FLIP ROBO

CARPRICE ANALYSIS

Submitted by:

SNEHA SANTRA

ACKNOWLEDGMENT

With the covid 19 impact in the market, we have seen lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make car price valuation model.

INTRODUCTION

- Business Problem Framing

  With the covid 19 impact in the market, we have seen lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make car price valuation model.

- Conceptual Background of the Domain Problem

  The project is sub-divided following section. These are:

  1. Loading necessary libraries

  2. Loading Dataset from a CSV file

  3. Summarization of Data to understand Dataset (Descriptive Statistics)

  4. Visualization of Data to understand Dataset (Plots, Graphs etc.)

  5.Processing the data for modeling

  6.skewness and outliers detection for better accuracy

  7.Build the model and select the right model and save it

- Review of Literature

  There are 14 columns including Name,Location,kilometer driven,fuel-type,engine and main feature is price.

  Most of them are object variable and some int variable and main features is float type.

- Motivation for the Problem Undertaken

  After collecting the data, you need to build a machine learning model. Before model building do all data pre-processing steps. Try different models with different hyper parameters and select the best model.

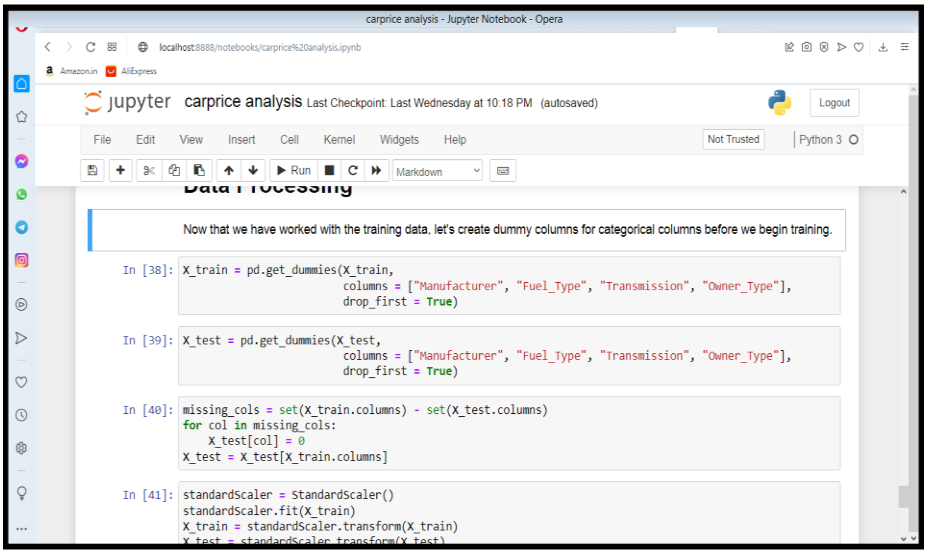
  Follow the complete life cycle of data science. Include all the steps like.

  1. Data Cleaning
  2. Exploratory Data Analysis
  3. Data Pre-processing
  4. Model Building
  5. Model Evaluation
  6. Selecting the best model

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

Data Processing: We have worked with the training data, let's create dummy columns for categorical columns before we begin training.

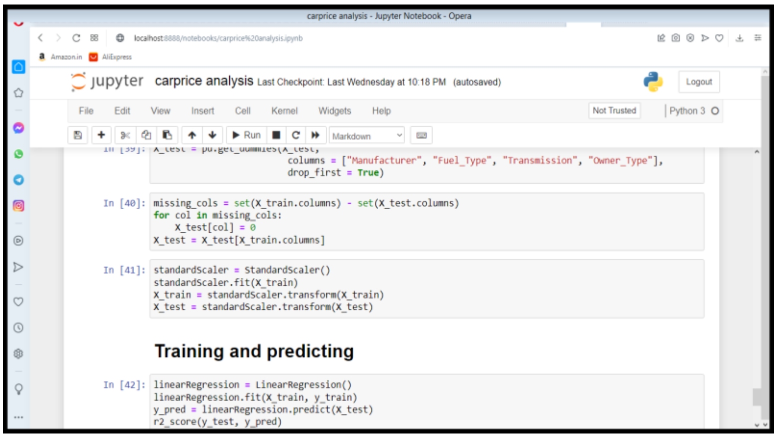

carprice analysis - Jupyter Notebook - Opera
localhost:8888/notebooks/carprice%20analysis.ipynb
Amazon.in
AliExpress
jupyter
carprice analysis Last Checkpoint: Last Wednesday at 10:18 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Markdown

