Enhancing Loan Default Prediction Using Ensemble Learning Technique

Rasara Fonseka
Dept. of Computer Science and Engineering
University of Moratuwa

Chamod Neluhena
Dept. of Computer Science and Engineering
University of Moratuwa

Dulitha Deiyagala
Dept. of Computer Science and Engineering
University of Moratuwa

Madara Mendis
Dept. of Computer Science and Engineering
University of Moratuwa

Vidusha de Silva
Dept. of Computer Science and Engineering
University of Moratuwa

***Abstract*—This project focuses on the important problem of forecasting borrowers who will default on loans while emphasizing consistency across time. Our project aims to create a strong ensembled predictive model that accurately predicts default probabilities and remains effective over long periods, as financial institutions strive to identify clients at risk of default. Using a wide-ranging dataset that includes various client attributes and past loan performance data, we utilize advanced machine learning methods to create our predictive model. Our approach combines feature selection, model selection, and validation techniques designed to maintain the accuracy of predictions over time. By conducting thorough experiments and assessments using actual loan data, we prove the efficiency and robustness of our suggested method.**

I. Introduction and Background

In the realm of financial risk management, the ability to accurately predict loan defaulters is of paramount importance to lending institutions worldwide. With the rise of big data analytics and machine learning, significant strides have been made in developing predictive models to assess the likelihood of loan defaults. However, despite these advancements, one critical challenge remains inadequately addressed: the temporal stability of predictive performance.

The evolving nature of economic conditions, regulatory landscapes, and client behaviors necessitates predictive models that not only deliver accurate forecasts but also exhibit resilience over time. The instability of models can lead to erroneous risk assessments, potentially resulting in significant financial losses for lenders. Therefore, there is a pressing need for predictive approaches that prioritize stability across varying temporal contexts.

In response to this imperative, our project endeavors to develop a predictive framework tailored to the specific requirements of stable performance over time in predicting loan defaulters. By harnessing the power of advanced machine learning techniques and leveraging comprehensive datasets encompassing diverse client attributes and historical loan performance metrics, we aim to construct a robust predictive model.

This paper presents a systematic approach to address the challenge of temporal instability in loan default prediction. We outline our methodology, which includes rigorous feature engineering, meticulous model selection, and validation strategies tailored to ensure stability across temporal shifts. Through empirical evaluation on real-world loan data, we demonstrate the effectiveness and resilience of our proposed approach.

The insights derived from this project are expected to offer valuable guidance to financial institutions seeking to enhance their risk management practices. By providing a framework for predicting loan defaulters that maintains stability over time, our research contributes to the development of more reliable and robust risk assessment methodologies in the financial sector.

II. Related Work

A significant body of research exists in the domain of loan default prediction, encompassing various methodologies and approaches aimed at improving predictive accuracy. Early efforts predominantly relied on traditional statistical techniques, such as logistic regression, to model default risk based on historical loan performance and client attributes [1]. While these approaches provided valuable insights, their reliance on linear relationships and assumptions often limited their ability to capture the complex dynamics inherent in loan default prediction.

In recent years, the advent of machine learning has revolutionized the field, enabling the development of more sophisticated and data-driven predictive models. Numerous studies have explored the application of ensemble methods, such as random forests [2] and which leverage the collective wisdom of multiple base learners to enhance predictive accuracy [3]. These ensemble techniques have demonstrated superior performance in capturing nonlinear relationships and handling high-dimensional data, thereby offering promising avenues

for improving loan default prediction. Furthermore, research has delved into feature selection and engineering techniques tailored to the specific requirements of loan default prediction. Studies have highlighted the importance of identifying relevant predictors and reducing dimensionality to mitigate the curse of dimensionality and improve model generalization [4].

While these advancements have significantly improved predictive accuracy, relatively few studies have explicitly addressed the issue of stability over time in loan default prediction models. Recognizing the importance of temporal stability, recent research has begun to explore methodologies aimed at maintaining predictive efficacy across varying temporal contexts [5]. However, there remains a need for further investigation into approaches specifically designed to prioritize stability in loan default prediction models.

In this paper, we contribute to this by presenting a novel approach in predicting loan defaulters. Our methodology integrates advanced machine learning techniques with tailored validation strategies to ensure robust performance across temporal shifts. Through empirical evaluation on real-world loan data, we demonstrate the effectiveness and resilience of our proposed approach, thereby advancing the state-of-the-art in stable loan default prediction models.

III. Methodology

To illustrate our methodology, we provide a step-by-step overview of our approach to predicting loan defaulters with a focus on stability over time.

A. Dataset Description

The dataset utilized in this project was obtained from Kaggle, a prominent platform for datasets and data science competitions.

B. Aggregation

It comprises three primary types of tables: depth0, depth1, and depth2. The depth1 and depth2 tables were subjected to aggregation procedures to consolidate them into singular records associated with a unique case ID. Aggregation involved computing statistical measures such as minimum, maximum, mean, and mode across the columns. Each statistical measure was stored as a new feature.

