CSE 343 : Machine Learning Final-Project Report
Title : Credit Risk Assessment

Nikhil Suri
IIIT - Delhi
Okhla, New Delhi
nikhil21268@iiitd.ac.in

Maanas Gaur
IIIT - Delhi
Okhla, New Delhi
maanas21537@iiitd.ac.in

Adish Jain
IIIT - Delhi
Okhla, New Delhi
adish21227@iiitd.ac.in

1. Abstract

This project delves into the intricate domain of credit risk assessment within the peer-to-peer lending landscape, spotlighting LendingClub as a focal point. By leveraging Exploratory Data Analysis (EDA) and Machine Learning, the study seeks to decipher the critical factors influencing loan default, offering a profound understanding of risk analytics in the financial sector. This endeavor not only aims to minimize financial loss for the company but also strives to redefine how lending institutions evaluate and mitigate risk. Through meticulous data analysis and predictive modeling, this research contributes to the advancement of prudent lending practices, ultimately fortifying the financial industry's resilience to risk.

2. Introduction

The lending landscape has undergone a significant transformation with the advent of platforms like LendingClub, the world's largest peer-to-peer lending platform. In response to a surge in loan applications, the importance of precise credit risk assessment has never been more pronounced. This report presents a comprehensive case study focused on LendingClub, a pioneering peer-to-peer lending company based in the United States. Our primary objective is to uncover the intricacies of risk analytics within the banking and financial services sector, particularly within the context of urban customers.

At the heart of this challenge lies the decision-making process upon receiving a loan application. Striking a balance between two types of risks is imperative:

1. Potential Loss of Business: Denying a loan to a credible applicant directly translates to a loss of business for the company.

2. Risk of Default: Conversely, approving a loan for an applicant likely to default could lead to financial repercussions.

The provided dataset encompasses historical data of past loan applicants and their respective repayment behaviors. Our aim is to unearth discernible patterns that act as indicators of default. These insights could play a pivotal role in making informed decisions, such as adjusting loan amounts, assigning higher interest rates to riskier applicants, or even in some cases, denying loans.

Through a thorough examination of the data using exploratory data analysis (EDA) and leveraging machine learning techniques, we aspire to construct a robust framework for identifying high-risk loan applicants. Ultimately, our goal is to empower lending institutions with the knowledge and tools to minimize credit loss and optimize their lending portfolios.

3. Literature Survey

In the realm of credit risk assessment, significant strides have been made, with researchers exploring various facets of lending risk prediction. This section provides an overview of seminal works that have paved the way for understanding and addressing challenges in this domain.

1. **Consumer Credit Risk Assessment**
   Jonathan N. Crook, David B. Edelman, and Lyn C. Thomas delve into the realm of classification algorithms and profit scoring, shedding light on their application in the prediction of lending risks [1].
2. **P2P Loan Acceptance & Default Prediction with AI**
   Kenneth Kennedy's thesis provides valuable insights into the challenges faced in the development of credit scorecards, with a particular focus on issues like class imbalance and low-default portfolios [2].
3. **Deep Learning Credit Risk Modeling**
   J. D. Turiel and T. Aste employ cutting-edge deep learning techniques to forecast loan acceptance and default risk in peer-to-peer lending credit risk models [3].

