FLIP ROBO

CAR PRICE PREDICTION PROJECT

SUBMITTED BY:-

RISHABH JOHRI

ACKNOWLEDGEMENT:-

It is a great pleasure to express my gratitude to Team Flip Robo, for giving me the opportunity to work on a interesting project, which helped me in improving my knowledge, coding skills and my analysation skills.

Team Flip Robo also gave me opportunity to build PowerPoint Presentation and Project Report, which will help me to share steps taken while building the entire model. It has helped me in deciding about the future prospects of various Data Science fields. Now, I will explain the understanding of the project through this report.

INTRODUCTION

Scraping 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 for 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 Extra Trees Classifier achieves the best accuracy score.

Because of new computing technologies, machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks, researchers interested in artificial intelligence wanted to see if computers could learn from data. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable & repeatable decisions and results thereof.

Business Problem Framing:-

Car has become a basic requirement of every person in this fast and dynamic world. It’s a science that’s not new – but one that has gained fresh momentum. While there is an end number of applications of machine learning in real life one of the most prominent application is the prediction problems. There are various topics on which the prediction can be applied. One such application is what this project is focused upon. Websites recommending items you might like based on previous purchases are using machine learning to analyse your buying history – and promote other items you'd be interested in. This ability to capture data, analyse it and use it to personalize a shopping experience (or implement a marketing campaign) is the future of retail.

The requirement of model building is to predict the price of car in accordance with variable and how the variable describe the car and help people to purchase used cars.

• Conceptual Background of the Domain Problem:-

- Used cars selling more than new cars in India these days !

More consumers are looking at second-hand entry-level cars as opposed to buying new cars, this has taken a major chunk out of new car sales. A trend that is expected to grow even more!

Estimating the sale prices of used cars is one of the basic project in Data Science.

Technical Requirements:

- Scraped data having various variables.
- Data contains Null values. Data treatment using domain knowledge, understanding.
- Extensive EDA has to be performed to gain relationships of important variable and price.
- Data contains numerical as well as categorical variable, handle them accordingly.
- Machine Learning models, apply regularization and determine values of Hyper Parameters.
- Here, there is a need to find important features which affect the price positively or negatively.

Review of Literature:-

There are various applications and methods which inspired us to build our project. We did a background survey regarding the basic ideas of our project and used those ideas for the collection of information like the technological stack, algorithms, and shortcomings of our project which led us to build a better project. There are various platforms like OLX , Car Dekho etc where seller can sell their used car. It is an Indian Start-up with a simplified user interface which asks seller parameters like car model, kilometres travelled, year of registration and vehicle type (Petrol, Diesel, Petrol/CNG). These allow the web model to run certain algorithms on given parameters and predict the price. This app predicts vehicle prices on various parameter like Location, Model, Price and kilometres travelled. This app uses a machine learning approach to predict the price of a car, on the basis of fuel type petrol, diesel, electric or hybrid vehicle. App can predict the price of any vehicle because of the smartly optimized algorithm and some with the help of AI also these days.

Motivation for the Problem Undertaken:-

Basically, thinking on whether a used car is worth the posted price or not. Several factors, including mileage, make, model, year, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma to price a used car, Based on existing data, **the aim is to use machine learning algorithms to develop models for predicting used car prices.**

Analytical Problem Framing

- **Mathematical/ Analytical Modelling of the Problem:-**

- I am working with the used car price dataset that contains various features and information about the cars and its sale price. Using the scraped data in the form of 'read_csv' function, which can import the data into our python environment. After importing the data, I have used the 'head' function to get a glimpse of the dataset.

- As used car are growing in the market these days and there is huge demand for it, because car has become neccessity for everyone nowadays.

- Dataset was cleaned and transformed and some explorations were done on it to answer some basic questions that anybody would like to ask about cars.

- The dataset here is with the right variables to be used in the algorithms, which results in improved model accuracy. Different ensemble algorithms were used on the dataset for predicting various values.

Data Pre Processing :-

As a first step I have imported required libraries and I have imported the dataset using csv file.

Then, did all the statistical analysis like checking shape, n unique- for checking unique values in the dataset, value counts, info and description, etc.

While checking for null values I found null values in 3-4 columns that I handled using Mean and Mode methods as per the parameters.

I dropped Unnamed:0 as it is not required for proceeding ahead because it is containing index numbers and that we have already have in it.

I have used distplot for understanding the relationship between Features and the Variable for numerical columns.

I have used countplot for understanding the relationship between Features and the Variable for categorical columns.

Then also used other plotting methods like scatterplot, stripplot, pieplot, etc, for understanding visualisations in a better way.

Data Cleaning:-

In the dataset, we have found null values also in 3-4 columns, and found some outliers and skewness levels in the dataset.

To remove outliers, i have used ZScore method. And for removing skewness I have used Yeo-Johnson method.

Then I have used Standard Scaler method to Scale the data.

In this, issue of Multicollinearity is not there as checked and confirmed using VIF also.

Then, after all Preparations of our data, we have moved to model selection phase.

