Loan Default Prediction

Zeynep Meriç Aşık
Artificial Intelligence Engineering
TOBB ETU
Ankara, Türkiye
zeynepmericasik@gmail.com

***Abstract*—In the world of banking, lending operations are crucial for revenue generation while carrying significant risks, primarily due to the possibility of loan defaults. This project utilizes a dataset detailing past loan recipients to develop a predictive model that classifies potential default risks associated with new borrowers. The dataset includes various deterministic factors such as income, gender, and purpose of the loan, which are instrumental in shaping the predictive analytics. Employing machine learning models such as Support Vector Machines, XGBoost, Linear Regression, Random Forest and Multi-Layer Perceptrons, the study aims to enhance decision-making processes in financial institutions by integrating and analysing known customer data. The outcomes are designed to be directly applicable to real-world scenarios, potentially reducing the financial risks associated with lending activities by enabling more informed and strategic decision-making in loan approvals.**

***Index Terms*—Machine Learning, Loan Default, Random Forest, Extra Gradient Classifier**

I. Introduction

In the financial sector, loan issuance is a significant aspect of banking operations, serving as both a primary revenue stream and a potential risk area. This dual nature stems from the uncertainty associated with loan repayments, where defaults can lead to substantial financial losses. Addressing this, the current project utilizes a comprehensive dataset featuring historical loan repayment records to develop a predictive model that discerns the likelihood of new borrowers defaulting. This dataset is inherently complex, characterized by high dimensionality and an imbalance in class distribution, with the minority class—default cases—being crucial for risk assessment.

The challenge of unbalanced data is prevalent in predictive modelling for financial applications, where the minority class often holds more significance due to its impact on risk and cost. In this project, a strategic approach was adopted to balance the dataset by reducing the number of non-default samples to match the default ones, thus enhancing the model's sensitivity to potential defaults.

Furthermore, the curse of dimensionality presented significant problems, typical of datasets with numerous attributes, which can dilute the effectiveness of many machine learning algorithms. To mitigate these effects, initial steps involved rigorous feature engineering to eliminate redundant or irrelevant predictors. Subsequently, a Random Forest algorithm was employed to identify and retain the most important features, reshaping the predictive model without sacrificing essential information.

These methodologies not only improved the model's performance but also significantly increased the recall rate, ensuring that the majority of actual default cases were correctly identified. This enhancement is crucial for financial institutions, as it directly supports more informed decision-making, reducing the likelihood of loan defaults and fostering a more robust financial environment. The success of these models demonstrates their applicability in real-world settings, providing a reliable tool for assessing borrower risk and aiding in the strategic allocation of loans.

II. Related Work

Effective data cleaning is pivotal in ensuring the reliability of machine learning models. In this project, the initial step involved addressing missing data by removing columns predominantly filled with null values and imputing the remaining missing values. For numerical data, missing values were imputed using the column's mean, and for categorical data, the mode was utilized. These methods are supported by the literature [1], which highlights their effectiveness in various contexts, including genomic data, where similar strategies are employed to handle missing not at random (MNAR) values by using a range of methods to maintain data integrity and support subsequent analyses.

The project also involved feature engineering, where Pearson correlation was used to identify and eliminate highly correlated variables. This practice is well-documented and is crucial for reducing model complexity and avoiding multicollinearity, which can obscure the interpretation of variable importance. The removal of near-zero variance predictors, similar to the approach outlined in the literature [2], helps in refining models by eliminating redundant data, thus enhancing the predictive performance and interpretability of the model.

Addressing class imbalance by downsizing the majority class to equal the minority class's size is a critical step in enhancing model sensitivity towards the minority class, which often represents the more crucial outcomes to predict. This method, known as random under sampling, is noted for its simplicity and effectiveness in balancing class distribution, by this way improving the recall for the minority class without complicating the modelling process. This approach is reflected in various studies [3] where balancing the dataset significantly improved model performance on minority class predictions. These methodologies provide a robust framework for addressing common issues in data science projects, ensuring that the

models developed are both reliable and applicable in real-world settings.

