run and Evaluate selected models



```python
for m in model:
    m.fit(x_train,y_train)
    print('mean_absolute_error of ',m ,'model', mean_absolute_error(y_test,m.predict(x_test)))
    print('Root mean_square_error of',m,'model' , np.sqrt(mean_squared_error(y_test,m.predict(x_test))))
    print('R2 Score of',m,'model', r2_score(y_test,m.predict(x_test) )*100)
    print('X' * 50, '\n\n')
```

We have calculated MAE, RMSE and R2 score for all ML Algorithms.

- Key Metrics for success in solving problem under consideration

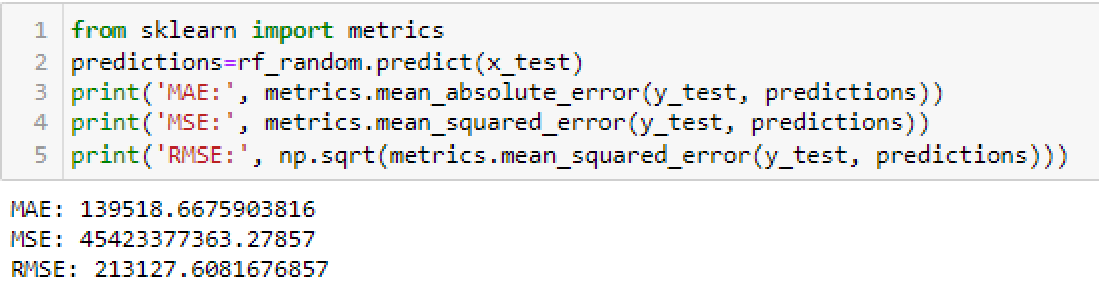
The regression model can be evaluated on following parameters:
1. Mean Square Error (MSE):

MSE is the single value that provides information about goodness of regression line. Smaller the MSE value, better the

fit because smaller value implies smaller magnitude of errors.
$\boldsymbol{MSE}= \mathbf{1}\boldsymbol{N}\Sigma|\boldsymbol{yi}-\boldsymbol{y}|\mathbf{2}\boldsymbol{Ni}=\mathbf{1}$
Equation 3 MSE equation
2. Root Mean Square Error (RMSE):

RMSE is the quadratic scoring rule that also measures the average magnitude of the error. It is the square root of average squared difference between prediction and actual observation.
3. Mean Absolute Error (MAE):

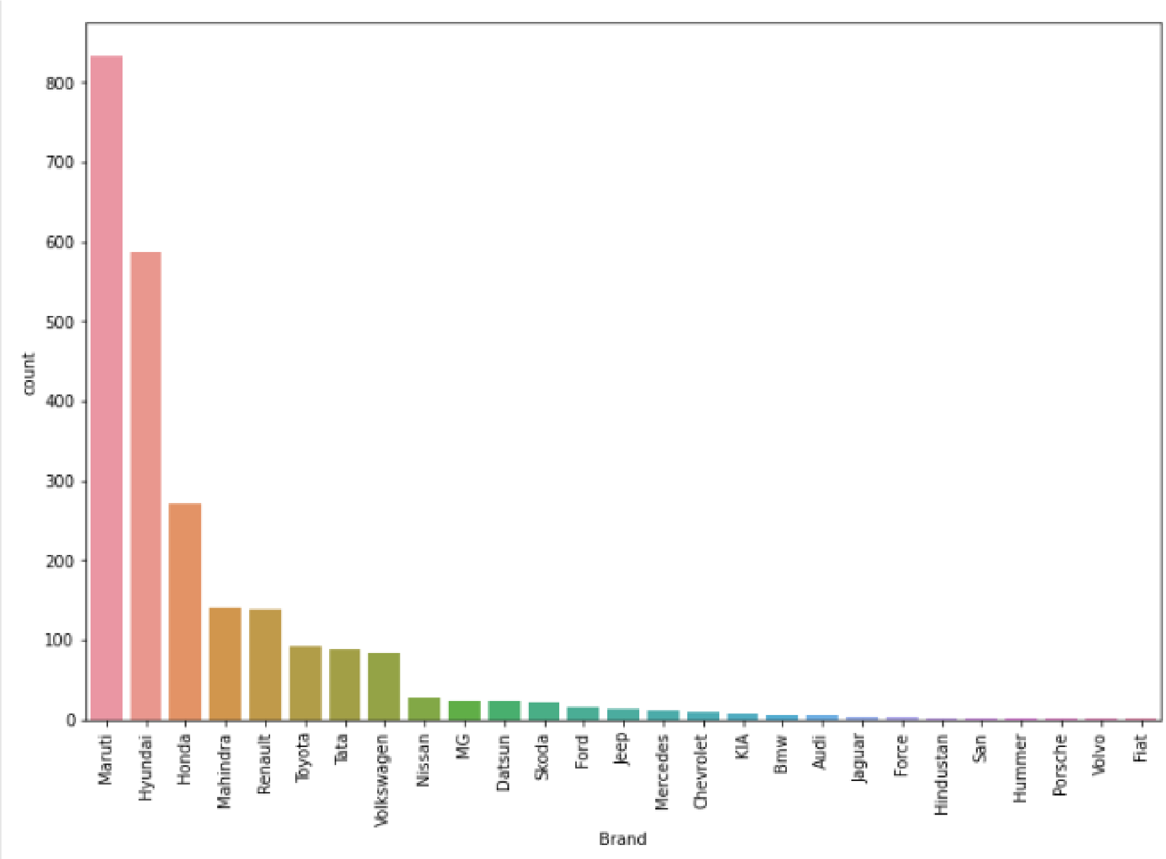
This measure represents the average absolute difference between the actual and predicted values in the dataset. It represents the average residual from the dataset. $\boldsymbol{MAE}= \mathbf{1}\boldsymbol{N}\Sigma|\boldsymbol{yi}-\boldsymbol{y}|$



```python
from sklearn import metrics
predictions=rf_random.predict(x_test)
print('MAE:', metrics.mean_absolute_error(y_test, predictions))
print('MSE:', metrics.mean_squared_error(y_test, predictions))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, predictions)))
```

