Machine Learning Algorithms in Small Business Credit Underwriting

Vinayak Menon

Abstract

Small business credit underwriting has undergone a massive transition in the last two decades with the rise of machine learning algorithms that enhance the accuracy and efficiency of credit risk assessments. In this paper, I study how machine learning models, including XGBoost, Random Forest, and k-Nearest Neighbors, can improve traditional credit scoring methods. By analyzing a dataset from the U.S. Small Business Administration, I evaluate how these models perform in predicting loan defaults, uncovering complex patterns in financial and behavioral data. In conclusion, the results demonstrate that machine learning models, particularly XGBoost, significantly outperform traditional methods, achieving higher accuracy in predicting credit risk. However, challenges such as data privacy concerns, algorithmic bias, and model transparency must be addressed to ensure ethical and reliable use in underwriting. The findings suggest that with responsible implementation, machine learning can optimize small business lending decisions and promote financial inclusion.

Introduction

Credit risk assessment is essential for maintaining stability, profitability, and trust. It uses complex procedures and checks to determine the likelihood of borrowers defaulting on loans. Many financial institutions have traditionally used simple scoring methods or basic statistical models for this process. This process generally considers various factors of an individual or a business, such as the borrower's credit history, financial health, collateral, and the macroeconomic climate at the time of the loan.

However, as society and technology advance, traditional credit risk assessment techniques pose many limitations. Simple scoring techniques may miss valuable insights hidden in larger, more complex data sets. Statistical models humans use are also limited in their ability to account for unique borrower situations or sudden shifts in economic conditions. Human labor poses a significant issue – manually reviewing detailed and complex financial data can be mentally taxing, laborious, and expensive. Humans handling larger datasets using only basic tools are bound to make mistakes, decreasing the accuracy and reliability of this method to evaluate credit risk loans and the underwriting process in general. Bias is also much more challenging to track and eliminate when applications undergo a solely manual process, leading to ethical and legal risks.

Machine learning models are designed to learn from data, allowing them to improve their accuracy over time as they are trained over more information and data evaluated by experts. Machine learning algorithms used in the credit risk assessment process can examine many types of data: spending patterns, transaction history, and even real-time economic factors. These can form a more complex and holistic view of a borrower's financial behavior and risk level. This helps lenders make decisions based on a broader range of information, which can lead to fairer and more accurate outcomes. For example, a machine learning model might spot positive spending habits or recent income stability that could support the borrower's loan application instead of looking at a borrower's credit score or past payment history. Although using these machine learning algorithms poses many benefits to the underwriting process for financial institutions, there are some drawbacks. These include experts for training and evaluating the models, access to customer and small business information on a detailed level (raising privacy concerns), lack of interpretability, and algorithmic bias.

This paper will explore using machine learning classifiers in credit risk assessment, focusing on how these models enhance and improve upon traditional approaches. By examining specific machine learning algorithms used in credit scoring, discussing their applications, and addressing both benefits and challenges, this paper aims to show how machine learning is transforming the field of credit risk assessment.

Literature Review

Credit risk assessment has been a vital part of finance for centuries. In its early days, the underwriting process was simple and manual. Lenders relied on personal relationships, reputation, and basic financial information to decide if someone could repay a loan. As borrowing increased, these methods became less effective, leading to the development of standardized credit scoring systems and statistical models in the 20th century (Bello, 2023). While these methods were more objective and scalable, they were still limited by how much data they could handle. They relied on rules that couldn't easily adapt to unique borrower situations or changing economic conditions (Jansen, Nguyen, & Shams, 2020).

With technological advances, underwriting evolved further. Financial institutions started using computers to process larger datasets and introduced more complex statistical models. However, these systems were still designed and used by humans and couldn't uncover hidden patterns in the data, leaving room for errors and biases (Tan & Zhang, 2023). This limitation became more apparent as the complexity of financial markets grew and borrowers' financial situations became more varied (Sahu, 2023).

