



NAME OF THE PROJECT

CAR PRICE PREDICTION PROJECT

Submitted by:

Pooja Rajpal

Internship batch 33

ACKNOWLEDGEMENT

I would like to express my appreciation to my institute

(Data Trained Education)

And to my respected teacher(Shankar sir) for his teachings and notes he provided to the students.I would like to thank my institute which is always available to answer our queries.

This project has been source of learning and bring our theoretical knowledge into real time projects .

I would really acknowledge my teacher's help and guidance.

INTRODUCTION

Business problem statement:

This problem statement is about predicting the price of car. Due to covid there are lot of changes in the car market. Cars in demand have high price while those having low demand have less price. There is a need for the Car Price Prediction to determine the worthiness of the car using lot of features .Due to the increased price of the new cars and the incapability of the customers to buy new cars due to lack of funds ,used cars sales are on a global increase .There is need for the used cars sales are on a global increase.

Conceptual background of the domain problem

To understand the problem it is necessary to understand what the project is about .As the prices of new cars are rising at high rates many individuals are interested in used cars ,so there is need for the effective system to determine the price of the cars

using variety of features such as model, kms run ,accidental etc.Different models and systems contributes on predicting the price of the used car.It is important to know their actual market value.

There are many websites having good prediction model.However having a second model may help them to get a better prediction.

Review of literature

For this project many used car data available on used car website is scrapped.and converted in the form of dataframe .The features available are model year ,kms run, name of the car, city, whether petrol or diesel.Deciding whether the posted price is worth is difficult .Several factors including modelyear, make etc influence the worth of the car.Before making the model it is necessary to understand about the features.

Motivation of the problem undertaken

Building project will improve my skills and make things better.Building successful projects gives the

feeling of pride .And for that I need to build more and more projects passionately and try to do it

with more accuracy. Based on the data, the aim is to use the machine learning algorithms to develop a model for predicting used car prices.

However, completing the project requires complex steps. Things get really difficult at some stages, but we need to adjust our vision. And try once again. Also, it's my passion to build a model with high accuracy and will definitely try to learn more and more for that.

Analytical Problem Framing

Mathematical/analytical modelling of the problem

The data is collected through web scrapping from various websites of used cars (OLX, cars 24 etc) and data is collected in a DataFrame and all the dataframes are concatenated to collect.

Statistical Modelling is necessary because it summarizes the results and can be observed by the evaluators. It is a relationship between variables.

Checking for nulls is another step in this dataset and impute them using imputers to clean the data.

Statistical techniques such as mean, median, standard deviation, interquartile ranges. These are simple and we can start it as starting point for EDA.

The goal is to collect the data and make predictions about real world.

**Two types statistics (Descriptive and inferential)
Eg(weather forecasting, stock market, loan approval and fraud detection, housing etc)**

Mean, median and mode are the measures of central tendency and is descriptive statistics.