Data Processing

Now that we have worked with the training data, let's create dummy columns for categorical columns before we begin training.

```
In [38]: X_train = pd.get_dummies(X_train,
                                   columns = ["Manufacturer", "Fuel_Type", "Transmission", "Owner_Type"],
                                   drop_first = True)

In [39]: X_test = pd.get_dummies(X_test,
                                   columns = ["Manufacturer", "Fuel_Type", "Transmission", "Owner_Type"],
                                   drop_first = True)

In [40]: missing_cols = set(X_train.columns) - set(X_test.columns)
         for col in missing_cols:
             X_test[col] = 0
         X_test = X_test[X_train.columns]

In [41]: standardScaler = StandardScaler()
         standardScaler.fit(X_train)
         X_train = standardScaler.transform(X_train)
         X_test = standardScaler.transform(X_test)
```

And then scalling the dataset for model building

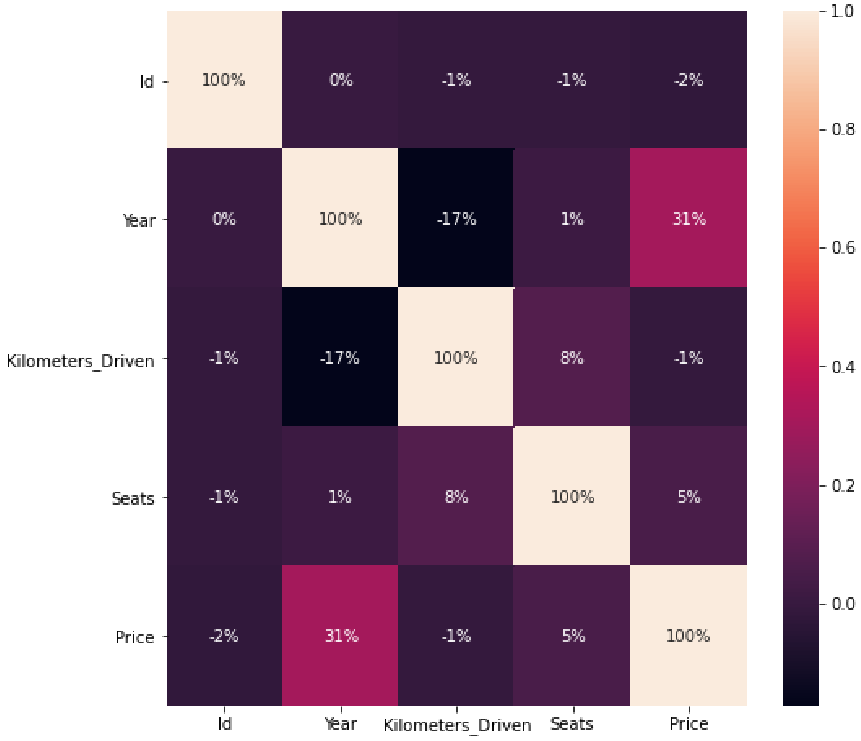

carprice analysis - Jupyter Notebook - Opera
localhost:8888/notebooks/carprice%20analysis.ipynb
Amazon.in
AliExpress
jupyter
carprice analysis Last Checkpoint: Last Wednesday at 10:18 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Markdown

```
In [39]: X_test = pd.get_dummies(X_test,
                                   columns = ["Manufacturer", "Fuel_Type", "Transmission", "Owner_Type"],
                                   drop_first = True)

In [40]: missing_cols = set(X_train.columns) - set(X_test.columns)
         for col in missing_cols:
             X_test[col] = 0
         X_test = X_test[X_train.columns]

In [41]: standardScaler = StandardScaler()
         standardScaler.fit(X_train)
         X_train = standardScaler.transform(X_train)
         X_test = standardScaler.transform(X_test)
```