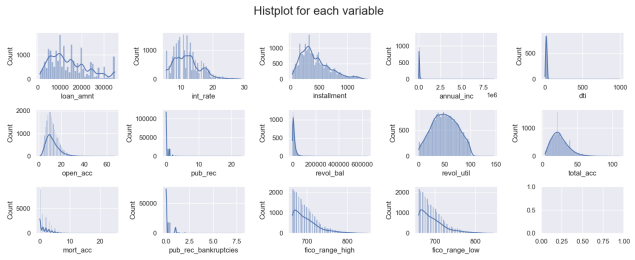

Figure 1. Histplot for each variable containing Numerical Data

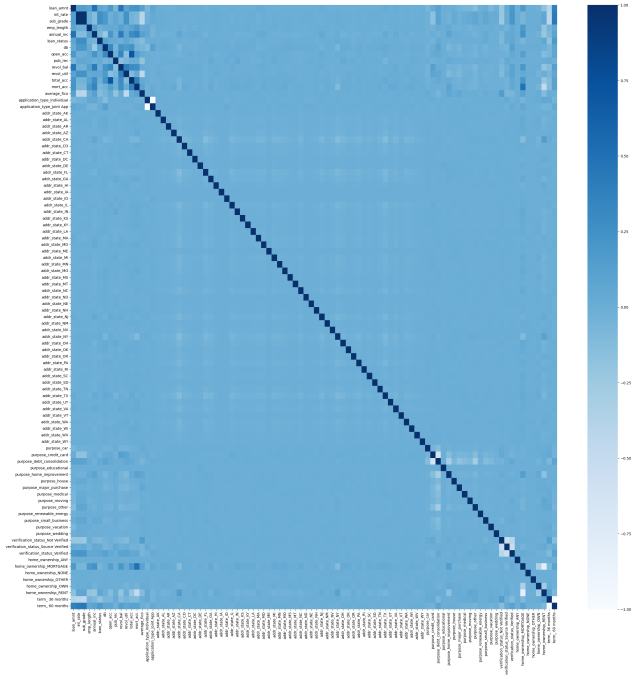

Figure 2. HeatMap

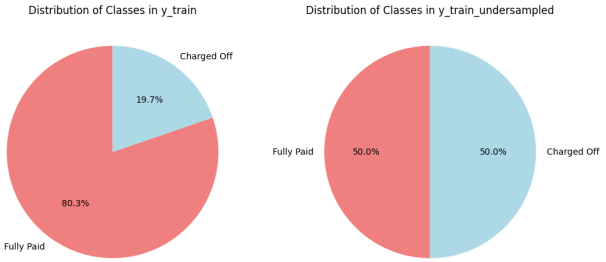

Figure 3. Y_Train Distribution

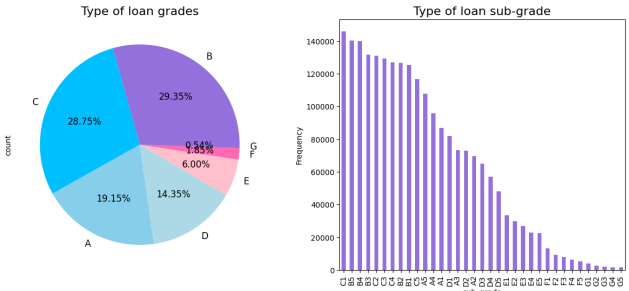

Figure 4. Loan Grade Analysis

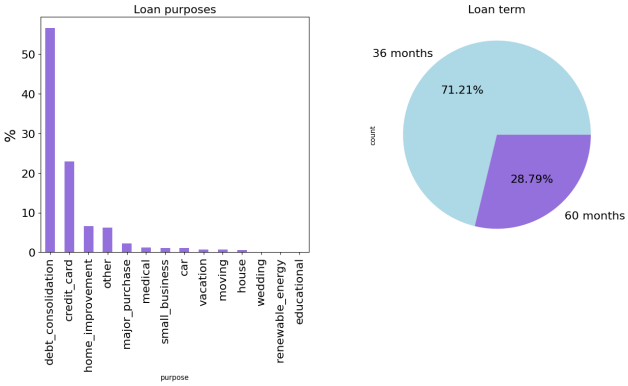

Figure 5. Loan Term

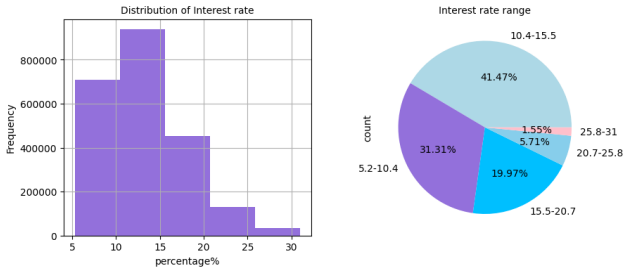

Figure 6. Interest Rate Analysis

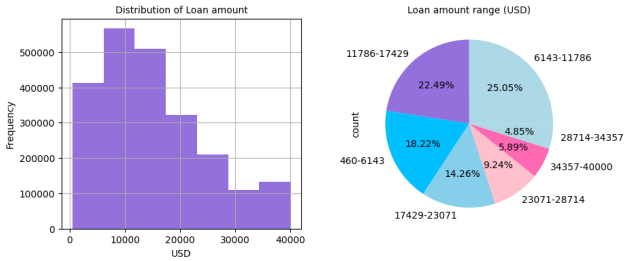

Figure 7. Loan Amount Analysis

4. Dataset with Pre-Processing Techniques

The dataset is loaded into a pandas DataFrame with about 1,00,000 rows being considered for analysis.

The dataset is initially explored to understand its structure.

