Capstone Project Submission

Instructions:

i) Please fill in all the required information.
ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1. Harshit Kumar - harshitgangwar427@gmail.com

Contribution –

i. Preview Data
ii. Check total number of entries and column types
iii. Check the null values
iv. Plot distribution of numeric data
v. Plot distribution of categorical data
vi. Remove the outliers
vii. Project Summary and Team Colab Building
viii. Confusion matrix, precision and recall
ix. Building the model
Logistic regression
XgBOOST
Random forest

2. Harshit Kumar – chauhanh8439@gmail.com

Contribution –

i. Data Cleaning
ii. EDA
iii. Feature engineering and Feature selection
iv. Distribution check for dependent and independent features numeric data
v. Outlier detection and elimination
vi. Confusion matrix
vii. PPT and Team Colab Building Contribution
viii. Building and evaluating the model
ix. Conclusion

Please paste the GitHub Repo link.

Github Link:- https://github.com/HarshitKumar-git/Credit-Card-Default-Prediction

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

This project is aimed at predicting the case of customers default payments in Taiwan. K-S chart can be used to evaluate which customers will default on their credit card payments.

The purpose of this project is to conduct quantitative analysis on credit card default risk by using interpretable machine learning models with accessible customer data, instead of credit score or credit history, with the goal of assisting and speeding up the human decision making process.

First of all we have done some exploratory data analysis to understand about the data. From the modeling, we are able to classify default risk with accessible customer data and find a decent model. using a Logistic Regression classifier, From all baseline model, Random Forest classifier shows highest test accuracy and F1 score and AUC. Baseline model of Random Forest and decision tree shows huge difference in train and test accuracy which shows overfitting. After cross validation and hyperparameter tuning, XG Boost shows highest test accuracy score of 87% and AUC is 0.874. Cross validation and hyperparameter tuning certainly reduces chances of overfitting and also increases performance of model.

This analysis uses 6 classification models - Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Machine, Gradient Boosting and XG Boosting. Since Random Forest and XGBoost are tree based on algorithms, rescaling is only performed on Logistic Regression, not on these 2 models. For each model, we first try the model's default parameters, train each model without SMOTE and with SMOTE samplings. Then tune each model's hyperparameters to find the optimal performance. As mentioned earlier, this dataset has imbalanced classes, therefore we use precision and recall, instead of accuracy as the performance metrics.

If the balance of recall and precision is the most important metric, then Random Forest is the ideal model.