Credit Card Default Prediction

Candidate Name: Sourav Chowdhury
Email: sourav.20497@gmail.com

Contributor Roles:

- Importing libraries
- Dataset Inspection
- EDA :
    1. Handling Missing Data
    2. Univariate Analysis
    3. Analysis on Limit Balance
    4. Analysis on Gender, Education and Marriage
    5. Analysis on Age
    6. Analysis on Payment History
    7. Correlation Matrix
- One Hot Encoding
- Handling Class Imbalance : Using SMOTE
- Data Transformation
- Train Test Splitting
- Model Implementation:
    1. Logistic Regression
    2. Decision Tree Classification
    3. Random Forest Classification
    4. Gradient Boosting
    5. XG Boosting
    6. Support Vector Machines
- Feature Importance On Random Forest Baseline Model
- Feature Importance On XG Boost Optimal Model

Credit Card Default Prediction is a Classification based Credit Default prediction project for the banking and financial sectors, the model is build on the data provided consisting of the case of customers default payments in Taiwan. The data contains 3000 non null entries with 25 columns.

As the first step, perform date wrangling over the raw data and select those columns which are important for analysis. The dataset was highly imbalanced, To handle Class Imbalance we used SMOTE(Synthetic Minority Oversampling Technique).
Once the data is prepared the machine learning algorithms were implemented they are Logistic Regression, Decision Tree Classifier, Random Forest classifier, Gradient Boost Classifier, XG Boost classifier, Support Vector Machines. After analysis on every algorithm the Evaluation metrics was calculated for each one of them with best parameters tuned using hyperparameter Tuning.

In the Analysis the more Focus was on the accuracy, F1 Score and ROC Score metrics. Tried to reach maximum value of it. And also the confusion matrix were Plotted for each of the Algorithm .for best models, feature importance graphs were drawn.

At the last a table of Model Comparison Summary Table for Baseline as well as Optimal Model was made. Where the evaluation parameter values were mentioned. Following are the Tables.

BaseLine Model Summary

SL NO.	Classifier	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	ROC Score
1	Logistic Regression	0.875	0.875	0.803	0.939	0.866	0.883
2	Decision Tree	1.0	0.811	0.826	0.802	0.814	0.811
3	Random Forest	0.999	0.877	0.834	0.912	0.871	0.879
4	Gradient Boost	0.871	0.868	0.802	0.923	0.859	0.874
5	XG Boost	0.87	0.867	0.801	0.924	0.858	0.874
5	Support Vector Machines	0.879	0.876	0.805	0.938	0.866	0.884

Optimal Model Summary

SL NO	MODEL_NAME	Test MSE	Test RMSE	Test R^2	Test Adjusted R^2
1	Linear Regression	169871.70934024057	412.15495792267325	0.5624656403341544	0.5584307413401177
2	Lasso Regression	169871.72597325977	412.1549781007865	0.5624655974928983	0.5584306981037839
3	Ridge Regression	169871.77465071727	412.1550371531534	0.5624654721155751	0.5584305715702432
4	ElasticNet Regression	183719.51037262668	428.6251396880807	0.5267982017652659	0.5224343811475394
5	DecisionTree Regressor	91626.47632392391	302.6986559664973	0.7639999514779179	0.7618235821543713
6	RandomForest Regressor	52363.04539891553	228.8297301464902	0.8651297992599828	0.863886039483706
7	Gradient Boost	67959.97028150507	260.69133142761973	0.8249571856578451	0.823342957975151
8	Xg Boost	68250.19807949613	261.2473886558412	0.8242096530978654	0.8225885317431483

The best Algorithm found to be XG Boost shows highest test accuracy score of 88.10% and ROC Score is 0.884.

GitHub link: https://github.com/SrvPioneer/Supervised-ML-Classification-Credit-Card-Default-Prediction-

Drive Link:
https://drive.google.com/drive/u/1/folders/1FleP-wIpcxhvVObJRfRo2QJrqbGng3Wj