Report: Car Class identification Project

Prepared by: Afees A A
Date:27-May-2023

Introduction

This project identifies the car class using classification models.. The dataset, 'cars_class.csv', has 719 samples and 18 numerical features. Our goal is to build an accurate classification model for car class identification.

To achieve this, we preprocess the data and handle outlier values. Different machine learning techniques like logistic regression and random forest are explored. Model performance is optimized through parameter tuning.

The 'final_model' is evaluated using metrics such as Accuracy, Confusion matrix and f1-score. Feature importance analysis is conducted to identify key predictors of car prices.

Through this project, we aim to develop an accurate classification model for car class identification and gain insights into important factors.

Data Loading and Preprocessing:

Initially, I imported the necessary libraries required for our project and then loaded our data set “cars_class.csv”. I have used only google colab for coding purposes.I have checked the dimensions of the data and conducted a descriptive statistical analysis.

There were no null or duplicate values in the data set. So I did not alter any values in the data.
Furthermore, I have dropped the Car ID feature from the data set, since it was not adding any value to our data and it may affect the final result.

Data Cleaning and Exploratory Data Analysis

I have visualized a box plot with all the features to identify the outlier and spread of values. It was found that there are a lot of outlier data. I split the data into 2 sets, to manage and replace the outlier values. Then I identified the outlier values and then replaced them by median.
Further, I replotted the box plot to recheck the outlier spread and confirmed that the data was cleaned. I then proceeded to visualize the entire features using scatter matrix for detailed insights.

Feature Engineering and Data Splitting

I selected features and labels for model testing and training. Then I have splitted the data into training and testing sets at 80 and 20% respectively. Furthermore, I applied feature scaling technique to both test and training data.

Machine Learning Techniques

I have performed various machine learning algorithms on the data. These include Logistic Regression, Decision Tree classification, Random Forest,K-NN, SVM, Kernel SVM and Naive Bayes models . I compared the performances of all the models using accuracy, Confusion matrix and f-1 Score values. It was found that the highest performance was given by Decision tree classification Model with around 79.8% accuracy. However, Kernel SVM and Random Forest Model performance were also good.
So we can use any of these three Models for predicting the price of a car in the future.

Model Tuning and Building 'final_model'

I have selected the Decision Tree classification Model as the final model and performed model tuning action using GridSearchCV function. It was observed that the accuracy and f1 scores remained unaffected. I then tried the possibility of ensembling technique using a random forest classifier. This resulted in improved accuracy with around 80.5%. We further explored the possibility of hyper tuning the parameters but did not give any positive results. so , we discarded it and selected the assembled model as our final model in identifying the class of a car.

Results and Interpretation

It was observed that the ensemble Random forest classifier performed the best compared to the other models. With tuning the parameters, it was identified that the performance was unaffected.

```
Accuracy: 0.8055555555555556

Confusion Matrix:
[[38  0  0  1]
 [ 0 18 11  3]
 [ 0 11 21  0]
 [ 1  1  0 39]]

Classification Report:
          precision    recall  f1-score   support

       0       0.97      0.97      0.97        39
       1       0.60      0.56      0.58        32
       2       0.66      0.66      0.66        32
       3       0.91      0.95      0.93        41

    accuracy                           0.81       144
   macro avg       0.78      0.79      0.78       144
weighted avg       0.80      0.81      0.80       144
```

Conclusion

- First we did the Basic Understanding of Data
- Then we performed Data Cleaning to make the raw data more usable while analyzing.
- Then we performed Exploratory Data Analysis to generate insights from the data.
- Then we performed Data Preprocessing to make data suitable for model training & testing.
- Then we trained our model using different Machine Learning Algorithms.
- In the end we came with 79.8% accuracy which was given by Decision Tree model.The accuracy of the model in predicting the car class is measured with accuracy, f-1 score.
- The model was further tuned to perform better. We have 80% accuracy with ensembling technique performed in a random forest algorithm. We tried to hyper tune the parameters but did not give any positive results.
- So we can use this Random Forest classification as Final model for identifying the class of a car in future.