CAR RESALE VALUE PREDICTION

1. INTRODUCTION

1.1 PROJECT OVERVIEW:

With difficult economic conditions, it is likely that sales of second-hand imported (reconditioned) cars and used cars will increase. In many developed countries, it is common to lease a car rather than buying it outright. After the lease period is over, the buyer has the possibility to buy the car at its residual value, i.e., its expected resale value. Thus, it is of commercial interest to sellers/financers to be able to predict the salvage value (residual value) of cars with accuracy.

1.2 PURPOSE:

To predict the resale value of the car, we proposed an intelligent, flexible, and effective system that is based on using regression algorithms. Considering the main factors which would affect the resale value of a vehicle a regression model is to be built that would give the nearest resale value of the vehicle. We will be using various regression algorithms and algorithm with the best accuracy will be taken as a solution, then it will be integrated to the web-based application where the user is notified with the status of his product.

2. LITERATURE SURVEY

Title: Used Cars Price Prediction using Supervised Learning Techniques **Author:**

Mukkesh Ganesh , Pattabiraman Venkatasubbu **Year:** December 2019

Abstract:

The production of cars has been steadily increasing in the past decade, with over 70 million passenger cars being produced in the year 2016. This has given rise to the used car market, which on its own has become a booming industry. The recent advent of online portals has facilitated the need for both the customer and the seller to be better informed about the trends and patterns that determine the value of a used car in the market. Using Machine Learning Algorithms such as Lasso Regression, Multiple Regression and Regression trees, we will try to develop a statistical model which will be able to predict the price of a used car, based on previous consumer data and a given set of features. We will also be comparing the prediction accuracy of these models to determine the optimal one.

Title: Used Cars Price Prediction and Valuation using Data Mining Techniques **Author:** Abdulla AlShared **Year:** December 2021 **Abstract:**

Due to the unprecedented number of cars being purchased and sold, used car price prediction is a topic of high interest. Because of the affordability of used cars in developing countries, people tend more purchase used cars. A primary objective of this project is to estimate used car prices by using attributes that are highly correlated with a label (Price). To accomplish this, data mining technology has been employed. Null, redundant, and missing values were removed from the dataset during pre-processing. In this supervised learning study, three regressors (Random Forest Regressor, Linear Regression, and Bagging Regressor) have been trained, tested, and compared against a benchmark dataset. Among all the experiments, the Random Forest Regressor had the highest score at 95%, followed by 0.025 MSE, 0.0008 MAE, and 0.0378 RMSE respectively. In addition to Random Forest Regression, Bagging Regression performed well with an 88% score, followed by Linear Regression having an 85% mark. A train-test split of 80/20 with 40

random states was used in all experiments. The researchers of this project anticipate that in the near future, the most sophisticated algorithm is used for making predictions, and then the model will be integrated into a mobile app or web page for the general public to use.

Title: Used Car Price Prediction

Author: Praful Rane, Deep Pandya, Dhawal Kotak

Year: April 2021

Abstract:

The price of a new car in the industry is 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 prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. Existing System includes a process where a seller decides a price randomly and buyer has no idea about the car and it's value in the present day scenario. In fact, seller also has no idea about the car's existing value or the price he should be selling the car at. To overcome this problem we have developed a model which will be highly effective. Regression Algorithms are used because they provide us with continuous value as an output and not a categorized value. Because of which it will be possible to predict the actual price a car rather than the price range of a car. User Interface has also been developed which acquires input from any user and displays the Price of a car according to user's inputs.

Title: Car Price Prediction Using Machine Learning

Author: Ketan Agrahari, Ayush Chaubey, Mamoor Khan, Manas Srivastava **Year:** June 2021 **Abstract:**

The demand for used cars has increased significantly in the past decade and it is prognosticated that with Covid-19 outbreak this requirement will augment considerably. Hence to enhance the reliability, with the expansion of the used car market, a model that can forecast the current market price of a used automobile on the basis of a variety of criteria. This analysis can be used to study the trends in the industry, offer better insight into the market, and aid the community in its smooth workflow. The aim of this research paper is to predict the car price as per the data set (previous consumer data like engine capacity, distance traveled, year of manufacture, etc.). The result of these algorithms will be analyzed and based on the efficiency and accuracy of these algorithms, the best one of them can be used for the said purpose.

Title: Vehicle Price Prediction using SVM Techniques

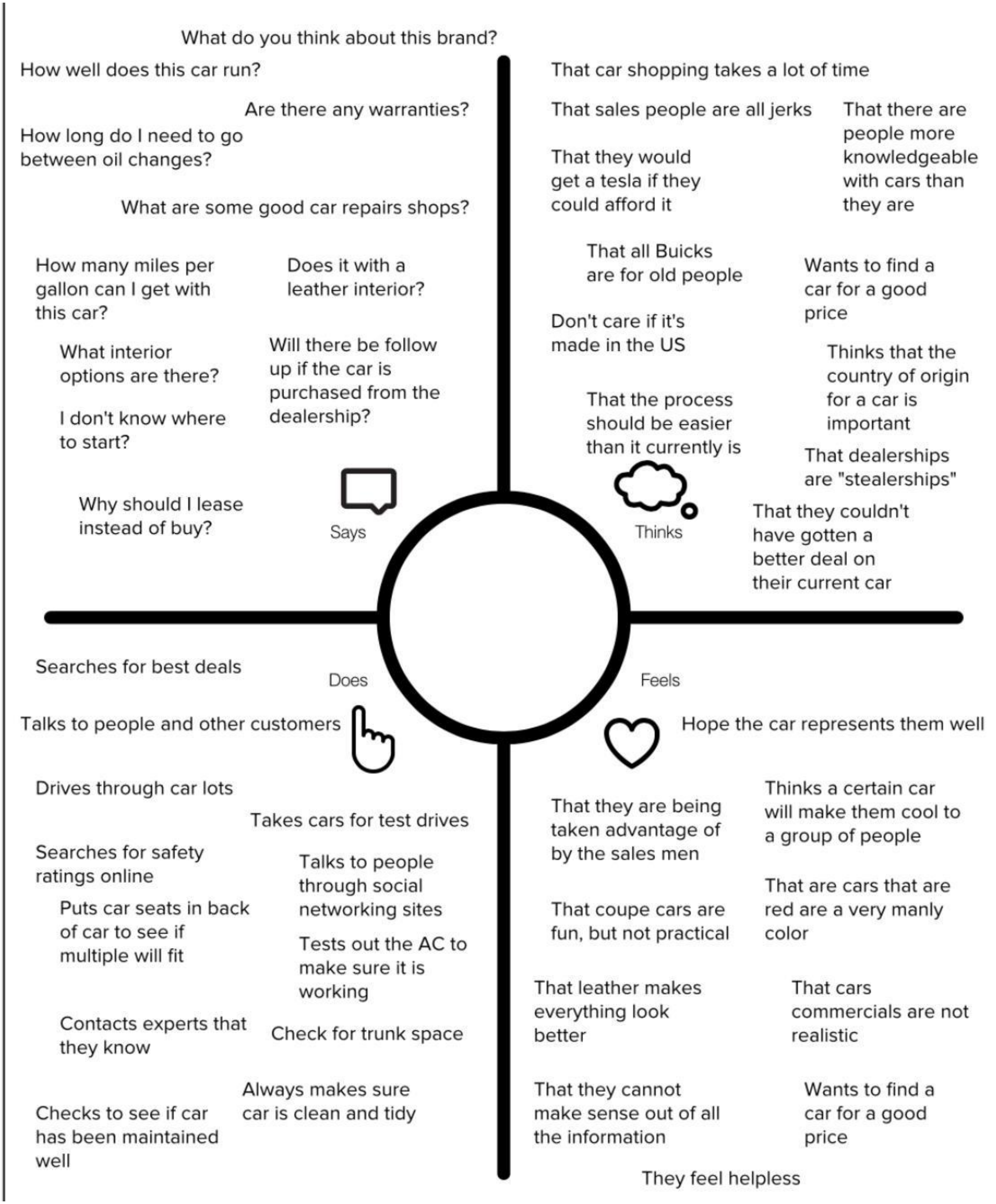
Author: S.E.Viswapriya, Durbaka Sai Sandeep Sharma, Gandavarapu Sathya kiran

