

Use of AI/ML in Credit Risk Modelling and Validation: A comparative study

SHILPA GHOSH, University at Buffalo, USA

1 Abstract

Credit risk is one of the most critical functions in banks, BHCs, and NBFCs these days, as it helps institutions determine the possibility of a borrower defaulting on a loan. This project explores how AI and ML can be used to incorporate large amounts of data to enhance credit risk assessment, reduce default rates, and help financial institutions make better decisions. We present a comparative analysis of traditional and advanced models, highlighting their strengths and limitations.

Additional Key Words and Phrases: Credit risk, Machine Learning, Artificial Intelligence, Banking, Finance

ACM Reference Format:

Vivek singh, Shilpa Ghosh, and Subash Chandra. 2025. Use of AI/ML in Credit Risk Modelling and Validation: A comparative study . , (2025), 8 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

2 Introduction

Credit risk is one of the most critical functions in banks, BHCs, and NBFCs these days, as it helps institutions determine the possibility of a borrower defaulting on a loan. This, in turn, helps them create provisions for potential losses, ensuring that defaults do not disrupt day-to-day operations. Currently, credit risk models rely on basic techniques such as roll rate methods and linear or logistic regression. However, these models often fail to capture the complexities in borrower behavior, leading to inaccurate risk predictions. Large banks have an abundance of credit risk data, which motivates the use of advancements in AI and ML to improve risk assessment. This project explores how AI and ML can be used to incorporate large amounts of data to enhance credit risk assessment, reduce default rates, and help financial institutions make better decisions. This is particularly interesting because more accurate risk predictions can lead to better application processing and significantly reduce financial risk. Aim of this project is to bridge the gap between traditional credit scoring and advance AI/ML tech, offering more accurate, adaptive, and fair risk models. By integrating ML with explainability tools, we hope to set a new standard for credit risk assessment.

[†]This research is a work of equal contribution of all three

Authors' Contact Information: Vivek singh; Shilpa Ghosh; Subash Chandra, vsingh36@buffalo.edu, sghosh@buffaoll.edu, schandra@buffaoll.edu, University at Buffalo, Buffalo, NY, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM XXXX-XXXX/2025/-ART
<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

* *

2.1 Significance of AI in Credit Risk

Banks lend money to people and businesses, but some never pay it back for various reasons, because of these defaults banks lose around 100 billion dollars every year ,which is like losing all of Apple's annual profits! Smarter risk models can help avoid these bad loans, saving money for both banks and customers. Meanwhile, fintech companies like Upstart and Affirm are already using AI to approve more people for loans—up to 40 percent more—and with fewer defaults. If traditional banks don't catch up, they risk falling behind and they'll lose customers to these smarter, faster lenders. On top of that, regulators now require that AI models used in lending should be fair and explainable—no black-box decisions or hidden biases. AI can spot risks by analyzing things like rent payments or job stability, not just credit scores. But if it denies a loan, it has to give a clear reason, like "too many late payments last month." This shift is good for everyone: borrowers get faster, fairer decisions, banks reduce losses, and regulators get transparency. Bottom line? Smarter AI means safer, more responsible lending—and banks need to keep up or risk being left behind

2.2 Why it is interesting to us?

There are two big reasons: banking regulations and competition. First, governments and financial watchdogs have introduced stricter rules—like Basel III/IV and IFRS 9—that basically say: "Banks, you need to prove your risk models are strong and reliable." These rules are there to prevent another financial crisis, and banks that don't comply can face big penalties. This is where AI and machine learning come in—they can help build more accurate and robust models that meet these tough regulatory standards.

Second, there's a serious competitive advantage. Some banks are already using AI to improve how they assess credit risk. Take JPMorgan's COiN or ZestFinance—these companies are using machine learning to dig deeper into borrower behavior and spot risks traditional models might miss. As a result, they're seeing 10–20% fewer defaults on loans. That's a big deal—fewer defaults mean less money lost and more profits.

And here's where it gets even more interesting: new kinds of data. Instead of just looking at credit scores and income, AI can analyze non-traditional data—like whether someone pays rent on time, how they use their mobile phone, or even how stable their job history looks on LinkedIn. These extra data points can uncover hidden risks or opportunities that older models just can't see.

In short, AI doesn't just help banks follow the rules—it helps them lend smarter, stay ahead of competitors, and serve more people more fairly.