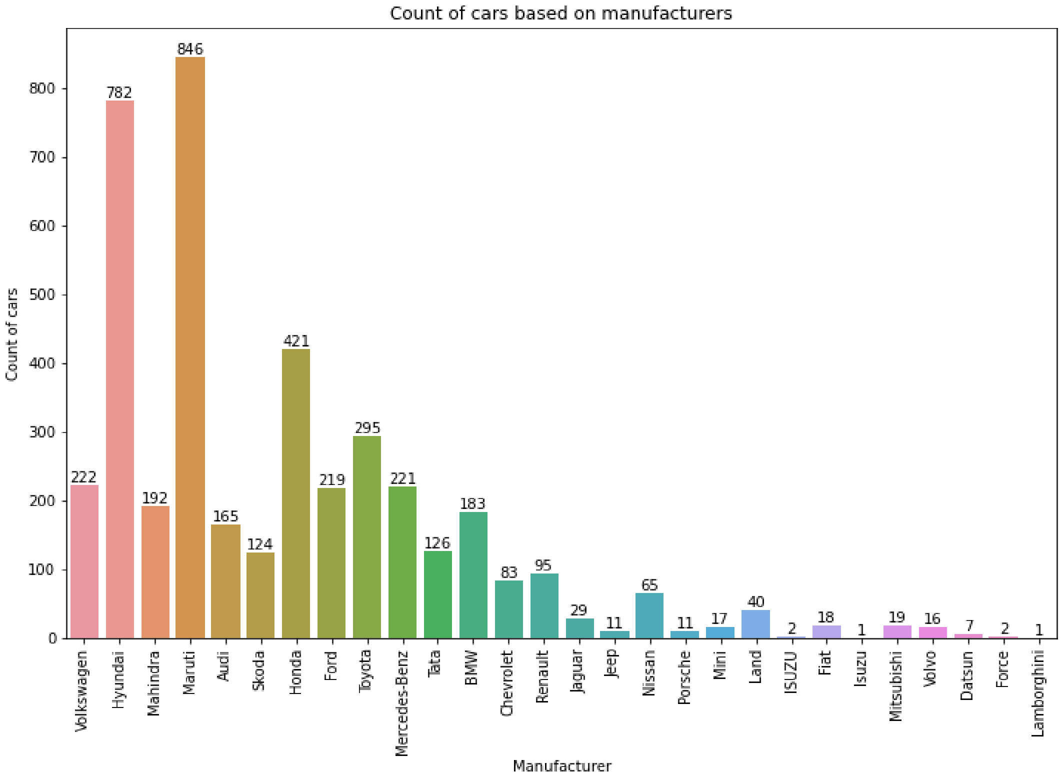
Training and predicting

```
In [42]: linearRegression = LinearRegression()
         linearRegression.fit(X_train, y_train)
         y_pred = linearRegression.predict(X_test)
         r2_score(y_test, y_pred)
```

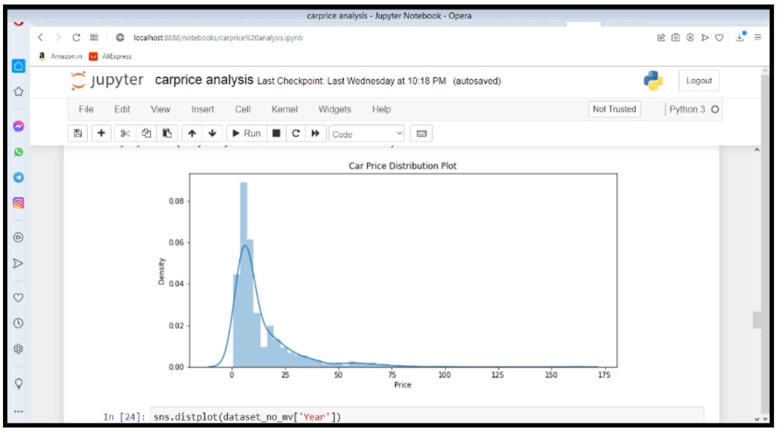
EDA:



