PROJECT REPORT

ON

CAR PRICE PREDICTION

in partial fulfillment for the award of the degree of

MASTER'S IN SCIENCE(MS)

IN

COMPUTER SCIENCE (CS)

At

PACE UNIVERSITY

Submitted To:

Prof. Shubham Sawant

(Adjunct Faculty)

Submitted by:

Tejaswini Kandyala

Manognya Raygir

Nikhil Katakam

Drona Gangarapu

Table of Contents

1 Introduction

When it comes to car selection, the customers consider different brands and features, so it is surely a topic of concern for them, as it is not easy to determine a car which fits into their financial ability. Depending on various parameters, the price of a car is decided, for instance, the car price can vary based on various parameters, such as the year of production, car model, brand, its features, type of fuel, mileage and so on.

1.1 Problem Statement

Each individual desires to own a car and as there are various models and features, which are being upgraded with time. So, it is challenging to predict the prices of the cars in the market. In order to help resolved this problem, a dataset containing all the details of the cars with their prices is collected for the analysis.

1.2 Aim and Objective

This data aims to help the customer to purchase the cars based on their financial capability.

The objective is to predict the car prices, for accomplishing this objective the car prices dataset is analyzed using the machine learning models with the help of Python programming. This data analysis involves data exploration, missing value handling, data splitting, creation of various machine learning models such as logistic regression, decision tree, MLP neural network and random forest, and evaluation of the created models (Kiran et al., 2022).

2 Dataset Description

The selected dataset is titled as "Car price dataset", which includes car price prediction (Mohanty, 2021). This dataset is sourced from a popular website called Kaggle and the source link is as follows-https://www.kaggle.com/datasets/sidharth178/car-prices-dataset

The increase in the car types and their varieties, features, and capabilities, the challenges of predicting car price is also increasing, and becomes a hiccup for the customers who wish to purchase their own car that matches their budget and features that are best at the time.

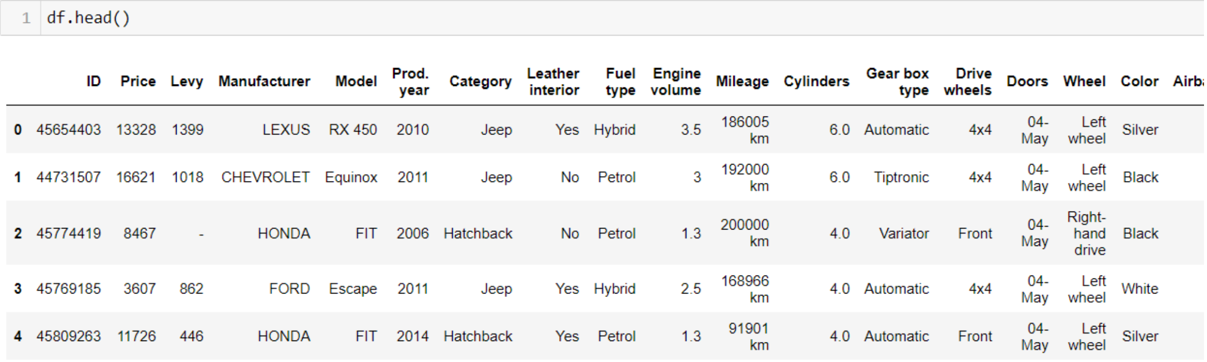
This dataset consists of the following variables - ID, Price: price of the car (Target Column), Levy, Manufacturer, Model, Prod. Year, Category, Leather interior, Fuel type, Engine volume, Mileage, Cylinders, Gear box type, Drive wheels, Doors, Wheel, Colour, and Airbags.

Using this dataset, machine learning models can be built to predict the car prices based on the car manufacturer, category, mileage, engine volume, cylinders, gear box type, doors, wheel, colour, etc. (Mohanty, 2021) (Maddali, 2022).

3 Exploratory Data Analysis (EDA)

This section demonstrates the exploratory data analysis on Python by following the below steps.

