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

Maria Teresa Luna
CS677 - Data Science with Python
Spring 2 Term 2024 / April-2024

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.

- A diverse set of supervised machine learning models will be assessed to determine which ones perform best in terms of accuracy and F1 score. Considering the dataset's imbalance, metrics like F1 score will be particularly relevant for evaluation.

The dataset

- This loan predictor dataset contains a set of features used to evaluate a prospect for a loan, based on variables such as gender, education, income, co-applicant income, credit history.

- 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

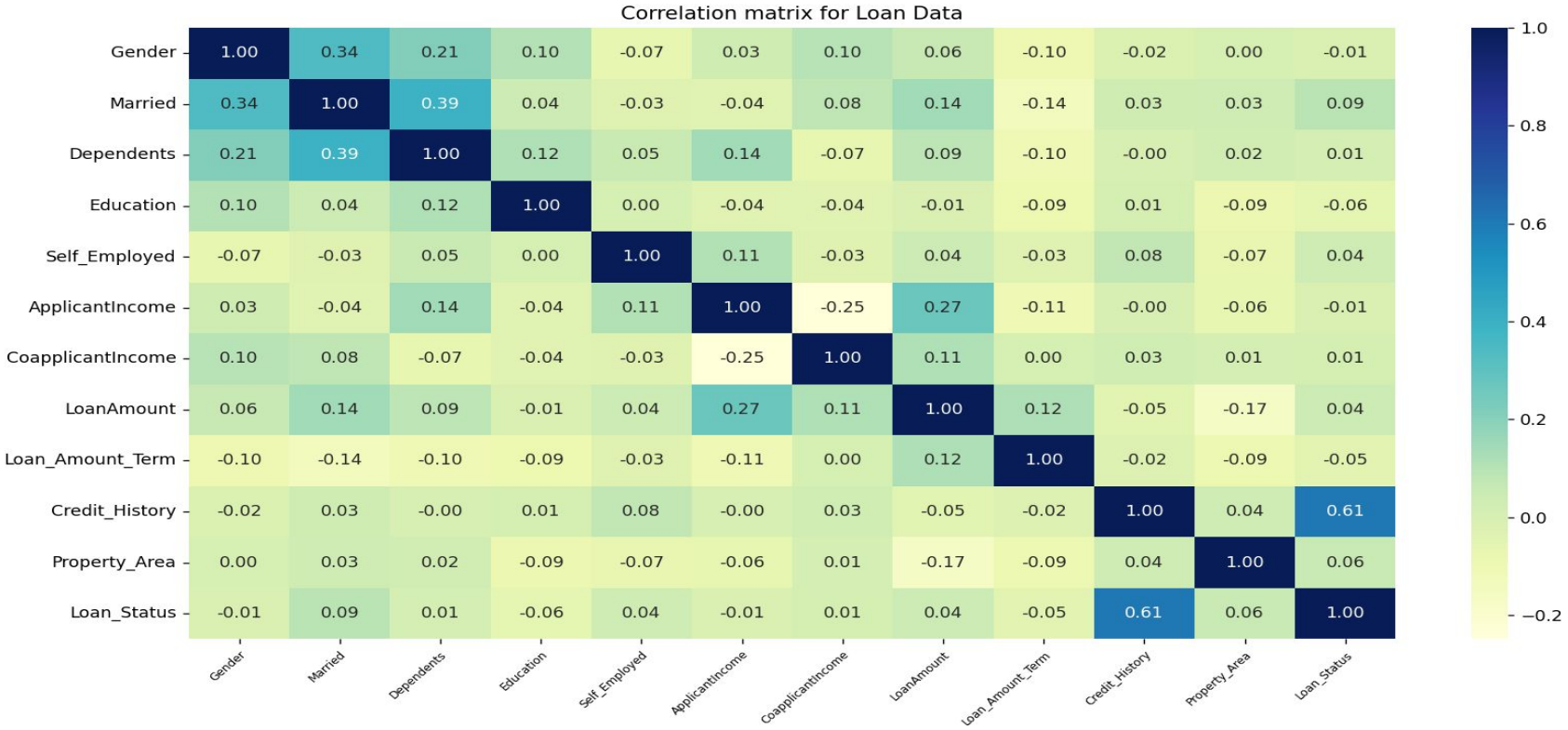
Pre-processing

Run an EDA (Exploratory Data Analysis) on the dataset to better understand the dataset characteristics, patterns, and potential issues.