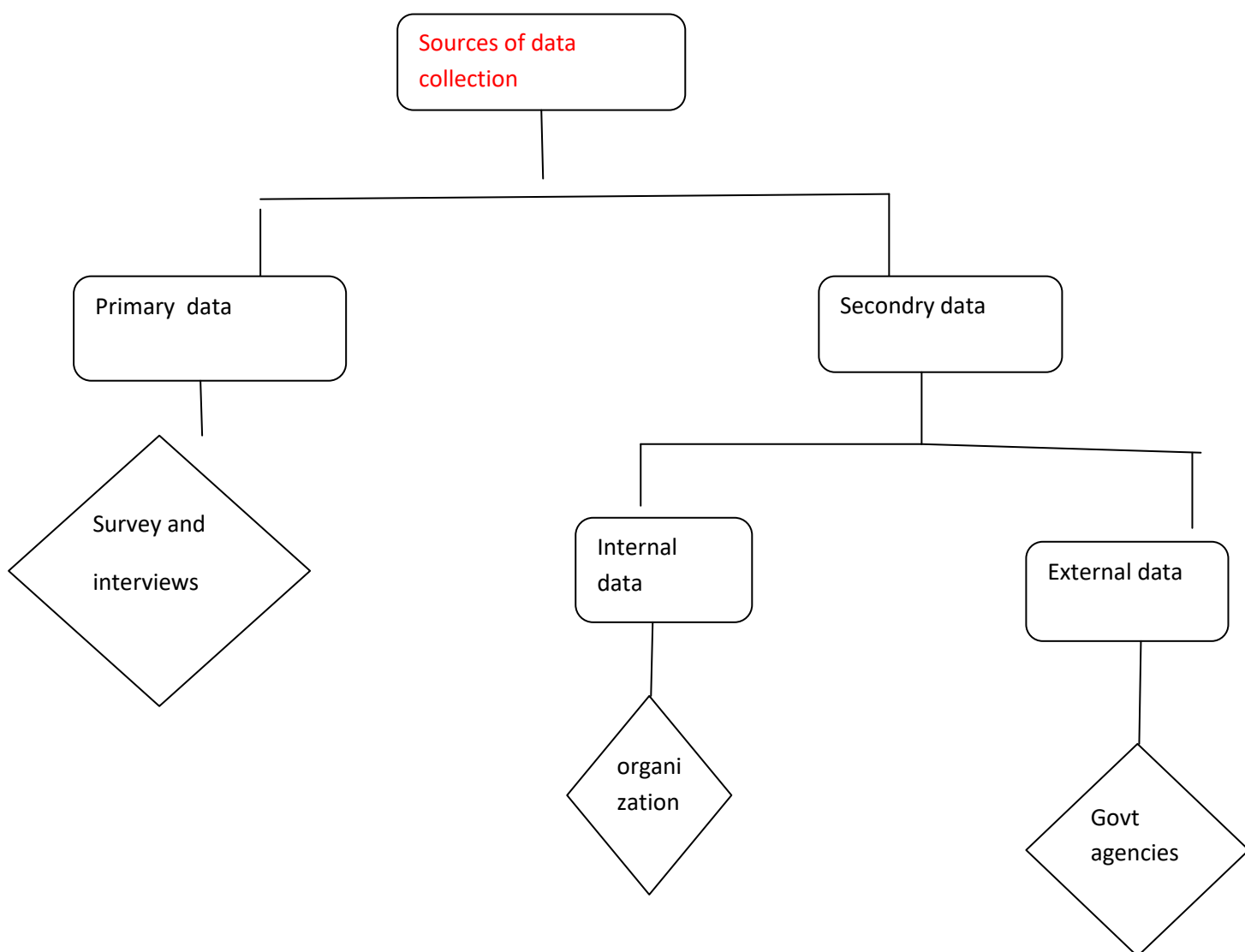
Sampling of data and infer the results to describe entire population.

Data sources and their formats

Data is collected through webscrapping. Web scrapping also known as data scrapping, is the process of importing the information from various websites into

jupyter notebook.Its one of the most efficient way to get the data from the web .

Data collection is the process of collecting data whether in structured format or unstructured format.



Web scrapping is one of the one of the online primary data collection.

Data preprocessing done

Step1:collection of data and load the dataset in jupyter notebook using pandas and for that we need to import pandas library .

Step2:data cleaning that include checking for nulls and treat them using imputers.

Step3:checking for datatypes and if object datatype convert them into integer as computers only understand numeric data

Step 4:data visualisation

Step 5:plotting boxplot for columns to check for outliers and treating them using outlier detection

Step6:checking for skewness and treat them using Power transformer

Step 7:when the data is cleaned we will check for collinearity using heatmap

At the end split the dataset into features and label.

Data inputs-logic-output relationship

There is a positive relationship between input and output. Or we can say that there are direct relationships between input and output. The input variables, i.e. names of the used cars to be sold, kms run, year etc are having the direct relationship with the output price as lots of features are to be seen by the customers to purchase the cars. Input variables are necessary for predicting the price (output) variable.