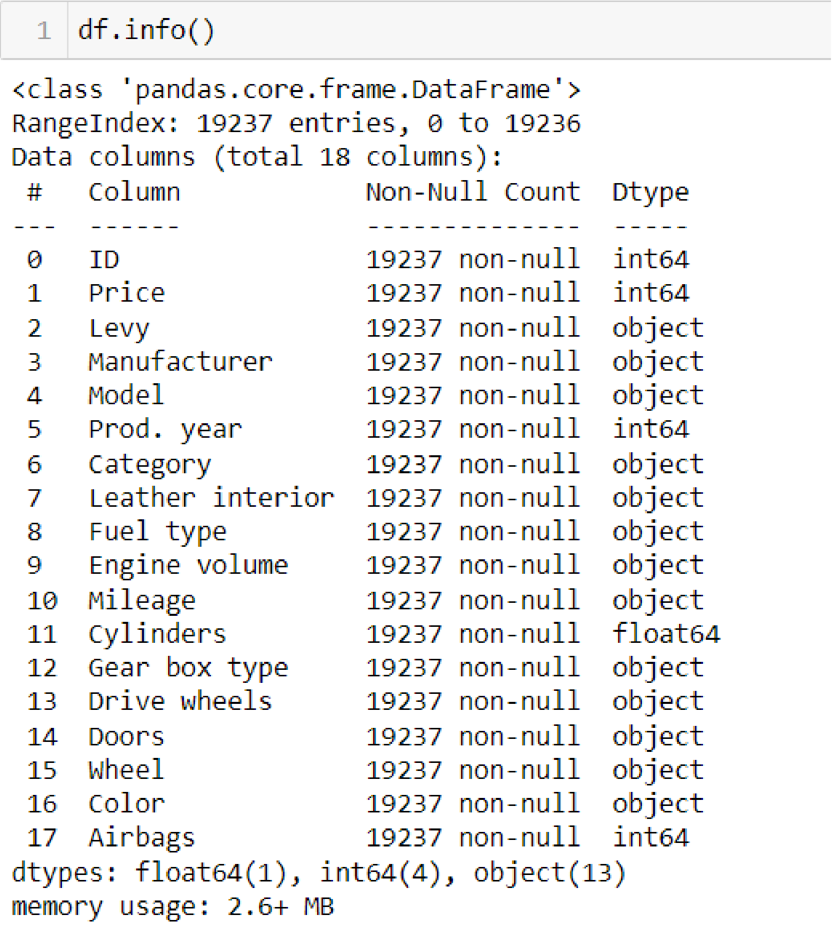
STEP 1: Start by dataset importing.



```
df.head()
```

	ID	Price	Levy	Manufacturer	Model	Prod. year	Category	Leather interior	Fuel type	Engine volume	Mileage	Cylinders	Gear box type	Drive wheels	Doors	Wheel	Color	Airb
0	45654403	13328	1399	LEXUS	RX 450	2010	Jeep	Yes	Hybrid	3.5	186005 km	6.0	Automatic	4x4	04-May	Left wheel	Silver	
1	44731507	16621	1018	CHEVROLET	Equinox	2011	Jeep	No	Petrol	3	192000 km	6.0	Tiptronic	4x4	04-May	Left wheel	Black	
2	45774419	8467	-	HONDA	FIT	2006	Hatchback	No	Petrol	1.3	200000 km	4.0	Variator	Front	04-May	Right-hand drive	Black	
3	45769185	3607	862	FORD	Escape	2011	Jeep	Yes	Hybrid	2.5	168966 km	4.0	Automatic	4x4	04-May	Left wheel	White	
4	45809263	11726	446	HONDA	FIT	2014	Hatchback	Yes	Petrol	1.3	91901 km	4.0	Automatic	Front	04-May	Left wheel	Silver	

The dataset information is displayed below.

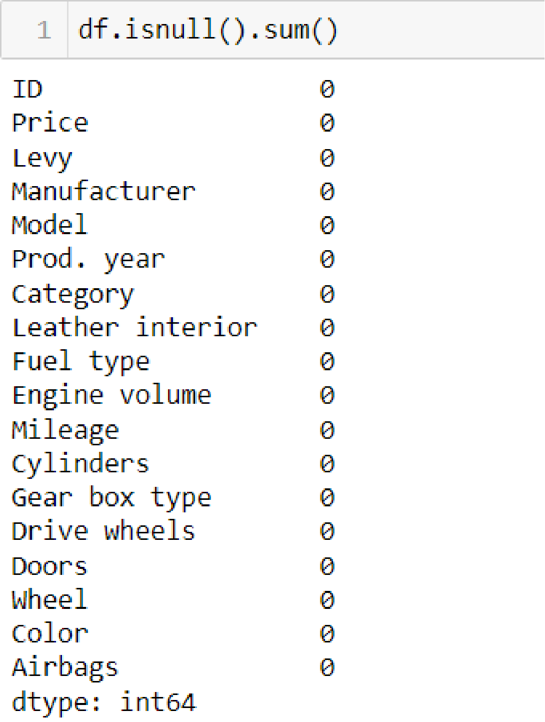


```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19237 entries, 0 to 19236
Data columns (total 18 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   ID                19237 non-null  int64  
 1   Price             19237 non-null  int64  
 2   Levy              19237 non-null  object 
 3   Manufacturer      19237 non-null  object 
 4   Model             19237 non-null  object 
 5   Prod. year        19237 non-null  int64  
 6   Category          19237 non-null  object 
 7   Leather interior  19237 non-null  object 
 8   Fuel type         19237 non-null  object 
 9   Engine volume     19237 non-null  object 
 10  Mileage           19237 non-null  object 
 11  Cylinders         19237 non-null  float64
 12  Gear box type     19237 non-null  object 
 13  Drive wheels      19237 non-null  object 
 14  Doors             19237 non-null  object 
 15  Wheel             19237 non-null  object 
 16  Color             19237 non-null  object 
 17  Airbags           19237 non-null  int64  
dtypes: float64(1), int64(4), object(13)
memory usage: 2.6+ MB
```

