Machine Learning-Based Loan Default Prediction in Peer-to-Peer Lending

Ruyi Yang*

Tongji University, Shanghai, China

*Corresponding author: 1953151@tongji.edu.cn

Abstract. The peer-to-peer (P2P) lending market has recently undergone significant growth, transforming traditional lending practices. However, this evolution brings with it unique challenges, particularly in managing credit risk and ensuring the reliability of loan approvals. Accurate prediction of loan defaults remains a pivotal aspect of risk management in this sector. This study introduces a comprehensive approach to improve bad loan prediction in peer-to-peer (P2P) lending, sourcing Lending Club data. In the face of challenges posed by imbalanced datasets and risk management in the loan industry, our methodology significantly enhances prediction accuracy, particularly in identifying bad loans. The study implements a comprehensive process that includes data cleaning, feature engineering, feature selection, balancing the dataset and machine learning models, achieving a noteworthy accuracy rate over 92% and recall rate above 87%. This research advances academic understanding of loan prediction and generates real-world impact. The accurate models identifying explanatory provide a valuable framework for improving decision-making and strategic planning in P2P lending platforms.

Keywords: Loan Default Prediction; Machine Learning; Feature Engineering.

1. Introduction

Loan prediction has become a pivotal aspect in the financial sector, especially in peer-to-peer (P2P) lending platforms like Lending Club. This importance is rooted in the need to accurately assess the risk of loan defaults, thereby safeguarding the interests of both lenders and borrowers. Accurate loan predictions not only ensure the financial stability of lending institutions but also enhance customer trust and market competitiveness.

In recent years, various methods have been employed to improve the accuracy of loan predictions. Novel methods in loan prediction have started incorporating algorithms like Random Forests, Neural Network, LightGBM, XGboostand Gradient Boosting, as evidenced by recent studies [1-2]. These advanced methodologies not only offer higher accuracy but also provide nuanced insights into the complex dynamics of loan default risks. Moreover, the CART decision-tree model [3], integration of LSTM neural networks [4] and a blend of algorithms that combine logistic regression, Random Forest, and CatBoost [5] has emerged, leveraging the strengths of novel models.

Gradient tree boosting implemented on the Tobit model (Grabit) [6] , the integration of support vector machines (SVM) with Dempster–Shafer theory [7], the development of heterogeneous stacking ensemble (HSE) models [8], and the application of spatial network distance analysis[9] represent the innovative strides in this field.

The first stage of loan prediction focuses on the likelihood of an application being denied, while the second stage, which is the focus of this research, scrutinizes the risk of default for approved loans [10].

This research paper offers significant contributions to the field of loan prediction, particularly using Lending Club data. Firstly, it showcases an exemplary recall rate of over 87%, a result of a comprehensive methodology that includes data cleaning, feature engineering, feature selection, and sophisticated modeling. This achievement is particularly notable given the challenge of imbalanced datasets in the loan industry, where defaults are rarer than non-defaults. Secondly, the study delves deep into the modeling process, identifying which steps are critical for enhancing recall in such imbalanced datasets. This investigation is not just academically significant but offers real-world

implications, providing valuable insights for financial institutions aiming to refine their predictive models for loan defaults. Lastly, the paper uncovers key factors that determine the quality of loans. This insight is immensely beneficial for the business sector, aiding in more informed decision-making and strategic planning.

2. Methodology

The methodology of this study systematically navigates through various stages of data analysis and model development using Lending Club's dataset [11].

The study encompasses extensive data preprocessing, including removing exclusions, removing outlier and missing value imputation, followed by a correlation analysis to generate overarching insights. Meticulous feature engineering techniques including binary encoding, target encoding, and normalization are applied. Feature selection processes are undertaken to streamline variables, utilizing both variance threshold filtering and stepwise selection wrapper methods to streamline variables. Stratified shuffle splitting in test-train set and undersampling are employed to address the class imbalance. The study explores various machine learning models, including Logistic Regression, Decision Trees, KNN, Random Forests, Gradient Boosting, Light GBM, XGBoost. In addition, a bagging and stacking approach was incorporated(as shown in the Fig. 1).