Model Selection:-

Since New_Price was our Target column and it was a Categorical column, so this particular problem was a Classification problem. And we have used all Classification algorithms to build our model. We Tried multiple models and to avoid the confusion of overfitting we went through cross validation. Below are the list of Classification algorithms I have used in my project. By looking into the least difference between accuracy score and cross validation score, I found Extra Trees Classifier as the best model for proceeding ahead. Other models which i have used were:

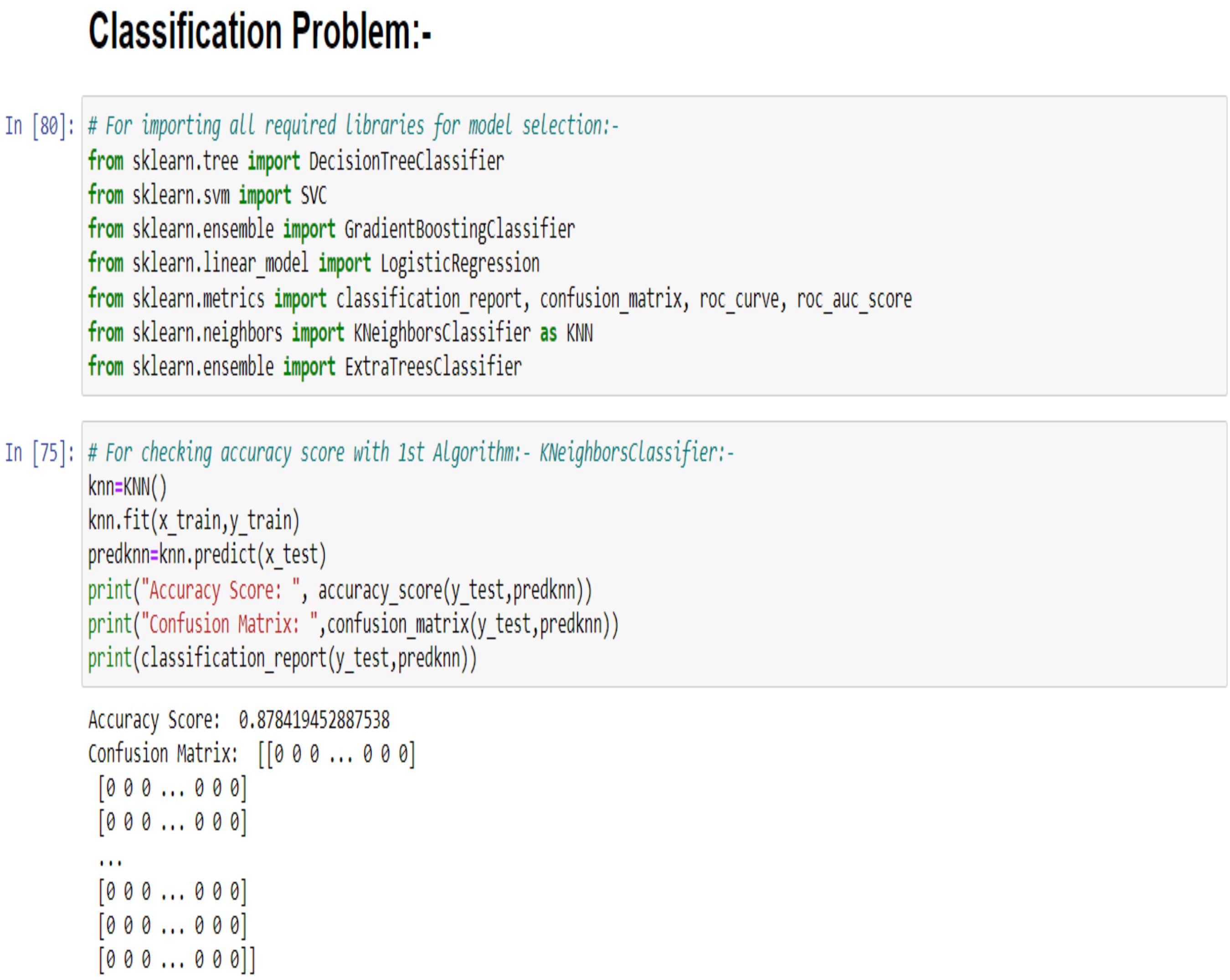
K Neighbors Classifier.

Decision Tree Classifier.

Gradient Boosting Classifier.

Random Forest Classifier.