As per the above result, the given dataset has 19237 of observations with 19 columns of attributes. The price variable is considered as the target variable, and the other variables are considered as the response variables.

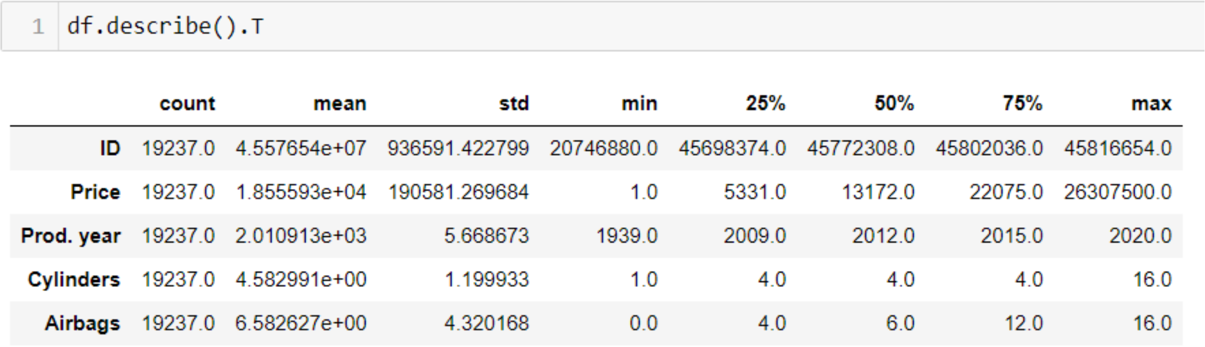
STEP 2: Check the missing values in the dataset, as demonstrated below. As per the below result, no missing values were found in this dataset.



```
df.isnull().sum()
```

```
ID                  0
Price               0
Levy                0
Manufacturer        0
Model               0
Prod. year          0
Category            0
Leather interior    0
Fuel type           0
Engine volume       0
Mileage             0
Cylinders           0
Gear box type       0
Drive wheels        0
Doors               0
Wheel               0
Color               0
Airbags             0
dtype: int64
```