Crucially, the effectiveness of these models is demonstrated by their performance metrics, as will be detailed, demonstrate the efficiency and accuracy of these models in predicting loan defaults.

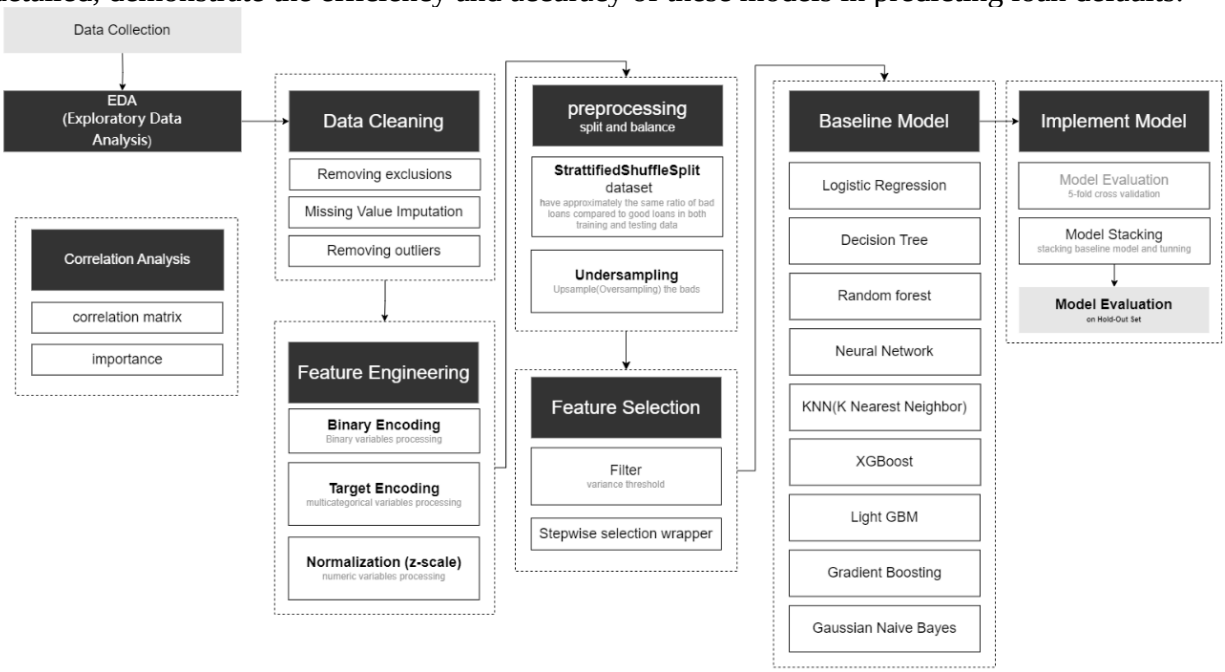

Fig. 1 Technical Route

2.1. Data Description and Data Cleaning

2.1.1. Data Description

This study's data was sourced from Lending Club, a peer-to-peer lending platform based in the U.S., which assesses loans and assigns credit grades, reflecting the risk and potential default rate.

The dataset employed in this analysis originally consisted of 100,000 records sampled from Lending Club data, encompassing 151 features. This includes 113 numerical variables and 38 categorical variables. Table 1 is a partial description of the characteristics.

Table 1. Data Variable Description Sample

feature	Description
last_fico_range_high	The maximum range within which the borrower's most recent FICO check. falls.
last_fico_range_low	The minimum range within which the borrower‘s most recent FICO check. falls.
annual_inc	The self-reported annual income provided by the borrower during registration.
acc_now_delinq	The quantity of accounts that are now past due by the borrower.
delinq_amnt	The amount of loan on which the borrower is delinquent.
last_pymnt_amnt	Last total payment amount received
total_rec_prncp	Principal received to date

2.1.2. Data Cleaning

In preparing the Lending Club dataset for analysis, a comprehensive data cleaning and preprocessing approach was undertaken, including removing exclusions, missing value imputation and removing outlier.

