```
"""
Maria Teresa Luna Plascencia
Class: CS 677
Date: 04/23/2024
Final Project
"""
```

Projet Name: Analyzing various ML models to treat unbalanced datasets for loan approval predictor

Data Science Problem

This project aims to construct a predictive model for loan approval or denial by utilizing a range of features including income, loan amount, loan duration, credit history, and personal attributes such as gender, education level, and number of family members. One notable characteristic of the dataset is its imbalance, with a greater number of approved loans compared to denials.

The dataset

The target variable in this analysis is the loan status, where approved loans are represented by the value 1, and 0 otherwise.

The dataset exhibits certain unbalance, with a total number of observations equals to 381 with 271 approved vs 110 rejected loans. Approximately 71% of the information represents approved loans.

Reference: https://www.kaggle.com/datasets/bhavikjikadara/loan-status-prediction

Pre-processing

Perform EDA to understand dataset patterns and issues. Encode categorical variables for model compatibility. Prioritize completing Credit History's missing values, leveraging loan status for inference and improved prediction accuracy.

Correlation Matrix

The strongest linear relationship was observed between Credit History and Status Loan. Missing values for datapoints were updated accordingly.

Metrics to analyze

Evaluation of supervised ML models will focus on accuracy and F1 score, offering insights into performance for approved and rejected loans. Accuracy measures overall correctness, while F1 score balances precision and recall, crucial for imbalanced datasets like loan approvals. These metrics provide a comprehensive understanding of model predictions.

Models to evaluate

Random Forest Classifier – Identification of optimal hyperparameters and evaluation.

Balanced Random Forest Classifier – Identification of optimal hyperparameters and evaluation.

KNeighborsClassifier – Identification of optimal hyperparameters and evaluation.

Logistic Regression – Evalluation.

Linear SVC (Simple Vector Machine) - Evalluation

Voting Classifier – Evalluation.

Results

Model	TP	FP	TN	FN	Accuracy	TPR	TNR	F1[0]	F1[1]
RandomForestClassifier	81	15	18	1	0.861	0.988	0.545	0.692	0.91
BalancedRandomForestClassifier	**81**	**14**	**19**	**1**	**0.87**	**0.988**	**0.576**	**0.717**	**0.915**
KNeighborsClassifier	81	15	18	1	0.861	0.988	0.545	0.692	0.91
LogisticRegression	80	15	18	2	0.852	0.976	0.545	0.679	0.904
LinearSVC	81	15	18	1	0.861	0.988	0.545	0.692	0.91
VotingClassifier	81	15	18	1	0.861	0.988	0.545	0.692	0.91
DecisionTreeClassifier	65	7	26	17	0.791	0.793	0.788	0.684	0.844

Conclusions

Best Performing Models: Random Forest (both balanced and regular), Linear SVC, and Voting Classifier displayed the highest accuracy and F1 score, indicating their effectiveness in capturing data patterns and making accurate predictions.

Balanced Random Forest Advantage: The Balanced Random Forest exhibited slightly better metrics for rejected loans compared to the regular Random Forest. This suggests that balancing techniques improve minority class classification, enhancing performance.

Lower Performance Models: Decision Tree's lower accuracy and F1 score may be due to its limitations in capturing complex data patterns effectively.

Limited Improvement with Voting Classifier: The Voting Classifier did not significantly enhance metrics, suggesting individual models performed well on their own.

Unbalanced Dataset Challenges: The low F1 score for rejected loans and high F1 score for approved loans highlight challenges with dataset imbalance. Biased models may favor accuracy on the majority class, necessitating exploration of sampling techniques or acquiring more rejected loan data.

Importance of EDA: Exploratory Data Analysis (EDA) is crucial for addressing missing values, outliers, and incorrect inputs. EDA enhances data quality, ensures model reliability, and guides informed decisions based on accurate data.