**PROPOSE CODE RESULT**

**Import Libraries:**  numpy, pandas, sklearn, matplotlib, seaborn,lightbgm etc.

# Dataset Name: Used Cars Price Prediction

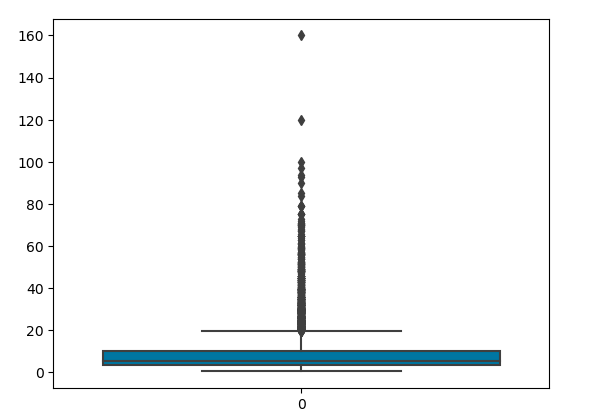
**Dataset Link:** <https://www.kaggle.com/datasets/avikasliwal/used-cars-price-prediction?select=train-data.csv>

**Data preprocessing:**

* We read our csv through “car\_price (1).csv”
* Then we have describe the data ,find the number of rows and columns then used to find the information of data
* Now, we have replace all the zero values of column ’seats’ with the median and drop the unnamed column.
* We use a lambda function taking x value applies to each value of the column mileage, Engine, Power and splits it using whitespace as delimiter
* After that we have replace the nan and null values of the column mileage, Engine, Power and convert it into float values, after that we have fill the null values with the median.
* We have draw the boxplots of the data to make it more clear
* At last, we have apply the Label Encoder to convert the categorical data into numerical data and provide the index to Owner\_Type column and use groupby for the location and price column
* After that we identify the outliers in the a and remove it
* Interquartile range (IQR) is calculated by inter\_qr as the difference between Q3 and Q1.
* The lower and upper boundaries for outlier detection (usually 1.5 times the IQR) are calculated by L\_boun and U\_boun using the IQR approach.
* The rows of the DataFrame df that fall within the determined bounds for the values of the 'Year' column are then kept.
* This code eliminates data points that the IQR technique identifies as outliers. Data points that fall outside of or outside of the bounds of L\_boun and U\_boun are considered outliers

**Data Visualization:**

**Boxplot**

****

**Model: -**

**Random Forest:** We have use the Random Forest model and Lgbm model to predict the R2score,mean squared error, root mean squared error and mean absolute error ,first of all we have divide the data into train and test after that we have fit the train and test and after that we have predict the data,first of all we have divide the data into 75% training and 25% testing and calculate the r2score,mse,rmse,and mae of the knn,then we have divide the data into 80% training and 20% testing and calculate the same ,at last we have divide data into 85% training and 15% testing .

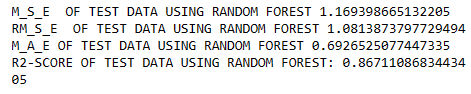
We have also apply the cross validation , fivefold and tenfold cross validation to predict the average score of the random forest and Lgbm model

**Testing Results:**

**1.Model-Random Forest**

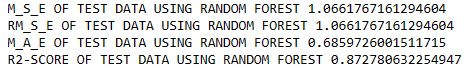
**For 25% Test Data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Forest | R2score | MSE | RMSE | MAE |
|  | 0.86 | 1.16 | 1.08 | 0.69 |

****

**For 20% Test Data**

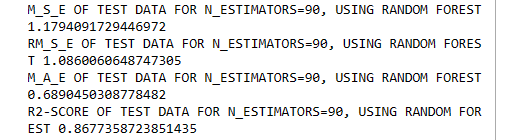
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Forest | R2score | MSE | RMSE | MAE |
|  | 0.87 | 1.06 | 1.06 | 0.68 |