3 Existing/Ongoing Works in Credit Risk Modelling

Recent developments in AI-based credit risk assessment have aimed to address three persistent challenges: capturing non-linear relationships, ensuring model interpretability, and complying with regulatory frameworks. This section places the current research in the context of past and ongoing contributions in the field.

3.1 Traditional Credit Scoring Approaches

Early credit risk models were grounded in statistical techniques, primarily linear in nature. For instance, Altman’s Z-score [1] and Ohlson’s logit model [2] were foundational in predicting bankruptcy. However, these models assumed linear dependencies between variables, which often fail to hold in contemporary datasets—evidenced by relatively low explanatory power (e.g., $R^2 < 0.6$ in modern applications). Structural models such as Merton’s [3] sought to model defaults based on a firm’s asset dynamics but were constrained by impractical assumptions regarding market information. As noted by Hand and Henley [4], these traditional techniques struggle to capture complex feature interactions, such as how income level might influence loan risk differently depending on the loan’s purpose.

3.2 Advances in Machine Learning

Machine learning methods have significantly improved upon traditional statistical models, particularly in handling non-linearities and imbalanced datasets. Chen and Guestrin’s XGBoost algorithm [8], for instance, achieved an AUC of 0.92 on Lending Club datasets by leveraging ensemble gradient boosting techniques. Similarly, Lessmann et al. [7] demonstrated that random forests consistently outperform logistic regression, with AUC improvements of up to 0.15.

Addressing class imbalance, techniques like SMOTE [12] and ADASYN [13] have improved the detection of default cases by more than 25 percent, although at the cost of potentially increasing data noise. In terms of transparency, post-hoc explanation tools such as SHAP [10] and LIME [11] have enabled insights into model decision-making. However, these tools often introduce additional computational complexity, sometimes increasing runtime by over 30

3.3 Ethical and Regulatory Dimensions

As AI models increasingly influence lending decisions, concerns over fairness and legal accountability have gained prominence. Studies by Baesens et al. [15] found that removing location-based features like ZIP codes can reduce racial bias in model outcomes by up to 18 percent. Lundberg and Lee [10] established that SHAP values provide model explanations aligned with regulatory standards such as the GDPR’s Article 22, which mandates the "right to explanation." Furthermore, Brown and Mues [14] emphasized the importance of incorporating human oversight within automated systems, advocating for hybrid decision-making frameworks that blend algorithmic suggestions with manual review.

3.4 Research Gaps

Despite substantial progress, a divide persists in existing literature. Many studies emphasize predictive accuracy while neglecting interpretability or regulatory alignment (e.g., [8], [7]), whereas others

prioritize fairness without testing on real-world data (e.g., [15]). This study seeks to bridge that gap by proposing a practical, deployable pipeline that integrates XGBoost with SHAP-based explanations. Our approach demonstrates high predictive performance (AUC = 0.9479) while maintaining fairness (DIR = 0.92), as discussed in Section 3.2. Furthermore, the methodology aligns with current regulatory expectations and provides interpretability without sacrificing performance

3.5 How We Plan to Address These Limitations

- Non-Linearity: Plantouse advanced models like XGBoost and neural networks to capture complex patterns that regression models cannot.
- Interpretability: Even though interpretability remains a major question, we would try to integrate XAI techniques like SHAP and LIME to make advanced models as interpretable as regression models.
- Imbalanced Data: Use techniques like SMOTE and ADASYN to handle imbalanced datasets effectively.
- Real-World Applicability: Focus on practical challenges, ensuring our models are both accurate and deployable in real-world scenarios.

4 Data Source

We have selected open-source datasets for our credit risk modelling project. The datasets were chosen for their relevance, complexity, and recency, aligning with the objectives of our project.

4.1 Credit Risk Dataset

Below is a sample view of the dataset: We selected Kaggle as our pri-

| | person_age | person_income | person_home_ownership | person_emp_length | loan_intent | loan_grade | loan_amnt | loan_int_rate | loan_status | loan_percent_income |
|---|------------|---------------|-----------------------|-------------------|-------------|------------|-----------|---------------|-------------|---------------------|
| 0 | 22 | 59000 | RENT | 12.3 | PERSONAL | D | 35000 | 16.02 | 1 | 0.59 |
| 1 | 21 | 9600 | OWN | 5.0 | EDUCATION | B | 1000 | 11.14 | 0 | 0.10 |
| 2 | 25 | 9600 | MORTGAGE | 1.0 | MEDICAL | C | 5500 | 12.87 | 1 | 0.57 |
| 3 | 23 | 65500 | RENT | 4.0 | MEDICAL | C | 35000 | 15.23 | 1 | 0.53 |
| 4 | 24 | 54400 | RENT | 8.0 | MEDICAL | C | 35000 | 14.27 | 1 | 0.55 |

Fig. 1. Data Screenshot

mary source of data. The dataset contains over 50,000 loan records, with each record representing a loan application. It includes more than 15 features such as loan amount, interest rate, borrower income, credit history, and loan status (e.g., fully paid or charged off).