Removing exclusions involved the removal of loans categorized as *'Current'* and *'Issued'*, which do not provide clear indications of loan quality. Loans labeled as *'Fully Paid'* and *'Does not meet the credit policy. Status: Fully Paid'* were classified as Good loans, while all others were categorized as Bad loans. After excluding loans with a *'Current'* loan status, many features exhibit significant differences in *'loan_condition'*, which is a positive signal (as shown in the Fig. 2). Variables exhibiting more than 80% missing values were eliminated. And repetitive and dynamically changing variables are deleted.

Furthermore, missing values in the remaining dataset were addressed through appropriate imputation methods. Finally, outliers in the dataset were also removed to enhance the models' robustness and accuracy. Custom thresholds were set for specific features, as the data did not adhere to a normal distribution. This nuanced approach to preprocessing was essential for refining the dataset, ensuring its suitability for in-depth analysis.

It is worth noting that a critical variable in this dataset is *'loan_condition'*, which categorizes the status of loans into 'Good *Loan'* (0) and *'Bad Loan'* (1). It is noted that *'Bad Loans'* constitute approximately 20% of the total loans in the dataset, a significant proportion that underscores the importance of accurate loan default prediction.

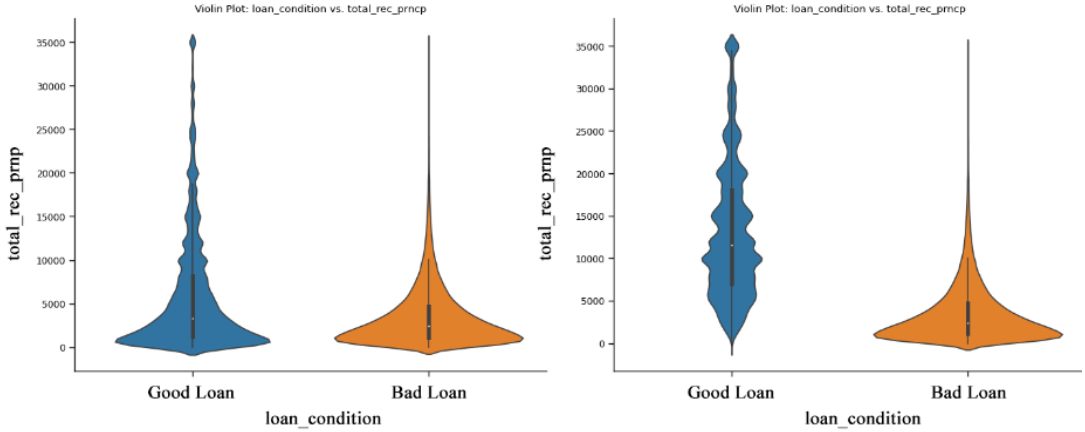

Fig. 2 Distribution of Principal Received by Loan Condition Before and After Data Cleaning (Left: Before, Right: After)

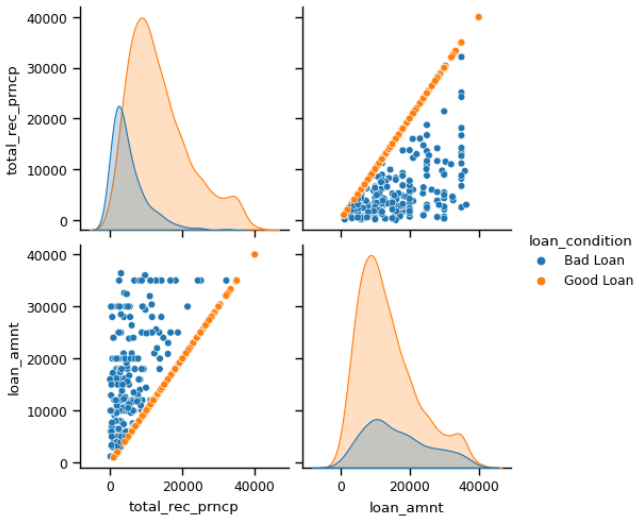

Fig. 3 Principal Received vs. Loan Amount: A Clear yet Risky Indicator of Account Status