Hardware and software requirements and tools used

Laptop and

Anaconda navigator is desktop graphical user interface that allows to launch applications

Jupyter notebook is open source software

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

Two main statistical methods are used in data analysis: descriptive (mean and median) and another is inferential that draws conclusion from data using statistical tests.

such as t –test .We used descriptive metod in solving the problem.

An analytical approach takes the problem and breaks it into elements to understand the problem .Analytical approach requires summarizes and finding the trend in the data .

- Testing of Identified Approaches (Algorithms)

As it is a regression problem linear regression and random forest regression are the algorithms used to build the model.

- Run and Evaluate selected models

Linear regression is a data analysis technique tha predicts the value of unknown data by using data value .It shows the linear relationship say between height and weight of a person.

```
In [334]: from sklearn.linear_model import LinearRegression
          from sklearn.metrics import r2_score
```

```
In [335]: lr=LinearRegression()
```

```
In [336]: for i in range(0,100):
          x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=i)
          lr.fit(x_train,y_train)
          pred_train=lr.predict(x_train)
          pred_test=lr.predict(x_test)
          print(f"at random state{i},the training accuracy is{r2_score(y_train,pred_train)}")
          print(f"at random state{i},the training accuracy is{r2_score(y_test,pred_test)}")
          print("\n")
```

```
at random state13,the training accuracy is0.4496188280683663
at random state13,the training accuracy is0.4514128980540183
```

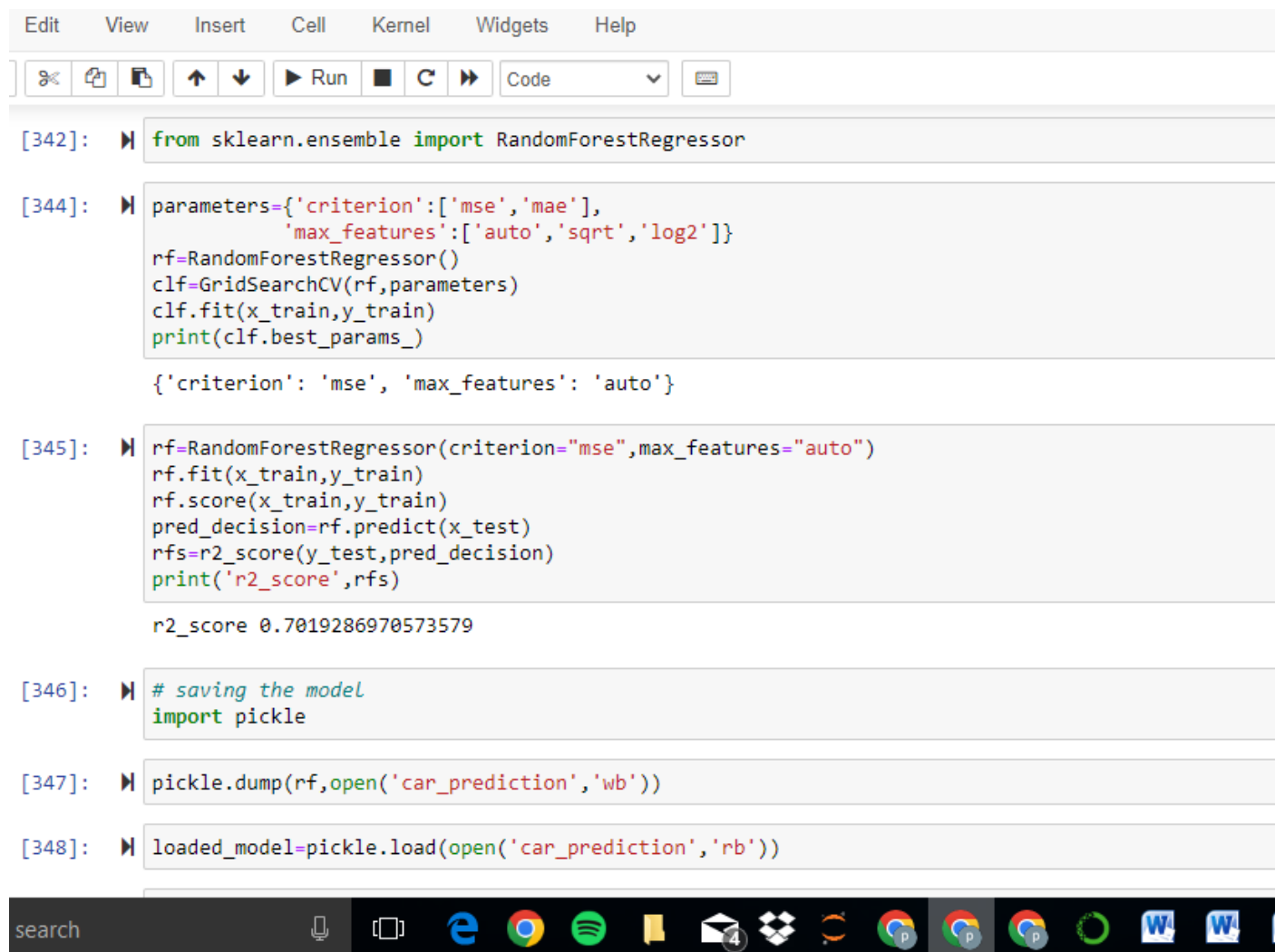
```
at random state14,the training accuracy is0.4564994496774091
at random state14,the training accuracy is0.4308377177334116
```