STEP 3: Do basic statistical analysis to understand the dataset by using the describe function as represented below (Kiran et al., 2022).



```
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
ID	19237.0	4.557654e+07	936591.422799	20746880.0	45698374.0	45772308.0	45802036.0	45816654.0
Price	19237.0	1.855593e+04	190581.269684	1.0	5331.0	13172.0	22075.0	26307500.0
Prod. year	19237.0	2.010913e+03	5.668673	1939.0	2009.0	2012.0	2015.0	2020.0
Cylinders	19237.0	4.582991e+00	1.199933	1.0	4.0	4.0	4.0	16.0
Airbags	19237.0	6.582627e+00	4.320168	0.0	4.0	6.0	12.0	16.0

STEP 4: Visualize the car prices over the years.

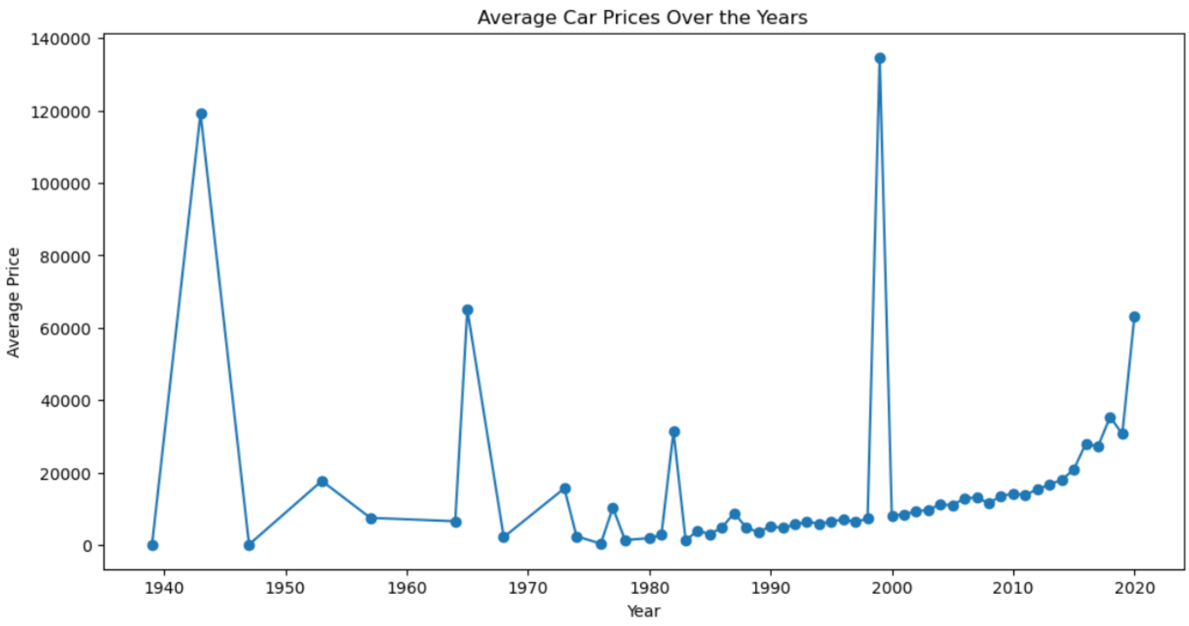


Figure 1: Average care prices over the years

As per the above visualization,

- The year 2000 had the highest car prices compared to other years.
- In contrast to other years, 1970 had the lowest car prices.

STEP 5: Visualize the car prices based on the manufacturer.

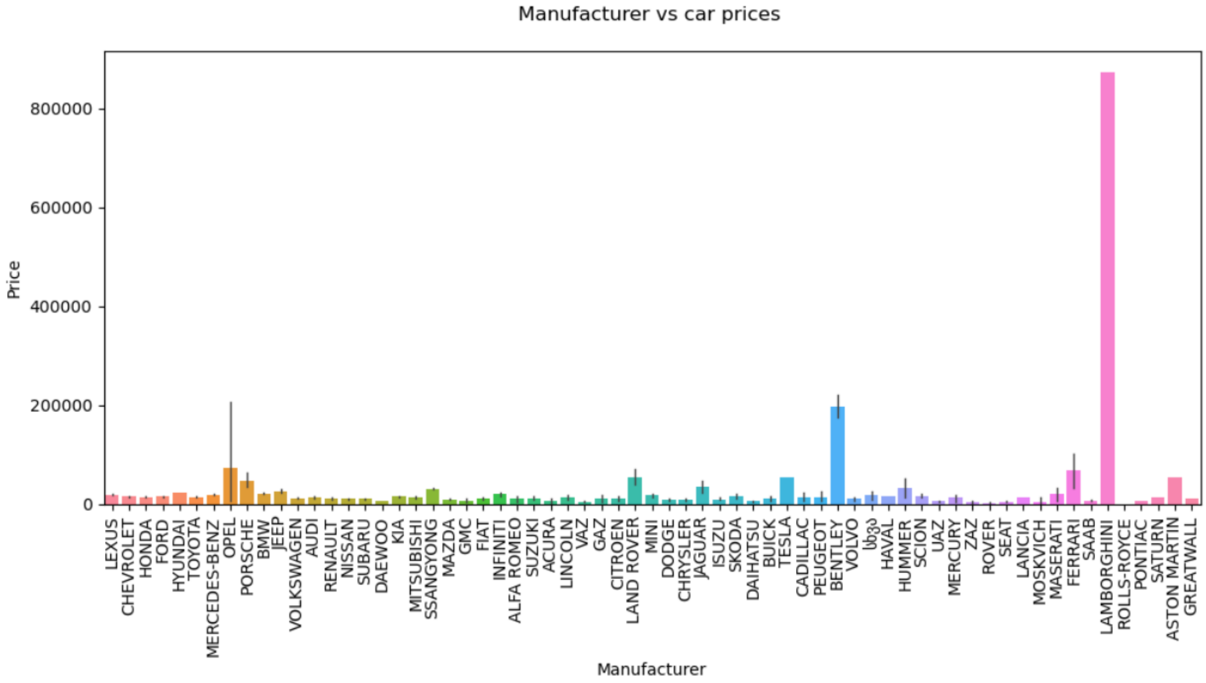


Figure 2: Manufacturer versus car prices

As per the visualization, Lamborghini manufacturer is shown to have the highest car price, whereas ROVER, ZAZ, and SEAT manufacturers are shown to have the lowest car prices compared to the other manufacturers.

STEP 6: Visualize the car prices based on the fuel type.

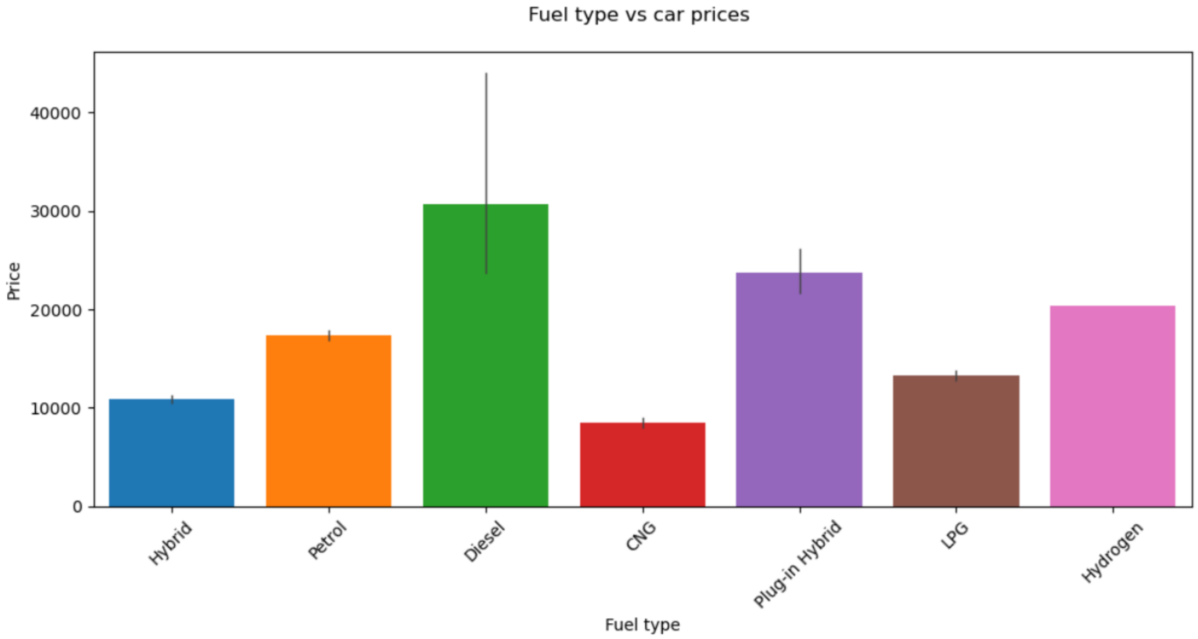


Figure 3: Fuel type versus car prices

As per the visualization, Diesel fuel type is observed to have the highest car price. But, CNG and hybrid fuel type are the lowest car prices than the other fuel types.

