-

USED CAR PRICE PREDICTION PROJECT

BACHELOR OF TECHNOLOGY
in
COMPUTER SCIENCE AND ENGINEERING

Rajiv Gandhi University of Knowledge Technologies

R.K.VALLEY

Submitted by:-

K.C.Nandini (R170021)

Under the Esteemed guidance of

Ms.S.Shabana
Assistant Professor

Department of Computer Science and Engineering

RGUKT RKValley

Rajiv Gandhi University of Knowledge Techonolgies

RK Valley, Kadapa (Dist), Andhra Pradesh, 516330

CERTIFICATE

This is to certify that the project work titled “**USED CAR PRICE PREDICTION**” is a bonafied project work submitted by **K.Dhana Lakshmi and K.C.Nandini** in **COMPUTER SCIENCE AND ENGINEERING** in partial fulfillment of requirements for the award of degree of **Bachelor of Technology** for the year **2022-2023**carried out the work under the supervision.

INTERNAL GUIDE
(S SHABANA,Asst.Professor)

HEAD OF THE DEPARTMENT
(N SATYANANDARAM)

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose constant guidance and encouragement crown all the efforts success.

I am extremely grateful to our respected Director,Prof. K. SANDHYA RANI for fostering an excellent academic climate in our institution.

I also express my sincere gratitude to our respected Head of the Department Mr SATYANANDARAM sir for his encouragement, overall guidance in viewing this project a good asset and effort in bringing out this project.

I would like to convey thanks to our guide at college S SHABANA for her guidance, encouragement, co-operation and kindness during the entire duration of the course and academics.

My sincere thanks to all the members who helped me directly and indirectly in the completion of project work. I express my profound gratitude to all our friends and family members for their encouragement.

INDEX

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

Today, the transportation industry is considered to be one of the backbones of the economy. Automobiles are referred to as the "Industry of Industries" in developed nations. According to industry professionals, the INDIA's automotive industry has seen remarkable growth. Besides being the fastest-growing nation in the automobile industry,it represents its global presence. In India, like most other countries, cars are gaining a great deal of popularity among the local population and the ex-pat community who work in the country. There are used cars for sale in the India of all makes. Almost everyone wants their own car these days, but because of factors like affordability or economic conditions, many prefer to opt for pre-owned cars. Accurately predicting used car prices requires expert knowledge due to the nature of their dependence on a variety of factors and features. Used car prices are not constant in the market, both buyers and sellers need an intelligent system that will allow them to predict the correct price efficiently. In this intelligent system, the most difficult problem is the collection of the dataset which contains all important elements like the manufacturing year of the car, its gas type, its condition, miles driven, horsepower, doors, number of times a car has been painted, customer reviews, the weight of the car, etc. It is clear that the price of the product is affected by many factors, but unfortunately, information about these features is not always readily available. Since this project primarily focuses on the Indian market,the benchmark dataset containing.

Purpose

The purpose of this study is to understand and evaluate used car prices in the India, and to develop a strategy that utilizes data mining techniques to predict used car prices.

Scope

This project aims to deliver price prediction models to the public,to help guide the individuals looking to buy or sell cars and to give them a better insight into theautomotive sector. Buying a used car from a dealer can be a frustrating and an unsatisfying experience as some dealers are known to deploy deceitful sale tactics toclose a deal. Therefore, to help consumers avoid falling victims to such tactics,this study hopes to equip consumers with right tools to guide them in their shopping experience.Another goal of the project is to explore new methods to evaluate used cars prices and to compare their accuracies. Considering this is an interesting research topic in the research community, and in continuing their footsteps, we hope to achieve significant results using more advanced methods of previous work.

Software and Hardware Requirements

HARDWARE:

Ram	:	16GB
Hardisk	:	512GB
Processor	:	2GHz

SOFTWARE:

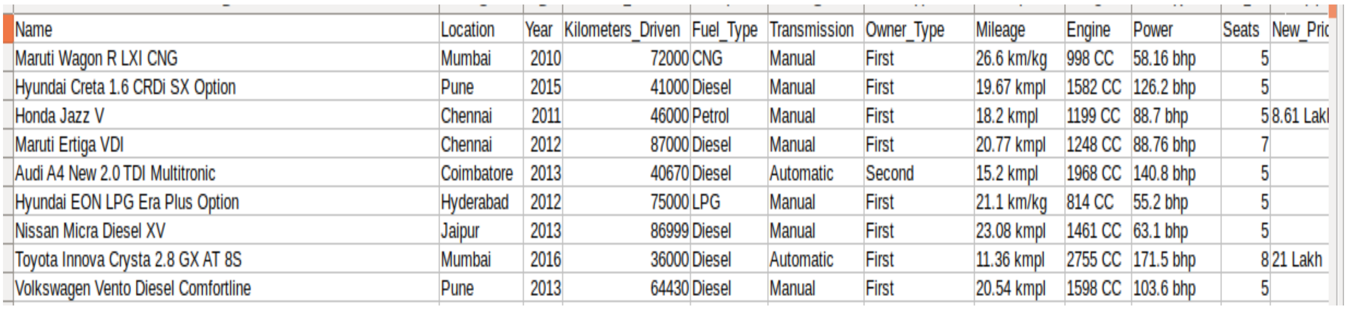
Language	:	Python3 ,HTML,CSS.
Tools	:	Latest Anakonda For Jupiter.
Additional Python Libraries	:	Numpy,Pandas,seaborn, matplotlib, scikit-learn,

