# Predicting Loan Defaulters using Machine Learning

**Mrunmai Patil #13253**
School of Computing
Dublin City University
Dublin, Ireland
*mrunmai.patil2@mail.dcu.ie*

**Atharva Patil #11866**
School of Computing
Dublin City University
Dublin, Ireland
*atharvarajkumar.patil2@mail.dcu.ie*

**Yash Katkamwar #11604**
School of Computing
Dublin City University
Dublin, Ireland
*yash.katkamwar2@mail.dcu.ie*

**Bhakti Khot #10542**
School of Computing
Dublin City University
Dublin, Ireland
*bhakti.khot2@mail.dcu.ie*

***Abstract*—Loan defaulter prediction represents a critical business need for minimizing financial risks present in lending operations. The research aims to build machine learning detection models for loan defaulters through solutions for the class imbalance problem. The research used three resampling techniques including Random Undersampling and SMOTE and ADASYN to handle class imbalance in defaulters and non-defaulters data from Kaggle's public database. Our study included models: Random Forest, Logistic Regression, XGBoost and Gradient Boosting Classifier (GBC). The XGBoost model combined with ADASYN resampling under GridSearchCV delivered remarkable recall at 0.99 along with excellent precision levels which was identically to GBC. The performance of the two models was outstanding while we observed potential overfitting issue. The implementation of GridSearchCV on Logistic Regression with ADASYN generated deployable outcomes because it reached a precision value of 0.87 together with recall of 0.78 while maintaining a clear output interpretation.**

***Index Terms*—Loan default prediction, Machine Learning, SMOTE, XGBoost, ADASYN, Random Forest, Gradient Boosting classifier, Undersampling, Credit Risk Assessment.**

## I. Introduction

With the rapid advancement of big data and digital financial services, online lending platforms have transformed the borrowing and lending landscape. Peer-to-Peer (P2P) lending has become an essential alternative to traditional banking, allowing individuals and small businesses to access loans quickly and efficiently. However, as the industry grows, so do the risks—particularly the risk of loan defaults, which can lead to financial losses for investors and instability for lending platforms [1], [2].

To address this challenge, our research focuses on predicting loan defaults using machine learning techniques. By leveraging Random Forest, XGBoost, Gradient Boosting Classifier, and Logistic Regression, we aim to build an accurate and reliable prediction model [1], [3], [4], [5]. A major issue in loan default prediction is the imbalance of data, where the number of default cases is significantly lower than non-default cases. To overcome this, we utilize advanced resampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique), ADASYN (Adaptive Synthetic Sampling), and Random Undersampling [3], [6], [10].

The dataset used for this study consists of loan application records from a financial institution. It includes various borrower attributes, loan details, credit history, and financial metrics that help predict whether a borrower will default on a loan. The dataset is sourced from the Kaggle repository and is a comprehensive collection of financial and demographic information on loan applicants. It contains over 400,000 records spanning from 2007 to 2018, across 25 columns, with both numerical and categorical variables [13]. Key features include loan amount, interest rate, debt-to-income ratio, recurring balance, and annual income, which are critical in identifying potential defaulters.

We follow the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology in this study, which offers a structured approach to data mining projects—from business understanding to model evaluation. This methodology aligns well with our research as it ensures systematic progression through data preprocessing, modeling, and validation phases, enhancing the reproducibility and reliability of results [5], [6], [11].

We provide a concise overview of related work, reviewing twelve research papers relevant to loan default prediction. nine of these are highly rated, offering strong theoretical and empirical grounding [1], [3], [5], [6], [2], [4], [7],[8], [9], while the remaining three serve as foundational references, [11],[10], [12]. Each paper is summarized with its contributions, methodologies, and models to help contextualize our approach within existing literature.

The methodology section outlines key steps: data preprocessing (including handling missing values and feature engineering), feature selection via Random Forest Feature Importance, and resampling techniques (SMOTE, ADASYN, and

Random Undersampling). It also covers model development and optimization using Logistic Regression, Random Forest, XGBoost, and Gradient Boosting Classifier, with tuning performed through GridSearchCV.