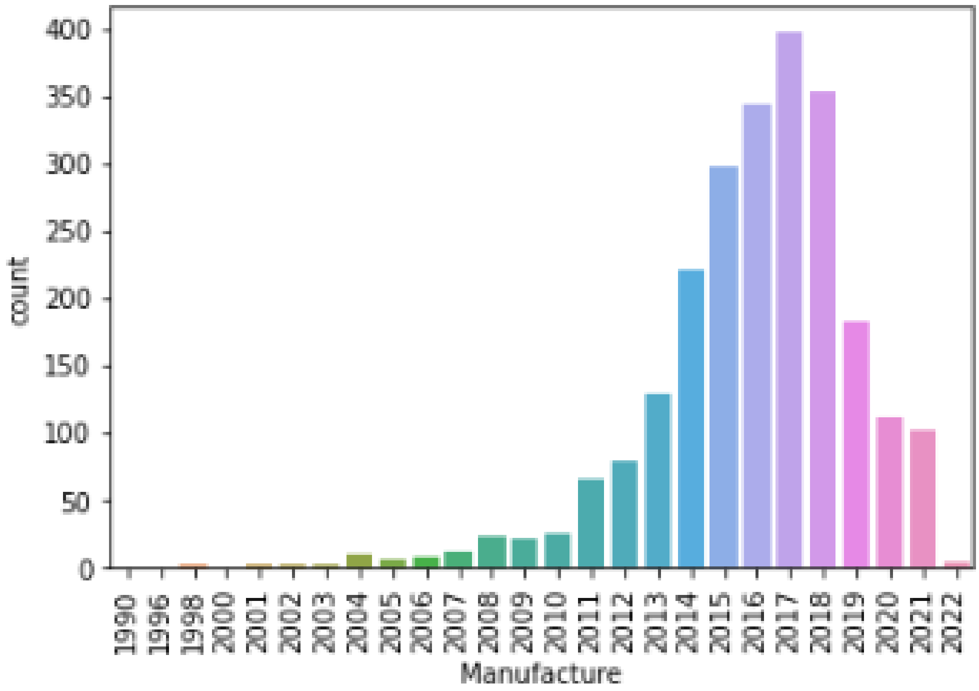
```
MAE: 139518.6675903816
MSE: 45423377363.27857
RMSE: 213127.6081676857
```

