MBAI 5600G

MBAI 5600G
Applied
Integrative
Analytics
Capstone Project

Enhanced Loan Default Prediction Using Machine Learning: Integrating Industry Segmentation and Sentiment Analysis

Hemasri Appavu Krishnaraju

Abstract

In the current financial landscape, accurate risk assessment methodologies are paramount for banks and lenders to prevent potential losses. This project aims to develop a risk assessment technique to predict loan defaulters by leveraging machine learning. Traditional methods often overlook industry-specific nuances, relying on broad classifications or subjective assessments. This project addresses these limitations by exploring the intricate relationship between employee titles, industry classifications, and default risk. The dataset used is from Lending Club, a US-based peer-to-peer lending company, encompassing details about previous loan applicants and their default status. Key features include loan amount, interest rate, employment length, and annual income, among others. The methodology involves data preprocessing, feature engineering, model selection, and evaluation. Various machine learning algorithms, including decision trees, random forests, support vector machines, and neural networks, will be employed to segment loan applicants and predict default probability within each segment. This project aims to empower financial institutions with robust tools for industry classification and default prediction, fostering more informed decision-making and risk management practices in the lending domain.

Keywords: Risk assessment; Loan default prediction; Industry classification; Machine learning; Sentiment analysis: Financial technology; LendingClub.

1. Introduction

In today's financial scenario, credit risk assessment is crucial for calculating the probability that a borrower may default, aiding in informed decision-making. Numerous studies emphasize the necessity of integrating machine learning and AI into credit risk assessment. This project will contribute to this body of knowledge by incorporating both traditional financial data and industry sentiment analysis to enhance the predictive accuracy of loan defaults. This study will explore the critical role of industry classification in credit risk assessment. Different sectors demonstrate varying levels of risk due to economic cycles, market conditions, and regulatory environments. By monitoring industry sentiment over time, it is possible to identify potential risk factors before they become evident in financial data. For example, a sudden decline in customer sentiment for a specific industry could signal upcoming economic challenges and potentially higher loan defaults.

2. Literature Review

The banking sector is evolving with technological advancements and the integration of data science, transforming the landscape of financial institutions (Maheswari and Narayana, 2020). However, there is a risk of significant capital losses if loans are approved without assessing default risk beforehand (Maheswari and Narayana, 2020). Therefore, accurate predictive systems are crucial for financial institutions (Maheswari and Narayana, 2020). The uncertainty of a customer making a loan payment in each period can be quantified by the Probability of Default (PD), which indicates the likelihood that a customer will default on a payment within a specified time frame (Singh et al., 2013). Predicting loan defaulters is critical, given the abundance of data banks possess, including customer data and transaction behavior (Maheswari and Narayana, 2020).

Loss-given default (LGD) indicates the severity of loss after a loan default and represents the percentage of the unrecoverable amount (Bessis, 2015; Bandyopadhyay, 2016). Collateral is used by banks to mitigate losses in the event of default. When a debtor defaults, the lender tries to recover the debt by selling the collateral, but the amount recovered is often less than the amount due to recovery expenses (Bessis, 2015).

Machine learning plays a significant role in credit risk assessment:

Pandey et al. (2010) emphasize that predicting loan defaulters is one of the most challenging tasks for banks, but it can significantly reduce losses and improve asset quality. They applied machine learning algorithms such as Logistic Regression, Decision Trees, Support Vector Machines, and Random Forests, finding Support Vector Machines most accurate in predicting loan acceptance (Pandey et al., 2010).

Chang et al. (2016) compared logistic regression, the Cox model, and decision tree models to predict short-term loan defaults. Their decision tree model had 81.9% specificity and 83.3% precision, outperforming other models in short-term default prediction.