One often-overlooked aspect is the necessity of excluding dynamic variables. For instance, in cases where loans have already been categorized as either good or bad, variables such as *'total_rec_prncp'* (Principal received to date) and *'loan_amount'* can directly indicate the loan's quality (as shown in the Fig. 3).

$$\begin{cases} Good\ Loan, & if\ total\_rec\_prncp = loan\_amount \\ Bad\ Loan, & if\ total\_rec\_prncp < loan\_amount \end{cases} \quad (1)$$

This implies that *'total_rec_prncp'* is a dynamic variable that is consistently lower than *'loan_amount'* within the repayment period. Consequently, these variables should not be included in the model training process.

Similar considerations apply to variables such as *'collection_recovery_fee,' 'last_pymnt_amnt,' 'out_prncp,' 'out_prncp_inv,'* 'recoveries,' *'total_pymnt,' 'total_pymnt_inv,' 'total_rec_int,'* and *'total_rec_late_fee'*. All of these variables should be removed from the analysis.

2.2. Correlation and Importance

A correlation matrix plot was employed to examine the interrelationships between variables in the Lending Club dataset, accompanied by an importance plot to specifically investigate the relationships of variables with the dependent variable specifically. Key observations include a notable negative correlation of *last_fico_range_high' and 'last_fico_range_low'* with dependent variable *'loan_condition_int'* (as shown in Fig. 4), while the two exhibiting a strong correlation between themselves. These correlations are crucial in understanding the factors influencing loan conditions.

Additionally, examining the relationships between variables proves beneficial in conducting exploratory analysis, aiding in the discovery of connections among variables that can be transformed into valuable business insights. As an illustration, the study has selected the top 20 variables (as shown in Fig. 5) based on the absolute values of their correlations with *'loan_condition'* to observe these relationships and coupling effects.