Visualizations



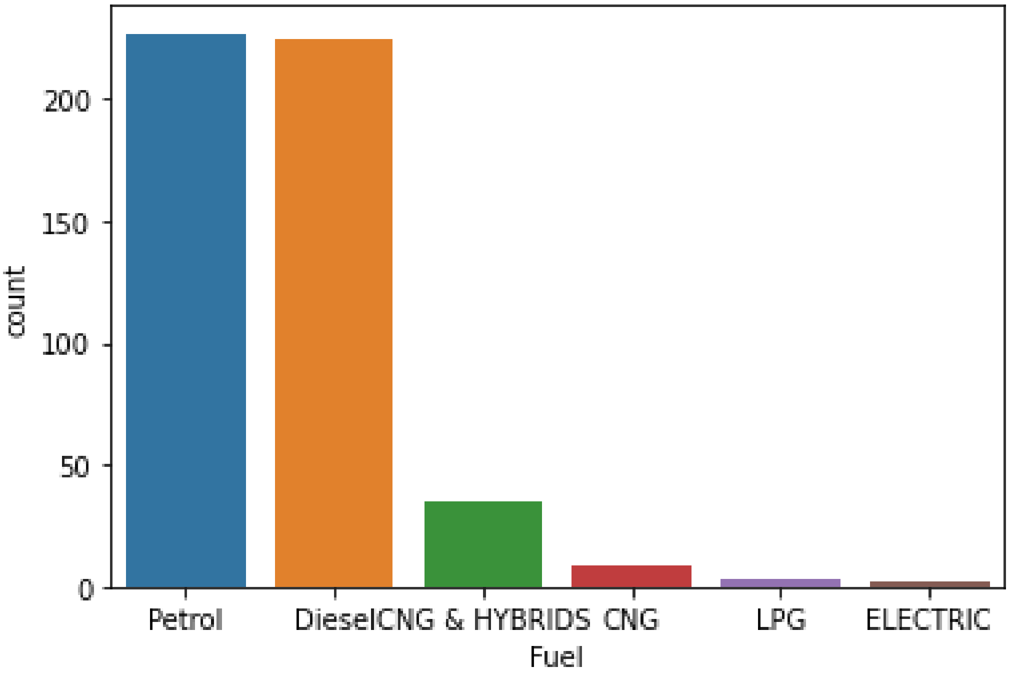
In our dataset: 34.6% cars are of Maruti and 24% are of Hyundai

11% are of Honda



Most of the Cars are available for resale are from the year 2014 to 2021. Absolutely, No one would buy very old car.

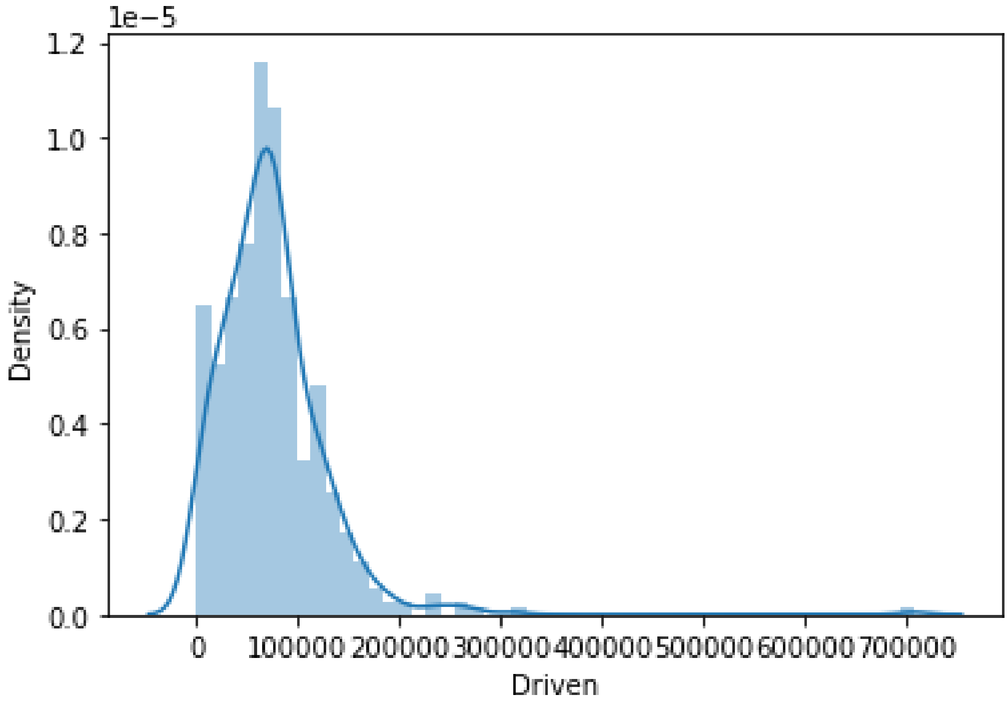
India also have Scrap law of 10years of Diesel car and 15 yrs for Petrol Car.