The selection and training of various machine learning models, including Logistic Regression, Naive Bayes, SVM, KNN, XGBoost, and Random Forest classifiers, are widely applied across different domains due to their ability to handle various types of data and prediction complexities. For instance, these models have been evaluated for their efficacy in healthcare applications [4], where they are used to predict disease outbreaks and patient outcomes based on diverse datasets. These classifiers are often chosen for their distinct handling of bias, variance, and their different assumptions about the data's underlying distribution.

Using Random Forest for determining feature importance is a recognized method to tackle the curse of dimensionality by identifying the most impactful features. This approach helps in reducing overfitting and improving model generalizability. For example, feature importance metrics derived from Random Forest have been successfully applied to select significant features that enhance model performance without compromising the accuracy [5].

Hyperparameter tuning is a critical step in optimizing machine learning models. GridSearch is commonly used for this purpose because it systematically tests a combination of parameters to find the most effective ones. This technique has been applied across various studies [5] to fine-tune models, ensuring that each model operates at its optimal performance level. The method is particularly valued for its thoroughness in exploring the parameter space, which can be crucial for achieving the best results from complex models like those used in fake news detection.

The emphasis on the recall metric in loan default prediction is driven by the critical need to identify all potential defaults accurately. In the domain of financial lending, missing out on detecting a likely default can result in significant financial losses. Therefore, prioritizing recall helps institutions minimize the occurrence of false negatives (failing to identify an actual default), which is vital for managing credit risk effectively.

In the literature, the significance of recall is highlighted in various studies where machine learning models are employed to enhance loan default predictions. For instance, research indicates that models optimized for high recall are particularly valuable in scenarios where the consequences of missing a default are more severe than misclassifying a non-default as a default. This approach is beneficial for financial institutions as it allows them to safeguard against potential financial instability caused by loan defaults.

Studies like those conducted by Zhou (2023) [6] utilize machine learning methods including Logistic Regression, Decision Trees, Random Forest, and XGBoost to focus on optimizing recall. These methods are chosen because they are capable of dealing with unbalanced datasets typical in default predictions, where defaults are a minority class. Zhou's research demonstrates that XGBoost, in particular, performs well with the highest recall, indicating its efficacy in identifying most of the actual defaults without a significant number of false negatives. This article implies the importance of focusing on recall in loan default prediction to ensure that financial institutions can effectively identify and mitigate risks associated with loan defaults.

III. Methodology

A. Data Cleaning and Pre-processing

The initial step involved finding data which is a comprehensive dataset detailing past loan recipients. The dataset includes various deterministic factors such as income, gender, and the purpose of the loan. This data underwent a series of pre-processing steps explained below.

Columns mostly filled with null values were removed to enhance the dataset quality. For the remaining missing values, numerical fields were imputed with the mean, and categorical fields were filled with the mode, adhering to common practices in handling missing data.

Features were assessed for multicollinearity using Pearson correlation, and highly correlated features were dropped. Additionally, binary-like features that predominantly contained a single value were also removed to simplify the model and improve computational efficiency.

There were also columns with single values such as the year of the loans were lent, which were all 2019, all unique values such as customer IDs, and also a column that has the possibility of the customer to default the loan which would cause data leakage if included. The columns carrying these properties are also dropped.

B. Addressing Class Imbalance

Given the unbalanced nature of the dataset with the minority class (default cases) being more critical, the class distribution was balanced by under sampling the majority class. This step was crucial to enhance the model's ability to detect the minority class, therefore improving the recall metric which is vital for identifying potential defaulters.

C. Model Selection and Training

Various machine learning models were employed to classify potential default risks. The models included Logistic Regression, Naive Bayes, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Extreme Gradient Boosting (XGBoost), and Random Forest.

These models were chosen for their diverse assumptions about data distribution and their ability to handle bias and variance effectively. Each model's performance was evaluated with a focus on optimizing the recall to capture as many true default cases as possible without a significant number of false negatives.