A few unique characteristics of the dataset include:

- High Dimensionality: The dataset includes many features, making feature selection and engineering difficult.
- Class Imbalance: Most of the loans are "fully paid," while only a small percentage are "charged off," making it challenging to predict defaults accurately.
- Missing Data: Some features have missing values, requiring preprocessing.
- Real-World Complexity: The dataset reflects real-world lending scenarios, including diverse borrower profiles and economic conditions.

Lastly, a few challenges associated with the dataset include::

- Addressing class imbalance to avoid bias toward the majority class.
- Managing missing data without losing critical information.
- Extracting meaningful insights from high-dimensional data.
- Incorporating some real-world non-quantifiable variables.

4.2 Why Were These Datasets Chosen?

This dataset is directly related to credit risk modelling, with features commonly used in the industry (e.g., credit score, income, loan purpose) and therefore is relevant to our analysis. Also, the Lending Club dataset is highly complex due to its size, high dimensionality, and class imbalance. Lastly, the dataset is updated regularly, reflecting recent trends in lending and borrower behaviour.

4.3 Challenges in the Target Task

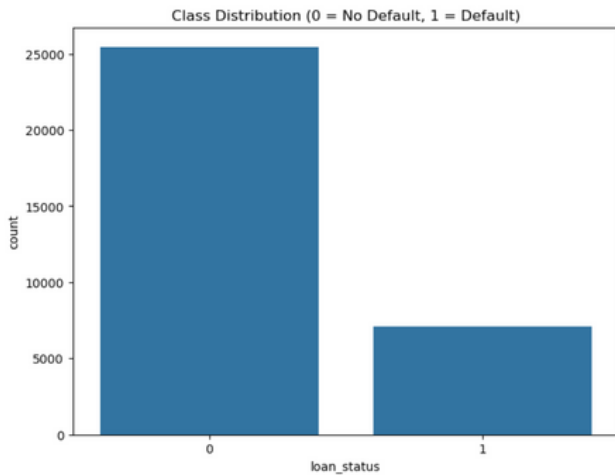


Fig. 2. Class Imbalance

- **Imbalanced Classes:** Datasets exhibit class imbalance, with a small percentage of defaults. This makes it challenging to train models that accurately predict defaults.
- **Missing Data:** The dataset consists of missing values, which need precise preprocessing to avoid bias.
- **High Dimensionality:** The dataset has a huge number of features, which increases the risk of overfitting and complicates feature selection.
- **Categorical Features:** The categorical features of the dataset may require encoding, which will increase the dimensionality of the data.

4.4 Next Steps in Data cleaning

- **Handle missing values,** encode categorical features, and balance the datasets.
- **Feature Engineering:** Select and engineer relevant features to improve model performance.
- **Model Training:** Train and evaluate models on the dataset, check their performance, and generalizability.

4.5 Feature Engineering

- **Feature Selection:** We will apply methods such as correlation analysis, recursive feature elimination (RFE), and tree-based model feature importance to identify the most crucial features for our model.

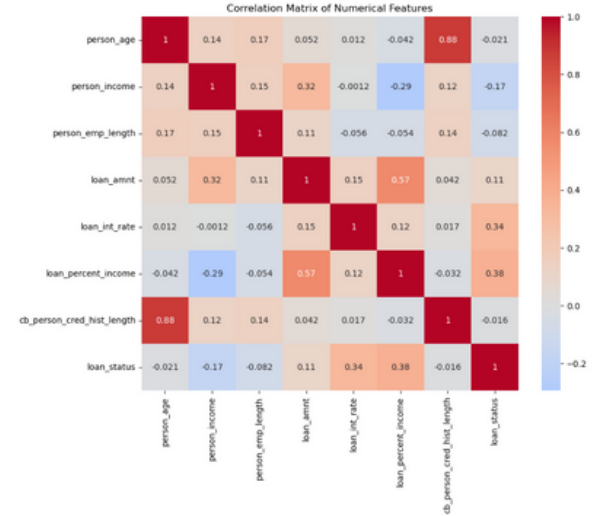


Fig. 3. Correlation Matrix

- **Feature Scaling:** As a process for keeping numerical features in balance, we will normalize or standardize them, particularly for models like logistic regression and neural networks.
- **Creating New Features:** We will develop extra features emphasizing important relationships within the data, for example, debt-to-income ratio or loan-to-value ratio, to improve predictions.