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem

Data Understanding and preparation is an essential part of building a model as it gives the insight into the data and what corrections or modifications shall be done before designing and executing the model,preliminary analysis of the data must be done to have deeper understanding into the quality of the data, in terms of outliers and the skewedness the figures,descriptive Statistics of categorical andnumerical variables was done for that to be achieved. As well as the ability to understand the main attributes that affect the results of the price. That was done through a correlation matrix for every attribute to understand the relations between the different factors.

Data Sources and theirformats



Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Pric
Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5	
Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5	
Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5	8.61 Lakl
Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7	
Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5	
Hyundai EON LPG Era Plus Option	Hyderabad	2012	75000	LPG	Manual	First	21.1 km/kg	814 CC	55.2 bhp	5	
Nissan Micra Diesel XV	Jaipur	2013	86999	Diesel	Manual	First	23.08 kmpl	1461 CC	63.1 bhp	5	
Toyota Innova Crysta 2.8 GX AT 8S	Mumbai	2016	36000	Diesel	Automatic	First	11.36 kmpl	2755 CC	171.5 bhp	8	21 Lakh
Volkswagen Vento Diesel Comfortline	Pune	2013	64430	Diesel	Manual	First	20.54 kmpl	1598 CC	103.6 bhp	5	

11 Features have been scrapped.

- 1. Name: Car complete name
- 2. Location: which state of India
- 3. Year: Car manufacturer year
- 4. Kilometers Driven: Total driven km by car
- 5. Fuel: Which fuel is being used in the car(Petrol/Diesel/CNG etc)
- 6. Driven: Total driven km by car
- 7. Transmission: if the car is Manual or Automatic
- 8. Owner: How many owners have been changed of the car
- 9.Mileage: Mileage of the car
- 10.Engine: Type of the engine.
- 11.Power: Power of the car
- 12.Seats: Number seats in car
- 13.New Price:Price of the car by average
- 14.Price: The actual price of the car by the features.

Data Preprocessing Done

After data collection the dataset was pre-processed to remove samples that have missing value,and remove non-numerical part from numerical attributes, converting categorical values into numerical (if needed), fix any discrepancies in the units, as well as removing attributes that doesn’t affect the price evaluations if needed to reduce the complexity of the Model Data Understanding and preparation is an essential part of building a model as it gives the insight into the data and what corrections or modifications shall be done before designing and executing the model,preliminary analysis of the data must be done to have deeper understanding into the quality of the data, in terms of outliers and the skewedness of the figures, descriptive Statistics of categorical and numerical variables was done for that to be achieved. As well as the ability to understand the main attributes that affect the results of the price. That was done through a correlation matrix for every attribute to understand the relations between the different factors.

Data Inputs- Logic- Output Relationships

Afterwards when the data is organized and transformed into a form that could be processed by the data mining technique. Different data mining models were designed to predict prices and values of used cars. In this study three models are proposed to be built using Logistic Regression model technique Random Forest Regressor and Bagging Regressor.Firstly, the data was portioned into section for training and the other part for testing, portioning percentage can be tested with different ratios to analyse different results.All three models were evaluated on four evaluation matrices known as model score, Mean Square Error (MSE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).From all, the Random Forest Regressor outperformed .Describe the relationship behind the data input, its format, the logic in between and the output. Describehow the input affects the output.

State the set of assumptions (if any) related

The problem under considerationIn the past year the world of automobiles has seen a drastic change with the semiconductor shortages after the pandemic, which led to spike in used car prices. Hence, there was fast change in car prices during this study which will affect the actual car pricing prediction future. As the current dataset will undervalue the cars in the market. Therefore, a model that is built on real time data can be best integrated into a mobile app for public use would be the idea solution.

Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods)

Pre-processing is a Data Mining technique that involves converting raw data into acomprehensibleformat. There is often a lack of specific activity or trend data, and many inaccurate facts are included inreal-world data. Consequently, this may result in poor-quality data collection, and, in turn, poor-qualitymodels constructed from the data. Such problems can be resolved by pre-processing the data. Pre-processing in Machine Learning is the process of modifying, or encoding, data so that the machine can parse it more easily.Thus, the algorithm can now properly interpret the data. In this project, following steps are preformed to pre- process the dataset.

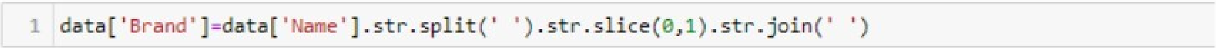
1. Dataset collection: we have collected the dataset from 3 leading websites which deals with used car sell-buy.

2. 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.

