USED CAR PRICE/QUALITY CLASSIFICATION

-ISHA SHAH -PRINCE KHENI -TANAY DANGAICH -TARUN KAUSHIK

INTRODUCTION

People experience a hard row to hoe in purchasing desired featured second-hand cars within the targeted budget, and they play it by ear multiple times. But often, it is hard to decide whether a used car is fraudulent or not. To be able to predict used cars' market value can help both buyers and sellers. Using different machine learning models, we are trying to broadly categorize the cars into three categories: Cheap(0) | Average(1) | High(2). Our project aims at finding out which machine learning model will be the best fit that we can use to predict the used car price/quality.

DATASOURCE

Source for the dataset: https://www.kaggle.com/datasets/gagasrock/car-quality-prediction.This dataset is manually collected from observations. It consists of 12 independent variables i.e., Year, Make, Model, Condition, Transmission, Cylinders, Fuel, Odometer, Engine Power, and Mileage. Generally, the Price Range or Quality of the Car depends on these parameters. These parameters play a vital role in the classification of the quality of the Car. Total observations: 12076

OVERVIEW

A Rule of thumb is that customers aim to buy a low to average-priced car having desired features they have slept on for nights. We will first do the data preprocessing and check for any null/noisy data. After this, the next question is, are all attributes important in predicting the price range? For this, we will be using the Random Forest algorithm for feature subset selection. Next, we will be running machine learning models to predict car prices. We have selected four machine-learning models for this:

1) K-nearest neighbor 2) Decision trees 3) Cat Boost classifier 4) Light Gradient Boosting classifier.

Model selection will be evaluated based on precision, Recall, and the F1 score. We will also make the precision-recall curve and compare the area under the curve for all the models.

DATA PREPROCESSING

We dropped the rows on the following constraints:

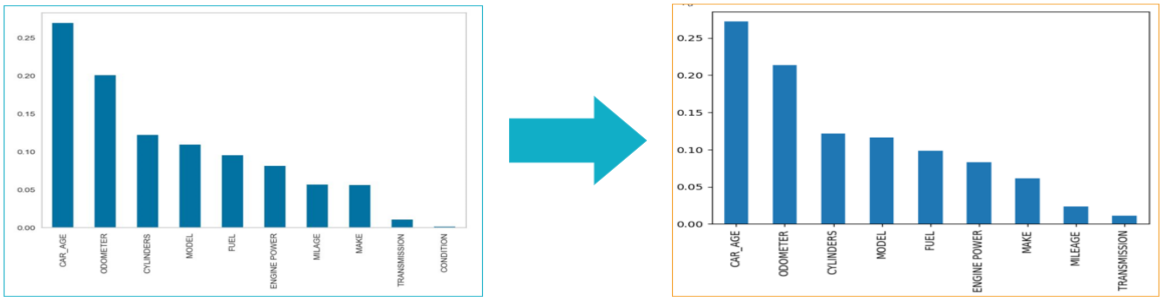
- If there were only 1 or 2 rows for a company.
- If the car age was very high or the car's odometer rating was very low.
- If the car age is > 2 and the odometer rating is < 100.

Total columns dropped: 85.

For the categorical data, we used the LabelEncoder() method to convert them into numerical values. The odometer reading was left-skewed; we normalized the data using the Box-Cox transformation and skew method.

FEATURE SELECTION

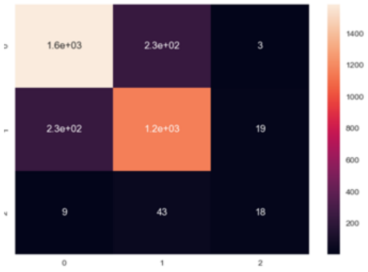
After data preprocessing, we split the data into 70% (training data) and 30% (test data).For feature selection, we used the Random Forest algorithm and dropped the ‘Condition’ feature as it was not showing any importance.



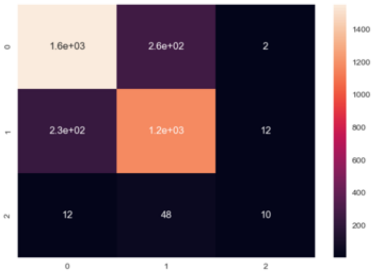
MODEL DEVELOPMENT

1) K NEAREST NEIGHBOUR

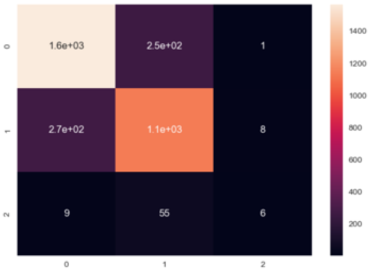
KNN helps to classify the new data points based on the similarity measure of the earlier stored data points.It can be developed for multiclass classifying dataset with comparative accuracy and precision.It is lazy learner which uses data with several classes to predict the classification of the new sample point. We have chosen K values as 3,5, and 7.