Year: June 2020

Abstract:

The prediction of price for a vehicle has been more popular in research area, and it needs predominant effort and information about the experts of this particular field. The number of different attributes is measured and also it has been considerable to predict the result in more reliable and accurate. To find the price of used vehicles a well defined model has been developed with the help of three machine learning techniques such as Artificial Neural Network, Support Vector Machine and Random Forest. These techniques were used not on the individual items but for the whole group of data items. This data group has been taken from some web portal and that same has been used for the prediction. The data must be collected using web scraper that was written in PHP programming language. Distinct machine learning algorithms of varying performances had been compared to get the best result of the given data set. The final prediction model was integrated into Java application.

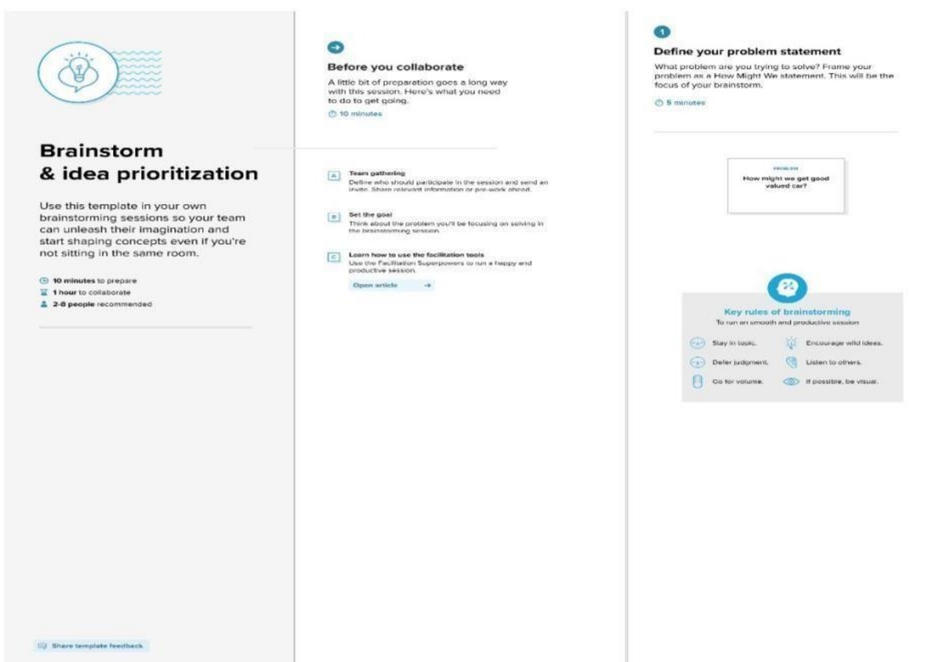
3.Ideation and Brainstorm



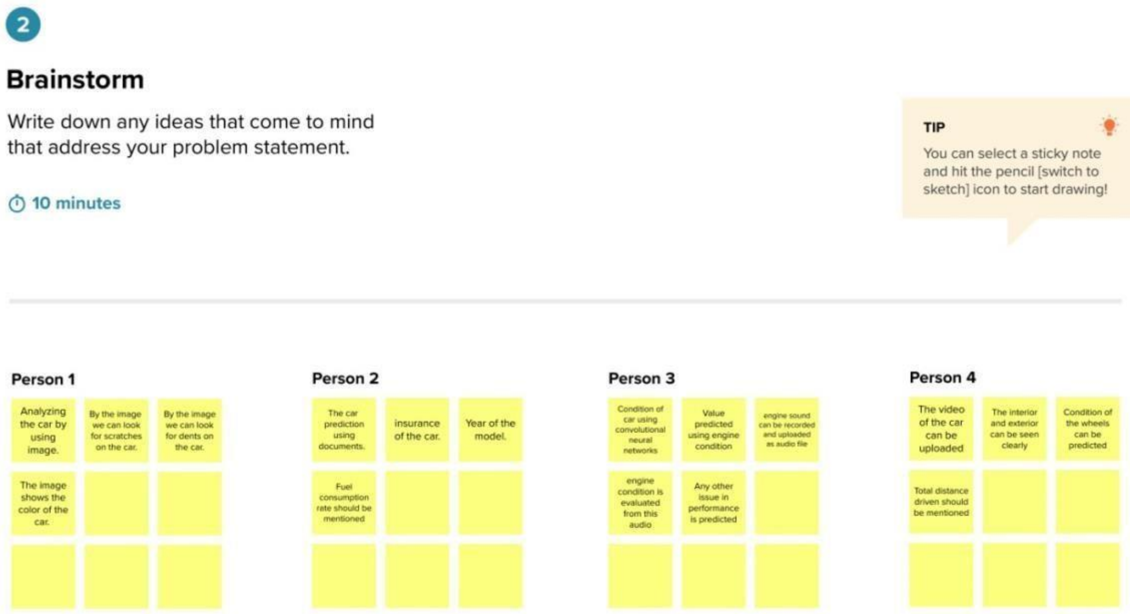
3.2 IDEATION & BRAINSTORMING:

Brainstorming provides a free and open environment that encourages everyone within a team to participate in the creative thinking process that leads to problem solving. Prioritizing volume over value, out-of-the-box ideas are welcome and built upon, and all participants are encouraged to collaborate, helping each other develop a rich amount of creative solutions. Use this template in your own brainstorming sessions so your team can unleash their imagination and start shaping concepts even if you're not sitting in the same room.