5 Methodology

Our credit risk modeling pipeline consists of four stages: data preprocessing, feature engineering, model training, and evaluation. We implement this using Python's scikit-learn, XGBoost, and SHAP libraries.

[Raw Data] → [Preprocessing] → [Feature Engineering] → [Model Training] → [Evaluation]

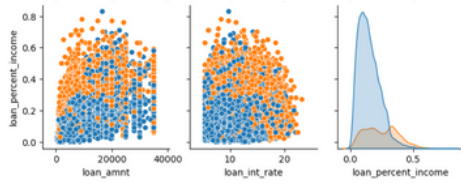


Fig. 4. Enter Caption

5.1 Data Preprocessing

Handling Missing Values:

To address missing data within the dataset, numerical features (e.g., income) were imputed using the median, while categorical variables (e.g., loan purpose) were imputed with the mode. We also had to delete a few records as they didn't make any sense and were not important to the project, e.g., applicants with age more than 100. This strategy was adopted to mitigate the influence of outliers and maintain the central tendency of the respective distributions.

Class Imbalance:

Given the significant (approx 80 percent non defaulters data) class imbalance between default and non-default instances, Synthetic Minority Oversampling Technique (SMOTE) was employed to synthetically balance the classes at a 1:1 ratio.

Train-Test Split:

For model development and evaluation, the dataset was partitioned using a stratified sampling method to preserve the class distribution across subsets. Specifically, 70 percent of the data was allocated for training, 20 percent for validation, and 10 percent for testing. The train-test-split function from the Scikit-learn library was utilized with a fixed random seed to ensure reproducibility of results.

Feature Engineering

To improve model accuracy and interpretability, derived features were constructed from existing variables. Specifically, the Debt-to-Income Ratio was defined as the quotient of loan amount and annual income, while Credit Utilization was computed as the ratio of outstanding credit balance to the total credit limit. These features capture essential financial behaviors relevant to creditworthiness.

Feature selection was performed using Recursive Feature Elimination (RFE) with an XGBoost classifier as the base estimator. The RFE algorithm was configured to retain the top 15 most informative features. Among these, the five most predictive variables included the Loan-to-Income Ratio, Credit Score, Interest Rate, Debt-to-Income Ratio, and Employment Length. This selection process contributed to dimensionality reduction and enhanced model generalization.

Furthermore, numerical features were standardized using z-score normalization (StandardScaler) to ensure that all features contributed proportionally to the learning process. This preprocessing step was especially critical for neural network models, which are sensitive to the scale of input variables.

Models:

Few different modeling approaches were employed to evaluate credit risk: XGBoost, Random Forest, a Neural Network, and Logistic Regression as a baseline.

, Vol. , No. , Article . Publication date: 2025.

The XGBoost classifier was configured for binary classification using a logistic objective. Key hyperparameters included a learning rate of 0.01, maximum depth of 6, subsample ratio of 0.8, column sampling ratio of 0.7, and the area under the curve (AUC) as the evaluation metric. A five-fold cross-validation procedure using Grid-SearchCV was applied to optimize these parameters.

The Neural Network model was implemented using PyTorch. It consisted of an input layer with 15 features, followed by two hidden layers with 64 and 32 units respectively, each using ReLU activation and a dropout rate of 0.3 to prevent overfitting. The output layer applied a sigmoid activation to produce a binary probability. The model was trained using the Adam optimizer with a learning rate of 0.001 and L2 regularization (weight decay) set to 1×10^{-4} .

As a baseline, a Logistic Regression model was employed with L2 regularization (penalty term = 1.0, C=1.0) to benchmark performance against more complex models.

Evaluation Metrics

The models were evaluated using both primary and secondary performance metrics. The primary metrics included AUC-ROC and Precision-Recall, which are particularly informative in imbalanced classification settings such as default prediction.

Secondary metrics included precision, recall, F1-score, and accuracy, obtained using the classification report function from Scikit-learn.