To evaluate our models, we prioritized metrics beyond simple accuracy due to class imbalance, focusing on precision and especially recall critical for minimizing false negatives. Our evaluation highlights Gradient Boosting and XGBoost as top performers in terms of recall (0.99), with ADASYN proving the most effective resampling technique. Despite the high performance, some models showed tendencies toward overfitting, underlining the need for careful deployment and consideration of generalizability in real-world settings.

## II. Related Work

Predicting loan defaults is a critical challenge in the financial industry, influencing lending decisions, risk assessment, and overall economic stability. Banks and financial institutions have increasingly turned to machine learning (ML) and deep learning (DL) models to enhance prediction accuracy and mitigate risks. The evolution of these models has led to significant advancements, but challenges related to data limitations, model generalizability, and feature selection still persist.

Different studies have tried to apply ML and DL for predicting loan default. Lakshmanarao et al. (2023) [1] used a combination of traditional ML models and ANNs for loan default prediction. Despite their study being of high accuracy (99.4), it was marred by the fact that it was not tested on other datasets, which questioned the model's robustness and generalizability. Similarly, Lai (2020) [2] contrasted ML approaches, observing their effectiveness in credit risk assessment. The study did not, however, involve a comparison with deep learning models, which are increasingly demonstrating higher efficiency in detecting complex patterns in financial information.

Another important contribution by Moscato et al. (2021) [3] contrasted various ML models for credit scoring. While this research provided a general comparison, it did not address imbalanced datasets—a big issue in loan default prediction where the number of defaulters is usually much smaller compared to non-defaulters. To this, Zhu et al. (2019) [4] used a random forest algorithm, showing its utility in prediction, but their study did not incorporate advanced preprocessing techniques like synthetic minority over-sampling (SMOTE) or adaptive synthetic sampling (ADASYN), which are instrumental in handling data imbalance effectively.

To enhance explainability, Xu Zhu et al. (2023) [5] suggested explainable ML models for loan default prediction. It made the financial institution's decision-making process clear. Explainability and accuracy don't go hand in hand, and the limitation that most highly interpretable models are less accurate than deep learning models. Similarly, Aruleba and Sun (2024) [6] used ensemble classifiers with model explanation and achieved improved credit risk prediction. Ensemble models however though continuously being enhanced are computer-hungrier and thus less trendy to execute in actual time.

Despite these advancements, there are still some limitations in loan default prediction studies. Nalić et al. (2020) [7] created a hybrid data mining approach, improving feature selection and classification effectiveness. But like in most studies, they didn't do testing on multiple datasets, so the generalizability of their result is limited. Similarly, Cheng et al.(2021) [8] also sought to keep low feature requirements to generate a lean model but powerful at the expense of sacrificing information.

A common limitation in most of the studies is the lack of external economic data in model estimates. Li and Wu (2024) [11] conceded this limitation, stating that due to data constraints, their work was not able to examine the impact of regional economies on defaults. This is one gap where there is a lack of research—models can predict defaults using borrower behavior but cannot incorporate macroeconomic variables, which are also important in ascertaining default risks.

Besides, the majority of the studies have classified loans as "fully paid" or "defaulted" and ignored risky loans that are not paid in full or defaulted but are exhibiting signs of distress. Owusu et al. (2023) [10] indicated that this category needs to be examined by future studies so that more will be known regarding loan risks.

While deep learning approaches, such as ANNs, have shown better performance, they are not transparent. Whereas decision trees or logistic regression-based models are interpretable and can tell us why a loan has been rejected, ANNs are black boxes that do not even enable financial institutions to give reasons for rejecting loans. Efforts directed at tackling this loophole, such as trying out SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-Agnostic Explanations), are yet to find their place in loan default prediction literature. To sum up, while existing research has made notable advancements in loan default prediction, there remain crucial challenges to be addressed. These involve model generalizability, data limitation, the exclusion of economic variables, and the need for interpretable models that need to be addressed to develop more stable and industry-specific solutions. Future research must prioritize multi-dataset validation, external economic variable utilization, and model development for handling borderline loan cases. With their completion, default prediction models on loans can prove to be stronger, more transparent, and useful for real-life application.