Step-1: Team Gathering, Collaboration and Select the Problem Statement



Step-2: Brainstorm



3.3 PROPOSED SOLUTION

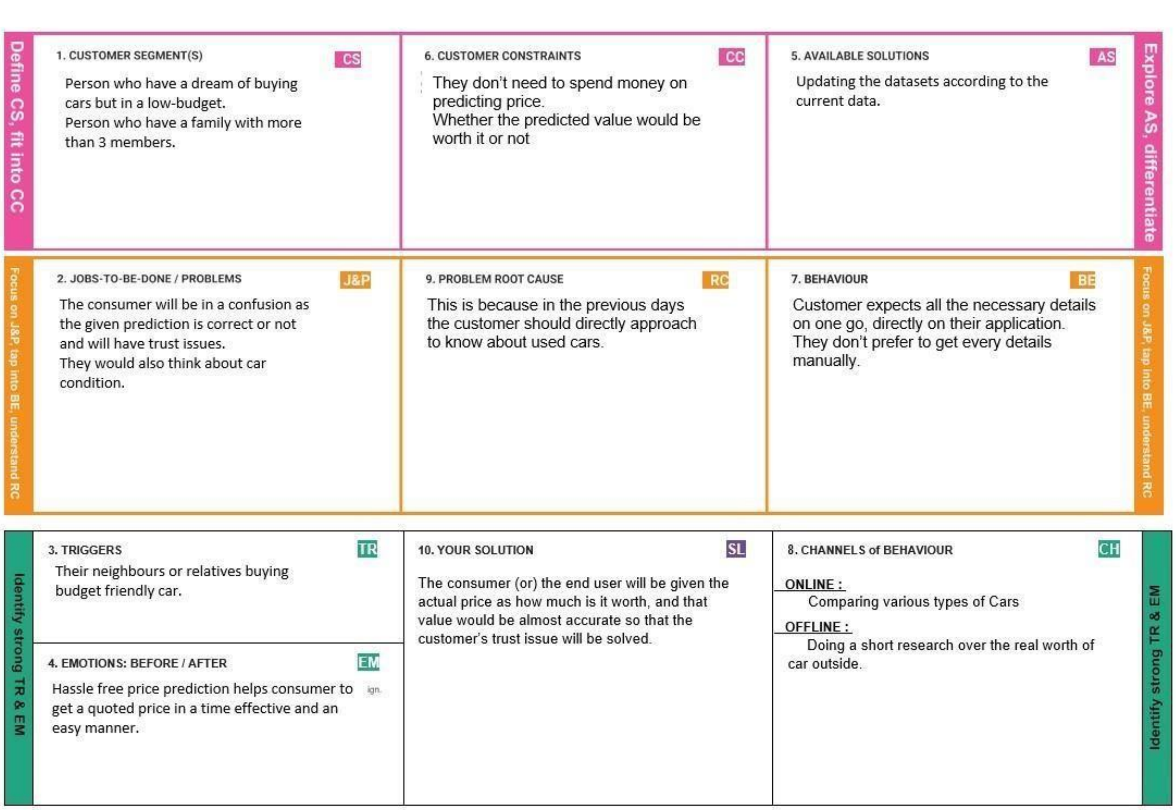
S.No.	Parameter	Descrip on
1.	Problem Statement (Problem to be solved)	**Car Resale Value Predic on** With difficult economic condi ons, it is likely that sales of second-hand imported (recondi oned) cars and used cars will increase. In many developed countries, it is common to lease a car rather than buying it outright. A er the lease period is over, the buyer has the possibility to buy the car at its residual value, i.e. its expected resale value. Thus, it is of commercial interest to sellers/financers to be able to predict the salvage value (residual value) of cars with accuracy.

2.	Idea / Solu on descrip on	Using regression algorithms, we proposed an intelligent, flexible, and effec ve system to predict the value of the car. By regression algorithms and other algorithms is used to predict the accuracy value of the cars. Depend on major parts and damages on the car will affect the price of the car.
3.	Novelty / Uniqueness	To predict the value, the most essen al elements for forecast are brand and model, period use of vehicle, mileage of vehicle, gear type and fuel type u lized in the vehicle just as fuel u liza on per mile profoundly influence cost of a vehicle because of con nuous changes in the cost of a fuel. By forecas ng the above details, AI can predict the value accurately.
4.	Social Impact / Customer Sa sfac on	Customer Sa sfac on plays a vital role in this, i.e for customer, he/she need to get profit from his car so customer expect that the predict value need to be good which gives him/her profit, but it is depend on the car condi on. Depend on the customer sa sfac on our applica on will create a social impact and may customer will increase.

5.	Business Model (Revenue Model)	A Revenue model is a framework for genera ng financial income. It iden fies which revenue source to pursue ,what value to offer ,how to price the value ,and who pays for the value.

6.	Scalability of the Solu on	The value of the car is predic ng by using different regression algorithms like linear regression , random forest regression , decision tree regression and so on. Thus the car will got accurate price. Those algorithms gives the results with the user given details about the car, but the best and approximate result is got by random forest algorithm. As random forest regression algorithm gives more as 15% then other algorithms.

3.4 PROBLE SOLUTION FIT:



4. REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMENTS: Operating system- Windows 7,8,10

Processor- dual core 2.4 GHz (i5 or i7 series Intel processor or equivalent AMD)

RAM-4GB

4.2 NON-FUNCTIONAL REQUIREMENTS:

Python Pycharm

PIP 2.7

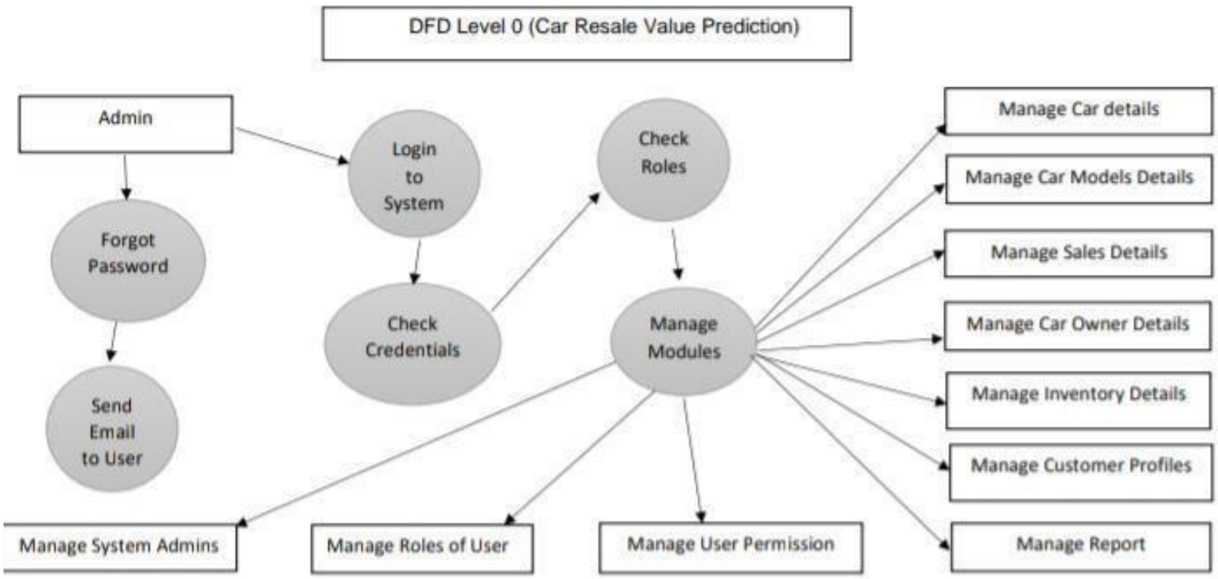
Jupyter Notebook

Chrome

5. PROJECT DESIGN

5.1 DATA FLOW DIAGRAMS:

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored. Example: (Simplified)