type here to search



Random forest Regressor

A random forest regressor works with data having a numeric or continuous output and they cannot be defined by classes.



```
[342]: from sklearn.ensemble import RandomForestRegressor

[344]: parameters={'criterion':['mse','mae'],
                 'max_features':['auto','sqrt','log2']}
rf=RandomForestRegressor()
clf=GridSearchCV(rf,parameters)
clf.fit(x_train,y_train)
print(clf.best_params_)

{'criterion': 'mse', 'max_features': 'auto'}

[345]: rf=RandomForestRegressor(criterion="mse",max_features="auto")
rf.fit(x_train,y_train)
rf.score(x_train,y_train)
pred_decision=rf.predict(x_test)
rfs=r2_score(y_test,pred_decision)
print('r2_score',rfs)

r2_score 0.7019286970573579

[346]: # saving the model
import pickle

[347]: pickle.dump(rf,open('car_prediction','wb'))

[348]: loaded_model=pickle.load(open('car_prediction','rb'))
```

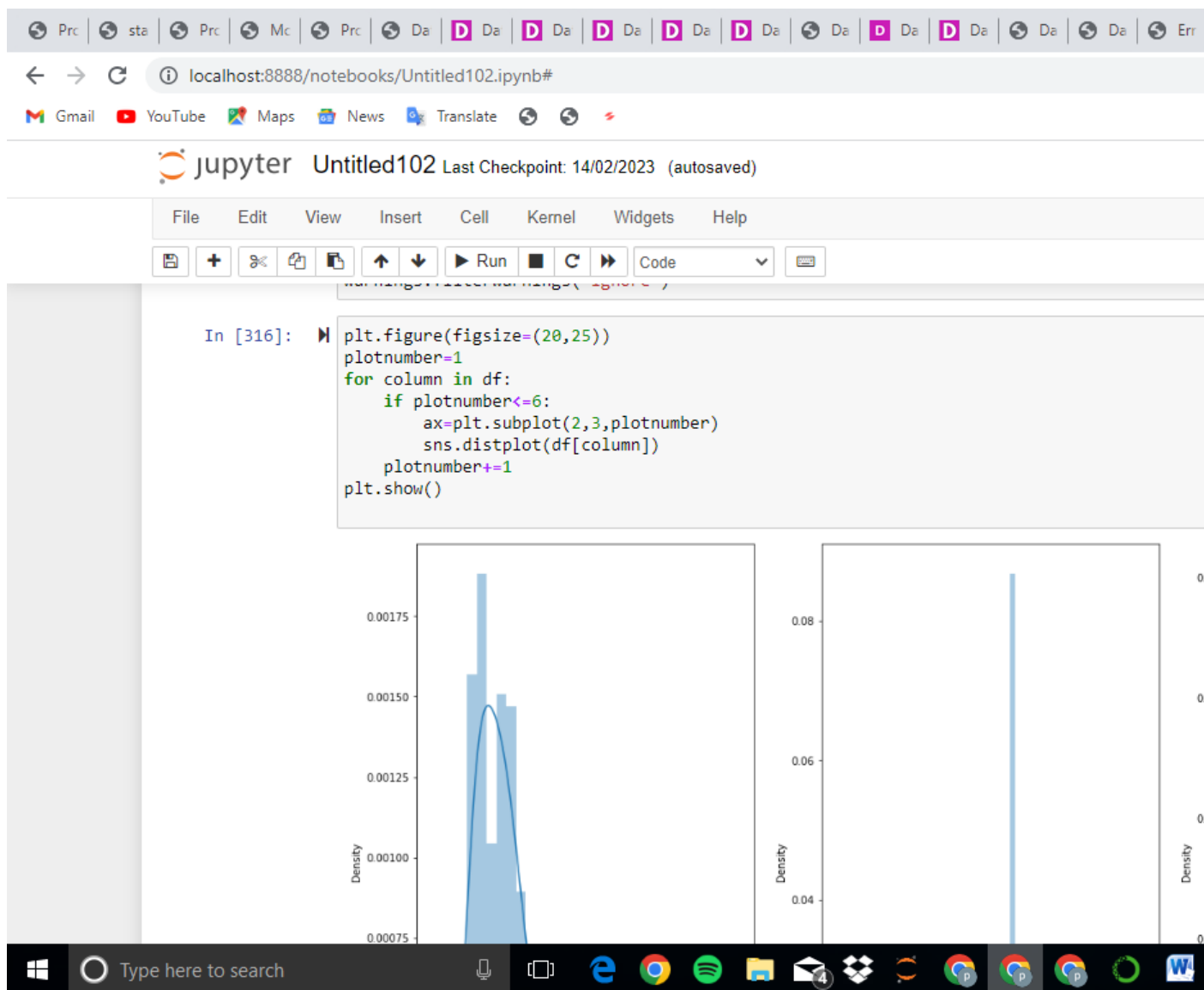
- Key Metrics for success in solving problem under consideration

Metrics are important as they enable you to understand the performance of your business . I used

the accuracy_score to find the accuracy of the model. because according to me it gives most reliable results and its easy to determine what is the level of accuracy of the model built.

- Visualizations

Distplot to visualise whether the data is right, left skewed to make it normalize.



- Interpretation of the Results

RandomForest Regressor works well with this project .As this is a regression problem because the label is continuous data .Preprocessing is necessary to clean the data

Visualization because it is easily understood by us to observe the data in graphical form.

Conclusion

This is a used car price prediction project in which data is collected by number of sites through web scrapping.

Data cleaning and Exploratory Data Analysis is necessary to analyse the data.After that checking the nulls and imputing them to build the model.Also it is necessary to deal with the skewness to increase the accuracy.Hyperparameter tuning is necessary to increase the accuracy

- Learning Outcomes of the Study in respect of Data Science

I faced lot of challenges in making this project
Also learnt lot of things. As I have collected the data through webscrapping I found bit difficulties in inspecting some of the elements..

In regression problem the output variable must be continuous and in classification it must be discrete.

- Limitations of this work and Scope for Future Work

I tried hard to increase the accuracy. But I think I should have worked harder to increase the accuracy .