To handle missing data, a thorough analysis is conducted. A bar plot visualizes the percentage of missing values for each column. The threshold for missing data is set at 20%. Columns with more than 20% missing data are identified and dropped. The DataFrame is updated to keep only columns with less than or equal to 20% missing data.

A list of selected features relevant for credit risk analysis

is defined. These features are chosen based on their importance and availability at the time of loan origination.

Table 1. Selected Columns Description

Column Name	Description
loan_amnt	Requested amount
term	Number of payments
int_rate	Interest rate
installment	Monthly payment
grade	Loan grade
sub_grade	Loan subgrade
emp_title	Borrower's job title
emp_length	Employment length
home_ownership	Home ownership status
annual_inc	Annual income
verification_status	Income verification status
issue_d	Loan issue month
loan_status	Loan status
purpose	Loan purpose
title	Loan title
zip_code	First 3 digits of zip code
addr_state	Borrower's state
dti	Debt-to-Income ratio
earliest_cr_line	Earliest credit line
open_acc	Open credit lines
pub_rec	Public records
revol_bal	Revolving balance
revol_util	Revolving line utilization
total_acc	Total credit lines
initial_list_status	Initial listing status
application_type	Application type
mort_acc	Mortgage accounts
pub_rec_bankruptcies	Bankruptcies
fico_range_high	FICO score upper bound
fico_range_low	FICO score lower bound

Next, the target variable (loan_status) is defined to include only 'Fully Paid' and 'Charged Off' statuses. This dataset is then prepared for further analysis, resulting in a dataset with two target statuses.

Categorical and numerical features are identified, and histplots and boxplots are generated to visualize the distribution of numerical features.

Further data preprocessing techniques are applied to clean and prepare the data for modeling. This includes steps such as mapping categorical variables, creating a new feature (fico_score_avg), and dropping unnecessary columns.

Lastly, missing values are removed from the dataset, and dummy variables are created for categorical features. The final preprocessed dataset is now ready for further analysis and modeling.

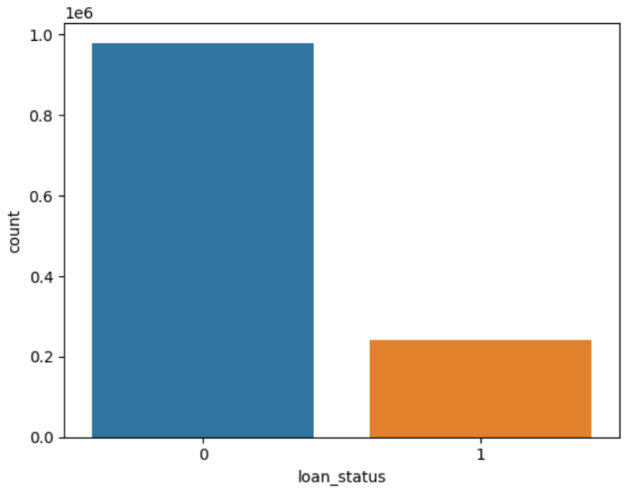

Figure 8. Bar Graph showing # of samples for each class
0 = Debt Paid Off, 1 = Debt Default (Insolvent)

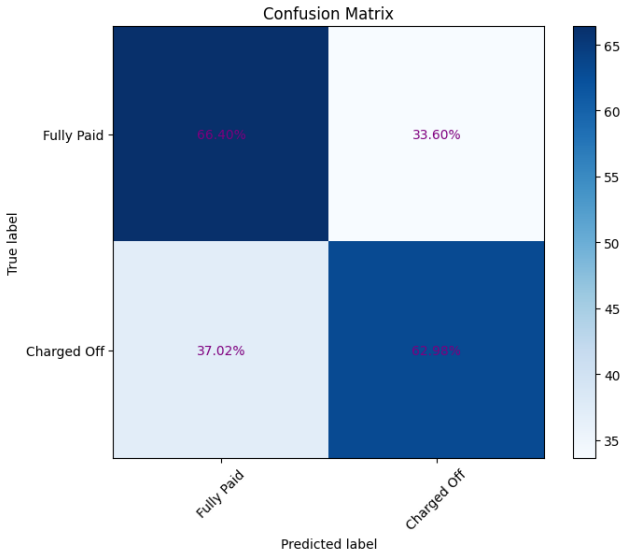

Figure 9. Confusion Matrix for Simple Logistic Regression

5. Methodology & Model Details

In terms of modeling, several algorithms have been explored. These include Logistic Regression, Random Forest, and Support Vector Machines (SVM). Each model is evaluated based on performance metrics like accuracy, precision, and recall. All the models are based on under-sampled data.