Encode categorical variables.

Complete missing values, especially for the Credit History feature, is essential as it shows a strong relationship with loan approval or rejection. The missing data can be inferred based on the loan status, as it provides valuable information for predicting the missing Credit History values.

Pre-processing - Correlation matrix



The only strong linear relationship between variables is between Credit History and Loan Status. Based on the correlation matrix, this is the strongest relationship, so pre-processing will pay special attention to any missing values in the dataset for this feature.

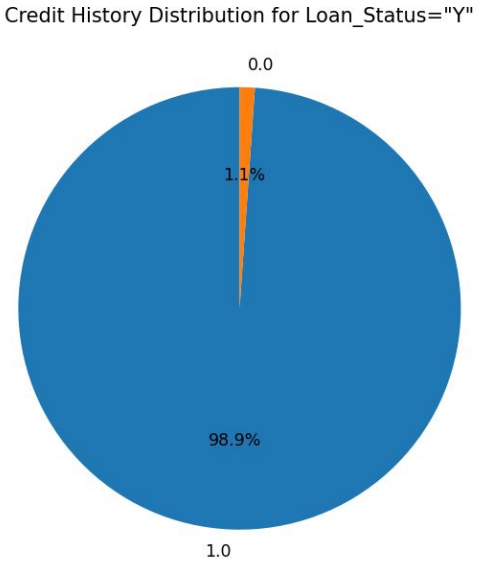
Filling missing data for Credit History feature

Missing data	
Gender	5
Married	0
Dependents	8
Education	0
Self_Employed	21
ApplicantIncome	0
CoapplicantIncome	0
LoanAmount	0
Loan_Amount_Term	11
Credit_History	30
Property_Area	0
Loan_Status	0

Interesting observation: When filling missing values with a 0 for this feature, the model's accuracy reached 0.75.

Based on the distribution plot, the optimal value for this feature appears to be 1, as 4 out of 30 observations with missing values were associated with a loan status of 1.

All the results presented here were obtained assuming a value of 1 for the Credit History feature for the missing data points.

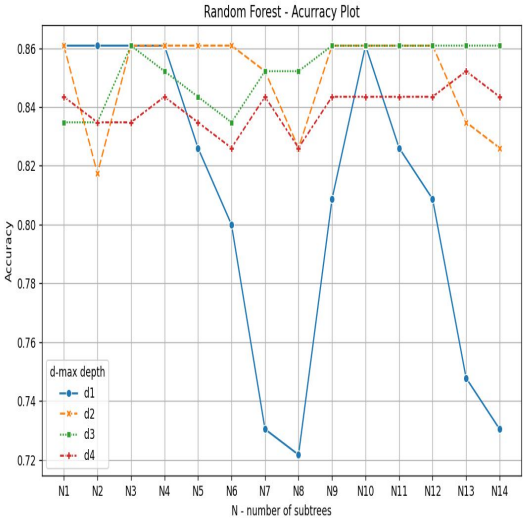


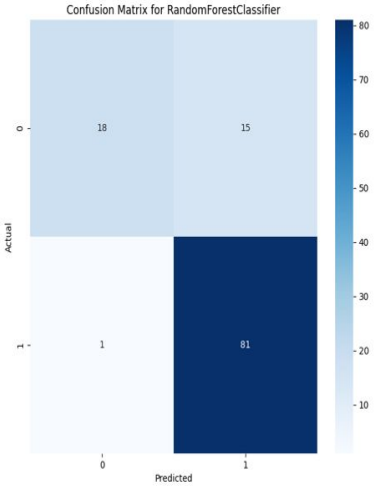
Models to evaluate

- Random Forest Classifier
- Balanced Random Forest Classifier
- KNeighborsClassifier
- Logistic Regression
- Linear SVC (Simple Vector Machine)
- Voting Classifier

Random Forest Classifier

- Finding the best hyperparameters, specifically number of trees, and max depth of the trees.
- Optimal number of trees: 1
- Max depth: 1