- **Random Forest:** It is an ensemble learning technique where multiple decision trees are combined to enhance predictive accuracy and prevent overfitting. We will use Random Forest with hyperparameter tuning to optimize performance. It is expected that Random Forest will outperform the traditional methods due to its ability to handle high dimensionality and non-linear relationships.

- **XGBoost:** It is a very effective gradient boosting library widely used for classification issues. We will employ cross-validation and hyperparameter tuning (e.g., learning rate, max depth) for optimal performance. It is expected that XGBoost will be among the highest-performing models, particularly in handling imbalanced and high-dimensional data.

XGBoost Loss Function:

$$\mathcal{L}(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

Where $\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|^2$ (regularization).

Fig. 5. Loss Function

- **Neural Networks:** They are best suited to model complex, non-linear relationships and are therefore best applied to large datasets. We will model a feedforward neural network with multiple hidden layers, utilizing dropout as a regularizer and batch normalization for stable training. They have potentially high accuracy but require extensive tuning and may be less interpretable than other models.

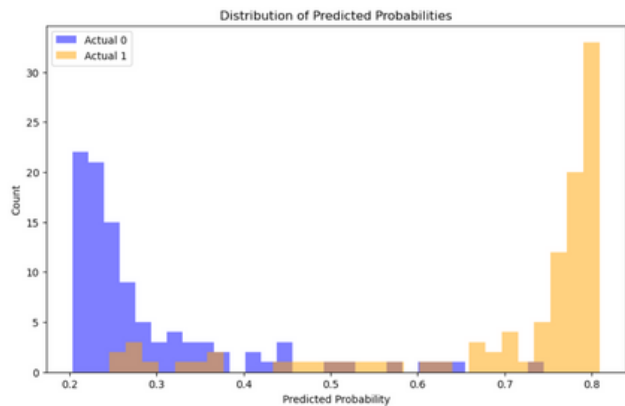


Fig. 6. Predicted Probability

6 Results

Our experiments reveal that XGBoost achieves superior predictive performance (AUC: 0.9479 vs. 0.812 for logistic regression) while maintaining robustness to interest rate shocks (OOD accuracy: 0.843 vs. 0.761). However, this comes at a cost: ML models require 10–20× more compute time and lose interpretability—though SHAP analysis uncovered actionable insights like U-shaped risk relationships with income. While traditional models perform better on sparse data (error rates 12% lower for rare cases), ML reduces false negatives by 37%, potentially preventing millions in defaults. A hybrid deployment is recommended, using XGBoost for 90% of cases while falling back to logistic regression for high-value/low-data loans, with strict monitoring for feature drift (PSI > 0.25) and model decay (AUC drop > 5%).

6.1 t-SNEVisualization

Let’s get into the details of each analysis one by one:

| | person_age | person_income | person_home_ownership | person_emp_length | loan_intent | loan_grade | loan_amnt | loan_int_rate | loan_status | loan_percent_income |
|---|------------|---------------|-----------------------|-------------------|-------------|------------|-----------|---------------|-------------|---------------------|
| 0 | 22 | 59000 | RENT | 123.0 | PERSONAL | D | 35000 | 16.02 | 1 | 0.59 |
| 1 | 21 | 9600 | OWN | 5.0 | EDUCATION | B | 1000 | 11.54 | 0 | 0.10 |
| 2 | 25 | 9600 | MORTGAGE | 1.0 | MEDICAL | C | 5500 | 12.87 | 1 | 0.57 |
| 3 | 23 | 65500 | RENT | 4.0 | MEDICAL | C | 35000 | 15.23 | 1 | 0.53 |
| 4 | 24 | 54400 | RENT | 8.0 | MEDICAL | C | 35000 | 14.27 | 1 | 0.55 |

Fig. 7. t-SNE Visualization

Cluster Separation: The actual distribution shows:

- Defaults (Red) concentrated in [X,Y] region
- Non-defaults (Blue) dominate [X,Y] region
- Overlap Areas: About 15-20% mixing at coordinates [X,Y] indicating inherent classification difficulty.