Chang et al. (2016) using Lending Club dataset implemented Logistic Regression initially achieved 92.9% test accuracy and 75.8% specificity, which improved to 77.1% after feature selection. Gaussian Naive Bayes outperformed other distributions with 80.4% specificity, though adding census data only marginally improved accuracy. For SVM, the linear kernel provided the best results, and feature expansion along with parameter tuning led to incremental performance gains.

Abid et al. (2018) used logistic regression and discriminant analysis on a Tunisian commercial bank's dataset of 603 loans, which had a 56.55% default rate. Their logistic regression model showed 99.41% sensitivity and 98.47% specificity, outperforming discriminant analysis.

Various studies have used metrics like classification accuracy, sensitivity, specificity, and ROC curves to evaluate prediction models. Lessmann et al. (2015) compared 41 learning algorithms and found that advanced classifiers like random forests outperformed logistic regression. Zhu et al. (2019) achieved high performance with random forests, showing 98% accuracy and 99% sensitivity.

Silva et al. (2020) examined default risk using logistic regression on a Portuguese credit dataset with 3221 individuals and a 10% default rate. Key variables included "Spread," "Term," "Age," "Credit cards," "Salary," and "Tax echelon," achieving a classification accuracy of 89.79%, with 0.94% sensitivity and 99.55% specificity.

Other research, including studies by Xia et al. (2017) and Tian et al. (2020), highlighted the effectiveness of models like XGBoost and gradient-boosting trees. These studies often lacked specificity or AUC data but generally supported the superiority of advanced machine-learning techniques over traditional methods.

Sheikh (2020) argues the importance of predicting loan defaulters in banking systems, highlighting that a bank's profitability heavily relies on loan repayment behavior. Their study using logistic regression suggests that incorporating personal attributes such as age, credit history, and wealth indicators like checking account information improves default prediction accuracy (Sheikh,2020).

Ndayisenga et al. (2021) worked with commercial banks to predict borrower behavior using data from the Bank of Kigali, demonstrating Gradient Boosting as the most effective model for predicting loan default, followed by XGBoosting, with Decision Trees, Random Forest, and Logistic Regression performing less effectively (Ndayisenga et al., 2021).

Orji et al. evaluate the classification accuracy of six computational intelligence methods across diverse datasets, highlighting the importance of selecting methods tailored to specific dataset characteristics (Orji et al., 2022).

Research such as "The sensitivity of the loss given default rate to systematic risk" (Caselli et al., 2008) has highlighted the connection between default rates and macroeconomic factors. Therefore, to develop a more comprehensive set of features, in this paper we will incorporate industry sentiments data, including information over time, to train our models.

3. Data Description

LendingClub, which is a US-based peer-to-peer lending company is among the largest marketplaces for personal loans, business loans, and medical financing globally. (All Lending Club loan data) Customers can easily access lower-interest-rate loans through a fast-online interface. (All Lending Club loan data)

When the company receives a loan application, the company has to decide for loan approval based on the applicant’s profile. Two types of risks are associated with the bank’s decision: (Sayah, 2023)

1) If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company. (Sayah, 2023)

2) If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company. (Sayah, 2023)

The provided dataset includes details about previous loan applicants and their default status. The goal is to detect patterns that suggest the likelihood of default, enabling actions like loan denial, reducing loan amounts, or offering loans to risky applicants at higher interest rates. (Sayah, 2023)

Like many other lending institutions, providing loans to "risky" applicants constitutes the primary source of financial loss, known as credit loss. Credit loss refers to the sum of money a lender loses when a borrower fails to repay or absconds with the borrowed funds. In simpler terms, borrowers labeled as "charged-off" represent defaulters and contribute significantly to the lenders' losses. (Sayah, 2023)

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. (Sayah, 2023)

3.1. Dataset/Features Description

Traditional Financial Data (Lending Club): There are 151 features with 2260701 rows. The dataset timeframe ranges from 2007-2018Q4. Some of the important features are FICO score, payment history. We have 113 numeric float columns and 38 categorical columns in the dataset.