STEP 7: Visualize the car prices based on the car category.

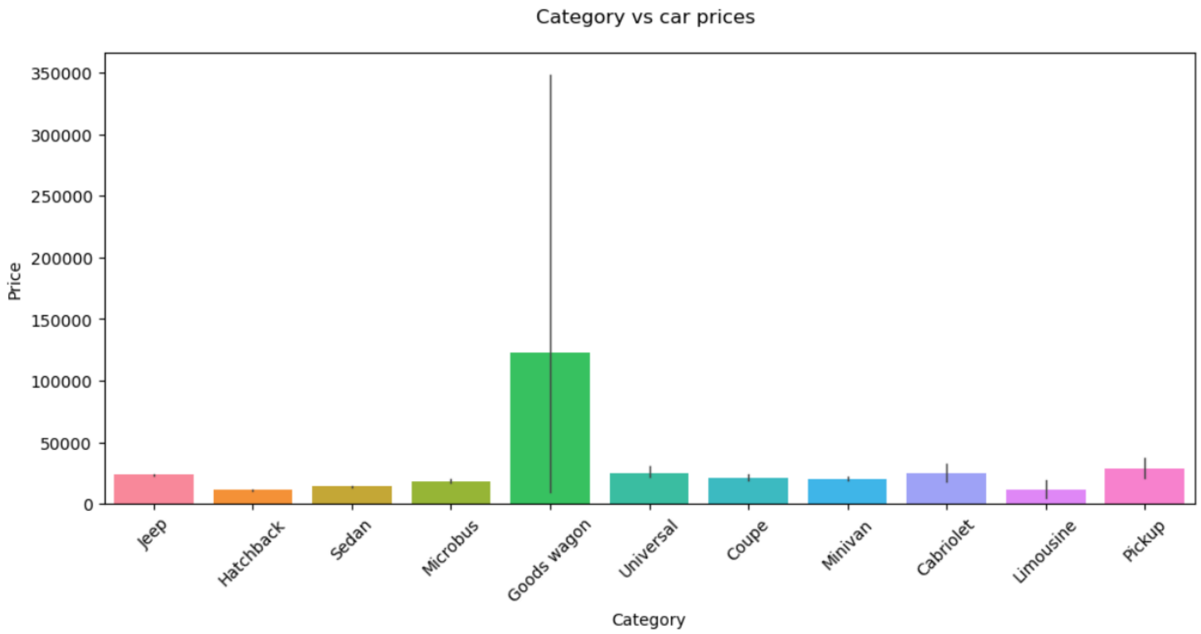


Figure 4: Category versus car prices

As per the visualization, Goods wagon category showed to have the highest car price, whereas the lowest ones are Hatchback, Limousine and Microbus.

STEP 8: Next comes, feature selection and dataset splitting into train and test sets. Before proceeding, cut the price variable into two categories such as below 30K and above 30K to help the user predict the prices based on the two categories like the cars above and below 30K. Then, set the target variable as “Price” and other variables are considered as response variables. Next, split the dataset into train and test set.

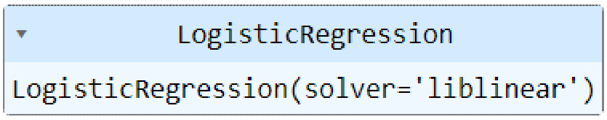
STEP 9: Finally, encode the categorical variables to build a machine learning model for predicting the car prices.

4 Machine Learning Model Development

This section is all about building a machine learning model to predict the car prices.

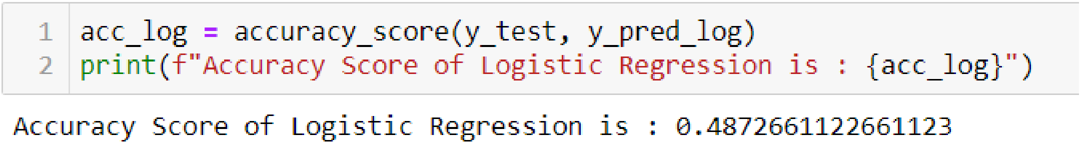
4.1 Logistic Regression

To begin with, create a logistic regression model with the help of sklearn libraries. To create logistic regression on Python, import the logistic regression from sklearn library and use lib solver parameter to build a logistic regression. Then, fit the created logistic regression model. The result as demonstrated below (Vermani, 2022).