In recent years, machine learning (ML) has transformed credit risk assessment. Unlike earlier models, ML algorithms can learn from data, uncover patterns, and adjust to new information. This shift allows lenders to move from rule-based systems to dynamic, data-driven models. ML can analyze vast amounts of data, including spending habits, transaction history, and real-time economic factors, to provide a complete background picture of a borrower's financial behavior (Nguyen, Jansen, & Shams, 2023). This improves the accuracy and fairness of lending decisions. For instance, Jansen et al. (2020) found that machine learning algorithms led to 15% more accurate predictions of risky borrowers and a 20% increase in profitability.

Additionally, studies by Sahu (2023) have shown that ML models can significantly reduce manual work and errors in credit risk assessment. They demonstrated that integrating ML algorithms into the underwriting process decreased processing time by 30% while maintaining high accuracy. This efficiency is particularly valuable for financial institutions handling large volumes of loan applications.

Despite its advantages, ML also introduces new challenges. Bias in training data can perpetuate discriminatory practices, even when the algorithms are designed to be neutral (Tan & Zhang, 2023). The lack of interpretability in some ML models makes it difficult for lenders to explain their decisions to borrowers and regulators (Nguyen, Jansen, & Shams, 2023). Privacy concerns also arise, as ML requires access to detailed personal and financial data (Bello, 2023). Addressing these issues is crucial to ensure that ML applications are both effective and ethical.

In summary, credit risk assessment has evolved from a basic, manual process to a data-driven approach powered by machine learning. ML offers greater accuracy, efficiency, and fairness, but challenges like bias, privacy, and transparency must still be addressed. Future work should focus on making ML models ethical and practical, ensuring they meet the needs of lenders and borrowers alike.

Data

The dataset used in this study is found on Kaggle and comprises small business loan data from 1987 through 2014 from the U.S. Small Business Administration. It includes 27 variables and has 899,164 observations. These variables include information about borrowers, including factors that can affect their credit risk. This data includes traditional indicators like credit scores, income, debt-to-income ratios, and data that reflect spending patterns, transaction history, and broader economic factors. This variety of information helps provide a more complete and detailed profile of each borrower, allowing machine learning models to be tested to identify patterns that might not be visible with basic and traditional credit scoring methods. By using both financial and behavioral data, this dataset is well-suited for training machine learning models to assess creditworthiness more accurately, while limiting bias.

Each entry in the dataset represents an individual borrower or applicant and includes the various details that help predict their credit risk. Traditional financial indicators, like prior credit history, current debts, and income stability, are essential parts of this data to help in the training split for a model. Additionally, information about payment habits, spending patterns, and transaction frequency adds complexity to calculating the risk posed by accepting a credit loan application from a borrower. This extra detail allows the model to go beyond a basic credit score usually calculated by traditional methods, forming a more comprehensive and detailed profile of each borrower's financial behavior. With this more complete data, the machine learning models can make credit risk assessments that are more accurate and fairer, helping lenders make better decisions backed by quantifiable accuracy.

Methods

The data was sourced from Kaggle's "Should This Loan Be Approved or Denied?" dataset. The data was loaded into a Google Colab notebook from the SBAnational.csv file from the dataset, and was extracted directly into the Colab environment. The CSV file was read into a DataFrame(sba) object using the pandas library for further processing and analysis.

The data cleaning process began by inspecting the DataFrame for missing values, incorrect data types, and redundant columns. Irrelevant columns that wouldn't be used to perform actions on the data and find results were dropped. These included columns such as location, bank information, and other irrelevant dates. Several columns also contained non-numerical data, which aren't directly usable in machine learning models. To simplify the analysis, these non-numerical columns were dropped rather than encoded. Additionally, empty and duplicated rows were identified and removed to reduce non-usable data in the dataset. Any missing values in critical columns were handled by removing rows, depending on their significance. Finally, outliers in numerical columns were reviewed and mitigated to avoid biasing the model. The final cleaned dataset was split into training and testing sets, ensuring a well-structured input for the machine learning model.