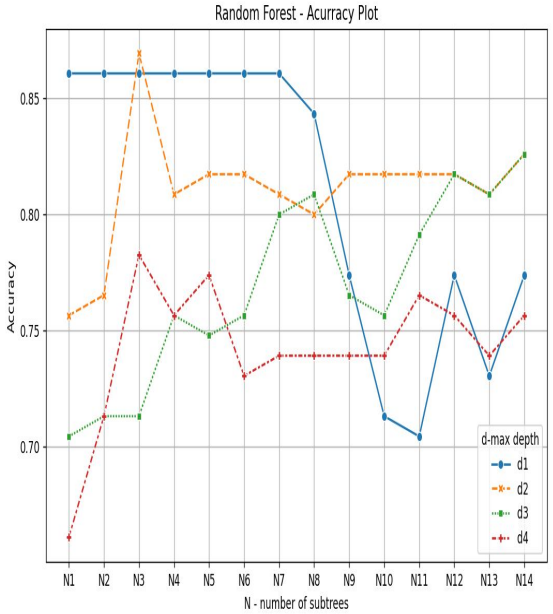


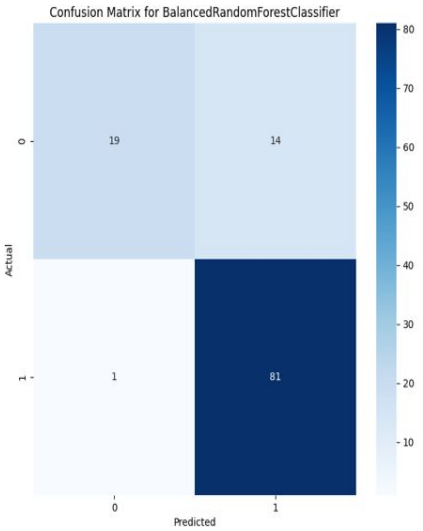


Model	TP	FP	TN	FN	Accuracy	TPR	TNR	F1[0]	F1[1]
RandomForestClassifier	81	15	18	1	0.861	0.988	0.545	0.692	0.91

Balanced Random Forest Classifier - Best Performing

- Finding the best hyperparameters, specifically number of trees, and max depth of the trees.
- It provided a better F1 score for rejected loans than the Random Forest Classifier without balance.
- Optimal values - Number of trees:3, max_depth: 2.

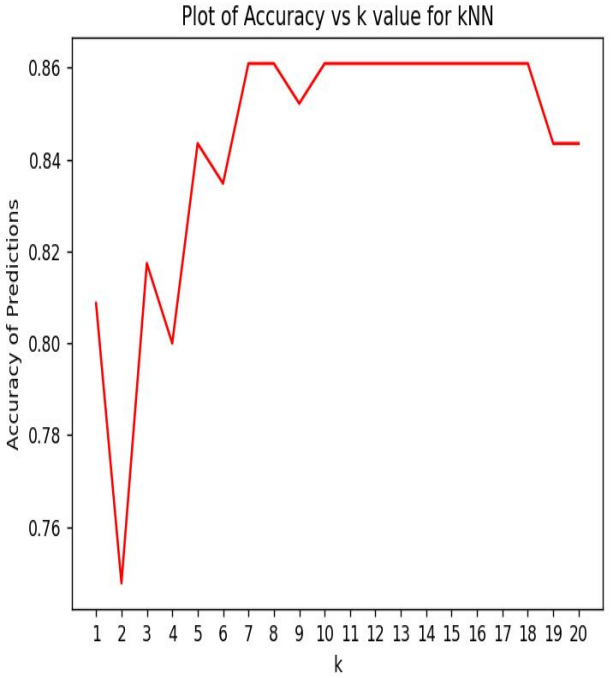


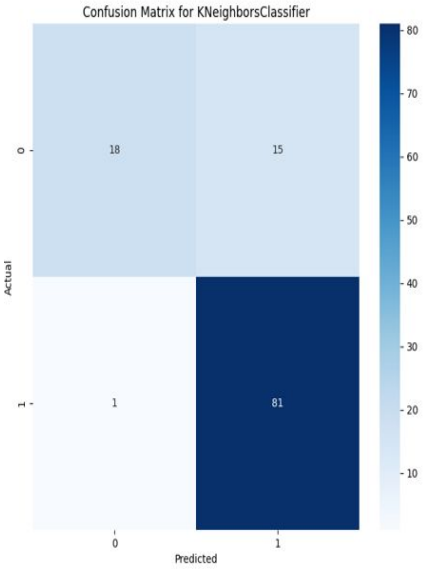


Model	TP	FP	TN	FN	Accuracy	TPR	TNR	F1[0]	F1[1]
BalancedRandomForestClassifier	81	14	19	1	0.87	0.988	0.576	0.717	0.915

K-Nearest Neighbor Classifier

- Finding the best hyperparameters, specifically number of trees, and max depth of the trees.
- It provided a similar F1 score for rejected loans than the other classifiers.