## III. METHODOLOGY

### A. Dataset Overview

This study analyzes financial institution loan application records as its dataset. The predictive model includes different characteristics of borrowers and loan specifics as well as their credit history and financial records to estimate the probability of default. The financial and demographic information about

loan applicants is obtained from the Kaggle repository collection. The database has 400,000 rows and 25 columns [13]. The available data contains both quantitative and qualitative features among its elements. The primary attributes of the dataset include loan amount and interest rate combined with debt-to-income ratio and recurring balance and annual income that determine loan default risk. The dataset features numerical variables with both continuous and discrete numerical quantities that deliver quantitative information regarding loan details and borrower financial standing as well as credit risk exposure. The assessment variables strongly influence the chances that borrowers will default on their loans. The dataset contains categorical variables that describe qualitative aspects which help explain borrower actions and financial standing. Machine learning models need data encoding procedures (label encoding and one-hot encoding) to work with these variables.

### *B. Exploratory Data Analysis*

The dataset is from 2007 to 2018 for over 400,000 records in 25 columns, which is mainly focused on finding the key attributes to recognize the loan defaulters.

Exploratory Data Analysis (EDA) is performed to achieve the shape of the dataset, assess data quality, and discover main patterns. First, the dataset was probed for missing values, outliers, and data distributions for various numerical and categorical features. It helps identify important variables, outlier detection, and hypothesis development regarding the data.

To have a clearer picture of the credit history and financial behavior of the applicants, some numerical features were plotted. These are interest rate, yearly income, amount funded, revolving balance, debt-to-income ratio (DTI), and FICO average score. Among them, the FICO Average Score was computed by averaging the provided fico-range-low and fico-range-high values in the dataset. This value serves as a proxy for the applicant's creditworthiness and is a strong indicator in loan approval decisions.

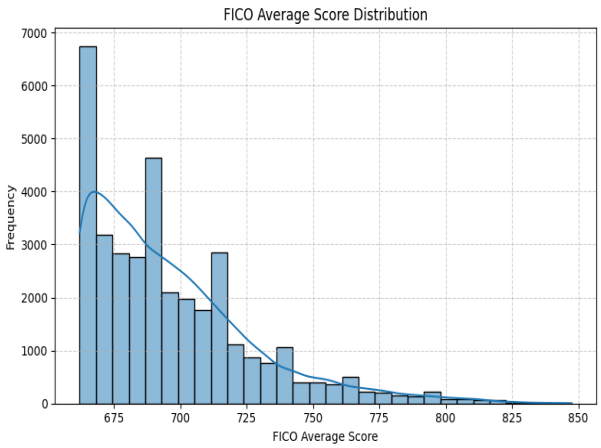


Fig. 1. Histogram of FICO Average Score of loan applicants

The distribution plot of the FICO Average Score (Figure 1) revealed that most applicants fall within the 660 to 750 score range, suggesting that the majority possess fair to good credit. This concentration in the middle-to-top bucket of the credit score could prove significant for default prediction since the higher-scoring applicants will have less risk to offer. The chart slopes gently to the left side, suggesting fewer of the poor-credit individuals.

### *C. Data Cleaning*

Before proceeding with analysis data cleaning functions as an essential preprocessing step which removes data inaccuracies as well as removes irrelevance and avoids misleading or redundant information. The original dataset contained many variables which did not relate to the predictive task and simultaneously presented data leakage risks. Several columns got removed during cleaning to improve model reliability along with generalization ability.

Unique identifiers such as member identification numbers and website URLs were discarded, as they held no predictive value. Similarly, geographical information including ZIP codes and state names, along with temporal features like the date of the borrower's earliest credit line and the number of months since the last delinquency, were excluded due to their potential to introduce bias or offer minimal contribution to model performance. The analysis dropped the credit utilization ratio and listing status features because they displayed minimal relationships to loan default.