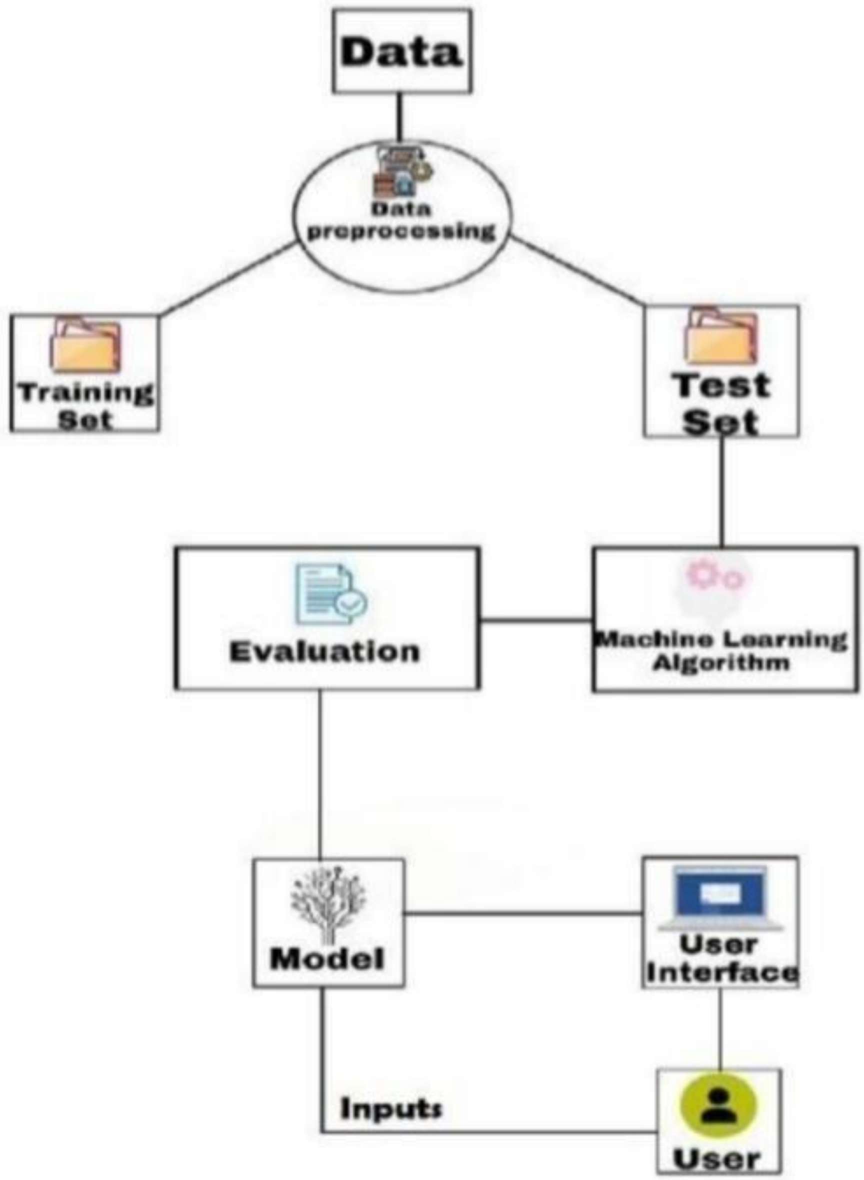


5.2 SOLUTION & TECHNICAL ARCHITECTURE:

Solution architecture is a complex process – with many sub-processes – that bridges the gap between business problems and technology solutions. Its goals are to:

- Find the best tech solution to solve existing business problems.
- Describe the structure, characteristics, behavior, and other aspects of the software to project stakeholders.

5.3 USER STORIES:



User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release

Customer (Mobile web user)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account / dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
		USN-3	As a user, I can register for the application through Google, Facebook.	I can register & access the dashboard with Google, Facebook Login	Low	Sprint-2
	Login	USN-4	As a user, I can log into the application by entering email & password	I can login using email and password	High	Sprint-1
	Dashboard	USN-5	As a user, I can access the dashboard after login	I can access the dashboard	High	Sprint-2

Customer (Web user)	Registration	USN-6	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account / dashboard	High	Sprint-1
		USN-7	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
		USN-8	As a user, I can register for the application through Google, Facebook.	I can register & access the dashboard with Google, Facebook Login	Low	Sprint-2
	Login	USN-9	As a user, I can log into the application by entering email & password	I can login using email and password	High	Sprint-1
	Dashboard	USN-10	As a user, I can access the dashboard after login	I can access the dashboard	High	Sprint-2
Customer Care Executive	Customer Support	USN-11	As a user, I can contact the customer care and chat with us	I can contact the customer care and chat with the person incharge	High	Sprint-2
Administrator		USN-12	As a user, my data is maintained by admin	Admin maintain customer data	High	Sprint-2

6. PROJECT PLANNING & SCHEDULING

6.1 SPRINT PLANNING & ESTIMATION:

Sprint	Function al Require ment (Epic)	User Story Numb er	User Story / Task	Story Points	Priori ty	Team Members
Sprint-1	Registration	USN-1	As a user, I can register for the application by entering my email, password.	10	High	Varshaa.H
Sprint-1	Confirmati on	USN-2	As a user, I will receive confirmation email once I have registered for the application	10	High	Sandhiya.S
Sprint-1	Login	USN-3	As a user, I can log into the applicati on by entering email & password	8	Medi um	, Varshaa.h
Sprint-2	Dataset	USN-4	Collect dataset, Import required libraries, Test and Train data.	10	High	Preetha Judy .B
Sprint-2	Algorithm	USN-5	Apply Regression algorithm and got the data (.pkl	10	High	Varshaah and Saranya.m

			file).			
Sprint-3	Dashboard	USN-6	HTML page contains Login, Details to be entered to predict the car price and a customer support.	10	High	Varshaa.h
Sprint-4	Building application	USN-7	Build python flask application	10	High	Saranya.m

6.2 SPRINT DELIVERY SCHEDULE:

Sprint	**Tot al Sto ry Poi nts**	**Duration**	**Sprint Start Date**	**Sprint End Date (Plann ed)**	**Story Points Complet ed (as on Planned End Date)**	**Sprint Release Date (Actual)**
Sprint-1	20	3 Days	08 Nov 2022	10 Nov 2022	20	11 Nov 2022
Sprint-2	20	3 Days	11 Nov 2022	13 Nov 2022	20	14 Nov 2022
Sprint-3	20	2 Days	14 Nov 2022	15 Nov 2022	20	16 Nov 2022
Sprint-4	20	2 Days	16 Nov 2022	17 Nov 2022	20	17 Nov 2022

7. CODING & SOLUTIONING

7.1 FEATURE 1:

Using the following code we have created the categories and included text boxes for getting input from the user, also in some categories the user will be able to select an option from the drop down list.

```
<tr>
<td><label for="month">Registration Month : </label></td>
<td><input id="month" maxlength="50" name="regmonth" type="text" />
<br>
<br>
</td>
</tr>
<tr>
<td><label for="year" padding:10px>Registration year : </label></td>
<td><input id="year" maxlength="50" name="regyear" type="text" />
<br>
<br>
</td>
</tr>
```

The following code is for getting kilometres driven information and power and other information about the car.

```
<tr>
<td><label for="kilometer">Kilometers driven : </label></td>
<td><input id="kilometer" maxlength="50" name="kms" type="text" />
<br>
<br>
</td>
</tr>
<tr> <td><label
for="power">Car
```

```
power in PS:
</label></td>
<td><input id="power" maxlength="50" name="powerps" type="text" />
<br>
<br>
</td>
</tr>

<tr>
<td><label for="geartype">Gear type : </label></td>
<td><input type="radio" name="geartype" value="manual"/> Manual
<input type="radio" name="geartype" value="automatic"/> Automatic
<input type="radio" name="geartype" value="not-declared"/> Not declared
<br>
<br>
</td>  </tr>
```

In the next code we have included a bunch of car models list and their brands list.

```
<tr>
<td><label for="model">Model Type : </label></td>
<td>
<select name="model" id="model">
<option value="" disabled selected hidden>Choose Model Name...</option>
<option value="golf">Golf </option> <option
value="grand">Grand </option>
………………………………………………………
………………………………………………………
<option value="serie_1">Serie 1 </option>
<option value="discovery_sport">Discovery Sport </option>
</select>
<br>
<br>
</td>
</tr>
```