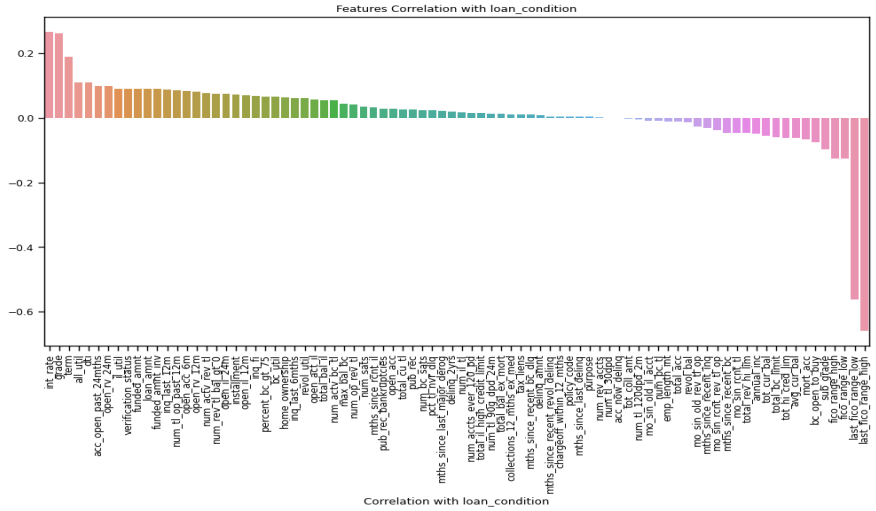

Fig. 5 Variable Importance to Loan Condition

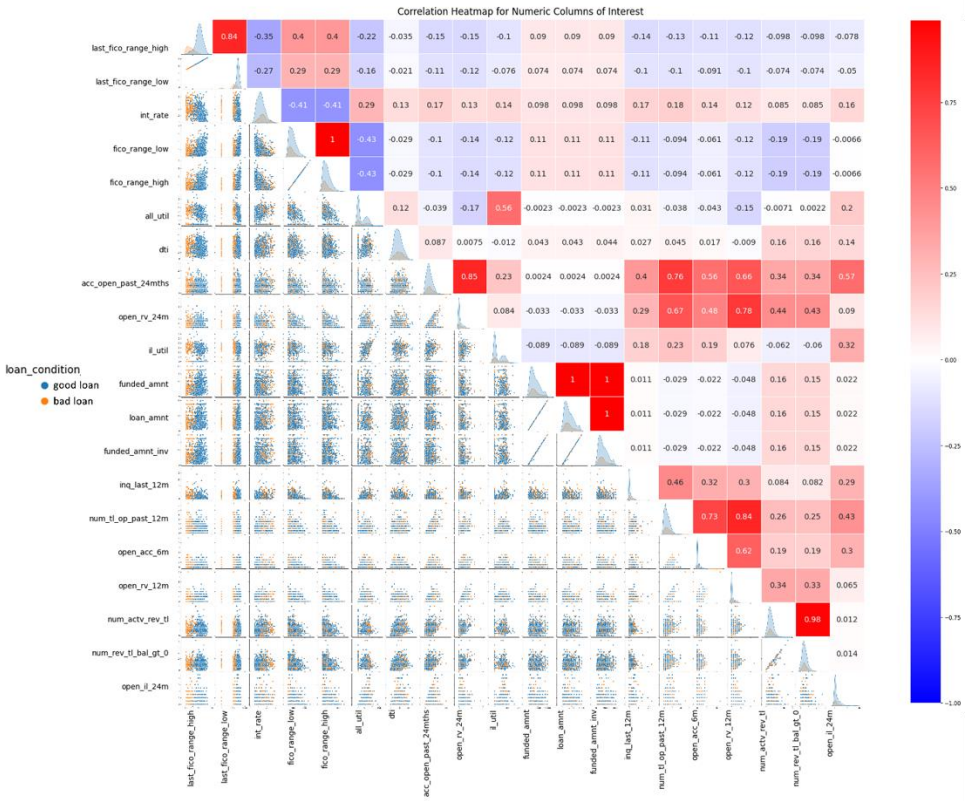

Fig. 6 Correlation Analysis (Top 20 variables)

2.3. Feature Engineering and Feature Selection

2.3.1. Feature Engineering

The process of transforming raw data into features that are appropriate for machine learning models is known as feature engineering. Feature engineering consists of creation, transformation, extraction, and selection of features. In the process of feature engineering, several essential steps are undertaken to enhance the quality of our data.

1) Binary Encoding for Binary (Boolean) Variables
2)Target Encoding for Multicategorical Variables.
3) Normalization for Numeric Variables.

When implementing target encoding, a smoothing method is employed to mitigate the risk of overfitting. This technique enhances the robustness of the encoding process and contributes to more accurate model performance. There are several approaches to normalizing, such as log transformation, z-score normalization, and min-max scaling. Z-score scaling is used in this case.

2.3.2. Feature Selection

1)Filter (Variance Threshold)
2) Wrapper (Forward Selection Wrapper)

Common filter feature selection methods include correlation coefficient, variance threshold, chi-square test, LDA (linear discriminant analysis), ANOVA, mutual information, etc. Variance threshold set was used, reducing the number of features from 51 to 26. A total of 26 features then underwent stepwise selection wrapper. Stepwise selection wrapper is a feature selection method commonly used in machine learning and statistical modeling, whose purpose is to optimize the performance of the model or simplify the model by gradually adding or removing features, with the model's performance varying as variables are introduced. Based on the process chart, the subsequent model construction will focus on the top 5 selected variables, being *'last_fico_range_high', 'last_fico_range_low', 'acc_now_delinq', 'delinq_amnt', 'num_tl_30dpd'.*

2.3.3. Dealing with Imbalanced Dataset

In the dataset, we encounter an inherent imbalance, with approximately 80% of the loans categorized as GOOD. Our primary focus lies in ensuring the accuracy of predicting BAD loans, underscoring the critical importance of balancing the data.

Balancing the data can be achieved through oversampling or undersampling techniques. Both random undersampling and SMOTE (oversampling) methods were tested out, and it was observed that the two methods had similar effects on the model performance. Given the substantial size of our dataset, a preference was given to undersampling, as it retains data authenticity while effectively addressing the class imbalance.