Industry Sentiment Analysis: Now, we are creating another dataset where we have utilized pretrained models from Hugging Face, from where we are getting the sentiment score for the industries with the content of review. We will utilize this once we have segmented the customers into their respective sectors to understand the sentiment of the industries over time.

3.2. Methodology

We are considering only the customers who are fully paid and are flagged as default (Charged Off) leading to the data shape of (1345310, 151).

a) Data Cleaning:

Handling the Missing values: The columns with more than 30% missing values are identified and removed. 58 columns with more than 30% missing values are removed making the shape of the data (1345310, 93). The other missing values are imputed by median for numerical columns and mode for categorical columns.

Identifying the Outliers: The columns with outliers are detected and the columns having more than 3% are removed from the dataset. We are utilizing the Z score value to calculate the outliers. Absolute z-scores for the specified column, indicating how many standard deviations away each value is from the mean. A subset of the column containing only the outlier values, where the absolute z-score exceeds the specified threshold is created. The percentage of values in the column that are considered outliers based on the z-score threshold are then calculated. The remaining outliers are not imputed with values. For this financial analysis, we need to consider extreme situations and values in the model building process.

b) Data Transformation: The date columns, which are object data types, are converted into datetime data types. In this dataset, we have the issue date, indicating when the loan was issued, and the last payment date, which are converted to datetime type.

c) Feature Engineering: One of the most crucial roles in this analysis is feature engineering, capturing the nuances of industry-specific data. This may involve extracting relevant information from employee titles, such as job roles, specialized skills and categorizing them into major industries, e.g. Education, Finance, IT, Healthcare, Manufacturing. A key aspect is this segmentation of loan applicants into industry-specific segments. This segmentation allows for building more targeted models that can capture the unique characteristics and risk profiles of different industries.

Assuming job titles such as Software Engineer, IT Specialist, Software Developer etc. to be in the IT sector. Similarly, assuming titles such as Teacher, Professor are from Education, Registered Nurse, Physician, Pharmacist is from Healthcare, Engineer, Quality control, Mechanical Engineer from Manufacturing and Accountant, Financial advisor, Banker from Finance and Banking.

Once we have the sectors, the sentiment score throughout the years 2007 to 2018 is merged to the data. Considering the loan issue date and the last payment date, we will analyze the sentiment scores of customers industries from the year the loan was issued until the year of the final payment. Now considering only these major sectors we are filtering the dataset, with the additional sentiment analysis, the dataset is in the shape (146231, 107)

d) Dimensionality Reduction: For this analysis, irrelevant or redundant columns in the dataset have been removed. Specifically, columns such as id, url, zip_code, title (since it overlaps with the purpose of the loan), and emp_title (since industry segmentation is available) were excluded. Following one-hot encoding, the dataset comprised approximately 158 features. Instead of applying dimensionality reduction techniques, all columns were utilized in the model-building process. Feature selection was conducted using a combination of correlation analysis and recursive feature elimination (RFE). Features with a high correlation to the target variable (loan status) were prioritized, while multicollinearity among features was minimized to enhance model interpretability and performance.

3.3. Exploratory Data Analysis

The target column for this analysis is loan status with values Fully Paid and Charged Off., borrowers labeled as "charged-off" represent defaulters and contribute significantly to the lenders' losses.
There is around 119467 Fully paid customers and 26764 Defaulters. There is a class imbalance in the target variable with 18% of the data being Defaulters, and instead of accuracy, we will evaluate the performance of the model using the F1 score. Evaluating the model using the F1 score is a proven approach, especially when dealing with class imbalance. F1 score balances precision and recall, which can be more informative than accuracy in such cases.