The Logistic Regression model is known for its simplicity and interpretability, making it an excellent starting point. Random Forest, a versatile ensemble method, is employed for its ability to handle complex relationships and reduce overfitting. SVM, known for its effectiveness in high-dimensional spaces, provides an alternative perspective.

To validate the models, a test set is reserved, and metrics are computed to assess their performance. Addition-

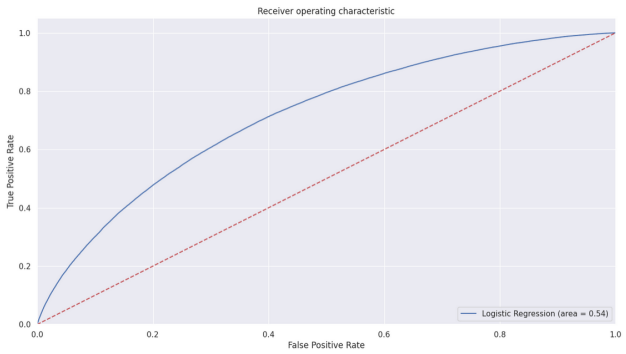

Figure 10. True Positive Rate v/s False Positive Rate for Logistic Regression

ally, techniques such as cross-validation and hyperparameter tuning are employed to optimize model performance.

The next steps involve fine-tuning the models to achieve the best possible balance between precision and recall. This is particularly critical in a credit risk assessment scenario, where correctly identifying potential defaults is of paramount importance.

Here's a brief summary for the steps taken, to carry-out the binary classification task:

5.1. Data Preparation

The dataset (`filtered_df4`) is divided into features (`X`) and the target variable (`y`), which represents the loan status. The features are then standardized using `StandardScaler`.

5.2. Train-Test Split

The dataset is split into training and testing sets using a 75-25 ratio.

5.3. Logistic Regression

A Logistic Regression model is trained on the standardized data. The model is evaluated using accuracy, a confusion matrix, and a ROC curve. Threshold adjustments are demonstrated to fine-tune classification.

5.4. Random Forest Classifier

A Random Forest Classifier is trained on the data without the need for standardization. The model is evaluated using ROC-AUC score.

5.5. K-Nearest Neighbors (KNN)

A K-Nearest Neighbors Classifier is trained after data imputation and standardization. The model is evaluated based on ROC-AUC score.

5.6. Hyperparameter Tuning

Hyperparameters of the selected model (Logistic Regression) are further tuned using a grid search with cross-validation.

5.7. Naive Bayes

A Gaussian Naive Bayes Classifier is trained on the data. The model is evaluated based on accuracy, a confusion matrix, and a classification report.

5.8. Support Vector Machine (SVM)

A Support Vector Machine Classifier with a linear kernel is trained on undersampled data. The model is evaluated using accuracy, a confusion matrix, and a ROC curve.

6. Results & Analysis

The credit risk analysis yields noteworthy insights into the dataset. After careful data preprocessing and model selection, the performance of various models is evaluated.

6.1. Logistic Regression

The Logistic Regression model, known for its simplicity and interpretability, provides a solid baseline. It exhibits a weighted Avg F1-score of 69%.

6.2. Random Forest

Moving on to the Random Forest model, its performance surpasses that of Logistic Regression. It is an ensemble learning method, that combines multiple decision trees to enhance predictive accuracy and control overfitting.

Table 2. Classification Report for Random Forest

Classification	Precision	Recall	F1-Score	Support
0	0.89	0.63	0.74	196,102
1	0.31	0.67	0.42	47,917
Accuracy			0.64	244,019
Macro Avg	0.60	0.65	0.58	244,019
Weighted Avg	0.77	0.64	0.68	244,019

Table 3. Classification Report for Logistic Regression

Classification	Precision	Recall	F1-Score	Support
0	0.88	0.66	0.76	196,102
1	0.31	0.63	0.42	47,917
Accuracy			0.66	244,019
Macro Avg	0.60	0.65	0.59	244,019
Weighted Avg	0.77	0.66	0.69	244,019

6.3. K - Nearest Neighbours

K-Nearest Neighbors (KNN) is employed to classify data points based on the similarity of their features to the ones in the training set.

