Loan Approval Prediction using Random Forest

1st Fidelis Prasetyo
Computer Science
California State Polytechnic University, Pomona
Pomona, USA
fprasetyo@cpp.edu

2nd Sheldin Lau
Computer Science
California State Polytechnic University, Pomona
Pomona, USA
srlau@cpp.edu

3rd Andrew Lau
Computer Science
California State Polytechnic University, Pomona
Pomona, USA
adlau@cpp.edu

4th Yaoqiang Lin
Computer Science
California State Polytechnic University, Pomona
Pomona, USA
yaolin@cpp.edu

5th Arze Lu
Computer Science
California State Polytechnic University, Pomona
Pomona, USA
arzelu@cpp.edu

***Abstract*—This project aims to build a machine learning model to predict the loan approval status based on the applicant attributes using a loan data set obtained from Kaggle. The data set consists of 45,000 samples with numerical and categorical characteristics, such as income, credit score, loan amount, and demographic information. Due to significant class imbalance (78% rejections, 22% approvals), an under-sampling technique will be applied to balance the training data. A random forest classifier is selected as the machine learning model for its ability to handle mixed data types without extensive preprocessing. The model yielded F1-score of 0.95 for the negative class, 0.8 for the positive class, and an accuracy of 0.92. The most influential feature in the dataset is the personal records of past loan defaults.**

***Index Terms*—loan approval, prediction, random forest, machine learning**

I. Introduction

Banks consider various factors before approving a loan application, ranging from clear metrics such as income, loan amount, and credit score to potentially more subjective attributes such as age or gender. By analyzing loan data, it is possible to build a machine learning model capable of predicting whether a loan application is likely to be approved or rejected based on these characteristics. The data set used in this project came from Kaggle titled 'Loan Approval Classification Dataset' by Ta-wei Lo [1]. The dataset will be split by 70-15-15 for training, validation, and testing, respectively. For the positive class, the classifier yielded a precision score of 0.91 and a recall score of 0.72, contributing to the F1-score of 0.8, lower than that of the negative class by roughly 0.15. The pattern of high precision and relatively lower recall rate reflects both the imbalanced dataset. After experimenting with weighted and SMOTE sampling methods, we found that no sampling yields the best result.

Loan approval is important to many people trying to buy a home or a car. Using machine learning to simplify and speed up this process can help increase the consistency and fairness of the decision being made to either approve or decline a person's loan request.

II. Dataset Details

The data set is about loan approvals and how different variables may affect a person's chances of getting approved for a loan. It is generated from different credit risk datasets and uses SMOTENC to generate additional data points.

There are 45,000 samples and 14 features. These features include the age, gender, education, income, years of employment, and home ownership status of a person, loan amount, loan intent, loan interest rate, etc. The class distribution for a loan that is declined is 78% and a loan being approved is 22%. There is a class imbalance, since there are more loans being declined than approved.

III. Methodology

Given that the dataset is heavily imbalanced, with 78% rejections and 22% approvals, an under-sampling approach will be employed to achieve a more balanced class distribution; otherwise, the model could potentially develop an unwanted bias toward the majority class.

This project will use the Random Forest [2], a supervised ensemble learning algorithm that constructs multiple decision trees and combines them using majority voting, as the classification model for predicting the loan approval status. Each tree is trained on a bootstrap sample of the dataset, and at each node, a random subset of features is considered for splitting. This randomness introduces diversity among the trees, reducing overfitting and improving generalization. The