To prevent data leakage, post loan information such as outstanding principal, total payment, and recovered amounts are eliminated, as these reflect outcomes that would not be known at the time of loan approval. Other low impact variables, such as policy codes and the type of application submitted, are removed due to their limited influence on predictions. Details related to joint applicants, like combined annual income and debt-to-income ratio, are not included, as the focus was on individual loan assessments. Hardship-related attributes, including the status of hardship and reasons for settlement, are excluded to maintain a clear emphasis on pre-approval factors. This comprehensive cleaning process helped establish a clean, unbiased dataset suitable for accurate loan defaulter prediction.

### *D. Data Pre-processing and Feature Engineering*

The machine learning model required target variable encoding and categorical feature encoding while handling missing data to achieve effectiveness.

*1) Encoding Loan Status for classification:* The loan status column contained multiple categories such as Fully Paid, Current, Charged Off, and Default. Two distinction groups were established for loans known as Good Loan for money still owed or being repaid without problems and Bad Loan for defaulted accounts and overdue payments. Multiple categories required separate classification when dealing with loans that were in grace periods or had ambiguous outcomes. A new

binary column was created for loan status, and the original categorical field was removed.

*2) Handling Missing Data:* A few columns contained missing data and Employment Length turned out to be one of them among those fields. The application of Rule-Based Data Transformation standardized categorical employment length values into numeric form to solve this issue. The values "10+ years" received the equivalent transformation to 10.5 years for representing extended work experience beyond ten years and "Less than 1 year" was changed to 0.5 years. The treatment converted the categorical values "5 years" to numeric form. The technique of Handling Missing Values combined with Parsing Textual Data was used to replace empty data entries with NaN while converting textual information into numerical values. The applied data transformation technique adopts the MAR (Missing at Random) assumption by showing a relationship of value absence probability to observed variables instead of the value being absent. The Employment Length column contains missing entries whose values depend on additional known attributes which include job title and industry sector as well as employment type. The missing data follows a Missing at Random (MAR) pattern that relies on registered features instead of missing employment duration records. Therefore rule-based transformations and text parsing methods prove both valid and practical to handle the missing data sections. By conducting this transformation, the dataset achieved better consistency and became more ready for machine learning model processing.

*3) Encoding Verification Status:* A set of rules determines the mapping process that examines particular loan status values to assign them numerical categories following predefined risk-related values. Numerical analysis of loan status requires conversion through a risk-based system which assigns numeric values to the categories. The values "Fully Paid" as well as "Current" received a mapping value of 1 as they demonstrated good loan characteristics. Status labels like "Charged Off" along with "Default" and "Overdue" received a numerical value of 0 which represented a bad loan. The transformation method matches Label Encoding (Ordinal Encoding) because the categorical loan statuses received integer labels (0, 1) through a defined algorithm.

*4) Dataset Splitting:* The dataset was split into training and testing sets using an 80/20 ratio, where 80% of the data was used for training the models and applying various sampling techniques such as SMOTE and ADASYN. The remaining 20% was reserved for testing to evaluate model performance on unseen data, ensuring reliable and unbiased validation.

*5) Feature Engineering:* The predictive model required distinctive transformed features which were generated from the initial dataset to boost forecasting accuracy. FICO average scores served as a crucial engineered feature which calculated the mean value between fico-range-low and fico-range-high scores for applicants. Lastly the model input became simplified because the creation of a single credit score representation through the FICO average score measure reduced redundancy.