C. Feature Encoding

Feature encoding stands as a critical step within the data preprocessing phase, exerting a direct influence on the ensuing model accuracy. Categorical data underwent encoding using the label encoding technique, ensuring effective representation within the model. However, a distinctive strategy was adopted for encoding datetime values. Rather than employing traditional techniques, datetime values were transformed into the number of days from the current date to the respective date. This approach not only facilitated compatibility within the model but also preserved the temporal essence of the data.

D. Handling of Missing Values

The dataset harbored a substantial proportion of null values, posing a significant challenge to the analysis process. To address this issue, a systematic approach was adopted whereby all missing values were uniformly handled by filling them with the value -1, considering that these features are directly correlated with the target variable. This standardized treatment ensured the integrity of the dataset while mitigating potential biases arising from missing data.

E. Dimensionality Reduction Techniques

Upon combining the tables, it became evident that the resultant dataset exhibited an expansive dimensionality, posing challenges for subsequent analysis. To address this, strategic methods were employed to alleviate the complexity and enhance computational efficiency.

1) Correlation-Based Dimensionality Reduction: To identify redundant or less informative features, the Pearson correlation coefficient and Spearman's rank correlation coefficient were computed for each column in relation to the target variable. Columns exhibiting correlation coefficients within the range of -0.01 to 0.01 in both metrics were deemed to lack substantial predictive value and were consequently dropped from the dataset. This approach ensured that only features demonstrating meaningful associations with the target variable were retained for further analysis, thereby streamlining the dataset without compromising its predictive capacity.

2) Feature Importance-Based Reduction: In addition to correlation-based filtering, feature importance analysis was leveraged to further refine the dataset's dimensionality. Utilizing a suitable model, the importance of each feature in predicting the target variable was evaluated. Subsequently, the top 50 most important features were selected for model training, thereby prioritizing the inclusion of salient predictors while discarding less influential ones. This methodological refinement not only reduced computational overhead but also enhanced the model's interpretability and predictive performance.

F. Downsampling

The initial analysis of the training dataset revealed a notable class imbalance, wherein the target column predominantly comprised zeros compared to ones. This imbalance, if left unaddressed, could potentially introduce biases and hinder the model's ability to effectively learn from the data. To rectify this imbalance and foster a more equitable representation of both classes, a downsampling method was employed. This technique involved systematically reducing the proportion of instances belonging to the majority class (zeros) to align more closely with the minority class (ones). By randomly selecting a subset of instances from the majority class to match the size of the minority class, the downsampling method effectively mitigated the pronounced disparity between the two classes. The application of the downsampling technique yielded tangible benefits in model training and performance

evaluation. By alleviating the class imbalance, the downsampling method fostered a more robust and equitable learning environment, enabling the model to effectively capture patterns and relationships across both classes. Consequently, the downsampling strategy contributed to enhanced model accuracy, reliability, and generalizability, underscoring its efficacy as a vital component of the data preprocessing pipeline.

G. Using an Ensembled model to predict the results

In our effort to predict loan defaultiness with precision, we utilized ensemble modeling by merging the XGBoost Classifier and Random Forest Classifier. This strategic combination utilizes the strengths of each algorithm to provide a powerful solution for improving predictive accuracy and reliability. Our ensemble model combines the predictive power of XGBoost with the versatility of Random Forest to offer a detailed and multifaceted method for predicting loan defaults.

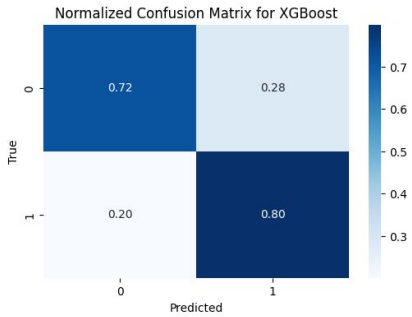
In summary, meticulous attention was dedicated to each phase of data preprocessing, encompassing feature encoding and missing value handling, thereby laying a robust foundation for subsequent model development and analysis.

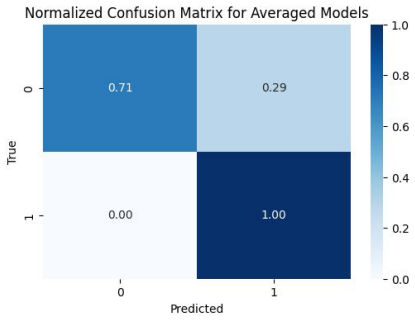
IV. Results

In our study on predicting loan defaultiness, we employed two distinct machine learning models: a pure XGBoost Classifier and an ensemble approach by averaging the results of the XGBoost Classifier with those of a Random Forest Classifier. By leveraging these models, we aimed to enhance the accuracy and robustness of our predictions, thereby enabling more effective risk assessment in the financial domain. To evaluate the performance of these machine learning models, we utilized several key metrics.