Table 4. Classification Report - KNN

	Precision	Recall	F1-Score	Support
0	0.81	0.98	0.89	244855
1	0.53	0.07	0.12	60168
Accuracy			0.80	305023
Macro Avg	0.67	0.53	0.51	305023
Weighted Avg	0.76	0.80	0.74	305023

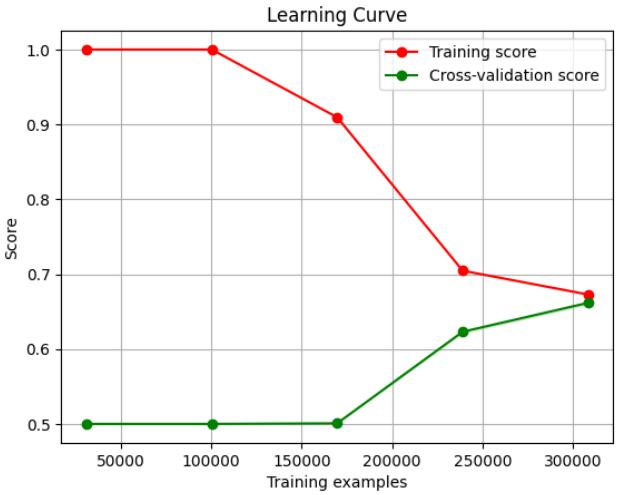

Figure 11. Learning Curve XG Boost

6.4. SVM

The Support Vector Machine (SVM) model, known for its effectiveness in high-dimensional spaces, offers yet another perspective, with a decent accuracy of 77%., which demonstrate SVM's strength in correctly classifying loans and capturing true positives.

Table 5. Classification Report - SVM

	Precision	Recall	F1-Score	Support
0	0.85	0.87	0.86	1701
1	0.39	0.35	0.37	398
Accuracy			0.77	2099
Macro Avg	0.62	0.61	0.62	2099
Weighted Avg	0.76	0.77	0.77	2099

6.5. XG Boost

XGBoost, short for eXtreme Gradient Boosting, is a powerful machine learning algorithm known for its efficiency in handling complex datasets and achieving high predictive accuracy through an ensemble of decision trees.

6.6. ANN

Artificial Neural Networks (ANNs) are a class of machine learning models inspired by the structure and function of the human brain, designed to learn and make predictions from complex data patterns.

Table 6. Classification Report for XGBoost

Classification	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.89	0.64	0.74	196,102
1	0.32	0.68	0.43	47,917
Accuracy			0.65	244,019
Macro Avg	0.60	0.66	0.59	244,019
Weighted Avg	0.78	0.65	0.68	244,019

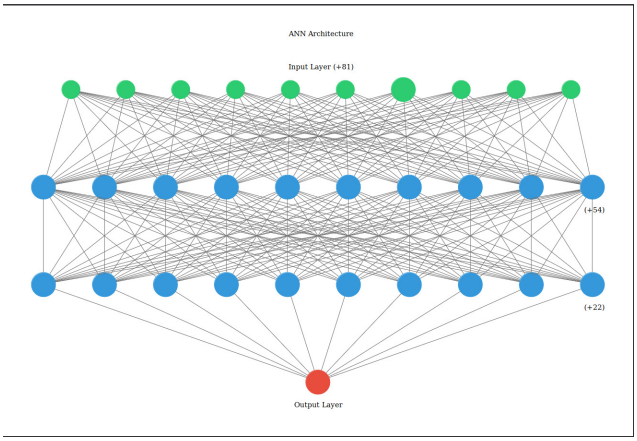

Figure 12. ANN Architecture

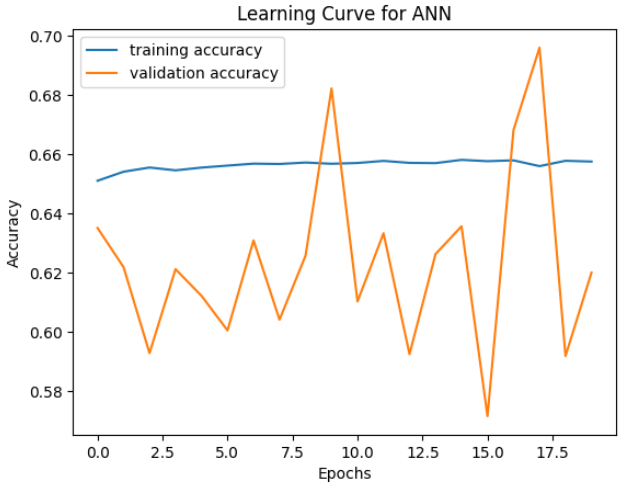

Figure 13. ANN Learning Curve

Table 7. Classification Report for ANN

Classification	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.90	0.60	0.72	196,102
1	0.30	0.72	0.43	47,917
Accuracy			0.62	244,019
Macro Avg	0.60	0.66	0.57	244,019
Weighted Avg	0.78	0.62	0.66	244,019