And For Brands:

```
<tr>
<td><label for="brand">Brand :</label></td>
```

```
<td>
<select name="brand" id="brand">
<option value="" disabled selected hidden>Choose Brand Name...</option>
<option value="volkswagen">Volkswagen </option>
<option value="land_rover">Land Rover </option>
<option value="lada">Lada </option>
</select>
<br>
<br>
</td>
</tr>
```

7.2 FEATURE 2:

In our project we have created an interactive design of car resale value prediction using advanced data science. We have Label encoded the categorical data.

```
from sklearn.preprocessing import LabelEncoder   labels =
['gearbox','notRepairedDamage','model','brand','fuelType','vehicleType']
mapper = {}  for i in labels:
    mapper[i] = LabelEncoder()      mapper[i].fit(new_df[i])      tr
= mapper[i].transform(new_df[i])
np.save(str('classes'+i+'.npy'), mapper[i].classes_)
print(i,":",mapper[i])      new_df.loc[:,i+'_labels'] = pd.Series(tr,
index=new_df.index)

labeled = new_df[['price','yearOfRegistration','powerPS','kilometer','monthOfRegistration']
+ [x+"_labels" for x in labels]]
print(labeled.columns)
```

8. TESTING

8.1 TEST CASES:

In order to understand what affects change in price of a used car, the relation between features available in the data sat will be examined by using inferential statistic methods. The primary assumption based on figures and tables is price must be affected by odometer and condition. There must be other features that affects price significantly. It will be investigated in the later phase of the study.

Checking normality: For checking normality, q-q plot helps us. Figure 9 tells that there is a violation of normality. This means that the data points that are used are not distributed normally. In addition, Shapiro-Wilk test was performed for checking normality.

Result:(0.9586305022239685, 0.0)

Here, the first value is W-test statistic and the second value is the p-value. For N > 5000, the W test statistic is accurate but the p-value may not be. By considering p-value of Shapiro-Wilk test, it can be concluded that the data is not normally distributed.

In this situation, we have problem with initial data points. May be, filtering data can solve this issue. For this purpose, the values of odometer and price that are two standard deviation away from mean were dropped and independent t-test applied.

Condition vs Price

The second hypothesis of this study focuses on effect of a car's condition on its price. In order to understand this relation, Table 6 and Figure 6 can be useful. By looking at Figure 10, it can be said that 'condition' effects median price of cars seriously. On the other hand, there are a lot of outliers in the condition values which is an expected result for such a lar dataset. We do not see outliers at the bottom of the Figure 10. This is mostly because during data cleaning, cars that lower than $750 price were dropped.

9. RESULTS

9.1 PERFORMANCE METRICS:

Performance metrics are a collection of data that employers evaluate against an established objective. It is important to note the difference between a performance metric and a key performance indicator.

We have used the Random Forest Regression method to evaluate the performance metrics.

We use the following code for it:

```
regressor = RandomForestRegressor(n_estimators=1000, max_depth=10,random_state=34)
#fitting the model
regressor.fit(X_train, np.ravel(Y_train,order='C'))
```

And we get:

```
RandomForestRegressor(max_depth=10, n_estimators=1000, random_state=34)
```

By considering all four metrics, it can be concluded that random forest the best model for the prediction for used car prices. Random Forest as a regression model gave the best.

10. ADVANTAGES & DISADVANTAGES

ADVANTAGES:

- It is very easy to use.
- It contains all the available models and their predictions.
- It gives almost 90% accurate prediction.
- No user is asked for their personal details.
- Can be very useful for people who are going to buy or sell used cars.

DISADVANTAGES:

- They ask for so many data about the cars.

- We have to know everything about the car precisely.

11. CONCLUSION

The increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. The proposed system will help to determine the accurate price of used car price prediction.

12.FUTURE SCOPE

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.

13. APPENDIX

Source Code:

In [339]:

```
import pandas as pd
import numpy as np
import matplotlib as plt
```

In [340]:

```
car=pd.read_csv('carresalevaluep.csv')
```

In [341]:

```
car.head()
```

Out[341]:

	Unnamed: 0	name	company	year	Price	kms_driven	fuel_type
0	0	Hyundai Santro Xing	Hyundai	2007	80000	45000	Petrol
1	1	Mahindra Jeep CL550	Mahindra	2006	425000	40	Diesel
2	2	Hyundai Grand i10	Hyundai	2014	325000	28000	Petrol
3	3	Ford EcoSport Titanium	Ford	2014	575000	36000	Diesel
4	4	Ford Figo	Ford	2012	175000	41000	Diesel

In [342]:

```
car.shape
```

Out[342]: (816, 7)

In [343]:

```
car.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 816 entries, 0 to 815
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  816 non-null    int64
 1   name        816 non-null    object
 2   company     816 non-null    object
 3   year        816 non-null    int64
 4   Price       816 non-null    int64
 5   kms_driven  816 non-null    int64
 6   fuel_type   816 non-null    object
dtypes: int64(4), object(3)
memory usage: 44.8+ KB
```

In [344]:

```
car
```

Out[344]:

	Unnamed: 0	name	company	year	Price	kms_driven	fuel_type
0	0	Hyundai Santro Xing	Hyundai	2007	80000	45000	Petrol
1	1	Mahindra Jeep CL550	Mahindra	2006	425000	40	Diesel
2	2	Hyundai Grand i10	Hyundai	2014	325000	28000	Petrol
3	3	Ford EcoSport Titanium	Ford	2014	575000	36000	Diesel
4	4	Ford Figo	Ford	2012	175000	41000	Diesel
...	...	...	...	...	...	...	...
811	811	Maruti Suzuki Ritz	Maruti	2011	270000	50000	Petrol
812	812	Tata Indica V2	Tata	2009	110000	30000	Diesel
813	813	Toyota Corolla Altis	Toyota	2009	300000	132000	Petrol
814	814	Tata Zest XM	Tata	2018	260000	27000	Diesel
815	815	Mahindra Quanto C8	Mahindra	2013	390000	40000	Diesel

816 rows × 7 columns

In [345]:

```
car['fuel_type'].unique()
```

Out[345]: array(['Petrol', 'Diesel', 'LPG'], dtype=object)

In [346]:

```
car['year'].unique()
```

Out[346]:

```
array([2007, 2006, 2014, 2012, 2013, 2016, 2015, 2010, 2017, 2008, 2018,
       2011, 2019, 2009, 2005, 2000, 2003, 2004, 1995, 2002, 2001],
      dtype=int64)
```

In [347]:

```
backup=car.copy()
```

In [348]:

```
car.head()
```

Out[348]:

	Unnamed: 0	name	company	year	Price	kms_driven	fuel_type
0	0	Hyundai Santro Xing	Hyundai	2007	80000	45000	Petrol
1	1	Mahindra Jeep CL550	Mahindra	2006	425000	40	Diesel
2	2	Hyundai Grand i10	Hyundai	2014	325000	28000	Petrol
3	3	Ford EcoSport Titanium	Ford	2014	575000	36000	Diesel
4	4	Ford Figo	Ford	2012	175000	41000	Diesel

In [349]:

```
car.reset_index(drop=True)
```

Out[349]:

	Unnamed: 0	name	company	year	Price	kms_driven	fuel_type
0	0	Hyundai Santro Xing	Hyundai	2007	80000	45000	Petrol
1	1	Mahindra Jeep CL550	Mahindra	2006	425000	40	Diesel
2	2	Hyundai Grand i10	Hyundai	2014	325000	28000	Petrol
3	3	Ford EcoSport Titanium	Ford	2014	575000	36000	Diesel
4	4	Ford Figo	Ford	2012	175000	41000	Diesel
...	...	...	...	...	...	...	...
811	811	Maruti Suzuki Ritz	Maruti	2011	270000	50000	Petrol
812	812	Tata Indica V2	Tata	2009	110000	30000	Diesel
813	813	Toyota Corolla Altis	Toyota	2009	300000	132000	Petrol
814	814	Tata Zest XM	Tata	2018	260000	27000	Diesel
815	815	Mahindra Quanto C8	Mahindra	2013	390000	40000	Diesel