D. Feature Selection Using Random Forest

Post-initial training, Random Forest was used to determine feature importance, which is a method to identify the most important features. Only the top seven features were used in subsequent model training sessions to mitigate the curse of dimensionality and to refine the model's performance.

E. Hyper parameter Tuning

The best-performing model (XGBoost) based on recall was selected for further optimization using GridSearchCV. This technique systematically tested various combinations of parameters to find the most effective settings, thus fine-tuning the model to achieve optimal performance in predicting loan defaults.

F. Evaluation Metrics

The primary metric for evaluating the model's performance was recall, due to the high cost associated with failing to identify default cases. However, other metrics such as accuracy, precision, and the F1-score were also monitored to ensure a balanced perspective on the model's overall performance. This comprehensive methodology ensures the development of a robust predictive model that can effectively assist financial institutions in predicting loan defaults, so that reducing potential financial risks to the minimal.

IV. Results

The dataset [7] utilized for this project involves 149,000 customers who have either taken or repaid loans. This dataset contains a total of 34 attributes, including 21 categorical, 12 numerical, and an additional ID column. Various processes such as encoding and normalization are applied to these attributes, followed by research to determine whether each attribute provides significant information, leading to feature selection. Some missing information in the dataset is filled in or removed, depending on the condition of the data. To balance the data different techniques are applied to check which one is the most suitable. Principal Vector Analysis (PCA) is applied to the data at some point but was not included in the project since it did not make a major difference on the result.

After data is ready to be trained, 5 different Machine Learning models are chosen: Gaussian Naive Bayes (GNB), Logistic Regression (LR), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), and XGBoost (XGB). These models are trained with their default parameters and evaluated mostly on recall and also on accuracy, precision, and f1-score metrics.

Figure 1 presents the recall values for all models trained after data balancing. The Random Forest and XGBoost Classifier models demonstrate the best performance, achieving near-perfect predictions on the test set. Despite its absence from the graph in Figure 1, Logistic Regression performs poorly in comparison, with a recall of 0.64 and accuracy of 0.63.

High dimensionality in the data may be complicating classification efforts. The Random Forest Classifier's importance metric identifies the most influential features, which predominantly relate to the financial aspects of the loans rather than customer-related attributes. The key features, ranked by importance, are: Loan Amount, Credit Score, LTV (Loan to Value ratio), Term, Upfront Charges, and Rate of Interest.

The dataset is reduced to include just these six columns plus the target variable, Status, and the models are retrained. Figure 2 displays the new test results based on recall values.

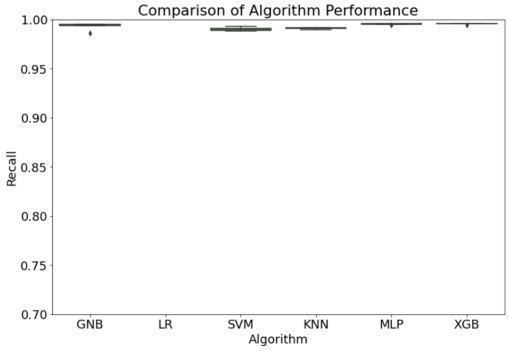

Fig. 1. Recall Values for Models

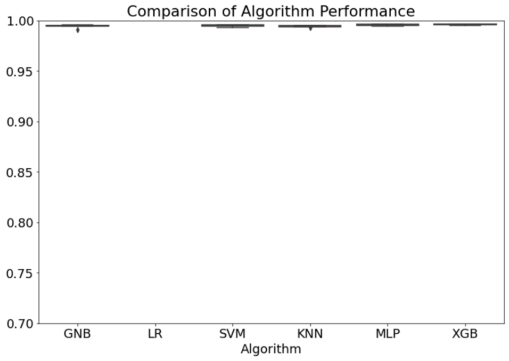

Fig. 2. Recall Values after Dimensionality Reduction

XGBoost continues to lead, closely followed by the Multilayer Perceptron (MLP) Classifier and Support Vector Machines (SVM). The models are highly accurate, making only about 10 errors per class, but there is still room for improvement.