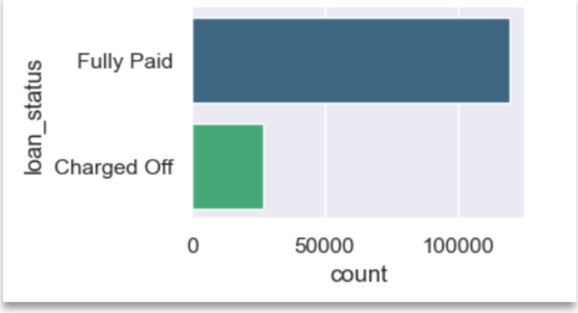

Fig. 1. Denoting an imbalance in the class of target variable.

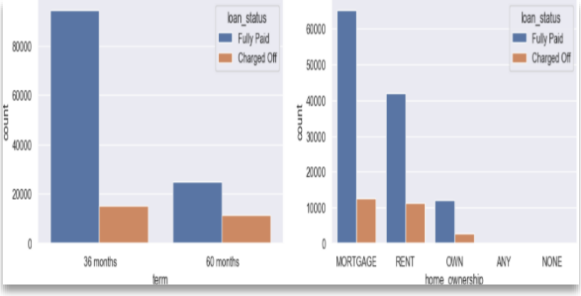

Fig. 2. Term, home ownership status underscores the popularity of fixed-term loans and the prevalence of mortgage holders seeking loan consolidation

In Figures 2 of the analysis, it is evident that most customers opt for a 36-month term for loan repayment. Mortgage is the predominant home ownership status among customers, and most applications have a verification status of verified. The primary purpose for taking loans across all loan statuses is debt consolidation, indicating a common financial goal among applicants. These findings underscore the popularity of fixed-term loans and the prevalence of mortgage holders seeking loan consolidation, with a significant proportion having their verification status confirmed.

We can see that in Fig 3 credit card, debt consolidation, home improvement, house, major purchase, and renewable energy are some of the purposes for loan requests with high loan amounts.

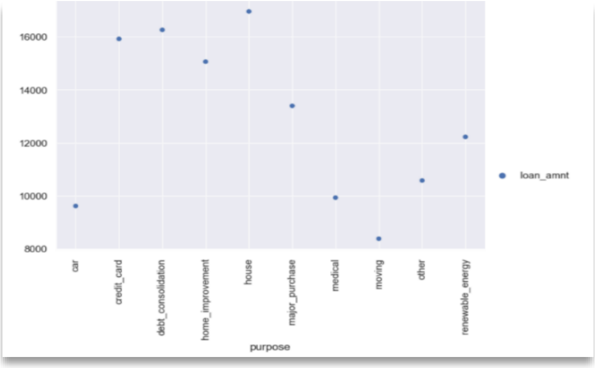

Fig. 3. Purposes for loan requests with high loan amounts.

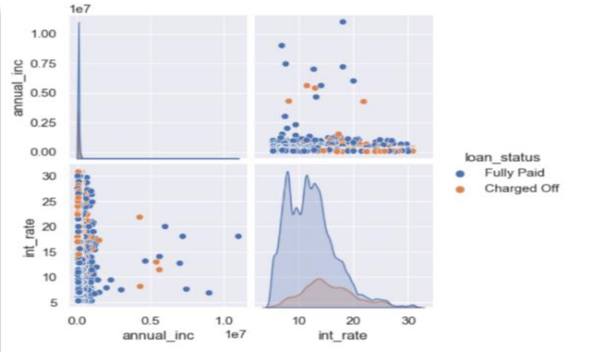

Fig. 4. The plot shows that there is a positive correlation between annual income and loan status.

The plot Fig 4 shows that there is a positive correlation between annual income and loan status. This means that as annual income increases, the likelihood of a loan being fully paid also increases. For example, in the plot, there are very few data points in the bottom left corner (where annual income is low and loan status is "Charged Off"). There are many more data points in the top right corner (where annual income is high and loan status is "Fully Paid").

From the plot Fig 5, we can see that both the class of the target has similar interest rates by grades.