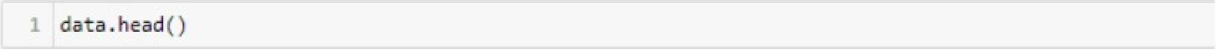
```
Data.dtypes

Name              object
Location          object
Year              object
Driven            object
Fuel              object
Transmission      object
Owner             object
Mileage           object
Engine            object
Power             object
Seats             object
New price         object
Price             object
```

Name

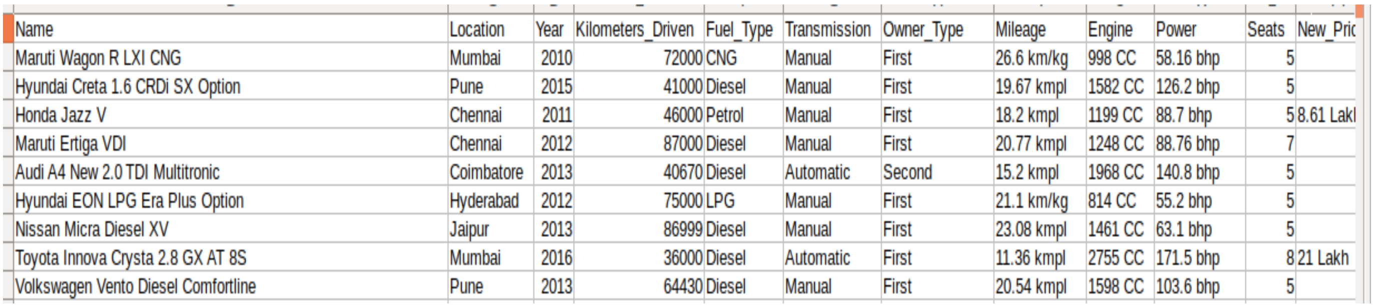
Name feature was having the complete name of the car, while we required the car manufacture company name so we extract this from the car name.



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



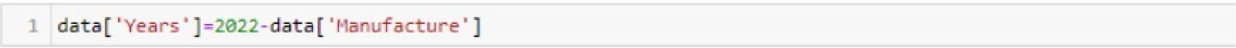
```
data.head()
```



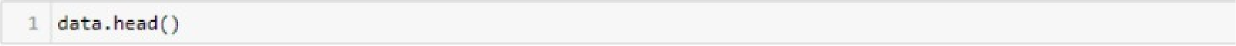
Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Pric
Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5	
Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5	
Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5	8.61 Lakl
Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7	
Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5	
Hyundai EON LPG Era Plus Option	Hyderabad	2012	75000	LPG	Manual	First	21.1 km/kg	814 CC	55.2 bhp	5	
Nissan Micra Diesel XV	Jaipur	2013	86999	Diesel	Manual	First	23.08 kmpl	1461 CC	63.1 bhp	5	
Toyota Innova Crysta 2.8 GX AT 8S	Mumbai	2016	36000	Diesel	Automatic	First	11.36 kmpl	2755 CC	171.5 bhp	8	21 Lakh
Volkswagen Vento Diesel Comfortline	Pune	2013	64430	Diesel	Manual	First	20.54 kmpl	1598 CC	103.6 bhp	5	

Manufacture

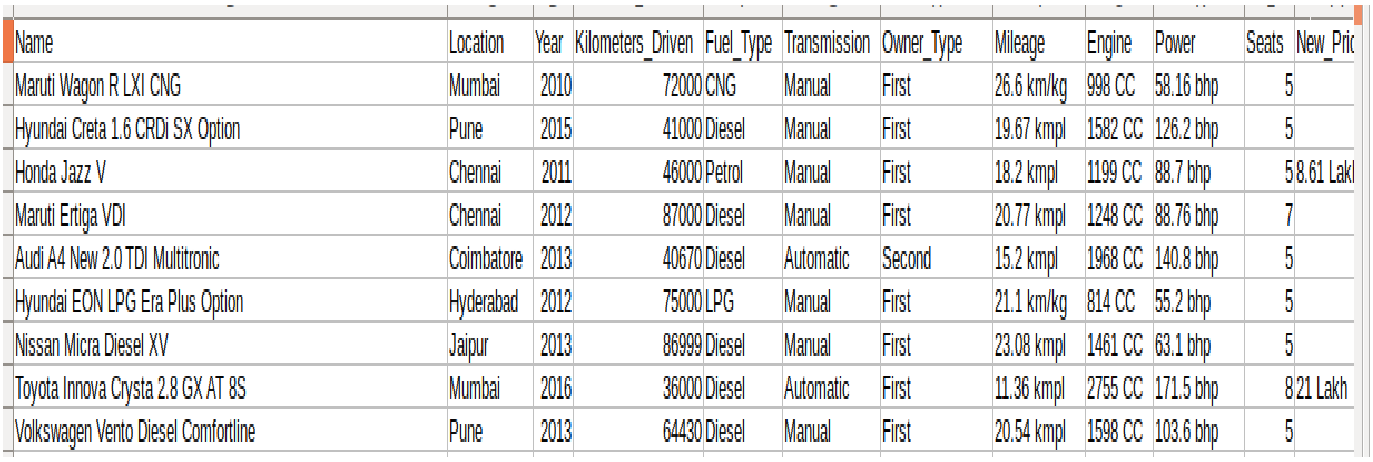
Feature provided the year of car manufacture which is required basically to know the age of that particular car at instance. So we calculated the Age of car at this moment.