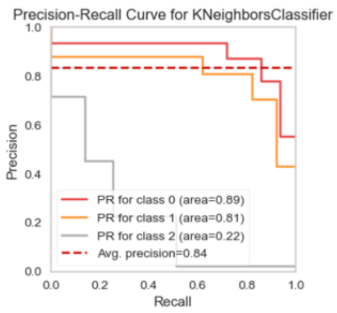
Confusion Matrix for K = 3

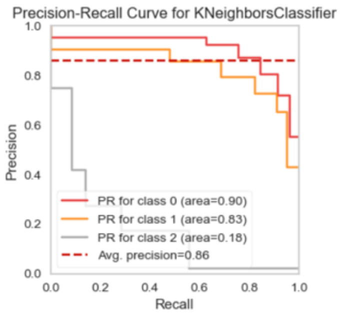


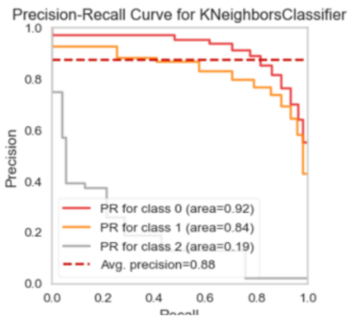
Confusion Matrix for K = 5



Confusion Matrix for K = 7





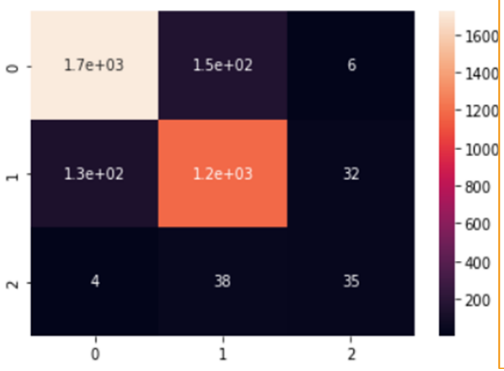


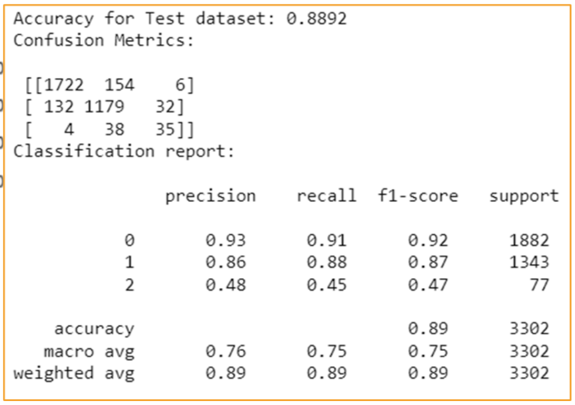
Precision – Recall Curve for K = 3 Precision – Recall Curve for K = 5 Precision – Recall Curve for K = 7

K = 3 has shown better results in terms of Accuracy and precision compared to other values of K. But AUC for KNN was quite low for ‘Class 2’.

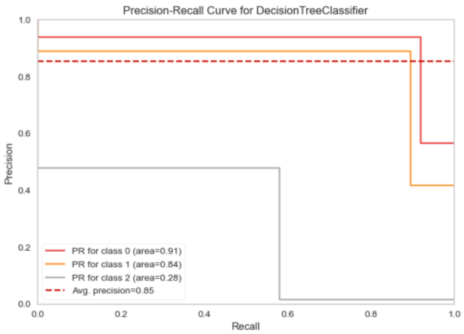
2) DECISION TREES

Decision trees easily visualize the prediction process in the flow chart format. They do not require feature scaling and can handle categorical features in the raw text format.While most models will suffer from missing values, decision trees are okay with them. Trees can provide the feature importance or how much each feature contributed to the model training results.





