Car Class Classification

1.Objective:

To build a car class classification model using machine learning algorithm. The aim is to maximize the accuracy of the model. The dependent variable in the given dataset is of multi class variable.

2.Data collection and Data Pre-processing:

- The given car class dataset consists of 719 rows and 20 columns.
- Data Cleaning: Find the missing values, outliers and inconsistencies in the data and should cleaned before giving to the model.
- **Exploratory Data Analysis (EDA):** Understand the characteristics, distributions and relationships among variables to get insights about the data. This was done through visualisation of the data through several plotting methods like boxplot, count plot, bar plot, scatter plot etc.

3.Feature Engineering:

- There 20 columns in the dataset it is very important to find the relevant features from the dataset. The features which have significant impact on car prices based on domain knowledge or statistical analysis are selected.
- **Feature transformation**: In this model, the Minmax Scaler to rescale the data was used. The Minmax scaler rescales the data in the range of 0 to 1.
- **Feature Extraction**: The Principal component analysis is used for the feature extraction. The most important features are extracted out. The explained variance is the percentage of variance that is attributed to each of the selected components.

4.Splitting the dataset:

- Divide the prepared dataset into training and testing data. Here in this model, I have used split 80-20 rule.

5.Model Selection and Training:

- In our dataset, the target variable is of categorical data. Hence, classification algorithms were used. The model is trained with KNN Classifier, Gradient Boosting Classifier, SGD Classifier, Random Forest Classifier, Decision Tree Classifier and Support Vector Classifier.
- Then, train the model using the Training dataset.

6.Model Evaluation and Fine-tuning:

- Evaluate the training model using the testing dataset to assess its performance
- The metrics used to evaluate the model were Accuracy, F1 score and confusion matrix. The aim is to maximise the accuracy of the model.
- The high performing model is further tuned by adjusting the hyperparameters to improve the model's performance.
- Random Forest Classifier was found to be the high performing model.

7.Prediction and Interpretation:

- The models such as Random Forest Classifier, Gradient Boosting Classifier, SGD Classifier are the outperformed individual models
- After hyperparameter tuning, the Random Forest classifier gave the highest accuracy of 81%.

Final Model: Random Forest Classifier

From the analysis, I suggest that Random Forest Classifier model can be used by car Rental companies, insurance companies, Automotive companies and many more users.

Language and libraries used:

Python - Language

Pandas - Data exploration

NumPy - for performing mathematical operations and for numerical computing

Matplotlib – Data Visualisation

Scikit-learn – Machine learning tasks.