Calculations performed on the data included computing correlations between numeric variables to identify relationships. A correlation matrix and a heatmap visualization were generated to highlight significant patterns. This analysis helped uncover how features such as loan amount and business type were related to default rates. Additionally, confusion matrices and classification reports for each model provided a detailed breakdown of prediction accuracy and errors.

Relationships between the data were explored using correlation analysis and visualizations. This analysis was useful for understanding patterns and statistics that could influence loan defaults. The correlation matrix was computed using numeric columns from the cleaned dataset. This matrix provided values showing how much one variable moved in relation to another. A heatmap was then created, where darker and lighter colors represented the strength and direction of these relationships. This visualization made it easier to spot key correlations, such as between loan amount and default status. Additionally, grouping data by industry and calculating default percentages for each sector provided insights into how business type affects loan risk. These techniques helped uncover connections between borrower characteristics and loan outcomes.

The raw results included several important findings from the model training. The XGBoost model demonstrated the highest accuracy among the models tested, achieving a 96% accuracy score. Its 5-fold cross-validation score averaging 0.95 indicates its strong predictive performance for loan defaults (reference Figure 1).

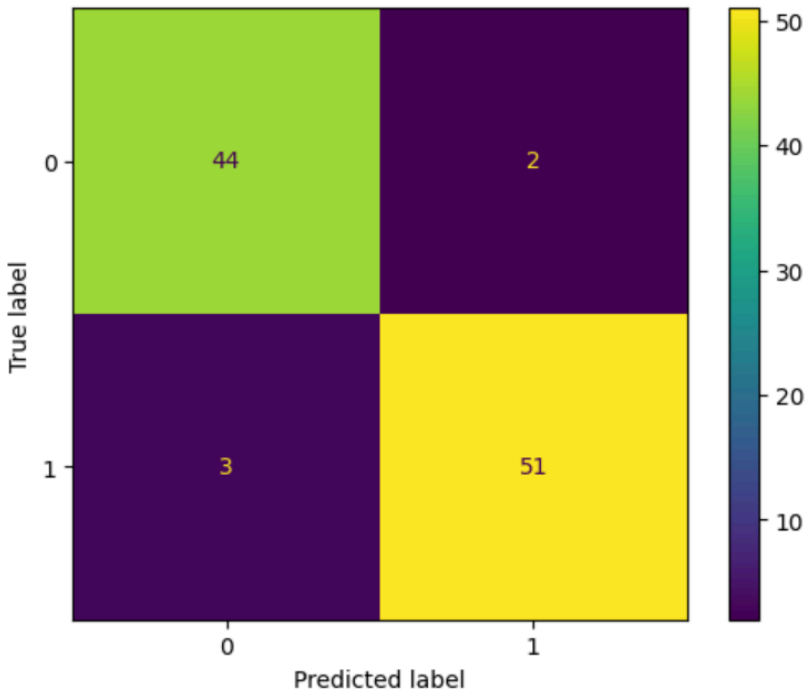

Figure 1. Confusion Matrix of XGBoost model.

The Random Forest model also performed well but was slightly less accurate. Its accuracy was 95%, and its 5-fold cross-validation score averaged 0.95, similar to the XGBoost model score (reference Figure 2).