2.4. Modelling and Stacking

Here are the machine learning models and ensemble algorithm utilized:

Logistic Regression：Logistic Regression is widely used in many fields and can be used as a baseline model for many classification tasks.

Decision Tree: Decision Trees, a method in non-parametric supervised learning, are employed for both classification and regression tasks. A model that forecasts the value of a target variable is created through the assimilation of simple decision-making rules, which are derived from the features of the data.

K Nearest Neighbors: K-Nearest Neighbors (KNN) is a straightforward supervised machine learning technique. In the KNN algorithm, data points are plotted in n-dimensional space where n is the number of features. To classify a new point, KNN identifies its K nearest neighbors and assigns the point to the class that appears most frequently within those neighbors.

Random Forest: To increase prediction accuracy and reduce over-fitting, Random Forest builds many decision trees using different dataset subsamples and averages the results.

Gaussian Naive Bayes: In Gaussian Naive Bayes, probabilities of each class are calculated using Bayes' theorem. The naive assumption made is that the features are conditionally independent given the class label. It is called Gaussian because it assumes that numerical features are distributed according to a normal or Gaussian distribution. It doesn’t work well in this case, as most variables are not distributed according to a normal distribution.

Light GBM: LightGBM is a high-performance gradient boosting framework based on decision tree algorithms. It is known for its speed and efficiency. Unlike traditional tree-based algorithms, LightGBM grows trees vertically, choosing the leaf with the maximum loss reduction.

XGBoost: XGBoost is an enhanced Gradient Boosting Machine Learning package that improves model generalization using sophisticated regularization approaches. The core XGBoost algorithm is parallelizable, meaning it can be parallelized within a single tree.

Gradient Boosting: Gradient Boosting builds a prediction model by combining a series of weak prediction models in an iterative manner, to create a strong, accurate model for regression and classification tasks.

Neural Network: Neural networks, often referred to as artificial neural networks (ANNs), consist of three main types of layers including input layer, hidden layers, and output layer.

Bagging: Bagging, or Bootstrap Aggregating, is an ensemble technique where multiple models are trained on varied subsets of data (created through bootstrap sampling) and their predictions are averaged or voted on for the final output. This approach reduces variance and overfitting, making it effective for high-variance models.

Stacking: Stacking is an ensemble learning technique where multiple base models (like decision trees, neural networks) are trained on the same data and their predictions are used as input for a meta-learner model. This meta-learner then learns how to optimally combine these predictions to improve overall accuracy. This approach leverages the strengths of various models, allowing for more complex pattern recognition and robust predictions

3. Result

In this paper, the experiment is carried out by using 5-fold cross Validation. Notably, the performance scores on the test set closely align with the cross-validated scores achieved during the training phase, showcasing a recall rate surpassing 87% and exceptional accuracy, F1 score, AUC, and KS score, all exceeding 87%. This congruence is indicative of a positive outcome, suggesting that the models exhibit robustness and do not succumb to overfitting.

In this study, the models that exhibited superior fitting performance included Decision Tree (bagging), Random Forest (bagging), Light GBM (bagging), XGBoost (bagging) (as shown in Table 2. In contrast, the Gaussian Naive Bayes model demonstrated instability compared to the others. The Gaussian Naive Bayes model exhibited instability among all models. This can be attributed to its inherent suitability for datasets that conform to a normal distribution, a condition not met by selected variables in the loan dataset. Consequently, this model is deemed unsuitable here.

In this study, stacking was implemented on the baseline models. The results indicated that bagging slightly outperformed stacking in enhancing model performance.