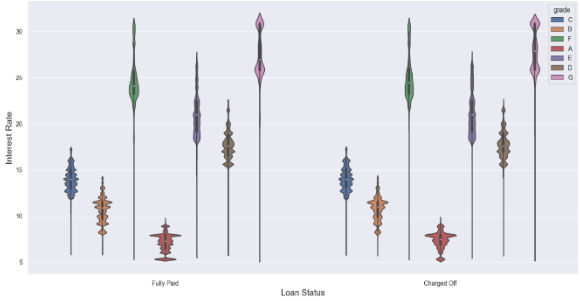

Fig. 5. Loan status classes with distribution of interest rate based on grades.

Employee titles such as Teacher, Nurse, Manager, Accountant, and Engineer are among the most common among loan customers from 2007 to 2018. We can see that Fig 5 most of the employees are in the IT sector with 27% as per our classification, followed by Manufacturing and Healthcare with 20% respectively.

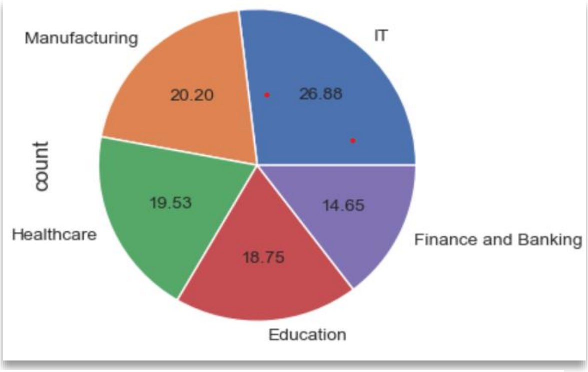

Fig. 6. Industry segmentation based on employee titles.

As we can see in the Fig 7, Healthcare has some of the highest sentiment scores, followed by IT and Finance over the period. All industries experienced a dip in sentiment scores in 2009, indicating a negative sentiment trend, likely due to the global economic crisis affecting all sectors.

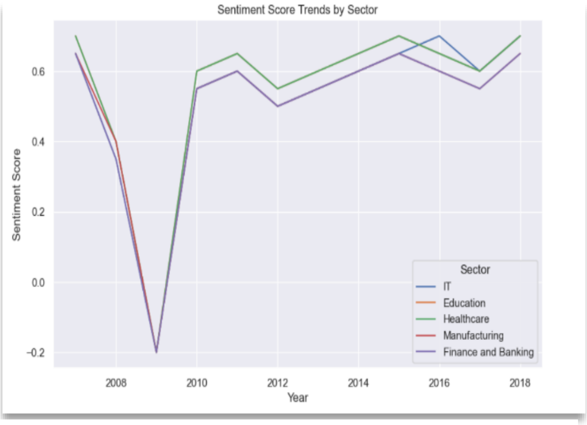

Fig. 7. Sentiment Score Trends over the year for the segmented sectors.

ıe

association between loan_status and industry segment. The p-value associated with the chi-square statistic is approximately 1.62. This extremely small p-value suggests strong evidence against the null hypothesis, indicating that there is a significant association between the two categorical variables.

Correlation between the features and target variable:
The last fico range high, last payment amount, total received principle, collection recovery fee, recoveries, and interest rate are some of the features positively and negatively highly correlated with the loan status target variable with a maximum 0.68 correlation coefficient.
Features positively correlated with loan status (up to 0.68 correlation coefficient):
Last FICO Range High: This indicates a borrower's creditworthiness. A higher FICO range suggests a better credit history, potentially implying a lower risk of defaulting on the loan. In other words, borrowers with a history of managing credit responsibly are more likely to repay their loans.
Features negatively correlated with loan status (up to -0.5 correlation coefficient):
Collection Recovery Fee: This fee being higher typically means past difficulty repaying loans, as this fee is typically applied when a borrower has missed payments in the past.
Recoveries: A borrower who has defaulted on a loan, the amount recovered is named as recoveries. A high recovery value might indicate past defaults, potentially making it more likely for the borrower to default again.