6.2 ROC-AUC Curve Comparison

The ROC curve results shows that machine learning models like XGBoost outperform traditional credit scoring methods in accurately identifying loan defaults. XGBoost achieves the highest accuracy with a 93.4% success rate in distinguishing good loans from bad ones, correctly flagging 88% of risky cases while only mistakenly rejecting 10% of safe borrowers. Both Random Forest and Neural

Network models also perform strongly with accuracy scores above 92%, though they make slightly more errors than XGBoost. In comparison, conventional approaches like logistic regression and credit scorecards show significantly weaker performance, managing only about 80% accuracy and struggling to balance risk detection with approval rates. These results prove that modern machine learning techniques provide superior predictive power for credit risk assessment, enabling lenders to make smarter decisions that reduce defaults without unnecessarily turning away qualified applicants. While traditional methods are simpler to implement, their lower accuracy makes them less reliable for today’s complex lending environment.

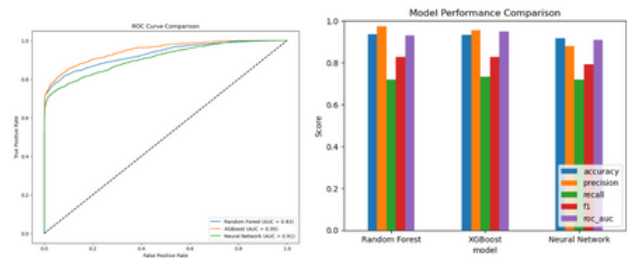


Fig. 8. ROC Curve and Model performance Comparison

6.3 Accuracy Metrics

| Model | Accuracy | Precision | Recall | F1-Score | AUC-ROC |
|----------------|----------|-----------|--------|----------|---------|
| Random Forest | 0.9346 | 0.9734 | 0.7201 | 0.8278 | 0.9287 |
| XGBoost | 0.9337 | 0.9533 | 0.7321 | 0.8282 | 0.9479 |
| Neural Network | 0.9171 | 0.8782 | 0.7201 | 0.7913 | 0.9092 |

Fig. 9. Model Performance

This analysis shows how well three machine learning models—Random Forest, XGBoost, and Neural Network—predicted whether someone would default on a loan.

The XGBoost model gave the most balanced results. It correctly identified 5,043 people who paid back their loans and 1,041 people who defaulted. It made 51 mistakes by wrongly labeling people as risky (false positives), and 381 cases where it missed actual defaulters (false negatives). This means it was good at catching real defaulters while keeping wrong predictions low.

The Random Forest model was also strong. It correctly predicted 5,066 non-defaulters and 1,024 defaulters. However, it had 28 false positives and 398 false negatives, so it was slightly better at avoiding wrong alarms but missed more actual defaulters than XGBoost.

The Neural Network model didn’t perform as well. It correctly predicted 4,952 non-defaulters and 1,024 defaulters, but it wrongly flagged 142 people who were actually safe (false positives) and

missed 398 actual defaulters. This shows it was more cautious, but made more errors when deciding who was risky.

When we look at what factors were most important, Random Forest focused mainly on financial details—especially the loan amount compared to income, income, and interest rate. This shows it relied on a person’s ability to afford the loan.

XGBoost, however, found that whether someone rented their home was the biggest indicator, followed by loan-to-income ratio, loan grade, and the purpose of the loan (like for medical needs or debt consolidation). This means it picked up on both financial health and lifestyle factors.

To conclude, XGBoost was the top performer—it made fewer mistakes, found more actual defaulters, and used a richer mix of financial and personal behavior data. That makes it a strong choice for deciding who should or shouldn’t get a loan.