### *E. CRISP-DM Framework: A Structured Approach to Loan Default Prediction*

The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology supplied us with a detailed project plan which maintained proper alignment between our tasks from beginning concept to model evaluation. Our core business understanding established the main goal of identifying loan defaulters to improve financial risk management. The analysis of the Kaggle dataset during Data Understanding allowed the team to identify essential variables that included loan amounts, FICO scores and debt-to-income ratios. During Data Preparation we cleaned the data and implemented FICO score averages and addressed missing points to boost our prediction capabilities. The Modeling phase included algorithm testing of Random Forest, XGBoost, and Gradient Boosting which integrated the SMOTE and ADASYN techniques to handle class imbalance. Recall proved more vital in Evaluation than precision because we wanted to prevent undetected defaulters due to financial risks. Our project remained focused with actionable outcomes by following the CRISP-DM process even though Deployment was not within scope.

### *F. Handling Class Imbalance in Dataset*

Class distribution imbalance was one of the major problems that we faced during defaulter identification, where defaulters (Class 1) were much more in number than non-defaulters (Class 0). Models became biased towards the majority class due to this imbalance, which negatively affected the defaulter identification, which is of utmost importance from the financial risk management point of view. We attempted several resampling techniques like Random Undersampling, SMOTE, and ADASYN in order to get rid of it. Random undersampling numercially balanced data but resulted in data loss and performance reduction. SMOTE performed a slight improvement by generating new instances for the minority class, but ADASYN performed optimally since it took more synthesis of samples in hard-to-learn defaulter instances, thereby making the model highly sensitive to class 1 without any chances of overfitting.

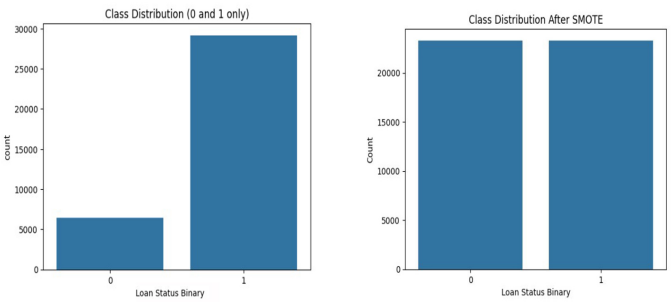


Fig. 2. Class 0 and 1 distribution before and after SMOTE Analysis

### G. Methods Overview

*1) Random Forest Classifier:* We began by employing Random Forest Classifier because it serves as an advanced ensemble learning tool which lowers overfitting and generates readable feature importance measures. We created the new fico-avg column from fico-range-high and fico-range-low because their high multicollinearity made them unsuitable for model input. Strong baseline performance emerged from our model yet the model failed to recall defaulting cases properly. SMOTE slightly enhanced the model output with a precision value of 0.84 and a recall value of 0.93, after adjustments with Random Undersampling offered better precision at 0.90 at the expense of lowered recall at 0.60. The evaluation results from ADASYN demonstrated maximum improvements because it achieved both the best recall level of 0.92 thereby illustrating the significance of resampling techniques for enhancing Random Forest models operating under imbalanced conditions.

*2) Logistic Regression:* We used Logistic Regression as our predictor model because it provides interpretability and complies with financial application regulations through its binary class probability classification nature. The model uses predictive feature linear combinations to compute default log-odds before generating outputs via the sigmoid function. Stakeholders could easily understand which features had most impact on loan default predictions because Logistic Regression works through linear mechanisms. The model failed to exhibit satisfactory results when analyzing unbalanced datasets. Even when using SMOTE and ADASYN resampling techniques the model showed average results where SMOTE achieved a precision rate of 0.87 and recall rate of 0.72 but ADASYN reached 0.87 precision and 0.78 recall. Despite its ease of explanation, Logistic Regression could not model complex nonlinear relationships in the data, which limited its effectiveness compared to ensemble models.

*3) XGBoost Classifier:* To overcome the linearity limitations of previous models, we utilized XGBoost because of its superior gradient boosting abilities and its ability to handle imbalanced data sets. This advanced technique shows outstanding complex feature detection and management of data distributions. The sequential operation of XGBoost utilizes weak learner improvements through identifying mistakes made during preceding rounds of execution. Application of ADASYN enabled the model to demonstrate superior performance with 0.96 recall rate and 0.84 precision. Our experimental data showed these results belonged to the highest values measured in this study. XGBoost showed lower susceptibility to overfitting when compared to Random Forest. The model required specialized tuning before training but achieved improved performance which made it stand as the best baseline for precision-recall trade-off.