```
▾ LogisticRegression
LogisticRegression(solver='liblinear')
```

Then, predict the created model and determine its accuracy to predict the car prices, as demonstrated below.

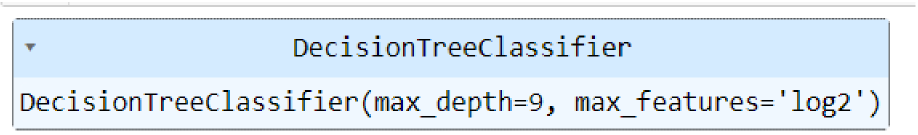


```
acc_log = accuracy_score(y_test, y_pred_log)
print(f"Accuracy Score of Logistic Regression is : {acc_log}")
```

```
Accuracy Score of Logistic Regression is : 0.4872661122661123
```

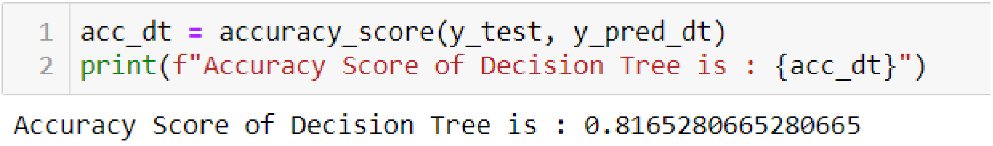
4.2 Decision Tree

The next model is decision tree by using same sklearn libraries. To create decision tree model on python, import the decision tree classifier from sklearn library and use 9 as the maximum depth, and maximum feature as "log2" parameters to build a decision tree classifier model. Then, fit the created decision tree classifier model. The result is shown below (Singh, 2023).