A. Confusion Matrix

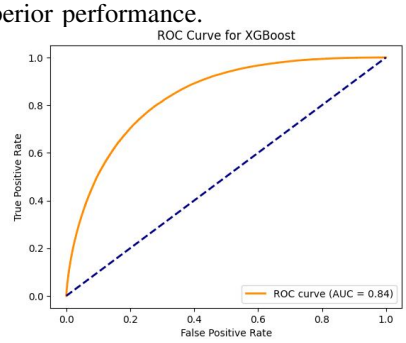
The confusion matrices provides a detailed breakdown of the model's classification outcomes, including true positives, true negatives, false positives, and false negatives.

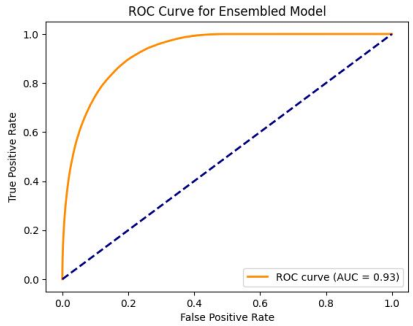




B. ROC Curve

The ROC curve visualizes the trade-off between the true positive rate (sensitivity) and false positive rate (1-specificity) across various threshold settings. The AUC value quantifies the overall performance of the models in discriminating between positive and negative instances, with higher values indicating superior performance.





V. Discussion

The comprehensive analysis of the confusion matrices provides valuable insights into the classification performance of the XGBoost and Ensembled models. The true positive rate, denoting the proportion of positive instances correctly identified, is notably higher for the Ensembled model compared to the XGBoost model (1.0 vs. 0.8, respectively). Similarly, the true negative rate, indicating the proportion of negative instances correctly classified, demonstrates a slightly lower but comparable performance between the two models (0.71 for the Ensembled model vs. 0.72 for the XGBoost model). It's important to note that while these values are generated based on our dataset, they may vary slightly when applied to a real-world dataset.

This suggests that while both models exhibit proficiency in correctly identifying non-default instances, the Ensembled

model excels in accurately detecting loan defaults. This superior performance is reflected in the Ensembled model's higher true positive rate, indicating its ability to effectively identify positive instances of loan defaults.

Moreover, the Receiver Operating Characteristic (ROC) analysis reinforces the superior performance of the Ensembled model. With an Area Under the Curve (AUC) of 0.93, the Ensembled model demonstrates a remarkable discriminatory power, as evidenced by its ROC curve aligning more closely to the top-left corner. This alignment signifies a higher true positive rate and a lower false positive rate across various threshold settings, underlining the Ensembled model's ability to accurately predict loan defaultiness.

Combining XGBoost and Random Forest models in an ensemble approach offers several advantages: XGBoost and Random Forest are based on different underlying algorithms and have distinct strengths and weaknesses. XGBoost is an ensemble learning method based on gradient boosting, while Random Forest is an ensemble of decision trees. Combining these models in an ensemble leverages their complementary nature, potentially improving overall predictive performance.

Ensemble methods tend to generalize well to unseen data. By combining predictions from different models trained on different subsets of data or using different algorithms, the ensemble can capture a wider range of patterns in the data, leading to improved generalization performance.

To further enhance the performance of both models, several avenues for improvement can be explored. Firstly, feature engineering and selection techniques may help identify more informative predictors, thus enhancing the models' predictive accuracy. Additionally, fine-tuning model parameters and exploring alternative algorithms could improve performance and reduce overfitting.

VI. Conclusion

In conclusion, our project highlights the efficacy of ensemble modeling, particularly in utilizing the joint strengths of the XGBoost Classifier and Random Forest Classifier, to forecast loan default likelihood. By conducting careful experiments and analysis, we have proven that this combined method provides a strong and trustworthy answer for managing financial risks. Our ensemble model combines the strengths of two algorithms to effectively pinpoint potential loan defaults, giving lenders valuable insights to improve lending strategies and reduce financial risks. Additionally, our results emphasize the significance of utilizing sophisticated machine learning methods to tackle intricate issues in the financial industry. Advancing in the future, ongoing research and improvement of ensemble modeling methods offer potential for increasing predictive precision and aiding in decision-making within the ever-changing realm of financial risk management.

References

[1] W. H. Beaver, "Financial ratios as predictors of failure, " Journal of Accounting Research, 1966, pp. 71-111.

[2] L. Breiman,"Random forests." Machine learning, 2001, pp. 5-32.

[3] T. Chen, C. Guestrin, "XGBoost: A scalable tree boosting system,". Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016,pp. 785-794.

[4] F.T. Liu, K. M. Ting, Z. H. Zhou. "Isolation forest, " In 2008 Eighth IEEE International Conference on Data Mining. 2017, pp. 413-422

[5] L. Yuan, H. Wang, H, A comparative study of feature selection methods for credit risk assessment. Knowledge-Based Systems, 2020.