Different Model Accuracy Scores:-

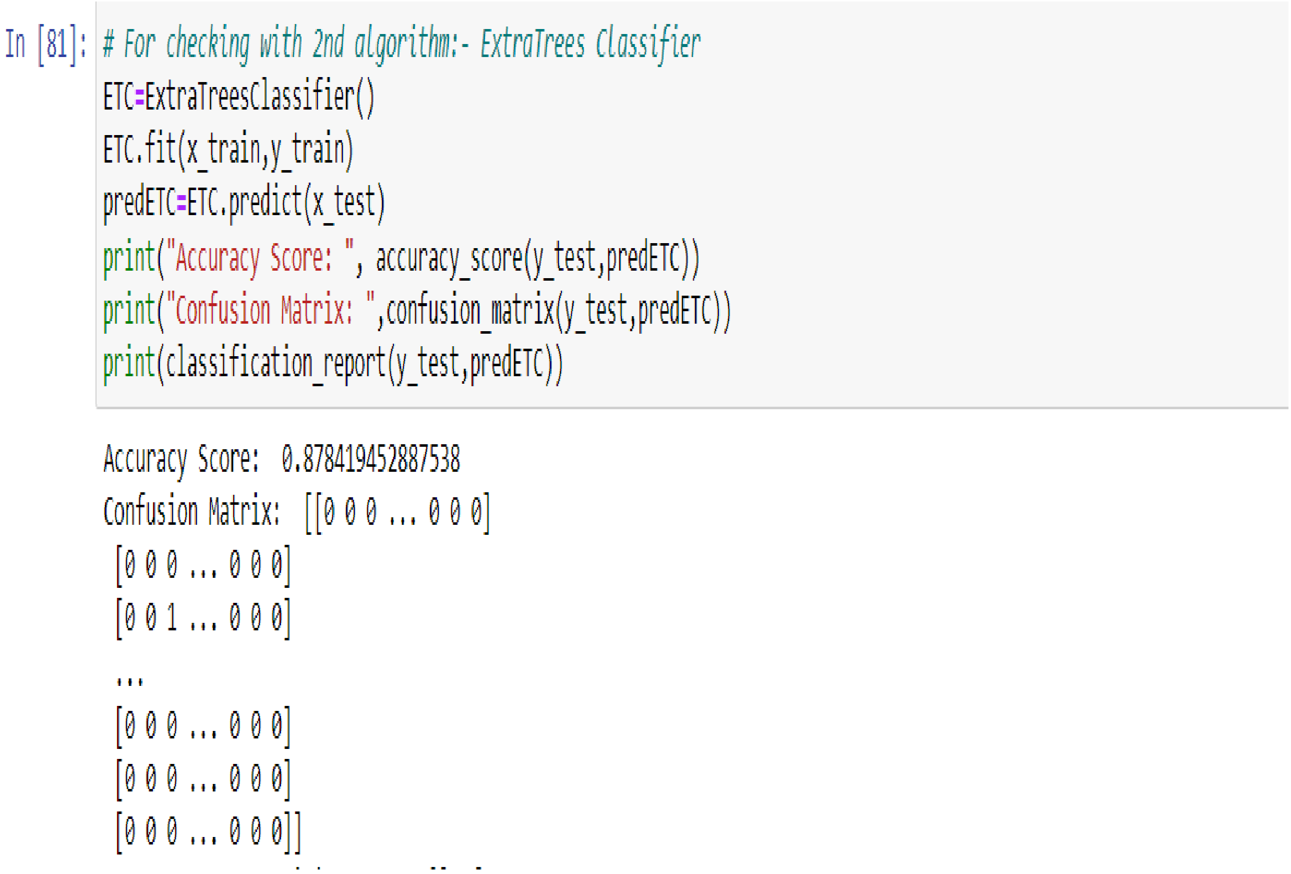


Classification Problem:-

```
In [80]: # For importing all required libraries for model selection:-
         from sklearn.tree import DecisionTreeClassifier
         from sklearn.svm import SVC
         from sklearn.ensemble import GradientBoostingClassifier
         from sklearn.linear_model import LogisticRegression
         from sklearn.metrics import classification_report, confusion_matrix, roc_curve, roc_auc_score
         from sklearn.neighbors import KNeighborsClassifier as KNN
         from sklearn.ensemble import ExtraTreesClassifier
```