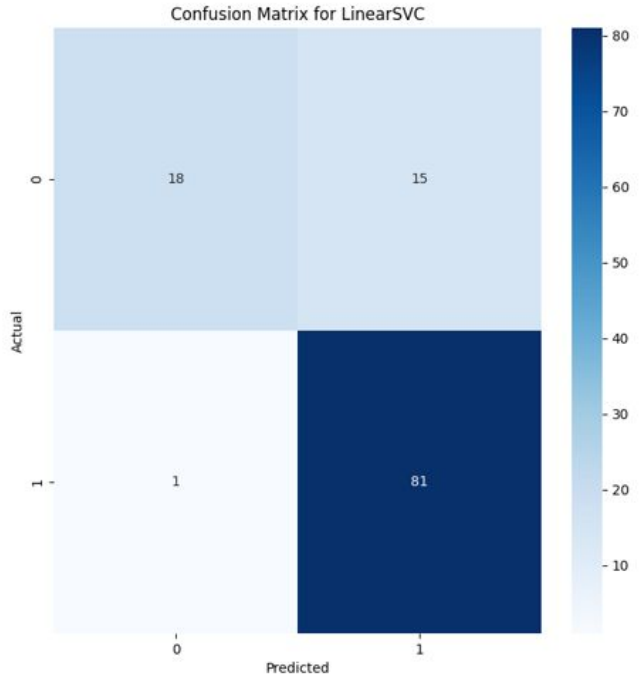




Model	TP	FP	TN	FN	Accuracy	TPR	TNR	F1[0]	F1[1]
KNN	81	15	18	1	0.861	0.988	0.545	0.692	0.91

Linear SVC

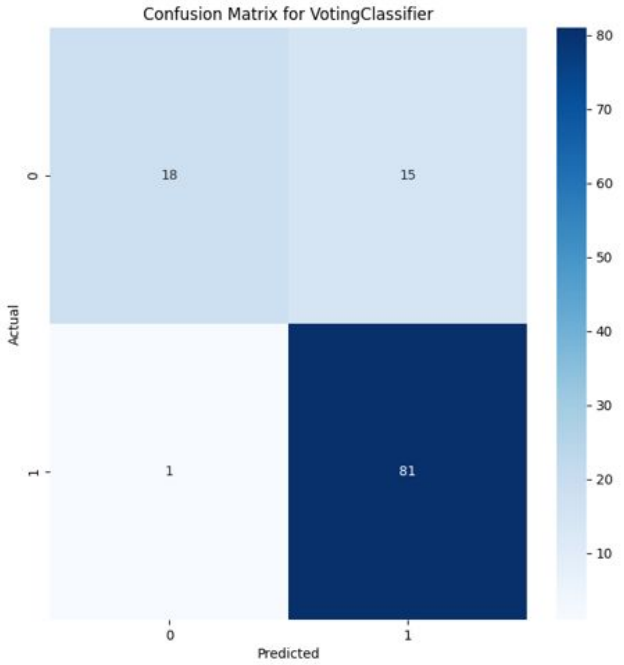
- Accuracy and F1 scores were identical to Random Forest and Voting classifiers.



Model	TP	FP	TN	FN	Accuracy	TPR	TNR	F1[0]	F1[1]
LInear SVC	81	15	18	1	0.861	0.988	0.545	0.692	0.91

Voting Classifier

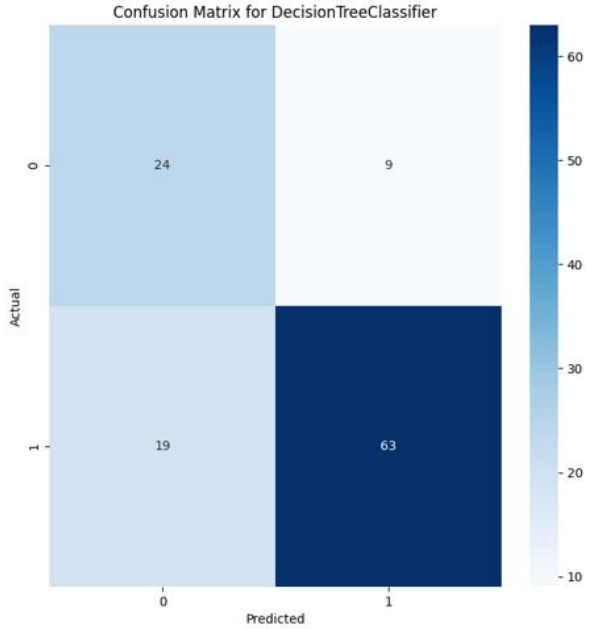
- Models used in voted classifier: Logistic Regression, Random Forest, SVM
- Accuracy nor F1 significantly improved.
- In addition to selecting a machine learning model, exploring other techniques could be beneficial to enhance these metrics.