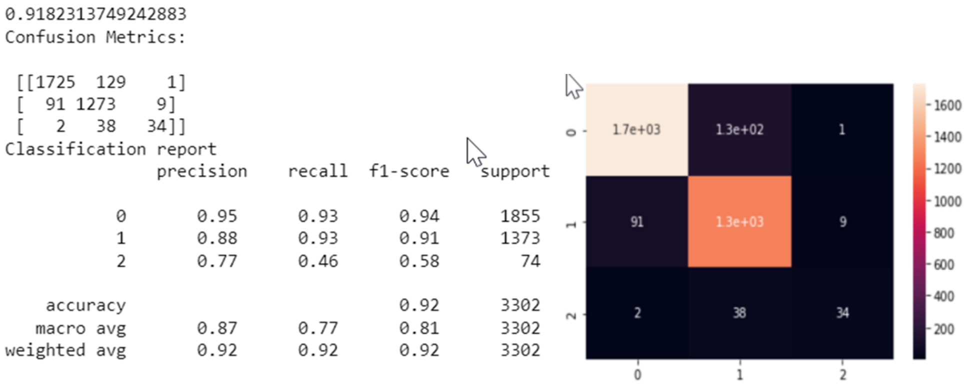
```
Accuracy for Test dataset: 0.8892
Confusion Metrics:

 [[1722  154    6]
 [ 132 1179   32]
 [   4   38   35]]
Classification report:

               precision    recall  f1-score   support

           0       0.93      0.91      0.92      1882
           1       0.86      0.88      0.87      1343
           2       0.48      0.45      0.47        77

    accuracy                           0.89      3302
   macro avg       0.76      0.75      0.75      3302
weighted avg       0.89      0.89      0.89      3302
```



Using Decision Tree Algorithms, we got an Accuracy of 88.92%. But AUC for this algorithm was low for ‘Class 2’, but better than KNN.

3) CAT BOOST CLASSIFIER

CatBoost is based on gradient boosted decision trees. During training, a set of decision trees is built consecutively. Each successive tree is built with reduced loss compared to the previous trees.Its training speed is fast and can handle categorical data really well.

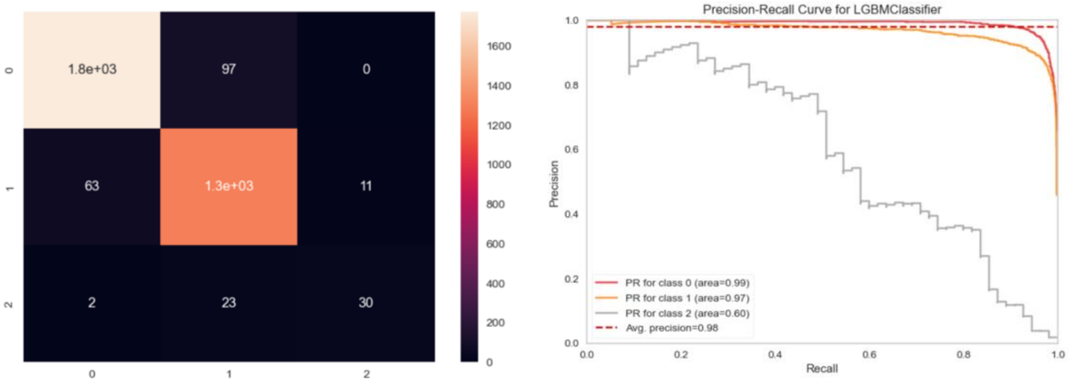


```
0.9182313749242883
Confusion Metrics:

 [[1725  129    1]
 [  91 1273    9]
 [   2   38   34]]
Classification report
              precision    recall  f1-score   support

           0       0.95      0.93      0.94      1855
           1       0.88      0.93      0.91      1373
           2       0.77      0.46      0.58        74

    accuracy                           0.92      3302
   macro avg       0.87      0.77      0.81      3302
weighted avg       0.92      0.92      0.92      3302
```

Using the Cat Boost Classifier, we got an Accuracy of 91.82%.

4) LIGHT GRADIENT BOOSTING CLASSIFIER

Light GBM grows trees vertically while another algorithm grows trees horizontally, meaning that Light GBM grows trees leaf-wise while another algorithm grows level-wise. It is one of the fastest and most efficient libraries for regression tasks.



Using the Light Gradient Boosting Classifier, we were able to classify quite well for all the classes. It generated better results than all other models.

DISCUSSION

1. While evaluating, we are taking the Average class as the True positives in the confusion matrix.
2. For the future scope of the project, we plan to implement deep learning network models and see if the model can be improved.
3. For better results, train on clusters of data rather than the whole dataset.

CONCLUSION

With this project's help, we can successfully classify used cars into three main categories: cheap, average, and high. Successful developments of different classification models have helped us to decide which model could be further used for a different dataset. Light Gradient Boost Classifier has given us maximum accuracy in the classification of cars, which can significantly impact the buyer in the market for a new, used car. There is room for further optimizing the model and trying deep learning models to check which would give the most accurate classifications.