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



```
data.head()
```



Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Pric
Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5	
Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5	
Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5	8.61 Lakl
Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7	
Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5	
Hyundai EON LPG Era Plus Option	Hyderabad	2012	75000	LPG	Manual	First	21.1 km/kg	814 CC	55.2 bhp	5	
Nissan Micra Diesel XV	Jaipur	2013	86999	Diesel	Manual	First	23.08 kmpl	1461 CC	63.1 bhp	5	
Toyota Innova Crysta 2.8 GX AT 8S	Mumbai	2016	36000	Diesel	Automatic	First	11.36 kmpl	2755 CC	171.5 bhp	8	21 Lakh
Volkswagen Vento Diesel Comfortline	Pune	2013	64430	Diesel	Manual	First	20.54 kmpl	1598 CC	103.6 bhp	5	

Fuel

Feature were having many garbage inputs sowe 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.

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 encodecategorical features into integer feature.

Mileage

Feature were having how much the mileage given by the car it is a major important thing to measure while buying a car.

Owner type

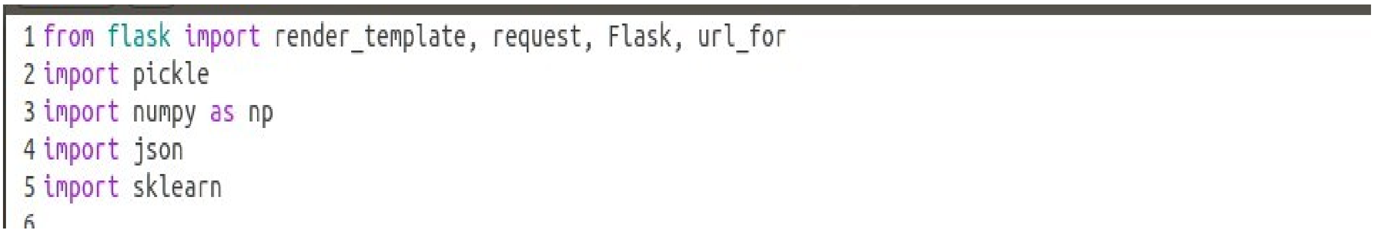
Fetures of knowing whether the car is new or used by someone else before like owner of type first or owner of type second etc.

Engine

Feature of knowing which type engine is of the car there are different type of engines are there engine also a main factor to predict the price of the car.

Testing of Identified Approaches (Algorithms)

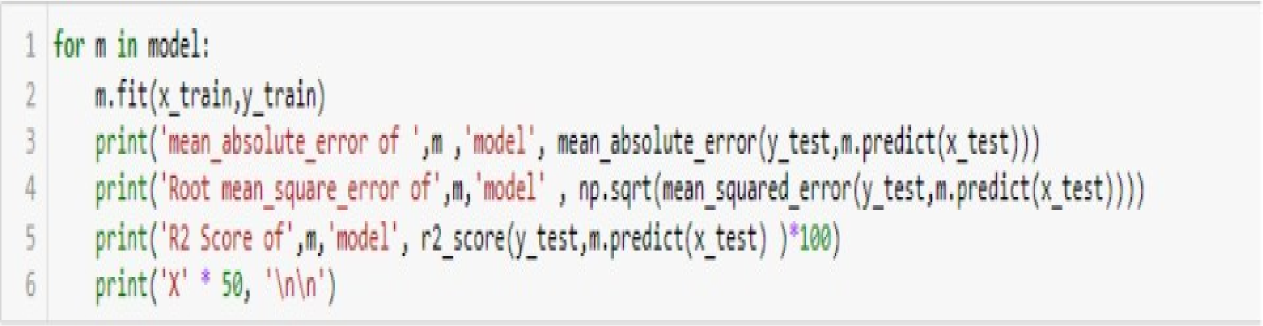
We have used several available Regressor Algorithms



```
from flask import render_template, request, Flask, url_for
import pickle
import numpy as np
import json
import sklearn
```

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

Run and Evaluate selected models



```
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')
```

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. MSE= 1NΣ|yi−y|2Ni=1yi−y|yi−y|2Ni=12Ni=1.