*4) Gradient Boosting Classifier:* The Gradient Boosting Classifier represents our approach for improving on XGBoost through enhanced interpretability alongside simplified training process. Every successive tree learns from previous errors to construct the model sequentially. The application of ADASYN transformed Gradient Boosting so it became more effective at default detection. The performance of the model reached its peak after executing hyperparameter tuning through GridSearchCV. The Gradient Boosting model reached both precision 0.83 and recall 0.93 for the best performance among the models we evaluated throughout our work. Our main goal focused on correct defaulter identification (Class 1) through effective risk prediction prevention in financial lending operations. The attractive relationship between model reliability and interpretability and recall performance established the model as an ideal answer to our loan default prediction issue.

## IV. EVALUATION

Since defaulters represent the biggest financial risk, our entire approach focused on identifying them as accurately as possible. The high imbalance in our dataset made accuracy measures insufficient so our team settled on precision and recall standards as our primary evaluation criteria. Our team deemed recall the most crucial metric among all the others we assessed. The identification of defectors represented our main concern because false negative errors would lead to significant financial loss. We handled dataset imbalance through the combination of SMOTE, ADASYN and Random Undersampling techniques. The majority effectiveness came from using SMOTE and ADASYN techniques.

The Gradient Boosting Classifier achieved better defaulter identification when we applied GridSearchCV tuning with those selected parameters thereby demonstrating significant performance enhancement. The updated model produced the most effective results because it accomplished a precision level of 0.83 and maintained the greatest recall score of 0.99 compared to other models. The model maintains an excellent ability to detect defaulters with practical precision levels thus becoming a dependable risk-aware selection.

The combination of GridSearchCV and XGBoost models using ADASYN generated parallel effective results which led to a 0.99 recall at high precision rate thus establishing this option as a strong GBC replacement for this classification task. After fine-tuning both Gradient Boosting and XGBoost algorithms demonstrated strong performance which might lead to overfitting problems affecting their generalization capabilities on new datasets in the future.

The tuned Logistic Regression model employing ADASYN produced satisfactory results for default prediction by reaching precision of 0.87 alongside recall of 0.78 and F1 score of 0.82 for Class 1. The model performed favorably regarding its identification of high-risk borrowers since it correctly identified non defaulters (Class 0) with precision at 0.34 and recall at 0.36 even though these metrics were not as high as the others. Random Forest provided average performance under different resampling approaches so it became less effective in aiding our prediction needs.

| Model | Method/Algorithm | Precision (P) | Recall (R) | Accuracy (A) | F1 Score (F) |
|---|---|---|---|---|---|
| Random Forest | Random Undersampling | 0.90 | 0.60 | 0.61 | 0.72 |
| | SMOTE | 0.84 | 0.93 | 0.79 | 0.88 |
| | ADASYN | 0.84 | 0.92 | 0.79 | 0.88 |
| Logistic Regression | Random Undersampling | 0.89 | 0.66 | 0.65 | 0.76 |
| | SMOTE | 0.87 | 0.72 | 0.68 | 0.79 |
| | ADASYN | 0.87 | 0.78 | 0.71 | 0.82 |
| | GridSearch | 0.86 | 0.85 | 0.71 | 0.86 |
| XGBoost | Random Undersampling | 0.88 | 0.60 | 0.60 | 0.72 |
| | SMOTE | 0.84 | 0.96 | 0.81 | 0.90 |
| | ADASYN | 0.84 | 0.96 | 0.81 | 0.90 |
| | GridSearch | 0.83 | 0.99 | 0.82 | 0.91 |
| Gradient Boosting | Random Undersampling | 0.90 | 0.61 | 0.62 | 0.73 |
| | SMOTE | 0.84 | 0.93 | 0.79 | 0.88 |
| | ADASYN | 0.84 | 0.91 | 0.79 | 0.88 |
| | GridSearch | 0.83 | 0.99 | 0.82 | 0.91 |