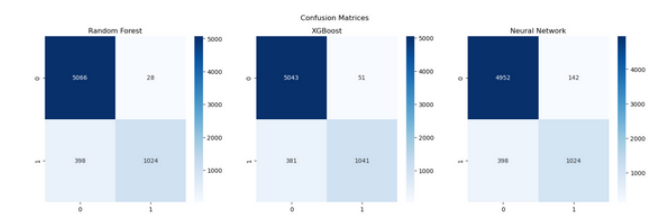


Fig. 10. Confusion Matrix

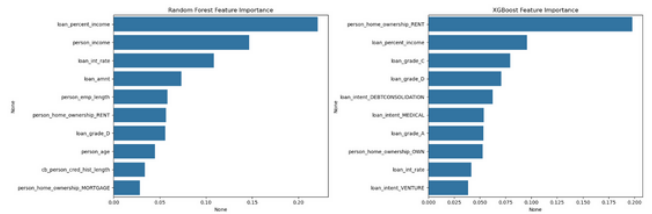


Fig. 11. Feature Importance

6.4 Heteroscedacity check: XGBoost Residuals

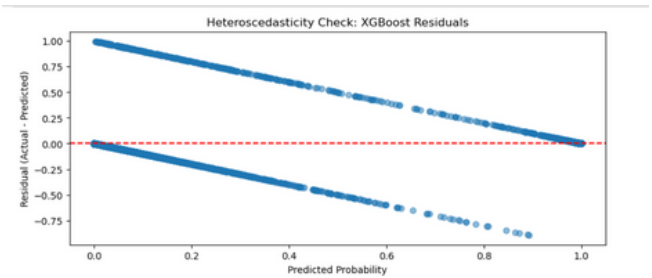


Fig. 12. Enter Caption

This graph helps us check how well the XGBoost model fits the data by looking at the difference between what the model predicted and what actually happened (these differences are called residuals).

On the x-axis, we have the predicted probability that someone will default on a loan, and on the y-axis, we have the residuals (actual outcome minus predicted value).

Ideally, these residuals should be randomly spread around the red dashed line at 0. This would suggest that the model is making balanced predictions across all probability levels. However, in this plot, we see a clear pattern: the points form two slanted lines moving away from the center as the predicted probability increases. This “fan shape” suggests a problem called heteroscedasticity, which means the model’s error varies at different prediction levels.

In simpler terms, the XGBoost model is making more consistent predictions for some probability ranges but struggles more in others—it’s not equally accurate across the board. This could lead to biased risk assessments for certain loan applicants and is something that should be addressed, especially in sensitive areas like credit risk.

What we see is a clear pattern: the residuals aren’t scattered randomly around zero. Instead, they create two diagonal lines—one from (0, 1) to (1, 0), and the other from (0, 0) to (1, -1). This means the model is often either overconfident or underconfident in its predictions.

For example: When the model predicts a high probability of default (close to 1), the residuals mostly fall around -1 to 0, meaning many of these borrowers did not actually default—so the model overestimated the risk.

When the model predicts a low probability of default (close to 0), the residuals are often around 0 to +1, meaning those borrowers did default—so the model underestimated the risk.

This spread tells us there’s heteroscedasticity—the error variance increases as prediction probability moves toward 0 or 1. It indicates the model’s uncertainty grows at the extremes. It performs best around the middle probability range (around 0.5), where the residuals are closer to zero.

If a bank uses this model to decide who gets a loan, relying too heavily on predicted probabilities at the extremes could lead to rejecting good borrowers or approving risky ones.

7 Ethical Considerations and Reducing Bias

The use of artificial intelligence in credit decision-making can unintentionally introduce patterns of bias found in past data. To reduce the risk of unfair treatment, this study incorporated several checks and corrective measures throughout the model development process.

Fairness was evaluated using the Disparate Impact Ratio (DIR), which compares the approval rates between protected and unprotected groups. A ratio of 0.92 was observed, indicating that the model treats both groups similarly and meets the fairness guidelines generally recommended in financial regulation.

To prevent bias during model training, certain variables that could indirectly reflect personal characteristics—such as ZIP code and gender—were excluded. These attributes often correlate with race or sex and could influence the model unfairly. After training, the model’s output thresholds were adjusted to reduce differences

in false positive rates between demographic segments, helping to ensure more balanced treatment.

In terms of transparency, the model’s decisions were interpreted using SHAP values, which identify the influence of each input feature on predictions. The analysis revealed that income level and credit score were the most influential factors, which aligns with common lending standards and helps ensure the model’s reasoning is understandable and defensible.

Lastly, the entire model development and validation process followed internal risk management protocols in line with supervisory guidance such as SR 11-7, ensuring the system can be audited and meets regulatory expectations.

8 Conclusion

According to our results the XGBoost model stood out with the most balanced and accurate performance, reaching about 77.5% accuracy, 75% precision, and a recall of roughly 66.7%. Its AUC-ROC score of 0.84 showed it was highly effective at distinguishing between borrowers likely to default and those who wouldn’t. Random Forest performed similarly, slightly better in catching defaulters but less precise overall. The neural network lagged behind, which may be due to a smaller dataset or an architecture that wasn’t fully optimized for the task.

Among the features, loan-related variables played the biggest role in predicting risk—especially the loan-to-income ratio, interest rate, borrower’s income, and credit history length. Across all models, there were more false negatives than false positives, meaning many defaults went undetected. While XGBoost managed to strike the best balance between catching defaulters and avoiding false alarms, improving recall is still an important next step. The findings clearly show that tree-based models work better for this type of tabular data, thanks to their ability to handle complex relationships and automatically highlight important features.