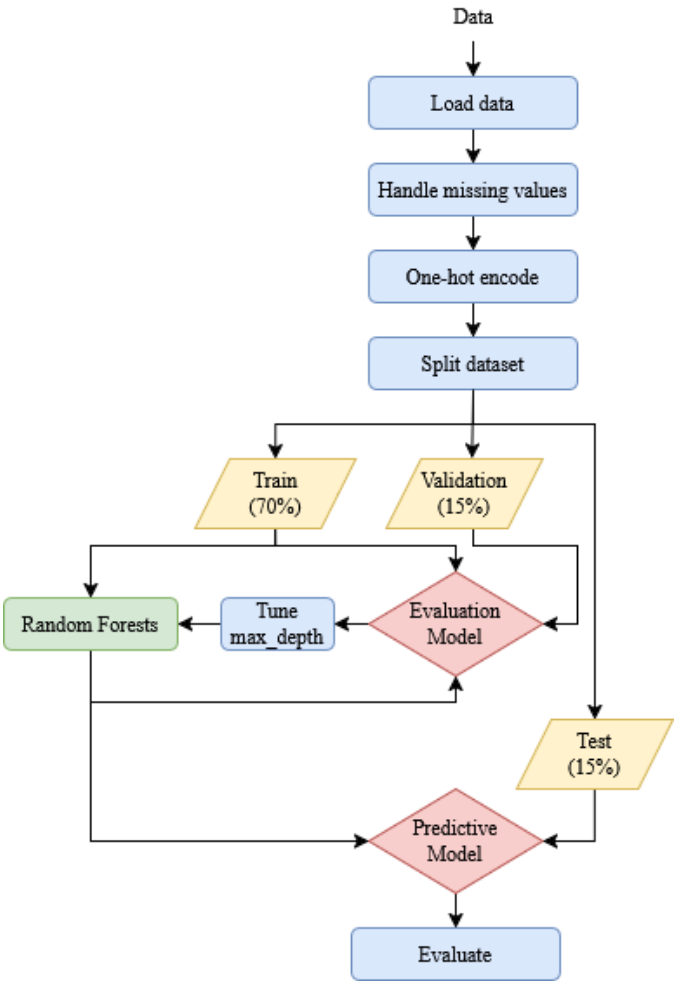

Fig. 1. Proposed model development pipeline

final prediction is made by aggregating the outputs of all trees, which makes the model robust to noise and capable of handling both numerical and categorical data. Given the nature of the dataset, Random Forest is a particularly suitable choice for this purpose.

To evaluate the model's performance, the dataset will be split into training and test datasets with an 70-15-15 splits, with 70% as the training set, 15% as the validation set, and 15% as the test set, ensuring no overlap between them. The validation set will be used to fine-tune the model's parameters in order to prevent both underfitting and overfitting during training. Finally, The test dataset will be applied to the trained model to evaluate the performance of the model.

Fig. 1 illustrates the proposed model development pipeline in this study. The preprocessing stage includes loading the dataset, handling missing values by removing instances with missing attributes, one-hot encoding categorical features to eliminate ordinality, and splitting the dataset into 70-15-15 training-validation-test datasets. The machine learning phase begins by training a evaluation model on the training set and then comparing F1-scores on both the training and validation

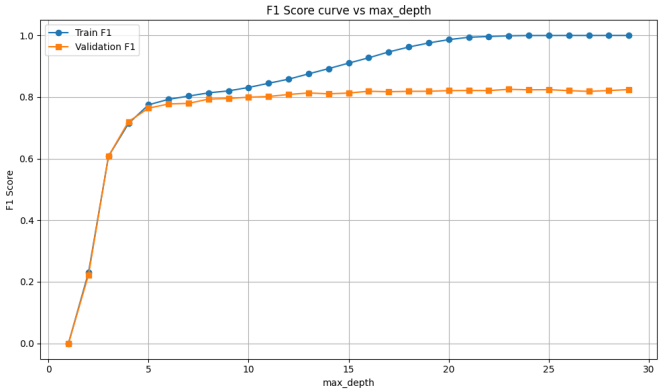

Fig. 2. F1-score curve with train and validation set

sets. This comparison helps determine whether the model is underfitting, well-fitted, or overfitting. The max_depth, parameter that adjusts the maximum depth of the tree, is fine-tuned to find the optimal value that achieves a well-generalized model. Once the optimal value of max_depth is selected, the final predictive model is trained and evaluated on the test set to assess its generalization performance.

Some evaluation metrics will be calculated to assess the performance of the model, including precision, recall, accuracy, F1-Score, confusion matrix, feature importances, and ROC curve. These metrics will provide a comprehensive understanding of the model's ability to correctly classify loan approvals and rejections.

Lastly, this project will reference similar works, such as the study by Rani and Gupta [3], who developed a home loan approval prediction model using the same Random Forest algorithm. Their work will serve as a useful reference and provide a basis for performance comparison.

IV. Results

Fig. 2 shows the f1-scores of the both training and validation datasets across various max_depth values, ranging from 1 to 30. From the curve, it is evident that the model starts to overfit when max_depth is around 10 to 15, as the training curve diverges, continue to rise, while validation curve stays relatively flat. Therefore, we selected 10 as our max_depth final value to be used to build the predictive model.