Fig. 3. Evaluation Results

## V. CONCLUSION

This project focused on building a machine learning model to predict loan defaulters represented the main objective of this project because it helps reduce financial risks. The loan-status column underwent conversion into a two-value system while we established recall as the fundamental goal to detect defaulting clients accurately. The project handled class imbalance by applying the SMOTE and ADASYN and Random Undersampling approaches to Logistic Regression, Random Forest, XGBoost, and Gradient Boosting classifications. Boosting models beat their counterparts and provided good recall performance after tuning yet indicated their vulnerability to overfitting that could affect future deployment applications. Logistic Regression proved beneficial for practical deployment since it provided more stable and interpretation friendly results in contrast to other methods. By including individual borrower features such as credit trends and external scores the model will demonstrate better performance in various loan applications. The use of stacking combined with cost-sensitive learning and deep learning methods could potentially enhance performance specifically for intricate and borderline loan situations.

**Github Link to the code and dataset.**

## REFERENCES


[1] A. Lakshmanarao, C. Gupta, C. S. Koppireddy, U. V. Ramesh and D. R. Dev, "Loan Default Prediction Using Machine Learning Techniques and Deep Learning ANN Model," 2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems , pp. 1-5, doi: 10.1109

[2] L. Lai, Loan Default Prediction with Machine Learning Techniques," 2020 International Conference on Computer Communication and Network Security (CCNS), Xi'an, China, 2020, pp. 5-9, doi: 10.1109

[3] Vincenzo Moscato, Antonio Picariello, Giancarlo Sperlí, A benchmark of machine learning approaches for credit score prediction, Expert Systems with Applications, Volume 165, 2021, 113986, ISSN 0957-4174,

[4] Lin Zhu, Dafeng Qiu, Daji Ergu, Cai Ying, Kuiyi Liu, A study on predicting loan default based on the random forest algorithm, Procedia Computer Science, Volume 162, 2019, Pages 503-513, ISSN 1877-0509,

[5] Xu Zhu, Qingyong Chu, Xinchang Song, Ping Hu, Lu Peng, Explainable prediction of loan default based on machine learning models, Data Science and Management, Volume 6, Issue 3, 2023, Pages 123-133, ISSN 2666-7649

[6] I. Aruleba and Y. Sun, "Effective Credit Risk Prediction Using Ensemble Classifiers With Model Explanation," in IEEE Access, vol. 12, pp. 115015-115025, 2024, doi: 10.1109.

[7] IJasmina Nalić, Goran Martinović, Drago Žagar, New hybrid data mining model for credit scoring based on feature selection algorithm and ensemble classifiers, Advanced Engineering Informatics, Volume 45, 2020, 101130, ISSN 1474-0346,

[8] Cheng, Yun-Chieh, et al. "Predicting credit risk in peer-to-peer lending: A machine learning approach with few features." 2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI). IEEE, 2021.

[9] Zhou, Lifeng, and Hong Wang. "Loan default prediction on large imbalanced data using random forests." TELKOMNIKA Indonesian Journal of Electrical Engineering 10.6 (2012): 1519-1525.

[10] Owusu, Ebenezer, et al. "A deep learning approach for loan default prediction using imbalanced dataset." International Journal of Intelligent Information Technologies (IJIIT) 19.1 (2023): 1-16.

[11] Li, Huan, and Weixing Wu. "Loan default predictability with explainable machine learning." Finance Research Letters 60 (2024): 104867.

[12] Uwais, Aiman Muhammad, and Hamidreza Khaleghzadeh. "Loan default prediction using spark machine learning algorithms." AIAI 29th Irish Conference on Artificial Intelligence and Cognitive Science. CEUR Workshop Proceedings, 2022.

[13] Dataset: https://www.kaggle.com/datasets/krishnamurthi1703/housing-datasetloan-accepted-and-rejected