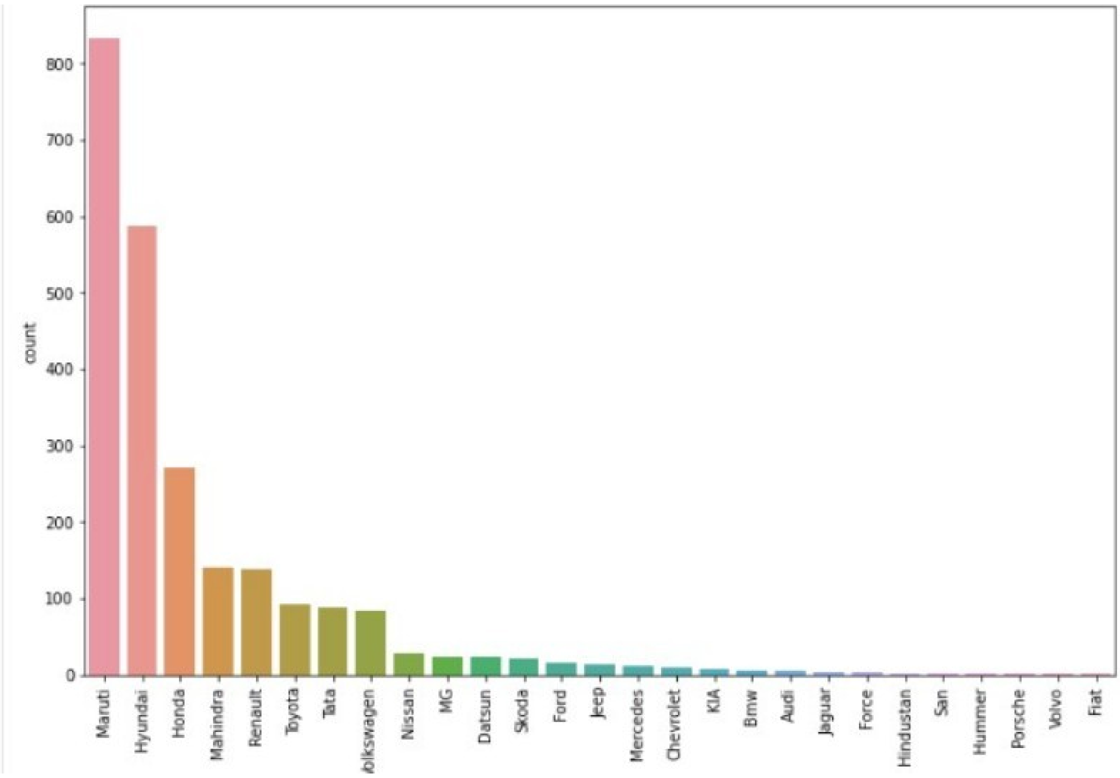
2. Root Mean Square Error (RMSE):

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. MAE= 1NΣ|yi−y|2Ni=1yi−y|yi−y|2Ni=1.