816 rows × 7 columns

In [350]:

```
car.describe()
```

Out[350]:

	Unnamed: 0	year	Price	kms_driven
count	816.000000	816.000000	8.160000e+02	816.000000
mean	407.500000	2012.444853	4.117176e+05	46275.531863
std	235.703203	4.002992	4.751844e+05	34297.428044
min	0.000000	1995.000000	3.000000e+04	0.000000
25%	203.750000	2010.000000	1.750000e+05	27000.000000
50%	407.500000	2013.000000	2.999990e+05	41000.000000
75%	611.250000	2015.000000	4.912500e+05	56818.500000
max	815.000000	2019.000000	8.500003e+06	400000.000000

In [351]:

```
x=car.drop(columns='Price')
y=car['Price']
```

In [352]:

```
x=car.drop(columns='Price')
y=car['Price']
```

In [353]:

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train, y_test=train_test_split(x,y, test_size=0.2)
```

In [354]:

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import make_column_transformer
from sklearn.pipeline import make_pipeline
```

In [355]:

```
ohe = OneHotEncoder()
ohe.fit(x[['name','company','fuel_type']])
```

Out[355]: OneHotEncoder()

In [356]:
```
ohe.categories_
```

Out[356]:
```
[array(['Audi A3 Cabriolet', 'Audi A4 1.8', 'Audi A4 2.0', 'Audi A6 2.0',
       'Audi A8', 'Audi Q3 2.0', 'Audi Q5 2.0', 'Audi Q7', 'BMW 3 Series',
       'BMW 5 Series', 'BMW 7 Series', 'BMW X1', 'BMW X1 sDrive20d',
       'BMW X1 xDrive20d', 'Chevrolet Beat', 'Chevrolet Beat Diesel',
       'Chevrolet Beat LS', 'Chevrolet Beat LT', 'Chevrolet Beat PS',
       'Chevrolet Cruze LTZ', 'Chevrolet Enjoy', 'Chevrolet Enjoy 1.4',
       'Chevrolet Sail 1.2', 'Chevrolet Sail UVA', 'Chevrolet Spark',
       'Chevrolet Spark 1.0', 'Chevrolet Spark LS', 'Chevrolet Spark LT',
       'Chevrolet Tavera LS', 'Chevrolet Tavera Neo', 'Datsun GO T',
       'Datsun Go Plus', 'Datsun Redi GO', 'Fiat Linea Emotion',
       'Fiat Petra ELX', 'Fiat Punto Emotion', 'Force Motors Force',
       'Force Motors One', 'Ford EcoSport', 'Ford EcoSport Ambiente',
       'Ford EcoSport Titanium', 'Ford EcoSport Trend',
       'Ford Endeavor 4x4', 'Ford Fiesta', 'Ford Fiesta SXi', 'Ford Figo',
       'Ford Figo Diesel', 'Ford Figo Duratorq', 'Ford Figo Petrol',
       'Ford Fusion 1.4', 'Ford Ikon 1.3', 'Ford Ikon 1.6',
       'Hindustan Motors Ambassador', 'Honda Accord', 'Honda Amaze',
       'Honda Amaze 1.2', 'Honda Amaze 1.5', 'Honda Brio', 'Honda Brio V',
       'Honda Brio VX', 'Honda City', 'Honda City 1.5', 'Honda City SV',
       'Honda City VX', 'Honda City ZX', 'Honda Jazz S', 'Honda Jazz VX',
       'Honda Mobilio', 'Honda Mobilio S', 'Honda WR V', 'Hyundai Accent',
       'Hyundai Accent Executive', 'Hyundai Accent GLE',
       'Hyundai Accent GLX', 'Hyundai Creta', 'Hyundai Creta 1.6',
       'Hyundai Elantra 1.8', 'Hyundai Elantra SX', 'Hyundai Elite i20',
       'Hyundai Eon', 'Hyundai Eon D', 'Hyundai Eon Era',
       'Hyundai Eon Magna', 'Hyundai Eon Sportz', 'Hyundai Fluidic Verna',
       'Hyundai Getz', 'Hyundai Getz GLE', 'Hyundai Getz Prime',
       'Hyundai Grand i10', 'Hyundai Santro', 'Hyundai Santro AE',
       'Hyundai Santro Xing', 'Hyundai Sonata Transform', 'Hyundai Verna',
       'Hyundai Verna 1.4', 'Hyundai Verna 1.6', 'Hyundai Verna Fluidic',
       'Hyundai Verna Transform', 'Hyundai Verna VGT',
       'Hyundai Xcent Base', 'Hyundai Xcent SX', 'Hyundai i10',
       'Hyundai i10 Era', 'Hyundai i10 Magna', 'Hyundai i10 Sportz',
       'Hyundai i20', 'Hyundai i20 Active', 'Hyundai i20 Asta',
       'Hyundai i20 Magna', 'Hyundai i20 Select', 'Hyundai i20 Sportz',
       'Jaguar XE XE', 'Jaguar XF 2.2', 'Jeep Wrangler Unlimited',
       'Land Rover Freelander', 'Mahindra Bolero DI',
       'Mahindra Bolero Power', 'Mahindra Bolero SLE',
       'Mahindra Jeep CL550', 'Mahindra Jeep MM', 'Mahindra KUV100',
       'Mahindra KUV100 K8', 'Mahindra Logan', 'Mahindra Logan Diesel',
       'Mahindra Quanto C4', 'Mahindra Quanto C8', 'Mahindra Scorpio',
       'Mahindra Scorpio 2.6', 'Mahindra Scorpio LX',
       'Mahindra Scorpio S10', 'Mahindra Scorpio S4',
       'Mahindra Scorpio SLE', 'Mahindra Scorpio SLX',
       'Mahindra Scorpio VLX', 'Mahindra Scorpio Vlx',
       'Mahindra Scorpio W', 'Mahindra TUV300 T4', 'Mahindra TUV300 T8',
       'Mahindra Thar CRDe', 'Mahindra XUV500', 'Mahindra XUV500 W10',
       'Mahindra XUV500 W6', 'Mahindra XUV500 W8', 'Mahindra Xylo D2',
       'Mahindra Xylo E4', 'Mahindra Xylo E8', 'Maruti Suzuki 800',
       'Maruti Suzuki A', 'Maruti Suzuki Alto', 'Maruti Suzuki Baleno',
       'Maruti Suzuki Celerio', 'Maruti Suzuki Ciaz',
       'Maruti Suzuki Dzire', 'Maruti Suzuki Eeco',
       'Maruti Suzuki Ertiga', 'Maruti Suzuki Esteem',
       'Maruti Suzuki Estilo', 'Maruti Suzuki Maruti',
       'Maruti Suzuki Omni', 'Maruti Suzuki Ritz', 'Maruti Suzuki S',
       'Maruti Suzuki SX4', 'Maruti Suzuki Stingray',
       'Maruti Suzuki Swift', 'Maruti Suzuki Versa',
       'Maruti Suzuki Vitara', 'Maruti Suzuki Wagon', 'Maruti Suzuki Zen',
       'Mercedes Benz A', 'Mercedes Benz B', 'Mercedes Benz C',
       'Mercedes Benz GLA', 'Mini Cooper S', 'Mitsubishi Lancer 1.8',
       'Mitsubishi Pajero Sport', 'Nissan Micra XL', 'Nissan Micra XV',
       'Nissan Sunny', 'Nissan Sunny XL', 'Nissan Terrano XL',
       'Nissan X Trail', 'Renault Duster', 'Renault Duster 110',
       'Renault Duster 110PS', 'Renault Duster 85', 'Renault Duster 85PS',
       'Renault Duster RxL', 'Renault Kwid', 'Renault Kwid 1.0',
       'Renault Kwid RXT', 'Renault Lodgy 85', 'Renault Scala RxL',
       'Skoda Fabia', 'Skoda Fabia 1.2L', 'Skoda Fabia Classic',
       'Skoda Laura', 'Skoda Octavia Classic', 'Skoda Rapid Elegance',
       'Skoda Superb 1.8', 'Skoda Yeti Ambition', 'Tata Aria Pleasure',
       'Tata Bolt XM', 'Tata Indica', 'Tata Indica V2', 'Tata Indica eV2',
       'Tata Indigo CS', 'Tata Indigo LS', 'Tata Indigo LX',
       'Tata Indigo Marina', 'Tata Indigo eCS', 'Tata Manza',
       'Tata Manza Aqua', 'Tata Manza Aura', 'Tata Manza ELAN',
```

```
       'Tata Nano', 'Tata Nano Cx', 'Tata Nano GenX', 'Tata Nano LX',
       'Tata Nano Lx', 'Tata Sumo Gold', 'Tata Sumo Grande',
       'Tata Sumo Victa', 'Tata Tiago Revotorq', 'Tata Tiago Revotron',
       'Tata Tigor Revotron', 'Tata Venture EX', 'Tata Vista Quadrajet',
       'Tata Zest Quadrajet', 'Tata Zest XE', 'Tata Zest XM',
       'Toyota Corolla', 'Toyota Corolla Altis', 'Toyota Corolla H2',
       'Toyota Etios', 'Toyota Etios G', 'Toyota Etios GD',
       'Toyota Etios Liva', 'Toyota Fortuner', 'Toyota Fortuner 3.0',
       'Toyota Innova 2.0', 'Toyota Innova 2.5', 'Toyota Qualis',
       'Volkswagen Jetta Comfortline', 'Volkswagen Jetta Highline',
       'Volkswagen Passat Diesel', 'Volkswagen Polo',
       'Volkswagen Polo Comfortline', 'Volkswagen Polo Highline',
       'Volkswagen Polo Highline1.2L', 'Volkswagen Polo Trendline',
       'Volkswagen Vento Comfortline', 'Volkswagen Vento Highline',
       'Volkswagen Vento Konekt', 'Volvo S80 Summum'], dtype=object),
 array(['Audi', 'BMW', 'Chevrolet', 'Datsun', 'Fiat', 'Force', 'Ford',
       'Hindustan', 'Honda', 'Hyundai', 'Jaguar', 'Jeep', 'Land',
       'Mahindra', 'Maruti', 'Mercedes', 'Mini', 'Mitsubishi', 'Nissan',
       'Renault', 'Skoda', 'Tata', 'Toyota', 'Volkswagen', 'Volvo'],
      dtype=object),
 array(['Diesel', 'LPG', 'Petrol'], dtype=object)]
```