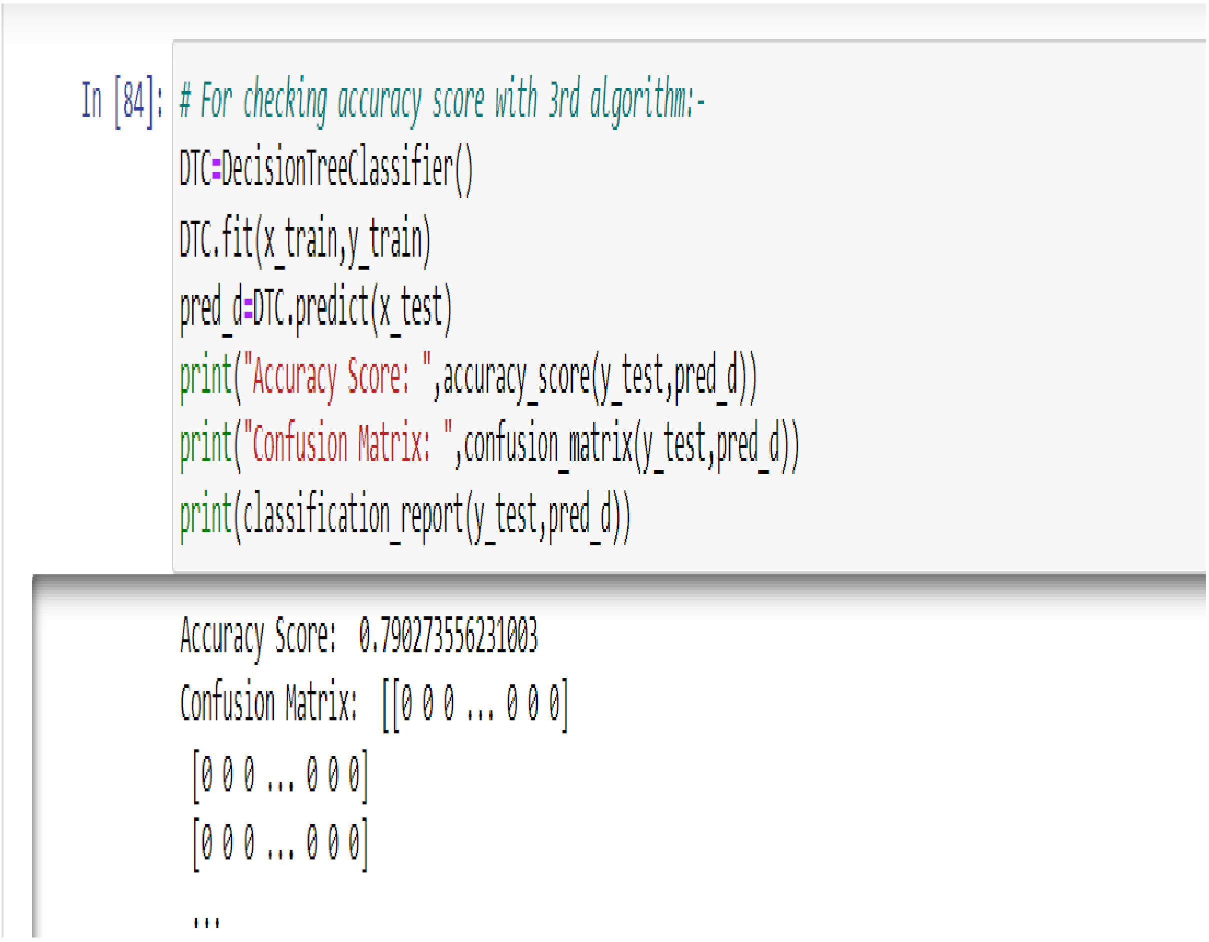
```
In [75]: # For checking accuracy score with 1st Algorithm:- KNeighborsClassifier:-
         knn=KNN()
         knn.fit(x_train,y_train)
         predknn=knn.predict(x_test)
         print("Accuracy Score: ", accuracy_score(y_test,predknn))
         print("Confusion Matrix: ",confusion_matrix(y_test,predknn))
         print(classification_report(y_test,predknn))
```

```
Accuracy Score:  0.878419452887538
Confusion Matrix:  [[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```



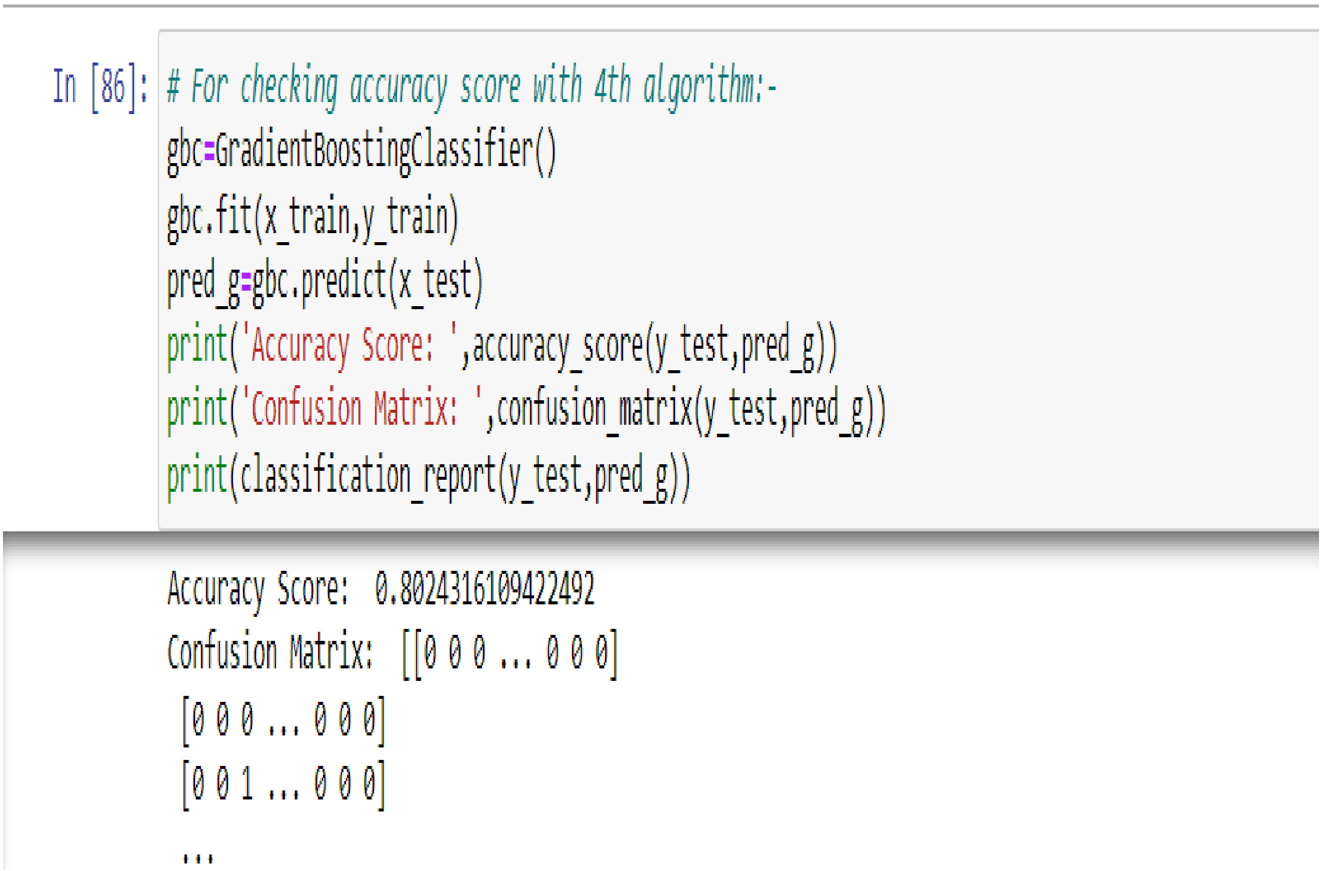
```
In [81]: # For checking with 2nd algorithm:- ExtraTrees Classifier
         ETC=ExtraTreesClassifier()
         ETC.fit(x_train,y_train)
         predETC=ETC.predict(x_test)
         print("Accuracy Score: ", accuracy_score(y_test,predETC))
         print("Confusion Matrix: ",confusion_matrix(y_test,predETC))
         print(classification_report(y_test,predETC))
```