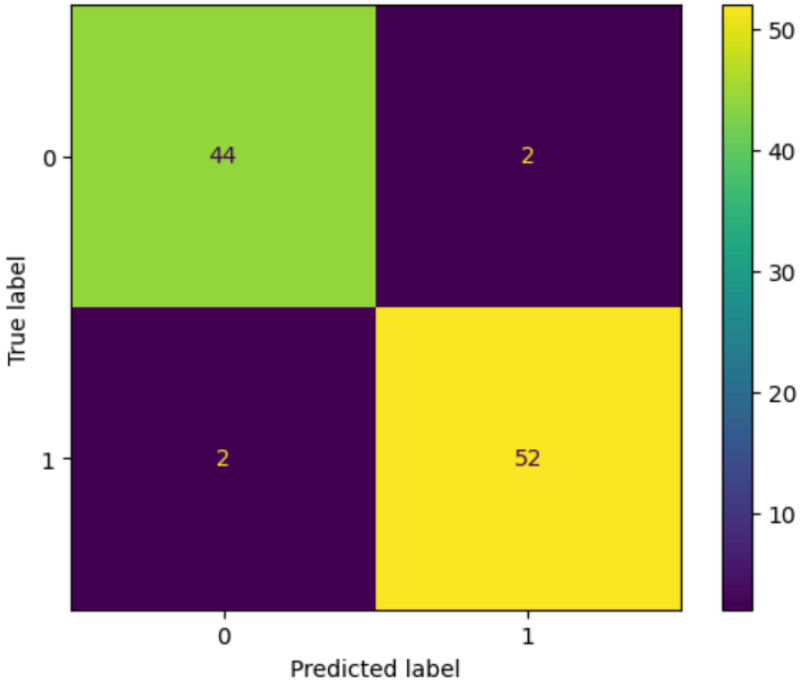

Figure 2. Confusion Matrix of Random Forest Classifier.

In contrast, the k-Nearest Neighbors (k-NN) model had a lower accuracy score at 91%, with a 5-fold cross-validation score averaging 0.91, likely due to the nature of the dataset and its reliance on distance-based classification. To further validate each model's performance, Stratified 5-Fold Cross-Validation was applied separately to XGBoost, Random Forest, and k-NN. This ensures that each fold maintains the same proportion of positive and negative loan outcomes, reducing bias in model evaluation. The graphs below (Figures 3-5) show the variation in accuracy across folds for each model, highlighting their consistency and potential overfitting or instability.

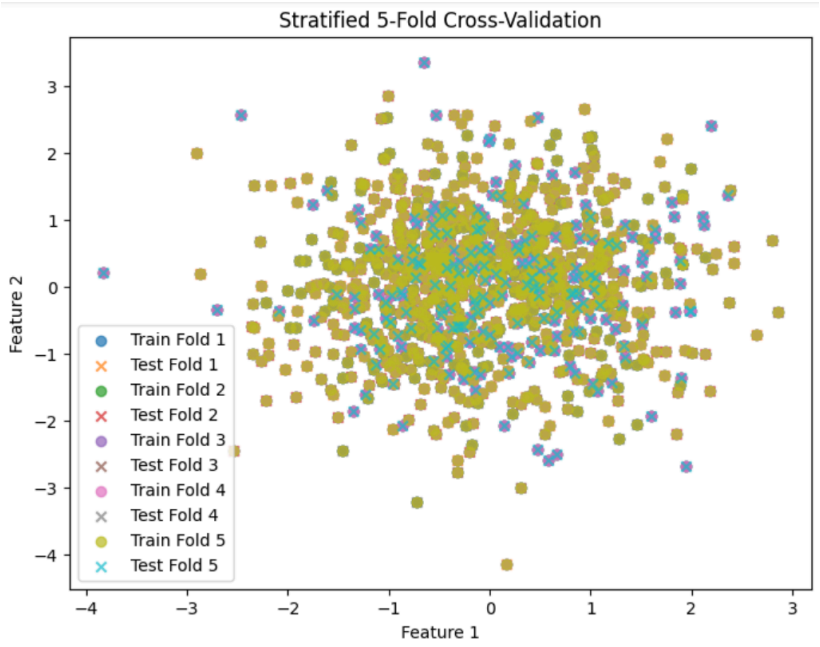

Figure 3. Stratified 5-Fold Cross-Validation graph of XGBoost model.