This is the heatmap of the dataset

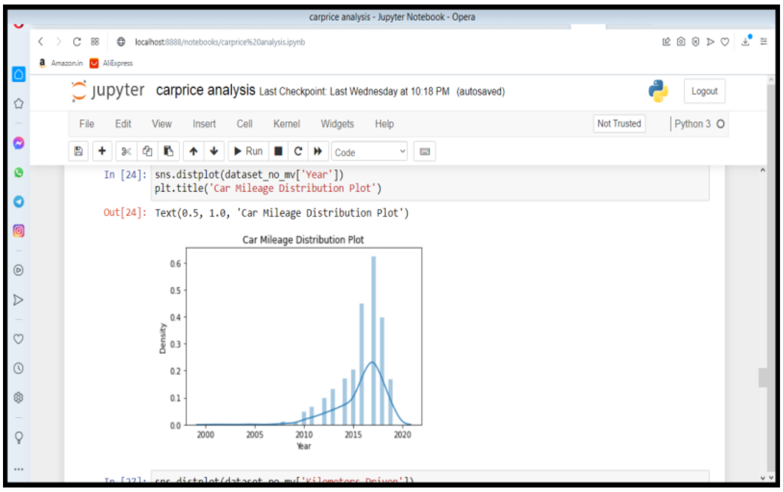


This is the relation between manufacturer and the count of cars


carprice analysis - Jupyter Notebook - Opera
localhost:8888/notebooks/carprice%20analysis.ipynb
Amazon.in
AliExpress
jupyter
carprice analysis Last Checkpoint: Last Wednesday at 10:18 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code
Car Price Distribution Plot
Density
0.00
0.02
0.04
0.06
0.08
0
25
50
75
100
125
150
175
Price

```
In [24]: sns.distplot(dataset_no_mv['Year'])
```

This is car price distribution plot


carprice analysis - Jupyter Notebook - Opera
localhost:8888/notebooks/carprice%20analysis.ipynb
Amazon.in
AliExpress
jupyter
carprice analysis Last Checkpoint: Last Wednesday at 10:18 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code
Car Mileage Distribution Plot
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
2000
2005
2010
2015
2020
Year

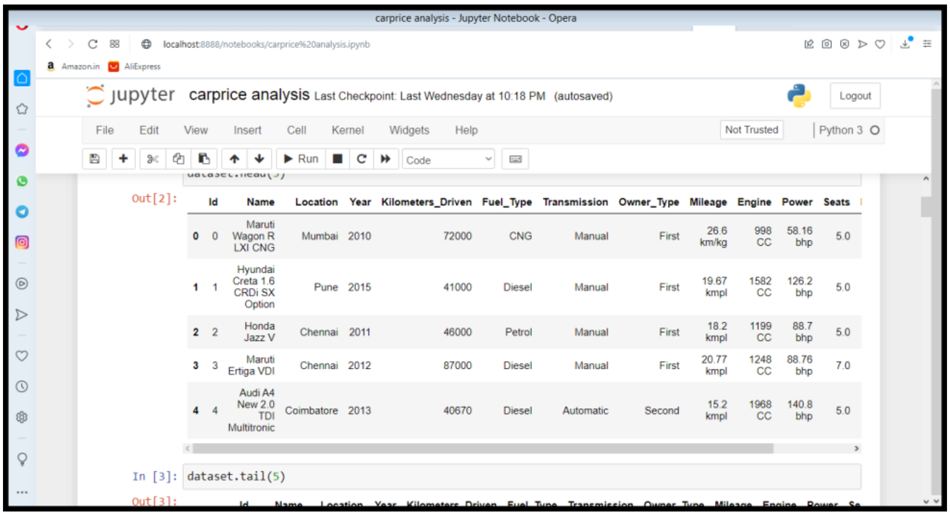
```
In [24]: sns.distplot(dataset_no_mv['Year'])
         plt.title('Car Mileage Distribution Plot')

Out[24]: Text(0.5, 1.0, 'Car Mileage Distribution Plot')

In [27]: sns.distplot(dataset_no_mv['Kilometers_Driven'])
```

Car Mileage distribution plot

- Data Sources and their formats

The data is collected from various site like cardekho,olx and scrapped the data and make a dataset. There are 14 columns and 6019 rows.

Target variable is price of the cars.



carprice analysis - Jupyter Notebook - Opera
localhost:8888/notebooks/carprice%20analysis.ipynb
Amazon.in
AliExpress
jupyter
carprice analysis Last Checkpoint: Last Wednesday at 10:18 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code

Out[2]:

	Id	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats
0	0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5.0
1	1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5.0
2	2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5.0
3	3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7.0
4	4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5.0

```
In [3]: dataset.tail(5)
```

- Data Inputs- Logic- Output Relationships

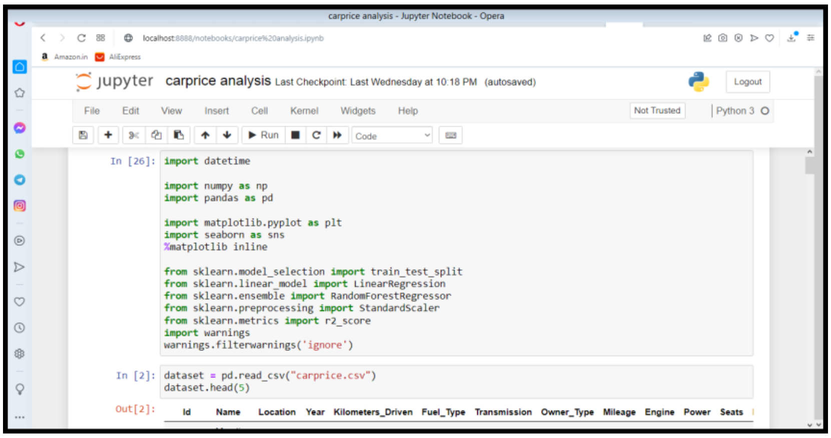
  Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

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

  Here, you can describe any presumptions taken by you.