```
Accuracy Score:  0.878419452887538
Confusion Matrix:  [[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 1 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```



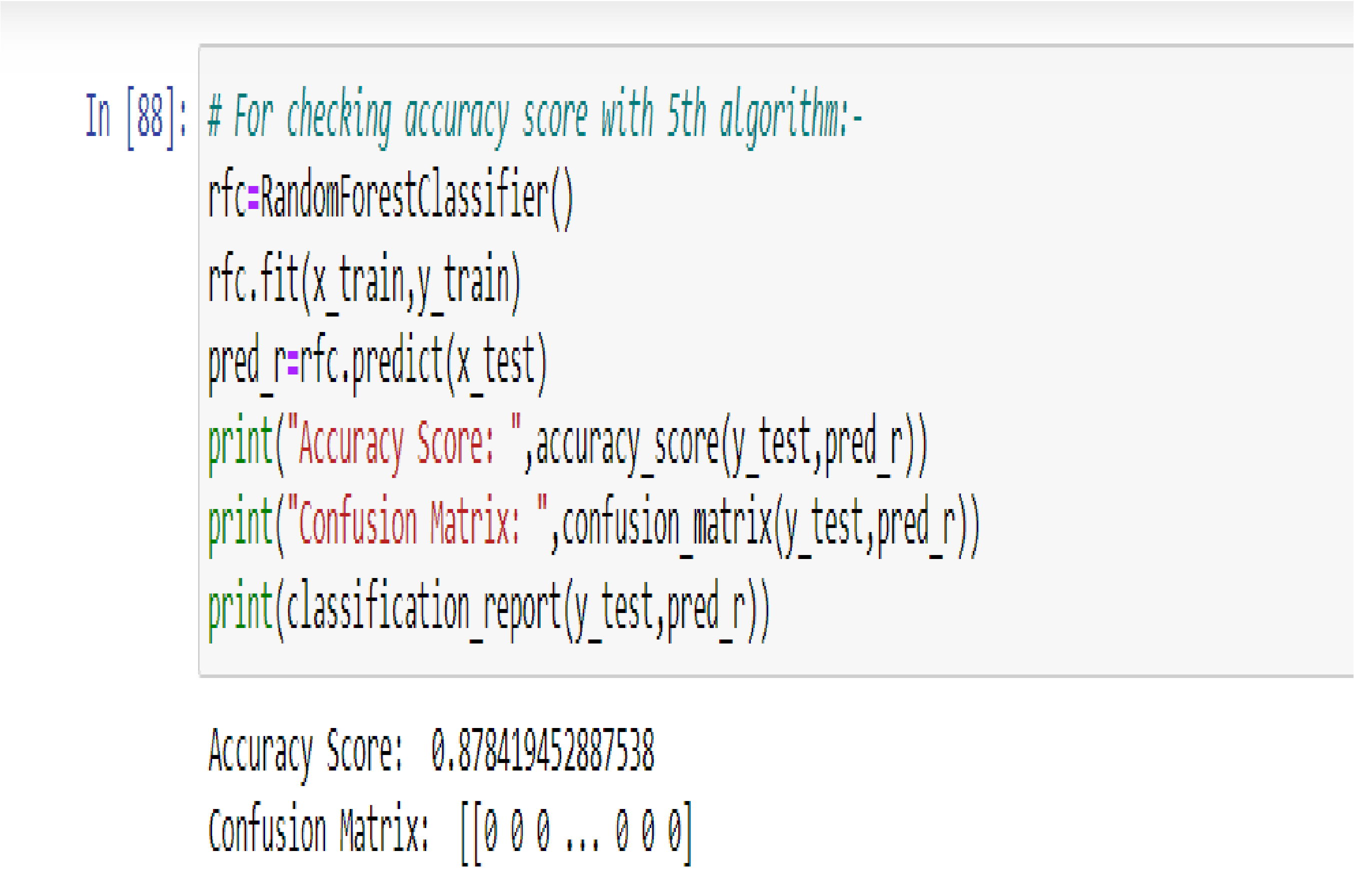
```
In [84]: # For checking accuracy score with 3rd algorithm:-
         DTC=DecisionTreeClassifier()
         DTC.fit(x_train,y_train)
         pred_d=DTC.predict(x_test)
         print("Accuracy Score: ",accuracy_score(y_test,pred_d))
         print("Confusion Matrix: ",confusion_matrix(y_test,pred_d))
         print(classification_report(y_test,pred_d))
```

```
Accuracy Score:  0.790273556231003
Confusion Matrix:  [[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
```



```
In [86]: # For checking accuracy score with 4th algorithm:-
         gbc=GradientBoostingClassifier()
         gbc.fit(x_train,y_train)
         pred_g=gbc.predict(x_test)
         print('Accuracy Score: ',accuracy_score(y_test,pred_g))
         print('Confusion Matrix: ',confusion_matrix(y_test,pred_g))
         print(classification_report(y_test,pred_g))
```