Model	TP	FP	TN	FN	Accuracy	TPR	TNR	F1[0]	F1[1]
Voting Classifier	81	15	18	1	0.861	0.988	0.545	0.692	0.91

Decision Tree

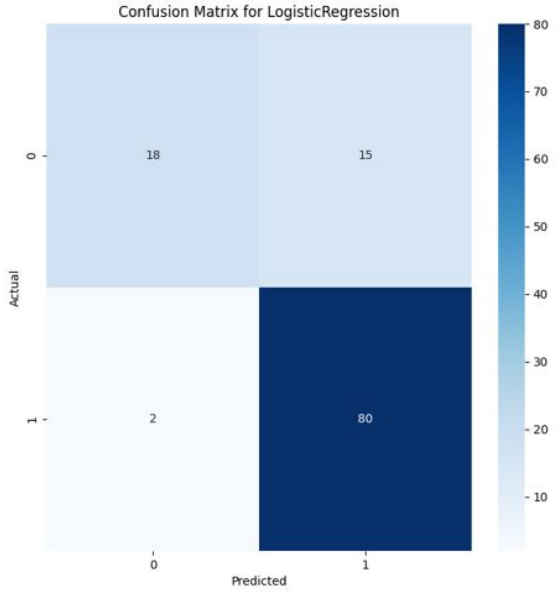
- The lowest performing model for both accuracy and F1 score for both rejected and approved loans was observed when using this classifier.



Model	TP	FP	TN	FN	Accuracy	TPR	TNR	F1[0]	F1[1]
Decision Tree	63	9	24	19	0.757	0.768	0.727	0.632	0.818

Logistic Regression

- It provided a similar accuracy compared to Random Forest and Balanced Random Forest, but F1 score was 4 points lower for rejected loans.



Model	TP	FP	TN	FN	Accuracy	TPR	TNR	F1[0]	F1[1]
LogisticRegression	80	15	18	2	0.852	0.976	0.545	0.679	0.904

Conclusions

1. In the chart below, there is a comparison amongst all the models utilized to process this dataset and predict the loan status based on input features.

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

1. Best Performing Models: Random Forest (both balanced and regular), Linear SVC, and Voting Classifier exhibited the best metrics in terms of accuracy and F1 score. This indicates that these models are effective in capturing the underlying patterns in the data and making accurate predictions.
2. Balanced Random Forest Advantage: The Balanced Random Forest provided a slightly higher accuracy and F1 score for rejected loans compared to the regular Random Forest. This suggests that the balancing technique used in the model helps in better classifying the minority class (rejected loans), leading to improved performance metrics for that class.
3. Poor Performance Models: Decision Tree did not perform as well in terms of both accuracy and F1 score. This could be due to the nature of the dataset and the inability of these models to capture the underlying patterns effectively.
4. Limited Improvement with Voting Classifier: The Voting Classifier did not provide any significant improvement over individual models. This indicates that the models included in the ensemble were already performing well individually, and combining them did not lead to a noticeable enhancement in performance. Exploring other techniques to further improve performance metrics would be beneficial.
5. Unbalanced Dataset Challenges: The low F1 score for rejected loans and high F1 score for approved loans highlight the challenge posed by the unbalanced nature of the dataset. This imbalance can lead to biased models that prioritize accuracy on the majority class while neglecting the minority class. Exploring sampling techniques or acquiring more data about rejected loans could potentially address this issue and improve model performance further. Other techniques could be explored but are outside of the scope of this project
6. Importance of executing exploratory data analysis (EDA) to determine the best course of action for missing values, outliers, incorrectly input values, and others. These EDA-driven actions are essential for enhancing data quality, ensuring model reliability, and making informed decisions based on accurate data.