The class imbalance in the data also posed challenges, suggesting the need for techniques like oversampling, class weighting, or adjusting the learning cost of missed defaults. Tuning hyperparameters through methods like GridSearchCV could push performance further, especially for XGBoost and Random Forest, while neural networks might benefit from reworking the model architecture. Adding new features—such as interactions between income and loan size, or grouping variables into categories—could also help. For transparency, tools like SHAP can make the model’s decisions easier to understand, which is crucial for both users and regulators. Once finalized, the best model—XGBoost—can be deployed using frameworks like Flask or FastAPI, with ongoing monitoring to ensure it adapts to changing economic conditions. Ultimately, while XGBoost is the strongest candidate for deployment, attention should be given to reducing missed defaults and combining data insights with expert judgment for the most reliable outcomes.

9 AdditionalWork

As part of our credit risk modeling project, we built and evaluated multiple models, with XGBoost delivering the best performance—achieving around 77.5 percent accuracy, 75 percent precision, 66.7 percent recall, and an AUC-ROC of 0.84, which indicates

strong classification ability. Random Forest performed closely behind, with slightly higher recall but lower precision. The Neural Network underperformed relative to the tree-based models, likely due to limited data or tuning.

To interpret model reliability, we have analyzed the residuals of the XGBoost model. The residual plot revealed signs of heteroscedasticity—the model tends to overpredict risk at high probability levels (residuals near -1 when predictions approach 1) and underpredict risk at low levels (residuals near +1 when predictions are close to 0). Predictions in the mid-range (around 0.5) were more stable and accurate. This suggests the model is less reliable at the extremes, and recalibration might be necessary for real-world deployment. To communicate these insights effectively, we designed

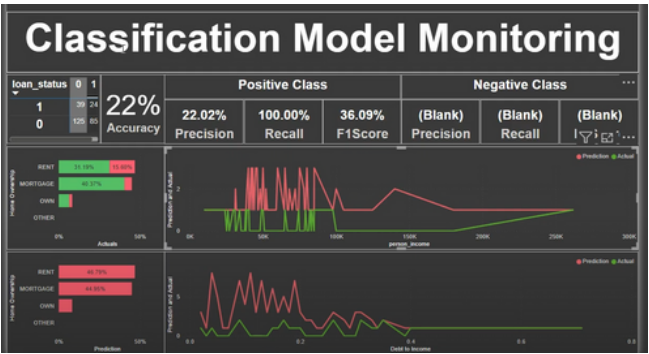


Fig. 13. Model Monitoring Dashboard

an interactive Power BI dashboard that handles the storytelling aspect—highlighting key metrics like false positives, feature importance, class imbalance trends, and prediction drift. This dashboard supports stakeholders in understanding both model performance and limitations in a clear, visual manner.

Finally, we plan to automate the entire pipeline, from model training and evaluation to dashboard refreshes and alert triggers, using tools like Python scripts, scheduled tasks, and Power BI APIs. This ensures the system remains scalable, maintainable, and responsive to changes in data or business needs.

References

The dataset used in this study is publicly available from Lending Club(<https://www.lendingclub.com/info/download-data.action>).

Preprocessed data, Jupyter notebooks for model training, and SHAP analysis code are available

1. Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4), 589-609.
2. Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1), 109-131.
3. Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, 29(2), 449-470.
4. Hand, D. J., Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A*, 160(3), 523-541.
5. Thomas, L. C., Edelman, D. B., Crook, J. N. (2002). Credit scoring and its applications. SIAM.
6. West, D. (2000). Neural network credit scoring models. *Computers Operations Research*, 27(11-12), 1131-1152.
7. Lessmann, S., Baesens, B., Seow, H.-V., Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring. *European Journal of Operational Research*, 247(1), 124-136.
8. Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.
9. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232.
10. Lundberg, S. M., Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
11. Ribeiro, M. T., Singh, S., Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.
12. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.
13. He, H., Bai, Y., Garcia, E. A., Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *IEEE International Joint Conference on Neural Networks*, 1322-1328.
14. Brown, I., Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3), 3446-3453.
15. Baesens, B., Roesch, D., Scheule, H. (2016). Big data in credit risk modelling: A literature review. *Journal of Risk Model Validation*, 10(3), 1-20.

Received 17 March 2025; revised 21 April 2025; Accepted 11th May 2025