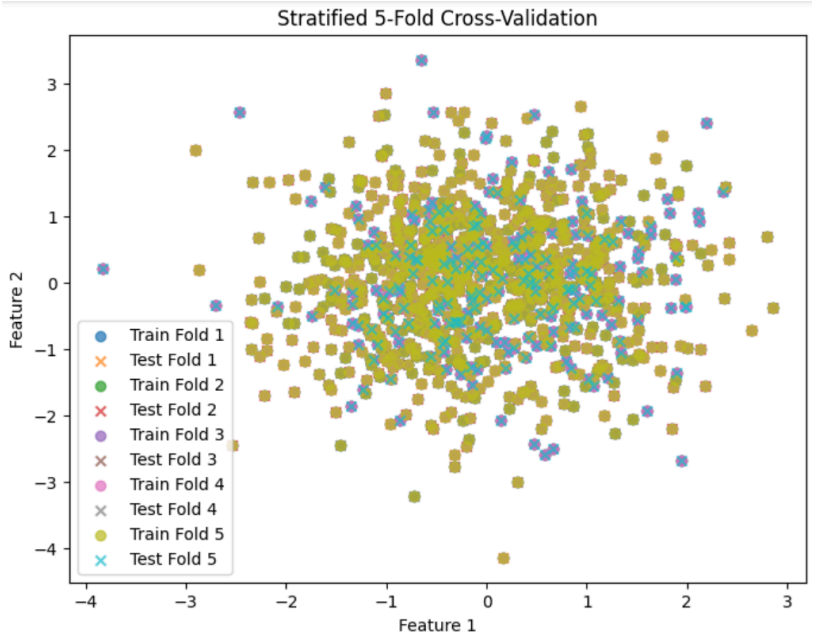

Figure 4. Stratified 5-Fold Cross-Validation graph of Random Forest Classifier model.

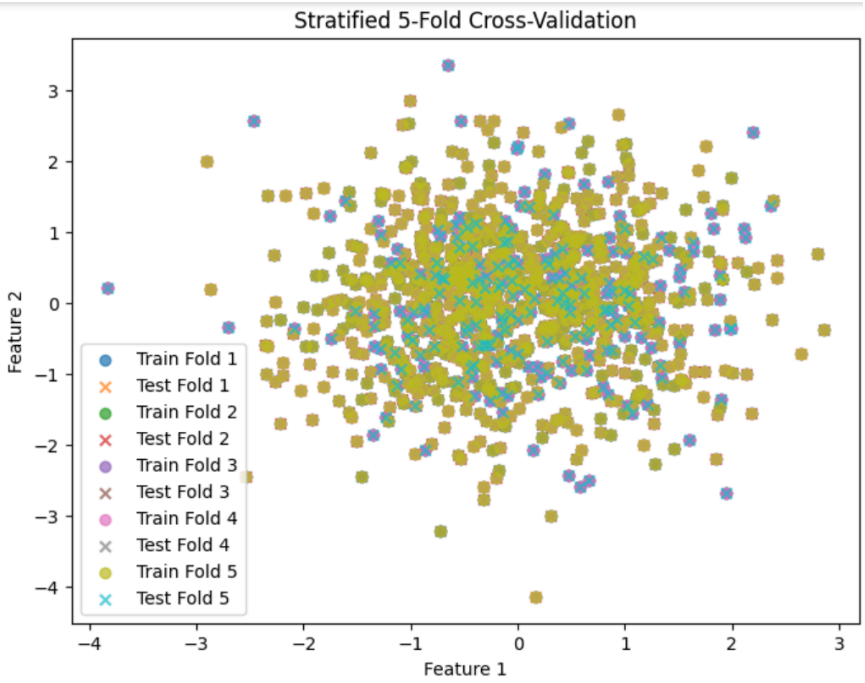

Figure 5. Stratified 5-Fold Cross-Validation graph of k-NN model.

The consistency between accuracy scores and cross-validation results suggests that the models generalize well to datasets similar to the one I used. However, slight variations in scores across folds may indicate sensitivity to certain features or imbalances in the dataset. Additionally, industry-specific default rates varied widely. Accommodation and Food Services had a much higher default rate than industries like Manufacturing and Finance, indicating greater financial risk in that sector. The correlation analysis also showed that larger loan amounts were associated with higher default likelihoods.