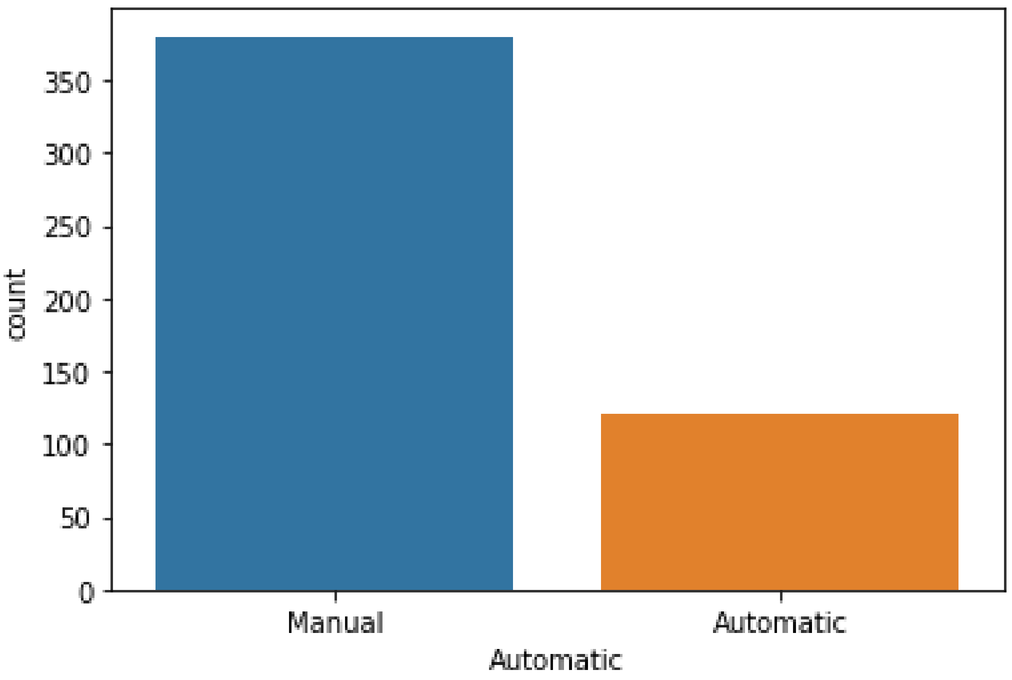
71.56% cars are on Petrol

25.49% cars are on Diesel

Majority of Cars in India are on Petrol and Diesel. However, India has started to manufacture Electric cars. So In coming years, we would see large number of electric cars in Indian market.