In [357]:
```
column_trans=make_column_transformer((OneHotEncoder(categories=ohe.categories_),['name','company','fuel_type']),re
```

lr=LinearRegression()

In [358]:
```
lr=LinearRegression()
```

In [359]:
```
pipe=make_pipeline(column_trans,lr)
```

In [360]:
```
pipe.fit(x_train,y_train)
```

Out[360]:
```
Pipeline(steps=[('columntransformer',
                 ColumnTransformer(remainder='passthrough',
                                   transformers=[('onehotencoder',
                                                  OneHotEncoder(categories=[array(['Audi A3 Cabriolet', 'Audi A4 1.8', 'Audi
A4 2.0', 'Audi A6 2.0',
       'Audi A8', 'Audi Q3 2.0', 'Audi Q5 2.0', 'Audi Q7', 'BMW 3 Series',
       'BMW 5 Series', 'BMW 7 Series', 'BMW X1', 'BMW X1 sDrive20d',
       'BMW X1 xDrive20d', 'Chevrolet Beat', 'Chevrolet Beat...
                                                                            array(['Audi', 'BMW', 'Chevrolet', 'Datsun', 'Fiat', 'Force', 'For
d',
       'Hindustan', 'Honda', 'Hyundai', 'Jaguar', 'Jeep', 'Land',
       'Mahindra', 'Maruti', 'Mercedes', 'Mini', 'Mitsubishi', 'Nissan',
       'Renault', 'Skoda', 'Tata', 'Toyota', 'Volkswagen', 'Volvo'],
      dtype=object),
                                                                            array(['Diesel', 'LPG', 'Petrol'], dtype=object)]),
                                                  ['name', 'company',
                                                   'fuel_type'])])),
                ('linearregression', LinearRegression())])
```

In [361]:
```
y_pred =pipe.predict(x_test)
```

In [362]:
```
r2_score(y_test,y_pred)
```

Out[362]:
```
0.4977698265719559
```

In [363]:
```
scores=[]
for i in range(1000):
    x_train,x_test,y_train, y_test=train_test_split(x,y, test_size=0.2)
    lr=LinearRegression()
    pipe=make_pipeline(column_trans,lr)
    pipe.fit(x_train,y_train)
    y_pred =pipe.predict(x_test)
    scores.append(r2_score(y_test,y_pred))
```

In [364]:
```
np.argmax(scores)
```

Out[364]:

```
577
```

In [365]:

```python
scores[np.argmax(scores)]
```

Out[365]:

```
0.7973804780567894
```

In [366]:

```python
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2, random_state=np.argmax(scores))
lr=LinearRegression()
pipe=make_pipeline(column_trans,lr)
pipe.fit(x_train,y_train)
y_pred =pipe.predict(x_test)
r2_score(y_test,y_pred)
```