Initially, the models were run with default settings. After identifying XGBoost as the most effective model, hyperparameter tuning was conducted. Figure 3 shows the confusion matrices for both the default and tuned versions of the model. The outcomes are very similar, suggesting that the hyperparameter adjustments had minimal impact on performance.

V. Conclusion

This project has demonstrated the effectiveness of applying a variety of machine learning techniques to predict loan defaults, utilizing a dataset enriched with diverse borrower characteristics. The methodology deployed provided a robust framework, beginning with meticulous data cleaning and

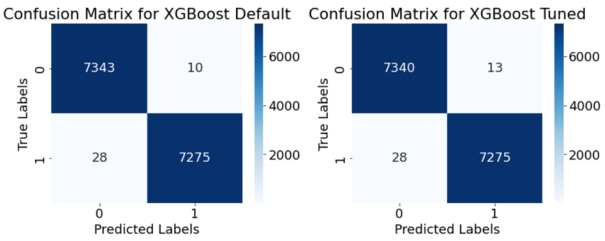

Fig. 3. Confusion Matrices for XGBoost Classifier

pre-processing to address missing values and reduce feature space dimensionality. By addressing class imbalance through strategic under sampling, the models were better balanced to identify the critical minority class—potential defaulters.

The ensemble of models—Logistic Regression, Naive Bayes, SVM, KNN, XGBoost, and Random Forest—was evaluated, with a particular emphasis on optimizing the recall metric. This focus was essential given the high cost associated with failing to predict actual defaults. Feature selection, operated by the Random Forest importance metric, was important in enhancing model performance by concentrating on the most influential variables and thus mitigating the curse of dimensionality.

Further refinement was achieved through hyper parameter tuning using GridSearchCV, which fine-tuned the models to their optimal performance settings. The project's outcomes have highlighted the potential of machine learning in transforming traditional credit risk assessment, offering predictions that are both reliable and scalable.

Finally, the project highlights the importance of continuous improvement and adaptation of the models to reflect new data and emerging trends in borrower behaviour. Future work may explore the integration of additional data sources, the application of more complex ensemble methods, or the deployment of these models in real-time lending decisions, thus enhancing their practical utility in dynamic financial environments. This study contributes to academic knowledge and also provides actionable insights for financial institutions aiming to enhance their risk assessment capabilities.

REFERENCES

[1] Petrazzini, B.O., Naya, H., Lopez-Bello, F. et al. Evaluation of different approaches for missing data imputation on features associated to genomic data. BioData Mining 14, 44 (2021).

[2] Adnan, F.A., Jamaludin, K.R., Wan Muhamad, W.Z.A. et al. A review of the current publication trends on missing data imputation over three decades: direction and future research. Neural Comput & Applic 34, 18325–18340 (2022).

[3] Lin, WC., Tsai, CF. Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev 53, 1487–1509 (2020).

[4] Goswami, M., Sebastian, N.J. (2022). Performance Analysis of Logistic Regression, KNN, SVM, Naïve Bayes Classifier for Healthcare Application During COVID-19. In: Raj, J.S., Kamel, K., Lafata, P. (eds) Innovative Data Communication Technologies and Application. Lecture Notes on Data Engineering and Communications Technologies, vol 96.

[5] Badhe, T., Borde, J., Thakur, V., Waghmare, B., Chaudhari, A. (2022). Comparison of Different Machine Learning Methods to Detect Fake News. In: Abraham, A., et al. Innovations in Bio-Inspired Computing and Applications. IBICA 2021. Lecture Notes in Networks and Systems, vol 419.

[6] Zhou, Yuran. "Loan Default Prediction Based on Machine Learning Methods." Proceedings of the 3rd International Conference on Big Data Economy and Information Management, BDEIM 2022, December 2-3, 2022, Zhengzhou, China. 2023.

[7] https://www.kaggle.com/datasets/yasserh/loan-default-dataset