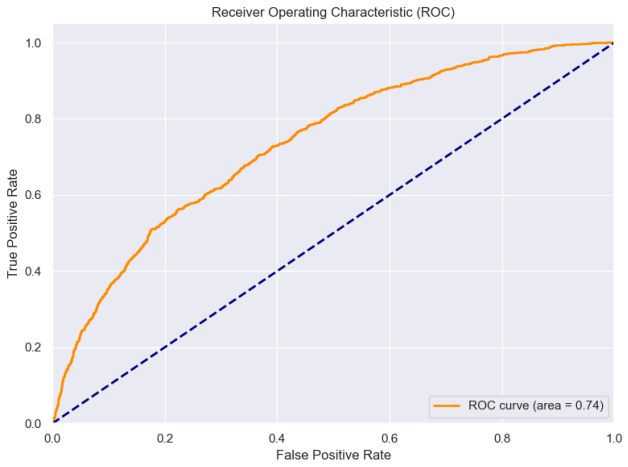

Figure 14. True Positive Rate v/s False Positive Rate for SVM

6.7. Naive Bayes

Gaussian Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem with an assumption of independence between features. The relevant classification report is mentioned below.

Table 8. Classification Report for Gaussian Naive Bayes

Classification	Precision	Recall	F1-Score	Support
0	0.81	0.94	0.87	196,102
1	0.28	0.10	0.14	47,917
Accuracy			0.77	244,019
Macro Avg	0.54	0.52	0.51	244,019
Weighted Avg	0.71	0.77	0.73	244,019

7. Conclusion

7.1. General Inference

Comparing these models, it becomes evident that Logistic Regression model, stands out as the most effective in this context. Its higher precision indicates a lower false positive rate, meaning it is better at correctly identifying actual defaults. This is crucial in credit risk assessment, where avoiding false negatives is of paramount importance.

The analysis also highlights areas for improvement. Fine-tuning hyperparameters and exploring advanced techniques such as gradient boosting may further enhance model performance.

In summary, the results showcase the effectiveness of machine learning models in credit risk assessment. The Logistic Regression model, in particular, exhibits promising performance and sets the stage for further refinements. This analysis not only aids in better understanding credit risk but also provides a foundation for building more sophisticated models in the future.

The models are evaluated based on their ability to predict loan statuses, with a focus on metrics like accuracy, precision, recall, ROC-AUC score, and visualizations to aid interpretation. The final choice of model for deployment depends on the specific requirements of the project and the trade-offs between different evaluation metrics.

This rigorous approach to data preprocessing and model selection forms the backbone of the credit risk analysis. The iterative nature of this process, coupled with a keen understanding of the underlying financial principles, enables the development of a robust and reliable credit risk assessment model.

7.2. Future Scope

The credit risk analysis undertaken in this study provides valuable insights into the lending landscape. Through rigorous data preprocessing and the application of machine learning models, we have gained a comprehensive understanding of loan performance.

Among the models evaluated, the Logistic Regression model, emerged as the most effective tool for credit risk assessment. With an Weighted Avg F1-score of 69%, it outperformed other ML models.

Additionally, this analysis has opened avenues for future research and improvement. Fine-tuning hyperparameters, exploring ensemble methods, and incorporating more extensive feature engineering could further enhance model performance. Moreover, examining the impact of macroeconomic factors and external variables on credit risk could yield additional valuable insights.

Ultimately, this study serves as a foundation for more sophisticated credit risk assessment models. By leveraging the power of machine learning, financial institutions can make more informed lending decisions, ultimately leading to a more stable and resilient financial ecosystem.

References

1 Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3), 1447-1465.
DOI: 10.1016/j.ejor.2006.09.100

2 Kennedy, K. (2013). Credit Scoring Using Machine Learning. Doctoral thesis, Technological University Dublin.
Available under a Creative Commons Attribution Non-Commercial Share Alike 4.0 International Licence.
DOI: 10.21427/D7NC7J

3 Turiel, J. D., & Aste, T. (2020). Peer-to-peer loan acceptance and default prediction with artificial intelligence. *Royal Society Open Science*, 7(6), 191649.
DOI: 10.1098/rsos.191649