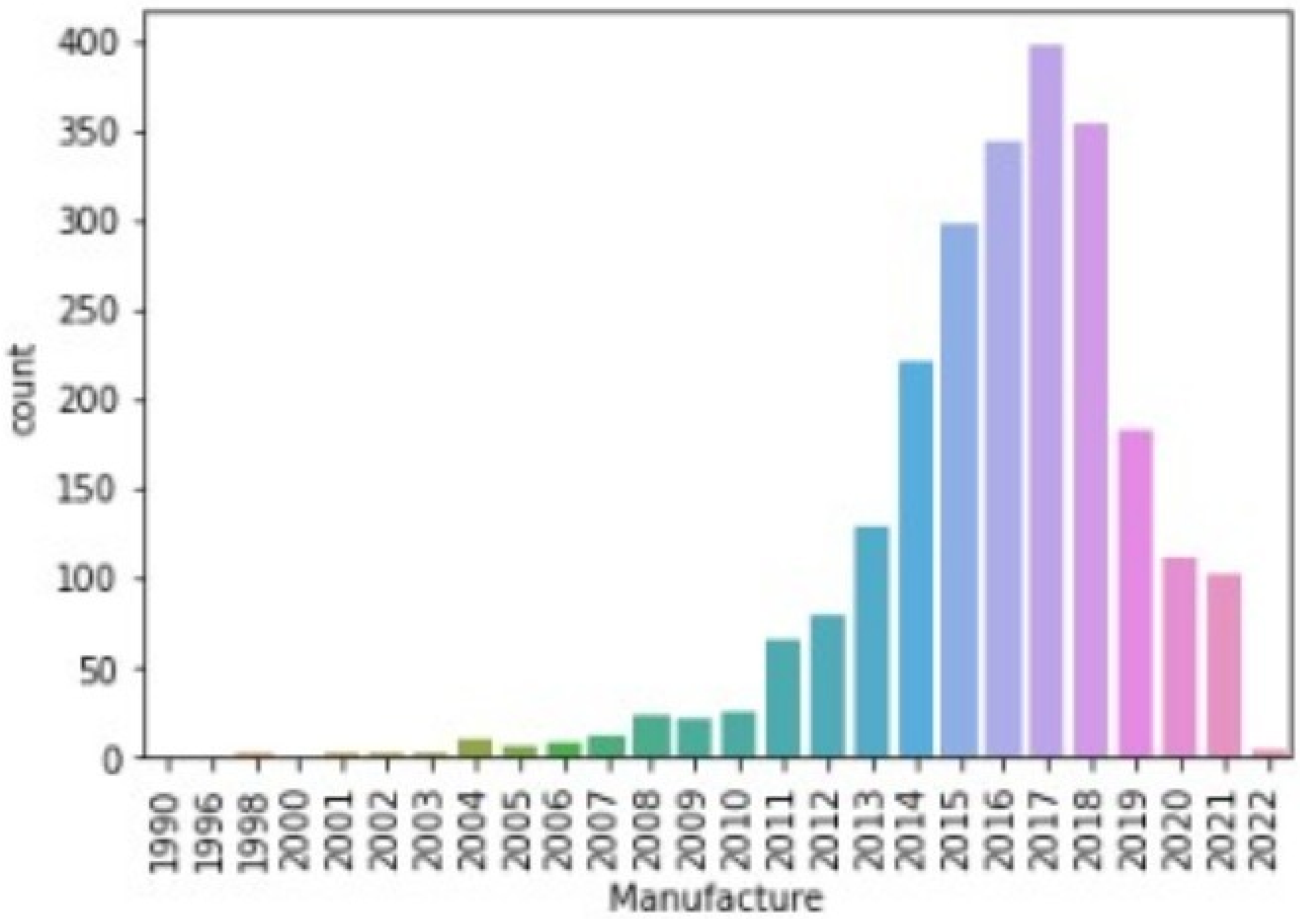
Visualizations



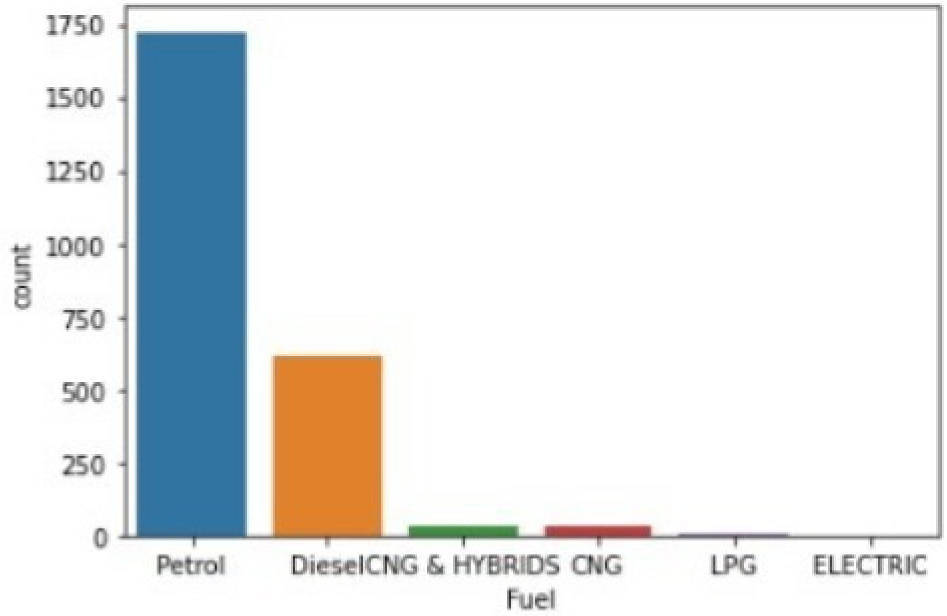
In our dataset: 34.6% cars are of Maruti and 24% are of Hyundai 11% are of Honda.

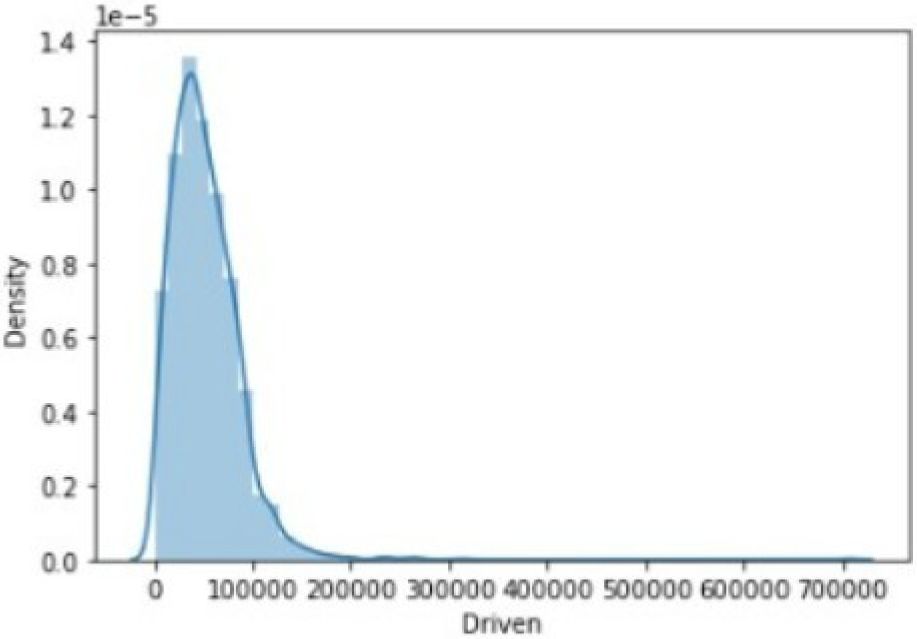
Most of the Cars are available for resale are from the year 2014 to 2021.Absolutely, No one would buy very old car.