TABLE I
Performance of the Model using different sampling methods

	Original		Weighted		With SMOTE	
Class	**0**	**1**	**0**	**1**	**0**	**1**
Precision	0.92	0.91	0.97	0.7	0.96	0.71
Recall	0.98	0.72	0.89	0.89	0.9	0.85
F1-Score	0.95	0.8	0.93	0.78	0.93	0.78
Accuracy	0.92		0.89		0.89	

*0 = Rejected, 1 = Approved

We also compared the performance of the model using different sampling strategies: Original, where no additional sampling technique is applied; Weighted, which automatically

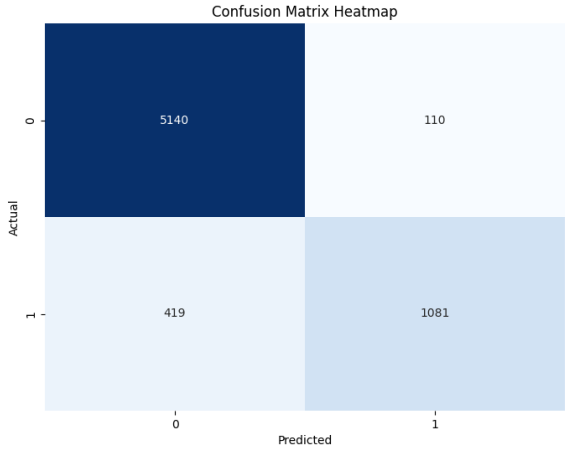

Fig. 3. Predictive model confusion matrix

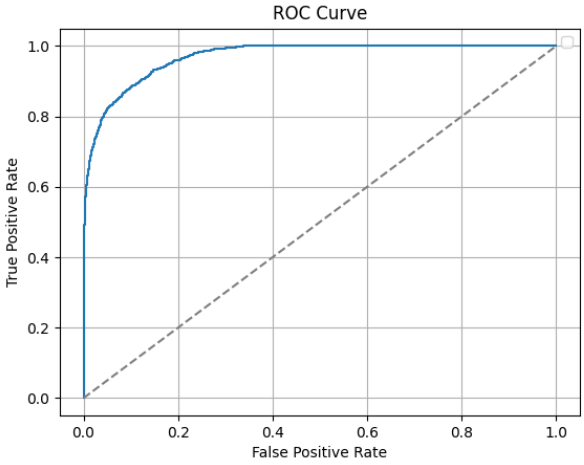

Fig. 4. ROC Curve

adjust class weights inversely proportional to class frequencies; and with SMOTE, which generates synthetic data points to balance the dataset. Table 1 presents those results. In terms of Accuracy and F1-score, the original model performs the best compared to the other two. Although the weighted and SMOTE approaches do improve the recall of the approved class compared to the original, they suffer a considerable drop in the precision. For loan approval prediction, it is more logical to prioritize precision, as low precision implies high false positives, meaning the bank may approve applicants who cannot pay their loans back. Therefore, we decided not to use any sampling techniques.

The confusion matrix of the final predictive model is shown in Fig. 3. As expected, the model performs well in predicting true negatives (rejections) and does reasonably well in predicting true positives (approvals). It achieves 72% recall, indicating that 28% actual approvals are missed; and 91% precision, which means that only 9% of the predicted approvals are incorrect.

Fig. 5 shows the rank of the feature importance based on impurity. Based on this dataset, the most determining factor is the applicant's loan default history, followed by loan amount is relative to their income, the interest rate of that loan, and applicant's income. Surprisingly, credit score ranks lower in importance than home ownership status. Demographics of the applicant (gender, age, education) barely affect the decision, meaning the model is more financially driven rather than demographically-biased.

Finally, Fig. 4 displays the ROC curve for the final predictive model. The curve illustrates the trade-off between the TPR (true positive rate) and FPR (false positive rate) across various classification thresholds. Compared to random guess, the model performs considerably better in distinguishing approved and rejected applications.