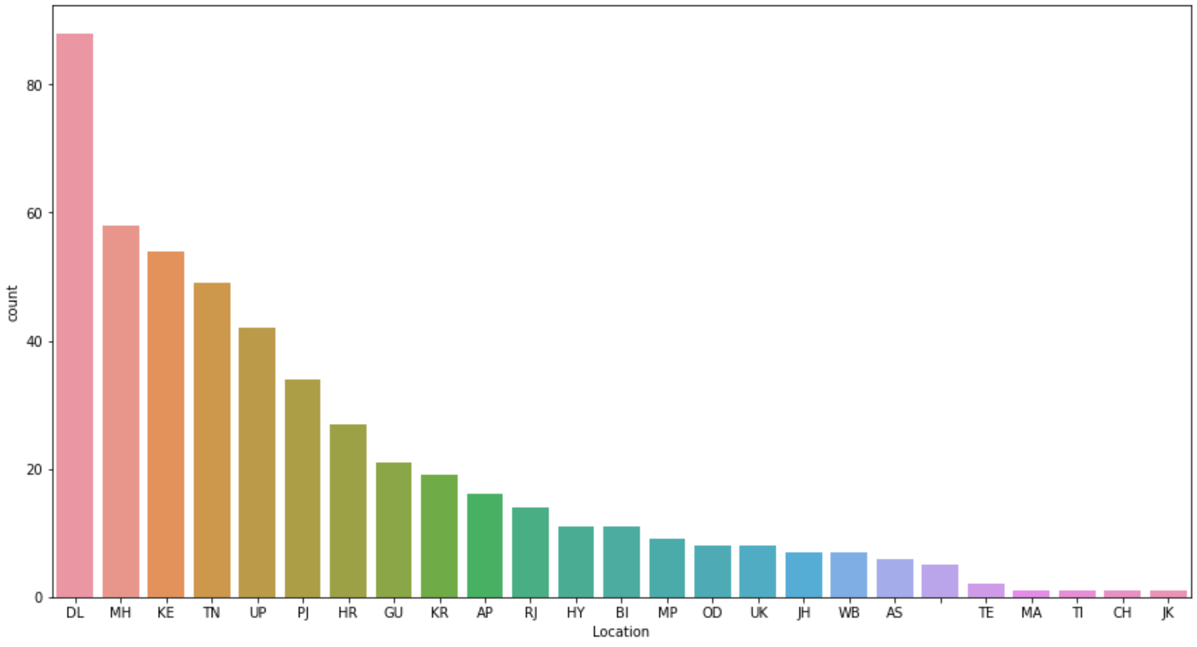


Distribution of KM driven by the old cars are right skewed but having a bell-curve shape

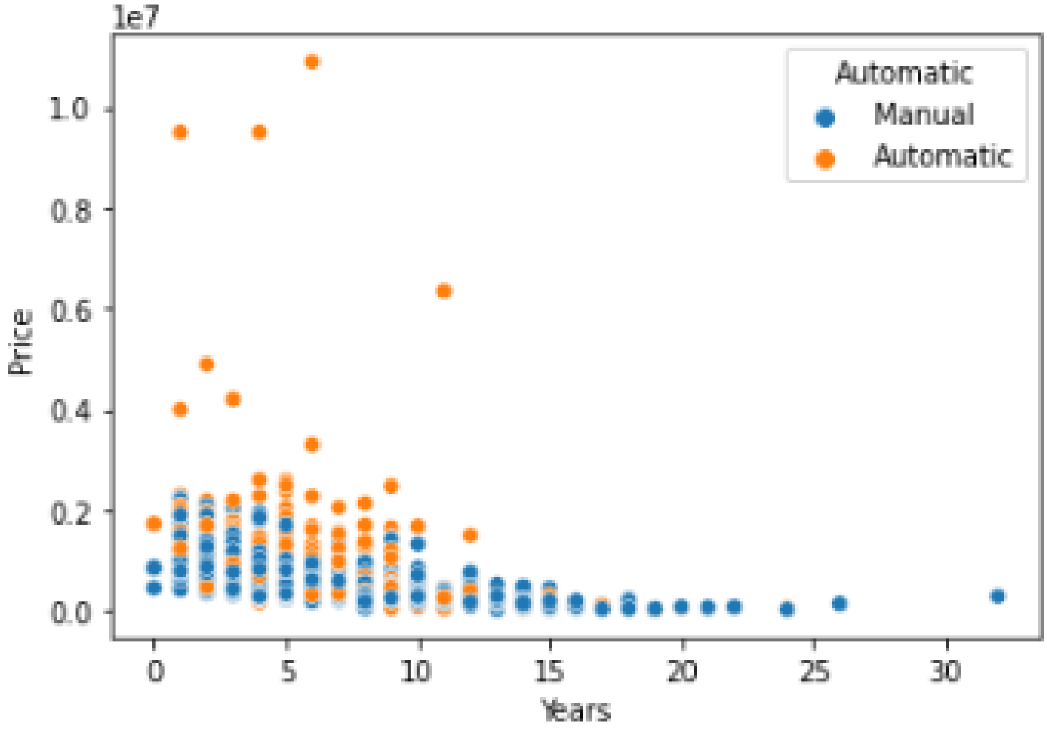


Most cars are manually operated

And remaining are Automatic.



Most of the Cars are from Maharastra and then Delhi



Before 15 years, India hardly had Automated car while in recents years we had many Automated Car’s.

This graphs also tells that Automated car price is relatively high than manual car.

CONCLUSION

- Key Findings and Conclusions of the Study

Using data mining and machine learning approaches, this project proposed a scalable framework for India based used cars price prediction. Car24.com and CarDekho websites was scraped using the Selenium scraping tool to collect the benchmark data. An efficient machine learning model is built by training, testing, and evaluating three machine learning regressors named Random Forest Regressor, Linear Regression, and Bagging Regressor. As a result of pre-processing and transformation, Random Forest Regressor came out on top with 69% accuracy.

Each experiment was performed in real-time Jupyter notebook.

- Limitations of this work and Scope for Future Work

  The intelligent model will then be included in online and mobile applications for general use. In addition, after the data gathering phase, the pandemic-related shortages of semiconductors caused a rise in automobile costs and had a significant impact on the used-car market. Therefore, it is necessary to periodically gather and analyse data; ideally, we would use a real-time processing tool.