****

**For 15% Test Data**

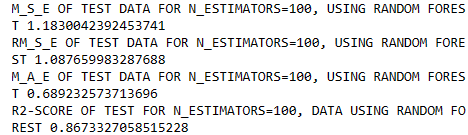
**1)for n-estimators=90**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Forest | R2score | MSE | RMSE | MAE |
|  | 0.8677 | 1.17 | 1.08 | 0.68 |

****

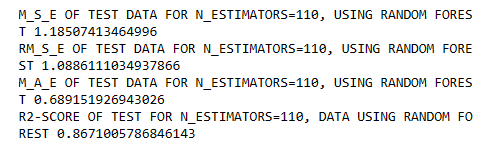
**2)for n-estimators=100**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Forest | R2score | MSE | RMSE | MAE |
|  | 0.86 | 1.18 | 1.08 | 0.68 |

****

**3)for n-estimators=110**

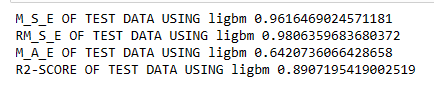
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Forest | R2score | MSE | RMSE | MAE |
|  | 0.867 | 1.185 | 1.088 | 0.689 |

****

**2.Model-Lgbm**

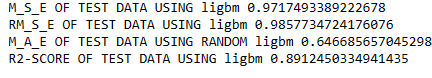
**For 25% Test data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LGBM | R2score | MSE | RMSE | MAE |
|  | 0.89 | 0.96 | 0.98 | 0.64 |

****

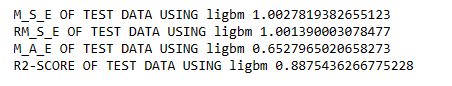
**For 20% Test data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LGBM | R2score | MSE | RMSE | MAE |
|  | 0.8912 | 0.97 | 0.98 | 0.6466 |

****

**For 15% Test data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LGBM | R2score | MSE | RMSE | MAE |
|  | 0.8875 | 1.00 | 1.00 | 0.65 |

****

**For 15% Test data**

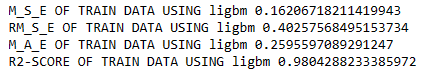
|  |  |  |
| --- | --- | --- |
| Cross validation | Models | Average score |
| 5 | Lgbm | 0.8759 |
| 10 | Lgbm | 0.8831 |
| 5 | Random Forest | 0.8685 |
| 10 | Random Forest | 0.8726 |

**Training Results:**

**1)Random Forest**

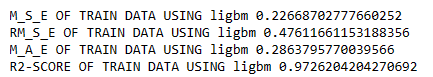
**For 25% Test Data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Forest | R2score | MSE | RMSE | MAE |
|  | 0.98 | 0.16 | 0.40 | 0.25 |

****

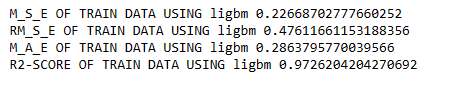
**For 20% Test Data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Forest | R2score | MSE | RMSE | MAE |
|  | 0.97 | 0.22 | 0.47 | 0.28 |

****

**For 15% Test Data**

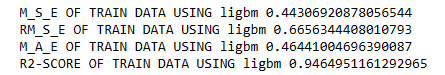
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Forest | R2score | MSE | RMSE | MAE |
|  | 0.97 | 0.22 | 0.47 | 0.28 |

****

**2) LGBM**

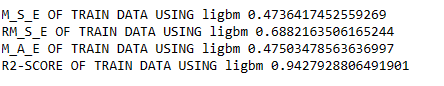
**For 25% Test Data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LGBM | R2score | MSE | RMSE | MAE |
|  | 0.94 | 0.44 | 0.66 | 0.46 |

****

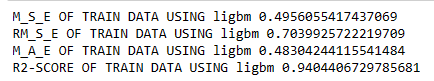
**For 20% Test Data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LGBM | R2score | MSE | RMSE | MAE |
|  | 0.94 | 0.47 | 0.68 | 0.47 |

****

**For 15% Test Data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LGBM | R2score | MSE | RMSE | MAE |
|  | 0.94 | 0.49 | 0.70 | 0.48 |

****