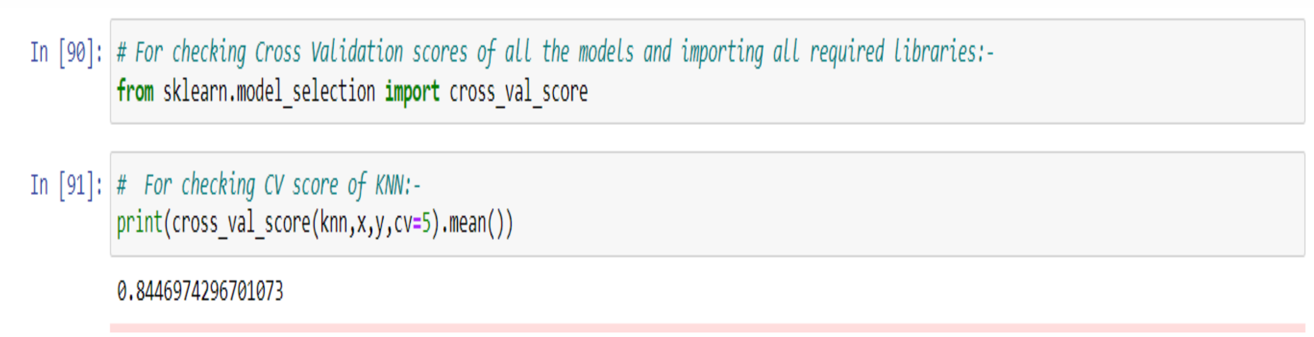
```
Accuracy Score:  0.8024316109422492
Confusion Matrix:  [[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 1 ... 0 0 0]
 ...
```



```
In [88]: # For checking accuracy score with 5th algorithm:-
         rfc=RandomForestClassifier()
         rfc.fit(x_train,y_train)
         pred_r=rfc.predict(x_test)
         print("Accuracy Score: ",accuracy_score(y_test,pred_r))
         print("Confusion Matrix: ",confusion_matrix(y_test,pred_r))
         print(classification_report(y_test,pred_r))
```

```
Accuracy Score:  0.878419452887538
Confusion Matrix:  [[0 0 0 ... 0 0 0]
```

CROSS VALIDATION PHASE:-



```
In [90]: # For checking Cross Validation scores of all the models and importing all required libraries:-
         from sklearn.model_selection import cross_val_score

In [91]: #  For checking CV score of KNN:-
         print(cross_val_score(knn,x,y,cv=5).mean())

         0.8446974296701073
```

For checking CV score of Extra Trees:-

print(cross_val_score(ETC,x,y,cv=5).mean())

```
0.8520980908048303
```

For checking CV score of DecisionTree:-

print(cross_val_score(DTC,x,y,cv=5).mean())

```
0.760912096066923
```

For checking CV score of RandomForest:-

print(cross_val_score(rfc,x,y,cv=5).mean())