Table 2. Model Performance on Testing Set (Baseline, After bagging or after Stacking)

index	accuracy	auc	ks	precision	recall	f1
Logistic Regression	0.8706	0.9281	0.7338	0.6611	0.8591	0.7472
Decision Tree	0.8610	0.9250	0.7307	0.6378	0.8691	0.7357
K Nearest Neighbors	0.8451	0.7823	0.4913	0.6977	0.5369	0.6068
Random Forest	0.8590	0.9262	0.7320	0.6320	0.8773	0.7347
Light GBM	0.8606	0.9293	0.7370	0.6350	0.8796	0.7375
XGBoost	0.8606	0.9287	0.7359	0.6351	0.8788	0.7374
Gradient Boosting	0.8606	0.9293	0.7361	0.6350	0.8784	0.7372
Neural Network	0.8638	0.9293	0.7368	0.6424	0.8755	0.7410
Logistic Regression-bagging	0.8708	0.9279	0.7340	0.6613	0.8598	0.7476
Decision Tre- bagging	0.8590	0.9271	0.7334	0.6315	0.8803	0.7354
K Nearest Neighbors-bagging	0.8583	0.9222	0.7294	0.6319	0.8706	0.7323
Random Forest-bagging	0.8602	0.9274	0.7346	0.6343	0.8788	0.7368
Light GBM-bagging	0.8606	0.9293	0.7375	0.6350	0.8788	0.7373
XGBoost-bagging	0.8609	0.9289	0.7366	0.6357	0.8788	0.7377
Gradient Boosting-bagging	0.8604	0.9292	0.7369	0.6346	0.8792	0.7371
Neural Network-bagging	0.8614	0.9292	0.7373	0.6366	0.8792	0.7385
Stacking Model (XGB)	0.8500	0.9113	0.7165	0.6159	0.8669	0.7201

4. Conclusion

The study covers the full machine learning pipeline from data preprocessing to feature engineering and to predictive modeling regarding Lending Club loans. The primary objective of this research, predicting loan quality, has been accomplished with a remarkable recall rate of over 87% on the test set, surpassing existing research benchmarks. This accomplishment empowers investors with the ability to make well-informed decisions.

From a business standpoint, pinpointing variables which exhibit strong explanatory power regarding loan default, is noteworthy. It has the potential to optimize credit grading system, reducing borrowing costs for borrowers, and supporting prudent investment choices for investors.

Furthermore, there are opportunities for improvement. For instance, time variables like issue_d could be used to construct time-series models or serve as auxiliary factors to potentially enhance model accuracy. Additionally, for critical variables with upper and lower limits, , exploring statistical models like Tobit or Heckman Selection models can offer further insights when dealing with truncated dependent variables. These refinements can enhance the model performnance and practicality of the research.

References

[1] Malekipirbazari, Milad, and Vural Aksakalli. Risk assessment in social lending via random forests. Expert Systems with Applications, 2015, 42(10): 4621-4631.

[2] Xu, Junhui, Zekai Lu, and Ying Xie. Loan default prediction of Chinese P2P market: a machine learning methodology, 2021, Scientific Reports 11(1): 18759.

[3] Galindo, Jorge, and Pablo Tamayo. Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. Computational economics, 2000, 15: 107-143.

[4] Liang, Longyue, and Xuanye Cai. Forecasting peer-to-peer platform default rate with LSTM neural network. Electronic Commerce Research and Applications, 2020, 43: 100997.

[5] Li, Xingyun, et al. Prediction of loan default based on multi-model fusion. Procedia Computer Science, 2022, 199: 757-764.

[6] Sigrist, Fabio, and Christoph Hirnschall. Grabit: Gradient tree-boosted Tobit models for default prediction. Journal of Banking & Finance, 2019, 102: 177-192.

[7] Kim, Aleum, and Sung-Bae Cho. An ensemble semi-supervised learning method for predicting defaults in social lending. Engineering applications of Artificial intelligence, 2019, 81: 193-199.

[8] Xia, Yufei, et al. Forecasting loss given default for peer-to-peer loans via heterogeneous stacking ensemble approach. International Journal of Forecasting, 2021, 37(4): 1590-1613.

[9] Lee, Jong Wook, and So Young Sohn. Evaluating borrowers' default risk with a spatial probit model reflecting the distance in their relational network. PloS one, 2021, 16(12): 1-11.

[10] Turiel, Jeremy D., and Tomaso Aste. P2P Loan acceptance and default prediction with Artificial Intelligence. arXiv preprint, 2019, arXiv:1907.01800.

[11] https://www.kaggle.com/datasets/wordsforthewise/lending-club.