Analysis and Discussion

These results show that machine learning models can improve credit risk assessment. Traditional methods might miss important patterns that machine learning can detect. This makes models like XGBoost valuable tools for helping lenders make better decisions, which could reduce losses and improve their overall success. By using these predictions, financial institutions can lower their risk of loan defaults, make better use of their money, and provide better services to borrowers.

The high accuracy of the XGBoost model can help financial institutions improve risk assessment by identifying high-risk borrowers, potentially reducing defaults, and maintaining a healthier loan portfolio. Adjusting interest rates based on risk can enable better loan terms and promote financial inclusion by expanding access to credit beyond traditional scoring methods. Additionally, XGBoost can enhance fraud detection and support regulatory compliance, contributing to more efficient and data-driven lending decisions.

One interesting pattern I found was that certain industries had higher default rates, with Accommodation and Food Services businesses defaulting more often than those in Manufacturing or Finance. This may be because the food and hospitality industry depends on seasonal demand, economic conditions, and consumer spending, making revenue less stable. These businesses also tend to have lower profit margins and higher operating costs, increasing financial risk. In contrast, Manufacturing often has long-term contracts and valuable assets, while Finance is more regulated and stable. Recognizing

these differences can help lenders assess risk better and adjust loan terms accordingly.

Limitations and Future Work

My project had a few limitations. One big one was that the dataset didn't have important information like borrowers' credit scores or financial histories, which would have made the predictions more accurate. The borrower's history would have been really beneficial to the training of the model, as you could increase the prediction accuracy using their comprehensive history of paying off credit loans. I also didn't have economic data, like interest rates or unemployment, which can affect loan defaults. If I had more time, I could have tested different parameters for my models to improve them.

With more time and abundant resources, I would add more features to the data and try more advanced models, like deep learning. I'd also use explainable AI tools to show why a model made certain predictions so that banks could trust the results more. Finally, I would explore real-time data to keep the models updated with current economic trends, making the predictions even more useful for lenders.

Conclusion

In this study, machine learning algorithms, particularly XGBoost, significantly improved credit risk assessment for small business underwriting. The XGBoost model achieved a high accuracy rate of 96%, outperforming other models like Random Forest and k-Nearest Neighbors. These results demonstrate that machine learning can uncover hidden patterns in borrower data that traditional methods might miss, leading to more accurate predictions and better decision-making. The findings also revealed that certain industries, like Accommodation and Food Services, had higher default rates than others, such as Manufacturing and Finance, offering valuable insights into sector-specific risks. Despite the promising results, challenges like data privacy, algorithmic bias, and model transparency remain, requiring further research to make these models more ethical and interpretable. Machine learning can improve financial inclusion and optimize risk management when deployed responsibly.

References

1. Bello, O.A. (2023). *Machine learning algorithms for credit risk assessment: An economic and financial analysis*. https://www.researchgate.net/publication/381548370_Citation_Bello_OA_2023_Machine_Learning_Algorithms_for_Credit_Risk_Assessment_An_Economic_and_Financial_Analysis

2. Jansen, M., Nguyen, H., & Shams, A. (2020). *Human vs. Machine: Underwriting Decisions in Finance* https://www.jbs.cam.ac.uk/wp-content/uploads/2020/08/2020-06-conference-paper-jansen-nguyen-shams.pdf

3. Nguyen, H., Jansen, M., & Shams, A. (2023). *Machine learning-based profit modeling for credit card underwriting - implications for credit risk* https://www.sciencedirect.com/science/article/abs/pii/S0378426623000213

4. Sahu, M.K. (2023). *Machine learning algorithms for automated underwriting in insurance: Techniques, tools, and real-world applications* https://dlabi.org/index.php/journal/article/view/104

5. Tan, Y., & Zhang, G. (2023). *The application of machine learning algorithms in the underwriting process* https://ieeexplore.ieee.org/abstract/document/1527552

6. Toktogaraev, Mirbek (2020). *Should this Loan be Approved or Denied?* https://www.kaggle.com/datasets/mirbektoktogaraev/should-this-loan-be-approved-or-denied/code