```
DecisionTreeClassifier
DecisionTreeClassifier(max_depth=9, max_features='log2')
```

Now, predict the created model and determine its accuracy to predict the car prices, as demonstrated below.

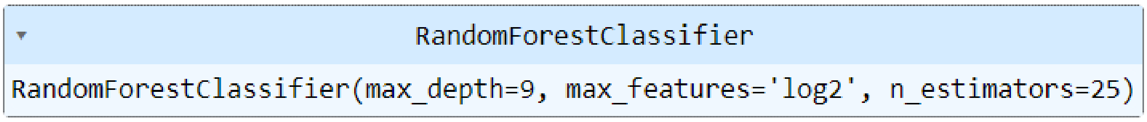


```
acc_dt = accuracy_score(y_test, y_pred_dt)
print(f"Accuracy Score of Decision Tree is : {acc_dt}")
```

```
Accuracy Score of Decision Tree is : 0.8165280665280665
```

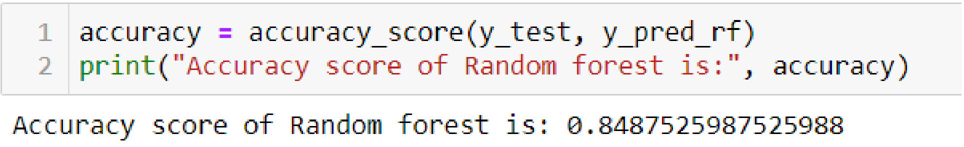
4.3 Random Forest

A random forest model is created using the sklearn libraries. To create random forest model on python, import the random forest classifier from sklearn library and use maximum depth as 9, maximum feature as "log2", and number of estimators as 25 parameters to build a random forest classifier model. Then, fit the created random forest classifier model. The result is demonstrated below.



```
RandomForestClassifier
RandomForestClassifier(max_depth=9, max_features='log2', n_estimators=25)
```

Like before, predict the created model and determine its accuracy to predict the car prices, as demonstrated below.



```python
accuracy = accuracy_score(y_test, y_pred_rf)
print("Accuracy score of Random forest is:", accuracy)
```