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

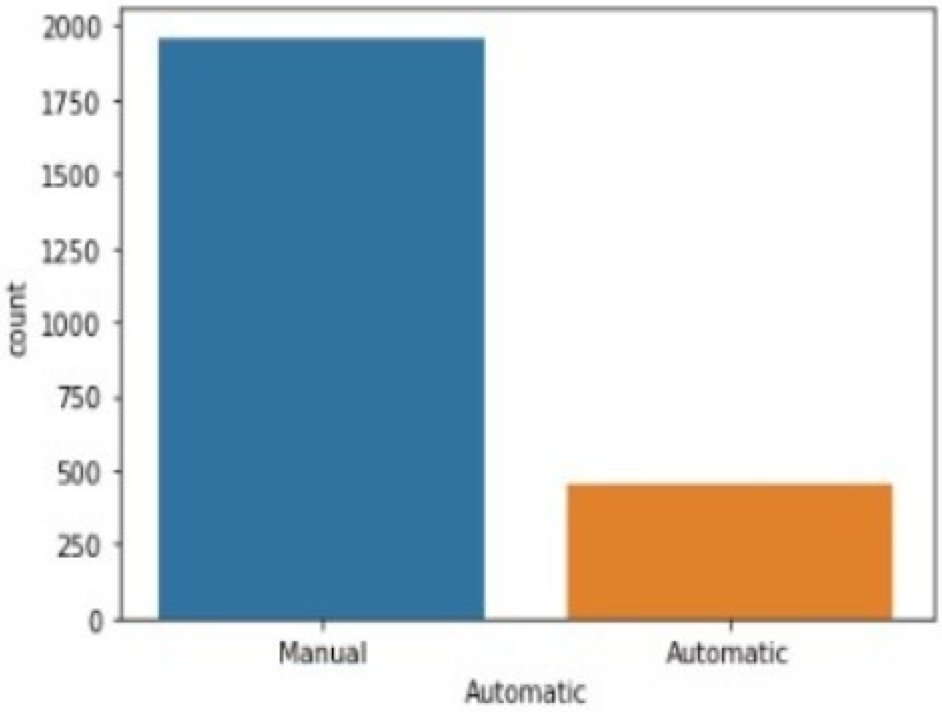


71.56% cars are on Petrol
25.49% cars are on Diesel
Majority of Cars in India are on Petrol and Diesel.
However, India has started to manufacture Electric cars. So In coming years, we would see large number of electric cars in Indian market.