3.4. Scope and Limitations

The scope of the project encompasses comprehensive data analysis, model development, and performance evaluation. Major deliverables include a detailed report outlining the classification of loan applicants into industry categories, the identification of defaulters within each industry segment, and a comparative analysis of machine learning algorithms' efficacy across different industries and the features having most impact.

Furthermore, the project aims to provide actionable insights for lenders, including recommendations for industry-specific risk mitigation strategies and potential enhancements to loan approval processes. Challenges anticipated during implementation include data quality issues, algorithmic complexity, and the interpretation of nuanced industry dynamics.

Incorporating industry sentiments into the analysis adds a layer of complexity due to the temporal nature of sentiment dynamics. Also considering the approach where we will be considering the sentiments of the industries there are challenges as to understating the sentiments of the industry with respect to the timeframe of the loan, as the sentiment of the industries fluctuates with time. The accuracy of the sentiment analysis depends on the quality of the review data and the chosen pre-trained model.

Furthermore, integrating sentiment data with traditional credit risk models requires careful calibration and validation to assess its predictive power and reliability. Ensemble modeling techniques, which combine sentiment signals with other relevant features, can help improve the robustness and accuracy of default prediction models.

In summary, this project endeavors to empower financial institutions with robust tools for industry classification and default prediction, ultimately fostering more informed decision-making and risk management practices in the lending domain.

The project faces several limitations that could impact the accuracy and generalizability of its loan default prediction models. First, the quality of the dataset, while data cleaning and preprocessing can mitigate some issues, they cannot eliminate them. Second, accurately classifying employee titles and industry sectors poses challenges, as misclassification can lead to errors in segmentation and model performance. Third, the dataset exhibits significant class imbalance, with only 18% representing defaulters. Although using the F1 score helps address this, it remains a challenge for model training and evaluation.

Temporal changes in economic conditions and lending practices over the dataset's timeframe (2007-2018) introduce variability that static models might struggle to capture. The model's generalization to other types of lending institutions or geographic regions may be limited, as it is trained specifically on LendingClub data. Furthermore, the model does not account for broader economic factors or individual borrower circumstances, such as changes in employment status or macroeconomic conditions, which can influence loan default risk.

Segmentation based on industry and job titles assumes homogeneity within each segment, potentially overlooking the diversity of risk profiles within industries. Finally, the extensive dataset and the use of multiple machine learning algorithms require significant computational resources, which may limit feasibility for some institutions. Despite these challenges, the project highlights the potential of machine learning to enhance default prediction accuracy, providing financial institutions with better tools for risk management and decision-making.

4. Prediction Models

The LendingClub dataset was preprocessed to handle missing values, categorical variables, and scaling of numerical features. Missing values were imputed using median values for numerical features and mode for categorical features. Categorical variables were encoded using one-hot encoding. All numerical features were standardized using standard scaler.

Feature selection was performed using a combination of correlation analysis and recursive feature elimination (RFE). Features highly correlated with the

target variable (loan status) were prioritized, while multicollinearity among features was minimized to improve model interpretability and performance.

Several Machine Learning models were considered for predicting loan defaulters:

1) Logistic Regression (LR): A linear model suitable for binary classification problems. Chang et al. (2016)
2) Random Forest (RF): An ensemble learning method based on decision trees, which improves predictive performance by reducing overfitting. Chang et al. (2016)
3) Decision Tree (DT): A simple and interpretable model that splits data based on feature values. Chang et al. (2016)
4) XGBoost (XGB): An optimized gradient boosting algorithm designed to be highly efficient, flexible, and portable. Chang et al. (2016)
5) Gaussian Naive Bayes (GNB): A probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between features. Chang et al. (2016)

The dataset was split into training (70%) and testing (30%) subsets. The models were trained using the training data and evaluated on the test data. Model performance was assessed using precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC).

5. Experimental Result

Model Evaluation: We will see the analysis where we have utilized sentiment values.