```
Accuracy score of Random forest is: 0.8487525987525988
```

5 Machine Learning Model Evaluation

This section is dedicated to evaluate the model by determining the accuracy and confusion matrix for each model. For logistic regression, the determined accuracy is 48.72%, which is less for predicting the car prices. The determined confusion matrix is depicted in the below represented matrix.

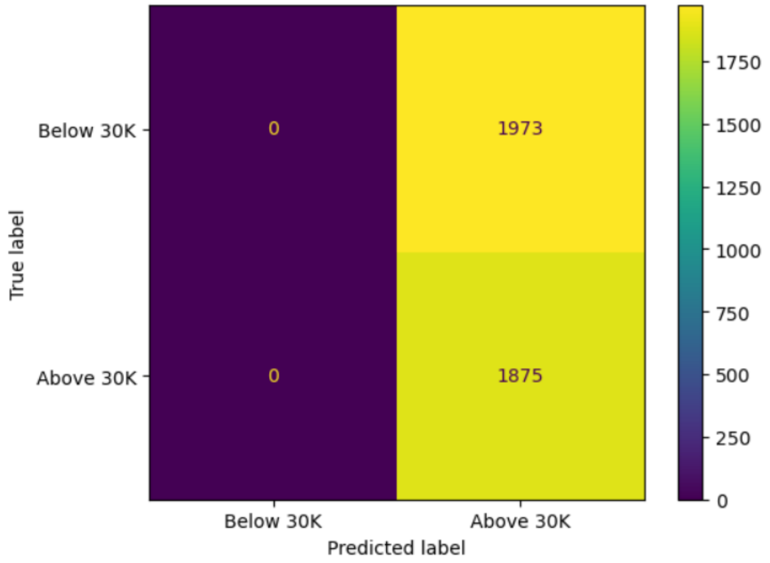


Figure 5: Confusion matrix of logistic regression

As per the figure 5, it correctly predicted that 0 cars are below 30K and 1973 cars are above 30K. And, no cars are above 30K, which is incorrectly predicted by the model, but 1875 cars are correctly predicted as above 30K.

For the decision tree classifier model, the determined accuracy is 81.65%, which is high for predicting the car prices.

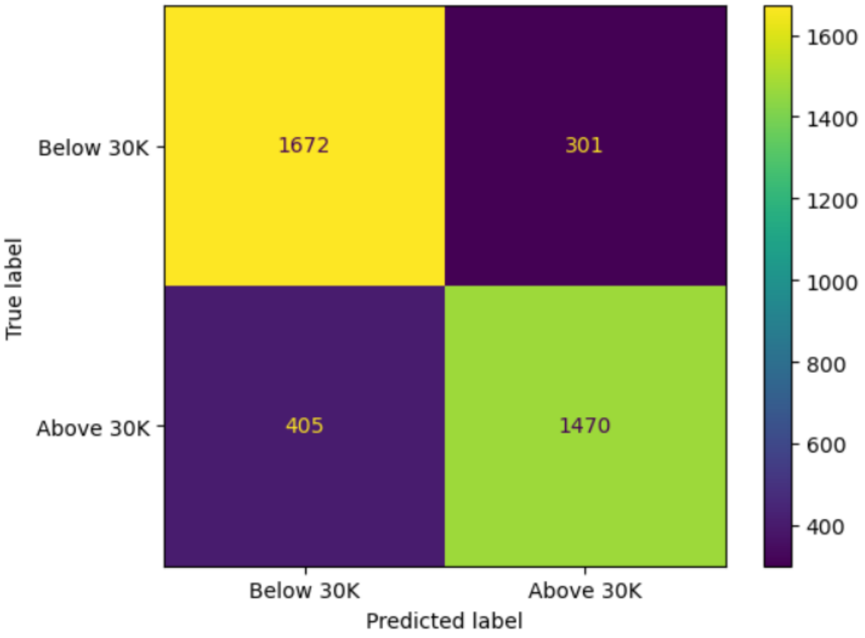


Figure 6: Confusion matrix of decision tree

As per the figure 6, 1672 cars are correctly predicted to be below 30K and 301 cars are above 30K, which was incorrectly predicted by the model. And there were 405 cars above 30K, which was incorrectly predicted by the model, but 1470 cars are correctly predicted to be above 30K. For random forest classifier model, the determined accuracy is 84.87%, which is high for predicting the car prices. The determined confusion matrix is demonstrated below.

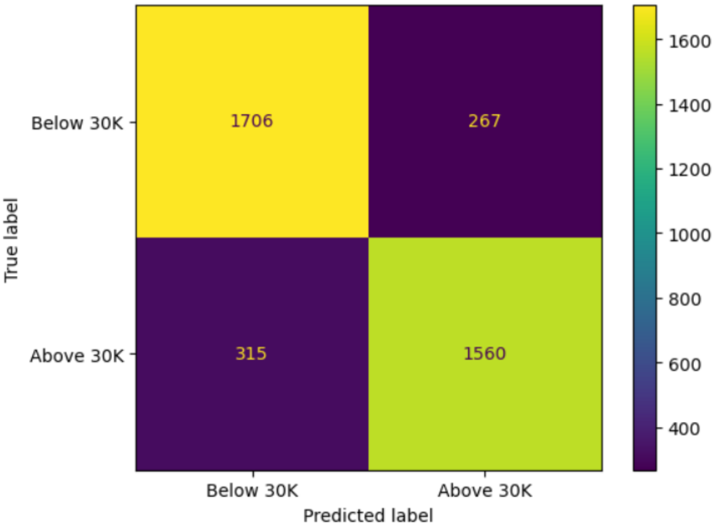


Figure 7: Confusion matrix of random forest

The figure 7 also correctly predicted 1706 cars are below 30K and 267 cars are above 30K, which was incorrectly predicted by the model. And, there were 315 cars above 30K, which was incorrectly predicted by the model, but 1560 cars are correctly predicted to be above 30K.

6 Conclusion

A dataset of car prices was considered for this assignment, which analyzed the selected car prices dataset by using machine learning models with the help of Python programming. From this report, it concludes that the random forest model is the best model for predicting car prices, because it showed high accuracy when compared with the other three machine learning models. Therefore, it correctly predicted that 1706 cars are below 30K and 1560 cars are correctly predicted to be above 30K.

References

Kiran, V.S. *et al.* (2022) 'USED CAR PRICE PREDICTION', *Journal of Engineering Sciences*, 13(06), pp. 387–398.

Maddali, S. (2022) *Predicting car prices using machine learning and Data Science*, *Medium*. Available at: https://medium.com/odscjournal/predicting-car-prices-using-machine-learning-and-data-science-52ed44abab1b (Accessed: 08 February 2024).

Mohanty, S.K. (2021) *Car prices dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/sidharth178/car-prices-dataset (Accessed: 08 February 2024).

Singh, A. (2023) *How: Implementing decision trees in python with Scikit-Learn (part 3)*, *Medium*. Available at: https://medium.com/@diehardankush/how-implementing-decision-trees-in-python-with-scikit-learn-part-3-29e5a787baaf (Accessed: 08 February 2024).

Vermani, G. (2022) *How to perform logistic regression in sklearn -*, *ProjectPro*. Available at: https://www.projectpro.io/recipes/perform-logistic-regression-sklearn (Accessed: 08 February 2024).