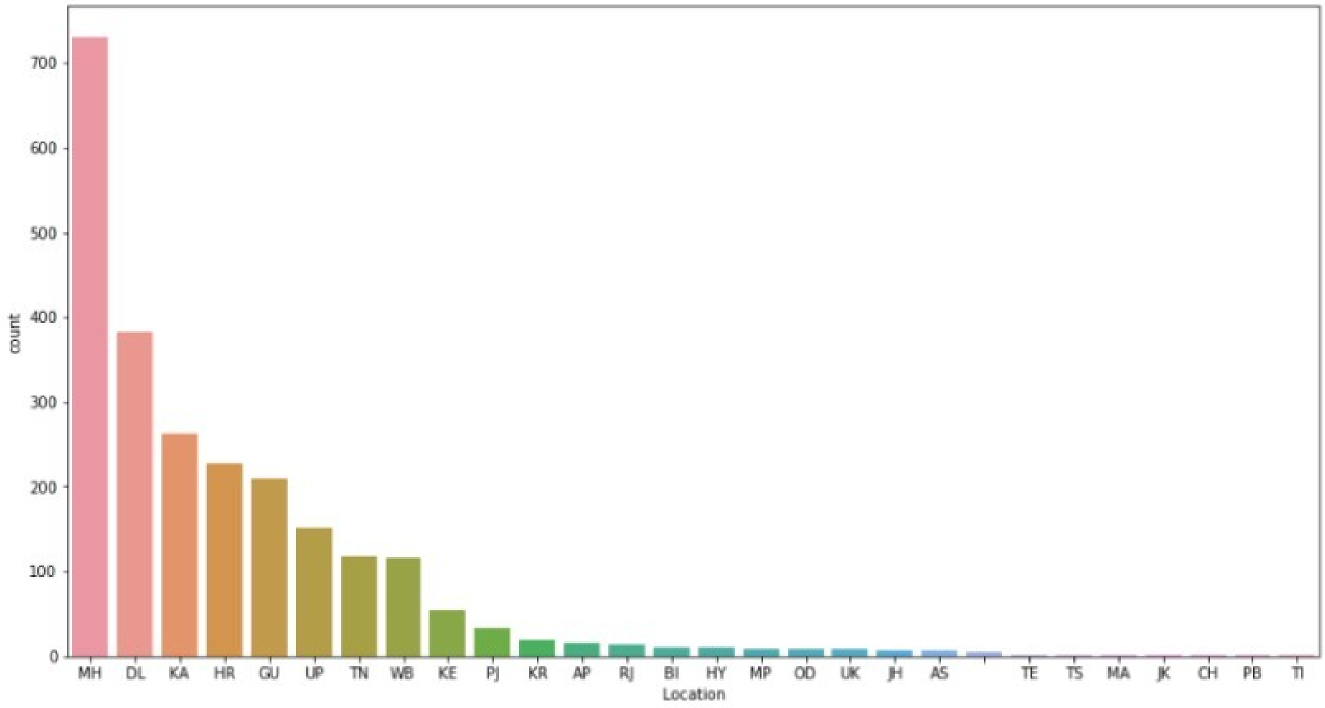


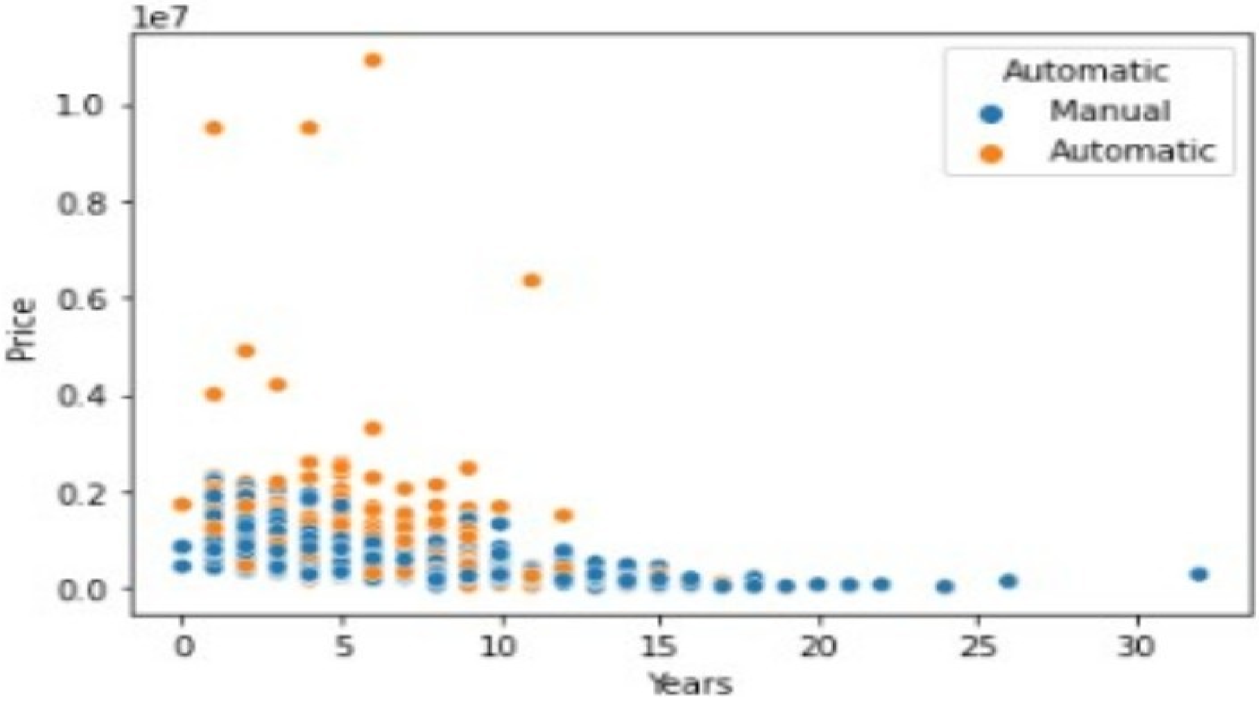


Distribution of KM driven by the old cars are right skewed but having a bell-curve shape 81.32% cars are manually operated And remaining are Automatic.



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, olx.com and cardekho websitees was scraped using the Selenium scraping tool to collect the benchmark data. An efficient machine learning model is built by training, testing, and evaluating two machine learning regressors named Random Forest Regressor,Linear Regression,. As a result of pre-processing and transformation, Random Forest Regressor came out on top with 64% accuracy.Each experiment was performed in real-time Jupyter notebook .

Limitations of this work and Scope for Future Work

In the future, more data will be collected using different web-scraping techniques, and deep learning classifiers will be tested.Algorithms like Quantile Regression, ANN and SVM will be tested.Afterwards, the intelligent model will be integrated with web and mobile-based applications for public use.Moreover, after the data collection phase Semiconductor shortages have incurred after the pandemic which led to an increase in car prices, and greatly affected the secondhand market. Hence having a regular Data collection and analysis is required periodically, ideally, we would be having areal time processing program.

References

https://www.autopijaca.ba/

http://paperpile.com/b/MhSbIM/caUu

http://www.bhas.ba/

https://digitalcc.coloradocollege.edu/islandora/object/coccc%3A1346

http://scikit-learn.org/stable/modules/generated sklearn.ensemble.RandomForestClassifier.html