ROC-AUC Score: This metric measures the model's ability to distinguish between defaulters and non-defaulters. Higher values (closer to 1) indicate better performance. Here, all models performed well, with XGBoost and Random Forest achieving the highest scores (0.999 and 0.998 respectively).

Confusion Matrix: This table shows how many data points were correctly and incorrectly classified by the model. All models achieved a high degree of accuracy, with XGBoost again performing the best (only 70 false positives and 1 false negative).

Classification Report: This report provides details like precision, recall, and F1-score for each class (defaulter/non-defaulter). All models have high values in these metrics, suggesting good performance in identifying both defaulters and non-defaulters.

Important Features: While KNN doesn't provide feature importance, the other models (LR, CART, RF, XGB) all highlight features that are informative for predicting loan defaults. Some of the most important include:

Collection Recovery Fee: This indicates past difficulty repaying loans and is a strong predictor of future defaults (important in all models).

FICO Range: This indicates a borrower's creditworthiness. In other words, both high and low FICO range borrowers have an impact on predicting the risk of defaulting on the loan.

Last Payment Amount: The last payment which is its recent activity can help us understand the ability of repayment.

Total Received Principal: The total amount of principal already repaid can influence the likelihood of defaulting on the remaining balance.

Sentiment Scores (2013-2018) appear as features in models in the top 15-20 ranges, but their importance is relatively low compared to financial factors. This suggests that borrower industry sentiment might not be a strong predictor of loan default in this dataset.

Roadmap: Hyperparameter tuning will be performed using grid search and cross-validation to identify the optimal set of hyperparameters for each model. This process will involve iterating over predefined parameter grids and evaluating model performance through k-fold cross-validation. Neural Network models and SVM along with feature reduction techniques will also be explored.

Technique	Model Name	Precision		Recall		F1 Score		ROC-AUC Score	Accuracy
		Default	Non-Default	Default	Non-Default	Default	Non-Default		
With Industry segmentation and Sentiment scores	LR	0.91	0.97	0.86	0.98	0.88	0.98	0.9828	0.96
	RF	0.99	0.99	0.97	1.00	0.98	1.00	0.9988	0.99
	CART	0.98	0.99	0.98	0.99	0.98	0.99	0.9849	0.99
	KNN	0.64	0.88	0.39	0.95	0.49	0.91	0.8280	0.85
	XGB	1.00	1.00	0.99	1.00	1.00	1.00	0.9998	1.00
	NB	0.91	0.98	0.90	0.98	0.90	0.98	0.9884	0.97

Table 1: Summary of data obtained systems classification algorithms to predict default or non-default.

6. Conclusion

The model result demonstrates significant enhancement in loan default prediction accuracy achievable through the integration of machine learning techniques and industry-specific insights. By leveraging the comprehensive LendingClub dataset, we explored the intricate relationships between borrower attributes, industry classifications, and default risk. Our analysis underscored the importance of industry sentiment data as a novel feature, providing valuable context that traditional financial metrics alone could not capture. However, the overall conclusions about XGBoost being the best performing model and the importance of financial factors over borrower sentiment conveys that the industry sentiment scores did not account well for the model performance. Industry-specific recommendations were not robust because the model's performance was consistent across all industries, resulting in non-comparative outcomes.

Among the models tested, XGBoost consistently outperformed others, achieving the highest ROC-AUC score of 0.9998, indicating superior predictive power. The integration of industry sentiment scores further refined the models, highlighting the relevance of sector-specific dynamics in assessing credit risk. Features such as collection recovery fee, FICO range, and last payment amount emerged as critical predictors of loan defaults.

The findings advocate for a multi-faceted approach to credit risk assessment, incorporating both traditional financial indicators and nuanced industry sentiments. This approach empowers financial institutions with more robust tools for risk management, enabling more informed and effective decision-making processes. Despite challenges such as data quality issues and class imbalance, the project underscores the potential of advanced machine learning algorithms to enhance predictive accuracy and provide actionable insights for the lending domain. Future research should focus on real-time sentiment analysis and the integration of broader economic indicators to further refine default prediction models.