- Hardware and Software Requirements and Tools Used

  Here we use lots of liaberies like pandas,numpy,matplot,seaborn, and we use python language for the coding purpose and import some other metrics liaberies also for model building like sklearn metrics ,Linear Regression,Random Forest Regressor ,accuracy report etc.



carprice analysis - Jupyter Notebook - Opera
localhost:8888/notebooks/carprice%20analysis.ipynb
Amazon.in
AliExpress
jupyter
carprice analysis Last Checkpoint: Last Wednesday at 10:18 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code

```
In [26]: import datetime

         import numpy as np
         import pandas as pd

         import matplotlib.pyplot as plt
         import seaborn as sns
         %matplotlib inline

         from sklearn.model_selection import train_test_split
         from sklearn.linear_model import LinearRegression
         from sklearn.ensemble import RandomForestRegressor
         from sklearn.preprocessing import StandardScaler
         from sklearn.metrics import r2_score
         import warnings
         warnings.filterwarnings('ignore')

In [2]: dataset = pd.read_csv("carprice.csv")
        dataset.head(5)

Out[2]:  Id  Name  Location  Year  Kilometers_Driven  Fuel_Type  Transmission  Owner_Type  Mileage  Engine  Power  Seats
```

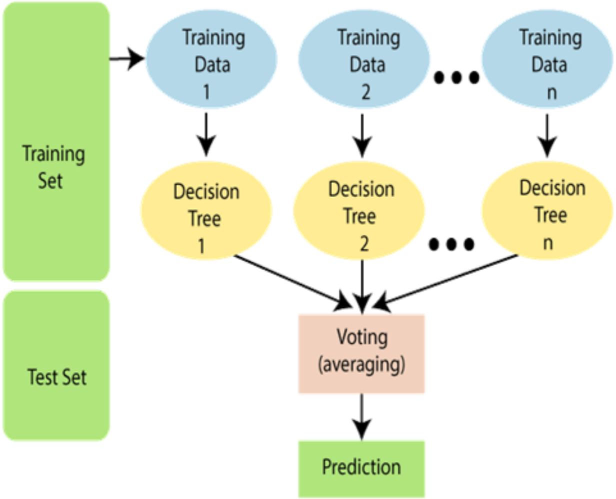
Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

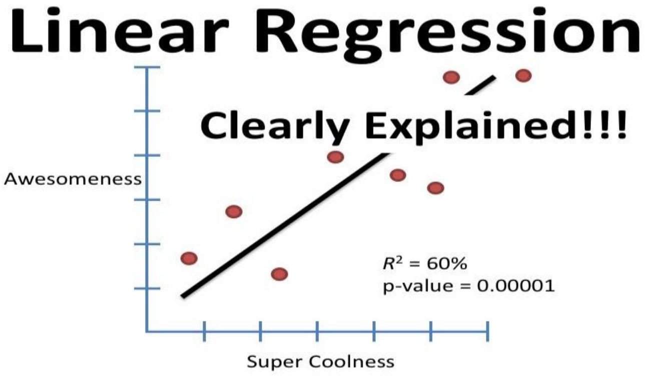
  We use randomforest classifier and Linear Regression for the model building as it is regression problem and finally we save the RandomForest Classifier model for gd accuracy.

- Testing of Identified Approaches (Algorithms)