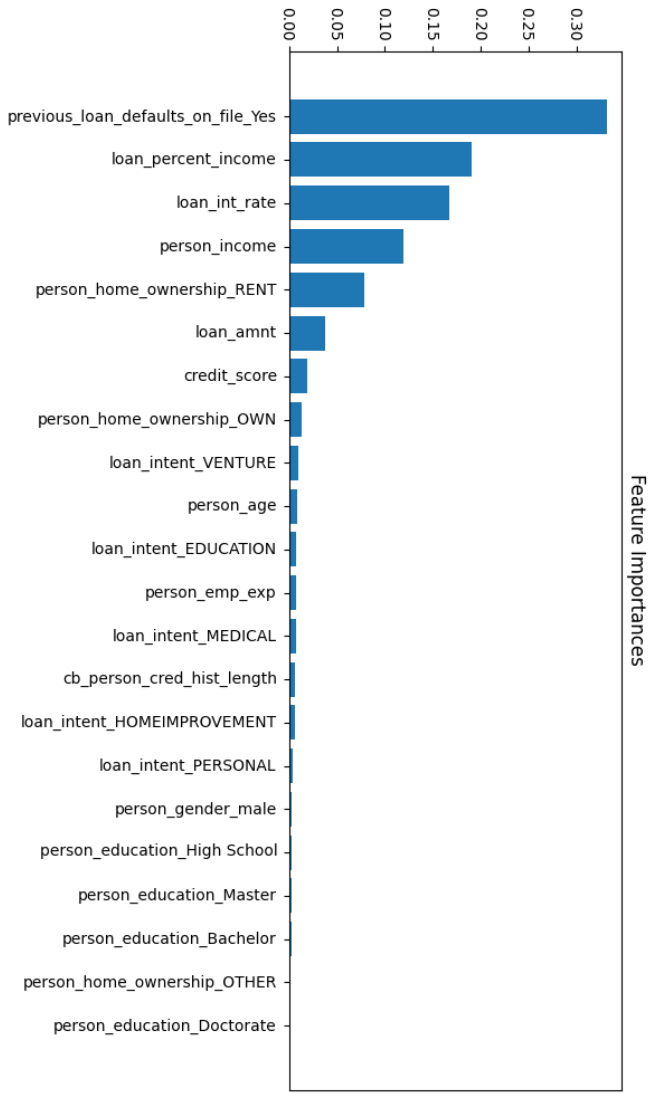

Fig. 5. Feature importance rank

V. Related Works

The study on Predicting Home Loans Using Random Forest Classifiers [3] shows the ability for Random Forest in classifying home loan approvals. Our study focuses on a more broad range of loans that range from student loans to personal loans. This report demonstrates the ability for Random Forest classifiers to perform well in difficult datasets such as those that have missing data, unbalanced datasets, and multi-dimensional financial features such as income and loan amount. The ability of Random Forest classifiers to not overfit the complex data also helps its performance. The main reason for Random Forest in loan prediction is due to the model being easy to interpret the results and how it got the prediction it got. The study also mentions how combining Random Forest with advanced techniques such as SMOTE for better class balance which would make this classifier perform well in our setting. Our model matches the results since we are getting very high accuracy in predicting a loan being declined correctly but not as well for predicting a loan being approved however it still performs reasonably well.

VI. Conclusion

This paper provides a way in training and evaluating a Random Forest classifier in loan classification. Through our testing we have found that the most important features are previous loans being defaulted, loan to income percentage, and loan interest rate. These factors were very important indicators in whether a loan would be declined such as when a person had previous loans being defaulted would guarantee that their loan would be declined. The ability for the Random Forest classifier to perform extremely well with such a complex dataset shows Random Forest's ability to handle complex data with many features and some features being correlated. The hyperparameter tuning has shown that the model can overfit very easily if the max depth is too high and that the weighted random forest does not perform well on this data since there is normally more loans being declined. Drawing the ROC curve shows that the Random Forest model is performing very well since it is closer to the top left corner. Random Forest classifiers being used in loan approvals would be very beneficial and would help give loan applicants a better idea of whether their loans would be approved or not.

References

[1] T. A. Wei Lo, "Loan Approval Classification Data," *Kaggle*, 2021. [Online]. Available: https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data. [Accessed: Apr. 14, 2025].

[2] T. K. Ho, "Random decision forests," in *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, vol. 1, pp. 278–282, IEEE, 1995.

[3] R. Rani and S. Gupta, "Predicting Home Loan Approvals Using Random Forest Classifiers: A Comprehensive Machine Learning Approach," in *Proceedings of the 2024 3rd International Conference for Advancement in Technology (ICONAT)*, pp. 1–4, IEEE, 2024. doi: 10.1109/ICONAT61936.2024.10775174.

VII. Supplementary

Latex file: https://www.overleaf.com/read/nfwccnwbjsmg#9ac97f
Source code: https://github.com/fidelisprasetyo/cs5990_project

TABLE II
Project member contributions

Name	Contribution
Fidelis Prasetyo	report, fine-tuning, sampling comparison, feature importance
Sheldin Lau	report, hyperparameter, visualizing data
Andrew Lau	slides, presentation
Yaoqiang Lin	preprocessing, training data, evaluations
Arze Lu	slides, presentation, report