7. References

1. Abid, L., Masmoudi, A., & Zouari-Ghorbel, S. (2018). The consumer loan’s payment default predictive model: An application of the logistic regression and the discriminant analysis in a Tunisian commercial bank. *Journal of the Knowledge Economy, 9(3),* 948-962. https://doi.org/10.1007/s13132-016-0382-8

2. Bandyopadhyay, A. (2016). *Managing portfolio credit risk in banks.* Cambridge University Press. https://doi.org/10.1017/CBO9781316550915

3. Bessis, J. (2015*). Risk management in banking (4th ed.).* John Wiley & Sons.

4. Caselli, S., Gatti, S., & Querci, F. (2008). The sensitivity of the loss given default rate to systematic risk: New empirical evidence on bank loans. *Journal of Financial Services Research, 34*(1), 1-34.

5. Chang, Y. C., Chang, K. H., Chu, H. H., & Tong, L. I. (2016). Establishing decision tree-based short-term default credit risk assessment models. *Communications in Statistics – Theory and Methods, 45*(23), 6803-6815. https://doi.org/10.1080/03610926.2014.968730

6. Chang, S., Kim, S. D., & Kondo, G. (2016). Predicting default risk of Lending Club loans. Stanford University.

7. Data Source: All Lending Club loan data. https://www.kaggle.com/datasets/wordsforthewise/lending-club

8. Lessmann, S., Baesens, B., Seow, H., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124-136. https://doi.org/10.1016/j.ejor.2015.05.030

9. Maheswari, P., & Narayana, C. V. (2020). Predictions of loan defaulter - A data science perspective. *2020 5th International Conference on Computing, Communication and Security (ICCCS),* Patna, India (pp. 1-4). IEEE. https://doi.org/10.1109/ICCCS49678.2020.9277458

10. Ndayisenga, T. (2021). Bank loan approval prediction using machine learning techniques (Doctoral dissertation).

11. Orji, U. E., Ugwuishiwu, C. H., Nguemaleu, J. C. N., & Ugwuanyi, P. N. (2022). Machine learning models for predicting bank loan eligibility. *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)* (pp. 1-5). IEEE. https://doi.org/10.1109/NIGERCON54645.2022.9803172

12. Pandey, N., Gupta, R., Uniyal, S., & Kumar, V. (2021). Analyzing loan approval prediction using machine learning algorithms. *Journal of Innovative Research in Technology, 8*(1), [Page numbers]. ISSN: 2349-6002.

13. Sayah, Fares. (2023). Lending Club Loan Defaulters Prediction.

https://www.kaggle.com/code/faressayah/lending-club-loan-defaulters-prediction

14. Sheikh, M. A., Goel, A. K., & Kumar, T. (2020). An approach for prediction of loan approval using machine learning algorithm. *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC),* Coimbatore, India (pp. 490-494). IEEE. https://doi.org/10.1109/ICESC48915.2020.9155614

15. Silva, E. C., Lopes, I. C., Correia, A., & Faria, S. (2020). A logistic regression model for consumer default risk. *Journal of Applied Statistics, 47*(13-15), 2879-2894. https://doi.org/10.1080/02664763.2020.1759030

16. Tian, Z., Xiao, J., Feng, H., & Wei, Y. (2020). Credit risk assessment based on gradient boosting decision tree. *Procedia Computer Science, 174,* 150-160. https://doi.org/10.1016/j.procs.2020.06.070

17. Xia, Y., Liu, C., & Liu, N. (2017). Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. *Electronic Commerce Research and Applications, 24,* 30-49. https://doi.org/10.1016/j.elerap.2017.06.004

18. Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science, 162,* 503-513. https://doi.org/10.1016/j.procs.2019.12.017