Out[366]:

```
0.7066410845475624
```

In [378]:

```python
pipe.steps[0][1].transformers[0][1].categories[0]
```

Out[378]:

```
array(['Audi A3 Cabriolet', 'Audi A4 1.8', 'Audi A4 2.0', 'Audi A6 2.0',
       'Audi A8', 'Audi Q3 2.0', 'Audi Q5 2.0', 'Audi Q7', 'BMW 3 Series',
       'BMW 5 Series', 'BMW 7 Series', 'BMW X1', 'BMW X1 sDrive20d',
       'BMW X1 xDrive20d', 'Chevrolet Beat', 'Chevrolet Beat Diesel',
       'Chevrolet Beat LS', 'Chevrolet Beat LT', 'Chevrolet Beat PS',
       'Chevrolet Cruze LTZ', 'Chevrolet Enjoy', 'Chevrolet Enjoy 1.4',
       'Chevrolet Sail 1.2', 'Chevrolet Sail UVA', 'Chevrolet Spark',
       'Chevrolet Spark 1.0', 'Chevrolet Spark LS', 'Chevrolet Spark LT',
       'Chevrolet Tavera LS', 'Chevrolet Tavera Neo', 'Datsun GO T',
       'Datsun Go Plus', 'Datsun Redi GO', 'Fiat Linea Emotion',
       'Fiat Petra ELX', 'Fiat Punto Emotion', 'Force Motors Force',
       'Force Motors One', 'Ford EcoSport', 'Ford EcoSport Ambiente',
       'Ford EcoSport Titanium', 'Ford EcoSport Trend',
       'Ford Endeavor 4x4', 'Ford Fiesta', 'Ford Fiesta SXi', 'Ford Figo',
       'Ford Figo Diesel', 'Ford Figo Duratorq', 'Ford Figo Petrol',
       'Ford Fusion 1.4', 'Ford Ikon 1.3', 'Ford Ikon 1.6',
       'Hindustan Motors Ambassador', 'Honda Accord', 'Honda Amaze',
       'Honda Amaze 1.2', 'Honda Amaze 1.5', 'Honda Brio', 'Honda Brio V',
       'Honda Brio VX', 'Honda City', 'Honda City 1.5', 'Honda City SV',
       'Honda City VX', 'Honda City ZX', 'Honda Jazz S', 'Honda Jazz VX',
       'Honda Mobilio', 'Honda Mobilio S', 'Honda WR V', 'Hyundai Accent',
       'Hyundai Accent Executive', 'Hyundai Accent GLE',
       'Hyundai Accent GLX', 'Hyundai Creta', 'Hyundai Creta 1.6',
       'Hyundai Elantra 1.8', 'Hyundai Elantra SX', 'Hyundai Elite i20',
       'Hyundai Eon', 'Hyundai Eon D', 'Hyundai Eon Era',
       'Hyundai Eon Magna', 'Hyundai Eon Sportz', 'Hyundai Fluidic Verna',
       'Hyundai Getz', 'Hyundai Getz GLE', 'Hyundai Getz Prime',
       'Hyundai Grand i10', 'Hyundai Santro', 'Hyundai Santro AE',
       'Hyundai Santro Xing', 'Hyundai Sonata Transform', 'Hyundai Verna',
       'Hyundai Verna 1.4', 'Hyundai Verna 1.6', 'Hyundai Verna Fluidic',
       'Hyundai Verna Transform', 'Hyundai Verna VGT',
       'Hyundai Xcent Base', 'Hyundai Xcent SX', 'Hyundai i10',
       'Hyundai i10 Era', 'Hyundai i10 Magna', 'Hyundai i10 Sportz',
       'Hyundai i20', 'Hyundai i20 Active', 'Hyundai i20 Asta',
       'Hyundai i20 Magna', 'Hyundai i20 Select', 'Hyundai i20 Sportz',
       'Jaguar XE XE', 'Jaguar XF 2.2', 'Jeep Wrangler Unlimited',
       'Land Rover Freelander', 'Mahindra Bolero DI',
       'Mahindra Bolero Power', 'Mahindra Bolero SLE',
       'Mahindra Jeep CL550', 'Mahindra Jeep MM', 'Mahindra KUV100',
       'Mahindra KUV100 K8', 'Mahindra Logan', 'Mahindra Logan Diesel',
       'Mahindra Quanto C4', 'Mahindra Quanto C8', 'Mahindra Scorpio',
       'Mahindra Scorpio 2.6', 'Mahindra Scorpio LX',
       'Mahindra Scorpio S10', 'Mahindra Scorpio S4',
       'Mahindra Scorpio SLE', 'Mahindra Scorpio SLX',
       'Mahindra Scorpio VLX', 'Mahindra Scorpio Vlx',
       'Mahindra Scorpio W', 'Mahindra TUV300 T4', 'Mahindra TUV300 T8',
       'Mahindra Thar CRDe', 'Mahindra XUV500', 'Mahindra XUV500 W10',
       'Mahindra XUV500 W6', 'Mahindra XUV500 W8', 'Mahindra Xylo D2',
       'Mahindra Xylo E4', 'Mahindra Xylo E8', 'Maruti Suzuki 800',
       'Maruti Suzuki A', 'Maruti Suzuki Alto', 'Maruti Suzuki Baleno',
       'Maruti Suzuki Celerio', 'Maruti Suzuki Ciaz',
       'Maruti Suzuki Dzire', 'Maruti Suzuki Eeco',
```

```
'Maruti Suzuki Ertiga', 'Maruti Suzuki Esteem',
'Maruti Suzuki Estilo', 'Maruti Suzuki Maruti',
'Maruti Suzuki Omni', 'Maruti Suzuki Ritz', 'Maruti Suzuki S',
'Maruti Suzuki SX4', 'Maruti Suzuki Stingray',
'Maruti Suzuki Swift', 'Maruti Suzuki Versa',
'Maruti Suzuki Vitara', 'Maruti Suzuki Wagon', 'Maruti Suzuki Zen',
'Mercedes Benz A', 'Mercedes Benz B', 'Mercedes Benz C',
'Mercedes Benz GLA', 'Mini Cooper S', 'Mitsubishi Lancer 1.8',
'Mitsubishi Pajero Sport', 'Nissan Micra XL', 'Nissan Micra XV',
'Nissan Sunny', 'Nissan Sunny XL', 'Nissan Terrano XL',
'Nissan X Trail', 'Renault Duster', 'Renault Duster 110',
'Renault Duster 110PS', 'Renault Duster 85', 'Renault Duster 85PS',
'Renault Duster RxL', 'Renault Kwid', 'Renault Kwid 1.0',
'Renault Kwid RXT', 'Renault Lodgy 85', 'Renault Scala RxL',
'Skoda Fabia', 'Skoda Fabia 1.2L', 'Skoda Fabia Classic',
'Skoda Laura', 'Skoda Octavia Classic', 'Skoda Rapid Elegance',
'Skoda Superb 1.8', 'Skoda Yeti Ambition', 'Tata Aria Pleasure',
'Tata Bolt XM', 'Tata Indica', 'Tata Indica V2', 'Tata Indica eV2',
'Tata Indigo CS', 'Tata Indigo LS', 'Tata Indigo LX',
'Tata Indigo Marina', 'Tata Indigo eCS', 'Tata Manza',
'Tata Manza Aqua', 'Tata Manza Aura', 'Tata Manza ELAN',
'Tata Nano', 'Tata Nano Cx', 'Tata Nano GenX', 'Tata Nano LX',
'Tata Nano Lx', 'Tata Sumo Gold', 'Tata Sumo Grande',
'Tata Sumo Victa', 'Tata Tiago Revotorq', 'Tata Tiago Revotron',
'Tata Tigor Revotron', 'Tata Venture EX', 'Tata Vista Quadrajet',
'Tata Zest Quadrajet', 'Tata Zest XE', 'Tata Zest XM',
'Toyota Corolla', 'Toyota Corolla Altis', 'Toyota Corolla H2',
'Toyota Etios', 'Toyota Etios G', 'Toyota Etios GD',
'Toyota Etios Liva', 'Toyota Fortuner', 'Toyota Fortuner 3.0',
'Toyota Innova 2.0', 'Toyota Innova 2.5', 'Toyota Qualis',
'Volkswagen Jetta Comfortline', 'Volkswagen Jetta Highline',
'Volkswagen Passat Diesel', 'Volkswagen Polo',
'Volkswagen Polo Comfortline', 'Volkswagen Polo Highline',
'Volkswagen Polo Highline1.2L', 'Volkswagen Polo Trendline',
'Volkswagen Vento Comfortline', 'Volkswagen Vento Highline',
'Volkswagen Vento Konekt', 'Volvo S80 Summum'], dtype=object)
```

GITHUB IBM-EPBL/IBM-PROJECT-42219-1660656410

PROJECT DEMO LINK https://youtu.be/iRCaMnR_bpA