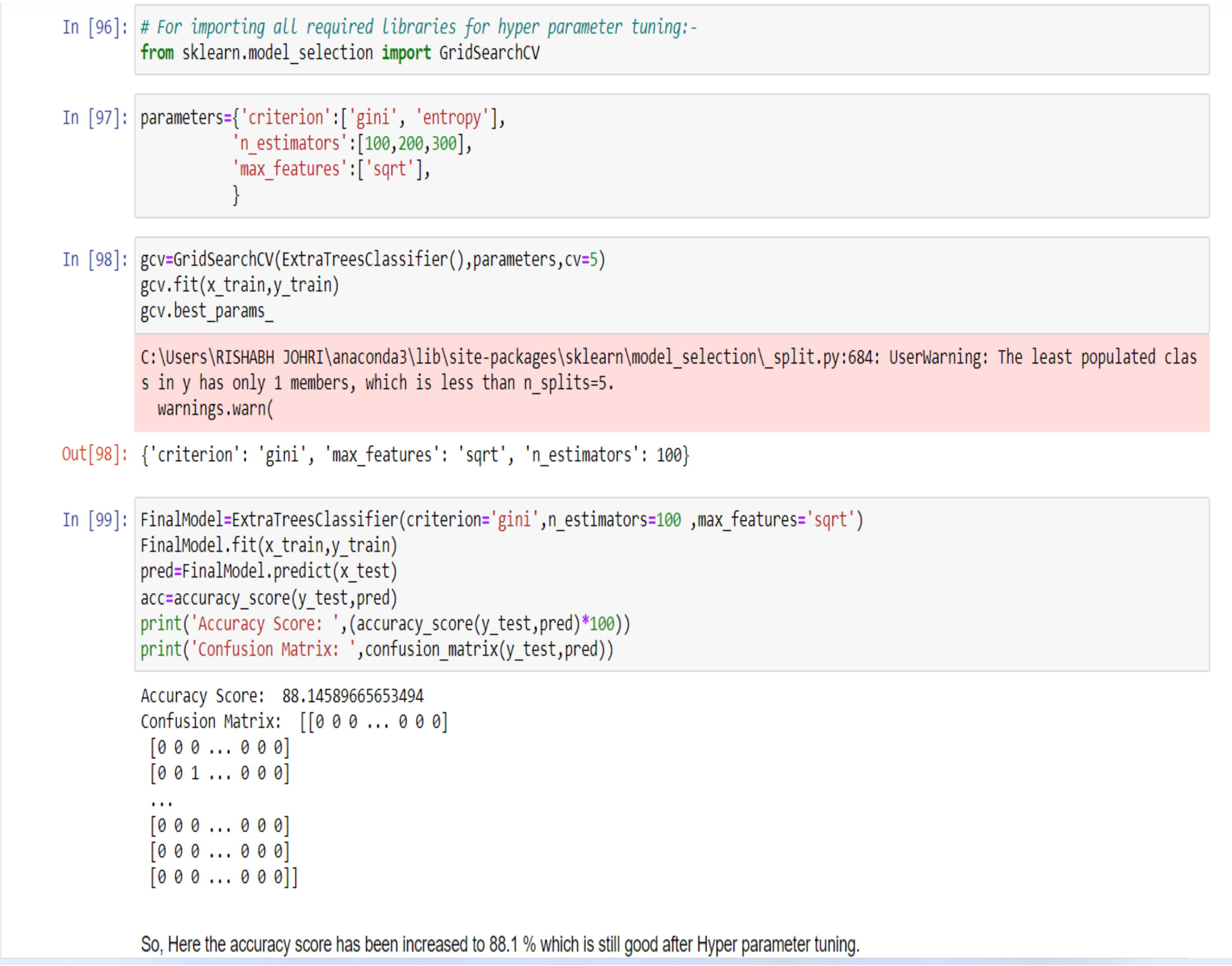
```
0.8537441813398097
```

For checking CV score of GradientBoost:-

print(cross_val_score(gbc,x,y,cv=5).mean())

```
0.7666363084395871
```

Hyper Parameter Tuning:-



```
In [96]: # For importing all required libraries for hyper parameter tuning:-
         from sklearn.model_selection import GridSearchCV
```

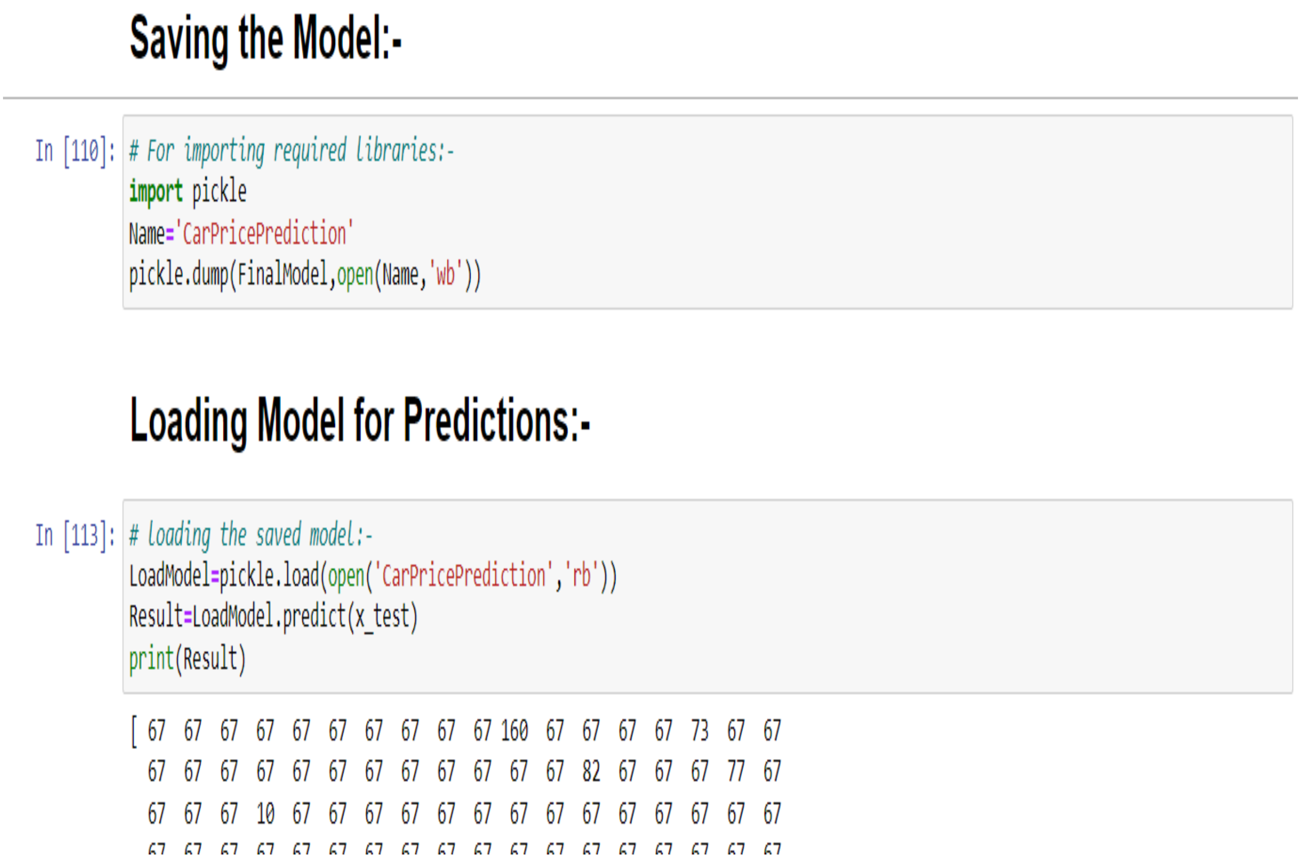
```
In [97]: parameters={'criterion':['gini', 'entropy'],
                    'n_estimators':[100,200,300],
                    'max_features':['sqrt'],
                    }
```

```
In [98]: gcv=GridSearchCV(ExtraTreesClassifier(),parameters,cv=5)
         gcv.fit(x_train,y_train)
         gcv.best_params_

         C:\Users\RISHABH JOHRI\anaconda3\lib\site-packages\sklearn\model_selection\_split.py:684: UserWarning: The least populated clas
         s in y has only 1 members, which is less than n_splits=5.
           warnings.warn(

Out[98]: {'criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 100}
```

```
In [99]: FinalModel=ExtraTreesClassifier(criterion='gini',n_estimators=100 ,max_features='sqrt')
         FinalModel.fit(x_train,y_train)
         pred=FinalModel.predict(x_test)
         acc=accuracy_score(y_test,pred)
         print('Accuracy Score: ',(accuracy_score(y_test,pred)*100))
         print('Confusion Matrix: ',confusion_matrix(y_test,pred))

         Accuracy Score:  88.14589665653494
         Confusion Matrix:  [[0 0 0 ... 0 0 0]
          [0 0 0 ... 0 0 0]
          [0 0 1 ... 0 0 0]
          ...
          [0 0 0 ... 0 0 0]
          [0 0 0 ... 0 0 0]
          [0 0 0 ... 0 0 0]]
```

So, Here the accuracy score has been increased to 88.1 % which is still good after Hyper parameter tuning.