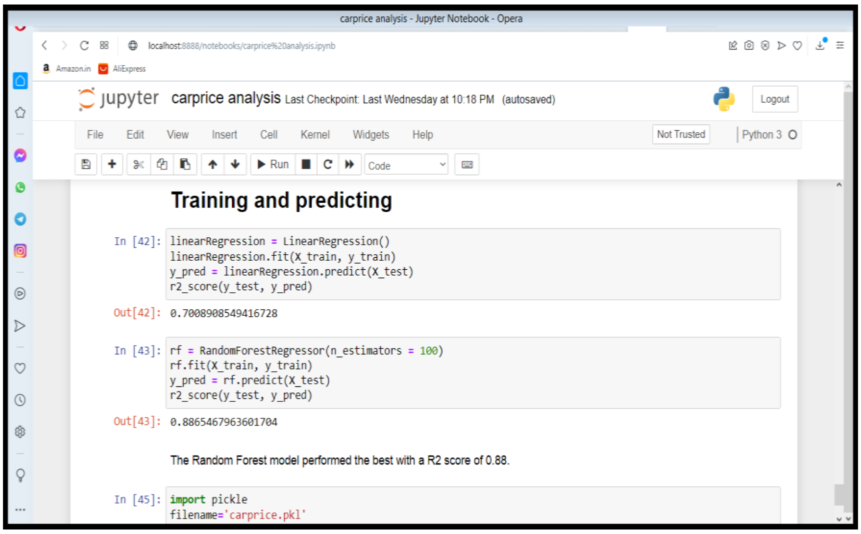
  Random Forest Classifier:



Linear Regression:



- Run and Evaluate selected models


carprice analysis - Jupyter Notebook - Opera
localhost:8888/notebooks/carprice%20analysis.ipynb
Amazon.in
AliExpress
jupyter
carprice analysis Last Checkpoint: Last Wednesday at 10:18 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code

Training and predicting

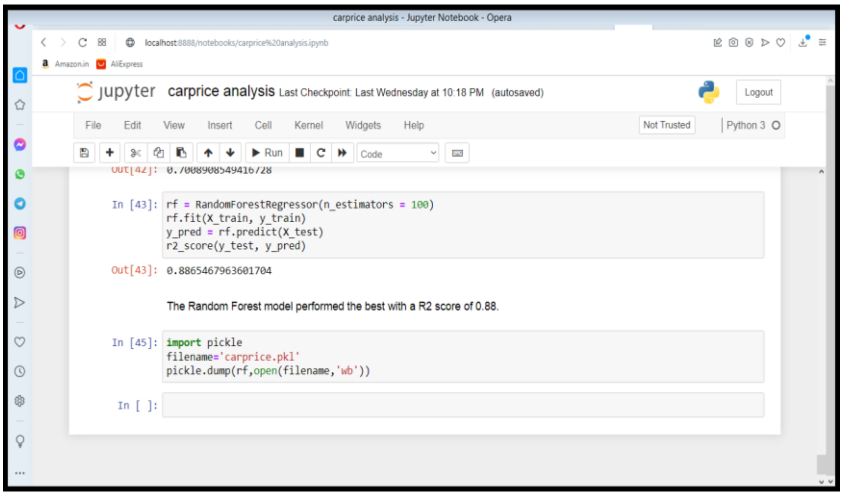
```
In [42]: linearRegression = LinearRegression()
         linearRegression.fit(X_train, y_train)
         y_pred = linearRegression.predict(X_test)
         r2_score(y_test, y_pred)

Out[42]: 0.7008908549416728

In [43]: rf = RandomForestRegressor(n_estimators = 100)
         rf.fit(X_train, y_train)
         y_pred = rf.predict(X_test)
         r2_score(y_test, y_pred)

Out[43]: 0.8865467963601704
```

The Random Forest model performed the best with a R2 score of 0.88.

```
In [45]: import pickle
         filename='carprice.pkl'
```

We save the best model and this is RandomForest Classifier model.

- Interpretation of the Results


carprice analysis - Jupyter Notebook - Opera
localhost:8888/notebooks/carprice%20analysis.ipynb
Amazon.in
AliExpress
jupyter
carprice analysis Last Checkpoint: Last Wednesday at 10:18 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code

```
Out[42]: 0.7008908549416728

In [43]: rf = RandomForestRegressor(n_estimators = 100)
         rf.fit(X_train, y_train)
         y_pred = rf.predict(X_test)
         r2_score(y_test, y_pred)

Out[43]: 0.8865467963601704
```

The Random Forest model performed the best with a R2 score of 0.88.

```
In [45]: import pickle
         filename='carprice.pkl'
         pickle.dump(rf,open(filename,'wb'))

In [ ]:
```

And we save Random Forest Classifier model and accuracy is so good.

CONCLUSION

- Key Findings and Conclusions of the Study

After the Final Submission of test data, my accuracy score was 88%

Feature engineering helped me increase my accuracy.

Amazingly Random Forest Classifier worked better than all other Ensemble models.

- Learning Outcomes of the Study in respect of Data Science

Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in car sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for car companies.