SAVING THE MODEL AND LOADING PREDICTIONS:-



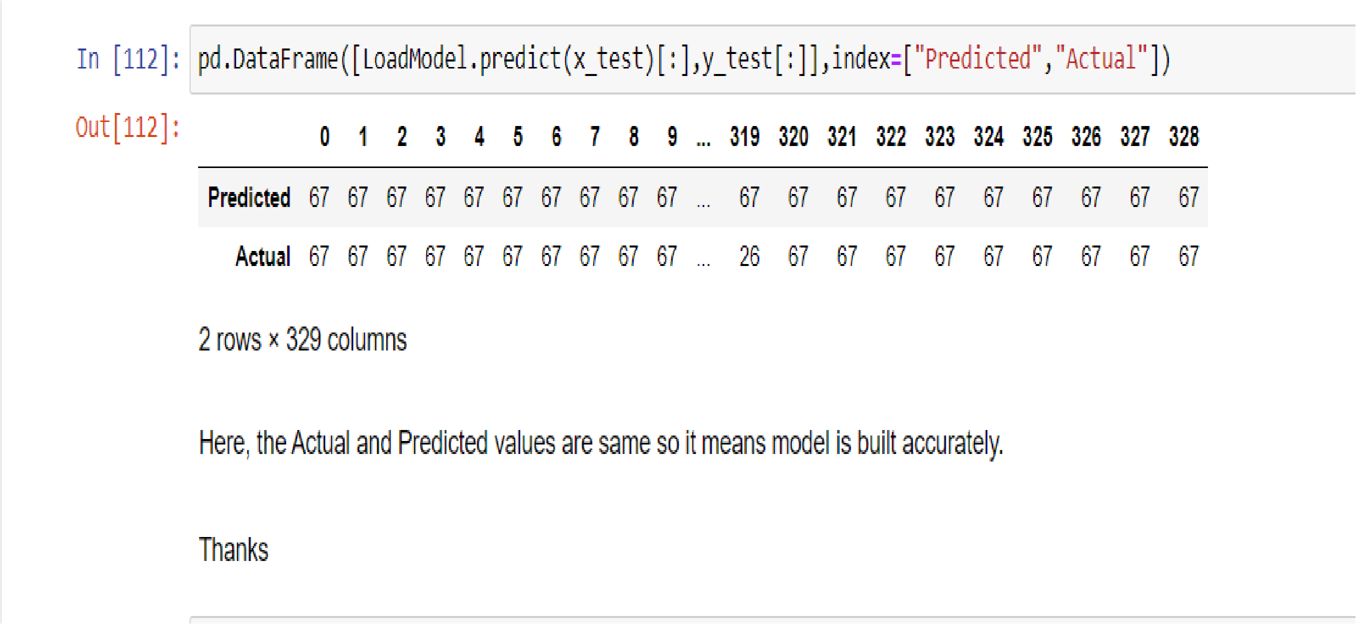
Saving the Model:-

In [110]:
```
# For importing required libraries:-
import pickle
Name='CarPricePrediction'
pickle.dump(FinalModel,open(Name,'wb'))
```

Loading Model for Predictions:-

In [113]:
```
# loading the saved model:-
LoadModel=pickle.load(open('CarPricePrediction','rb'))
Result=LoadModel.predict(x_test)
print(Result)
```

```
[ 67  67  67  67  67  67  67  67  67  67 160  67  67  67  67  73  67  67
  67  67  67  67  67  67  67  67  67  67  67  67  82  67  67  67  77  67
  67  67  67  10  67  67  67  67  67  67  67  67  67  67  67  67  67  67
  67  67  67  67  67  67  67  67  67  67  67  67  67  67  67  67  67  67
```



In [112]:
```
pd.DataFrame([LoadModel.predict(x_test)[:],y_test[:]],index=["Predicted","Actual"])
```

Out[112]:

	0	1	2	3	4	5	6	7	8	9	...	319	320	321	322	323	324	325	326	327	328
Predicted	67	67	67	67	67	67	67	67	67	67	...	67	67	67	67	67	67	67	67	67	67
Actual	67	67	67	67	67	67	67	67	67	67	...	26	67	67	67	67	67	67	67	67	67

2 rows × 329 columns

Here, the Actual and Predicted values are same so it means model is built accurately.

Thanks

Limitations of this work and Scope for Future Work:-

- Used car industry- It is the huge industry and the data given is quite less but for getting a start and for the decision making the data is sufficient. In this industry, there are always open opportunities to get started, but There are different raw data available for real estate also and applying data science to it, will move the industry at a next level.

- We can get the understanding of past, present and the future market of used cars. Today in fast moving world this could be a great investment for business also if taken step at a suitable period.

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.

- Future Analysis : In future, machine learning model may merge 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. For better performance, we plan to design Deep Learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset and can filter or add data as per the costumer’s requirements/needs.

India will be the biggest market for the used cars, as used car market in India has been the center of attraction in the slow growing automotive industry in India as per the current scenarios.