Proceedings of the 3rd International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/2025.31254

Enhancing Loan Approval Systems in Digital Banking: A Data-Driven Framework for High-Recall Predictions

Yichen Liu[1*], Xiaoyue Ni[2]

[1]*College of Agricultural and Environmental Sciences, University of California, Davis, USA*
[2]*Wuxi Dipont School of Arts and Science, Wuxi, China*
**Corresponding Author. Email: axy1680@gmail.com*

Abstract. The study proposes a machine learning framework that optimizes the approval processes of Neo Banks loan applications and overcomes the limitations of traditional credit scoring models. Using feature engineering, ensemble methods, and hyperparameter optimization on customer demographics, financials, and transaction details to determine loan acceptance. The results show that the improved models have the accuracy of 97.87% and 88.59% recall and are far better compared to traditional methods (logistic regression: 63.76% recall). Significant predictors are education level (importance: 0.35), income (0.30), and family size, with the probability of approval among high-income, high-education customers being 45% higher. The framework reduces the false negative rate, allowing the Neo Banks to focus on the cream of the crop applicants but avoiding risks. Examples of practical strategies are customized marketing and dynamic pricing. This work talks about limitations such as data imbalance, and future research suggests integrating real-time behavioral data and fairness-aware modeling. It fills a gap between tech innovation and operational requirements, providing a scalable method for updating credit risk evaluation in digital banking.

Keywords: Machine Learning, Loan Approval Optimization, Digital Banking

1. Introduction

The first Neo Banks was First Direct, launched in the UK in 1989. It grew quickly because they can provide a more convenient and user-friendly banking experience than traditional banks. Neo Banks do not have the same overhead costs as traditional banks, so they are able to offer lower fees and better interest rates. Loan approval is an extremely critical and time-consuming process for banks [1]. However, this transformation has brought challenges as the rapid development of Neo Banks and the expansion of market size, especially in areas such as credit risk assessment and loan approval. The evaluation process for loan approval entails assessing a borrower's financial credibility, usually relying on credit scores, proof of income, and historical financial data. However, these traditional approaches may prove inadequate [2]. Accurate prediction of credit choices is critical for a number of reasons. Initially, it has a direct effect on the monetary stability of emerging banks. The lack of prediction models may lead to significant economic losses due to rising default

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2025.31254

rates, while an overly cautious approach may also lead to negligence of creditworthy borrowers' income prospects [3].

Traditional financial models predict the likelihood of a borrower defaulting on their loans by using linear regression and similar statistical techniques. The old system relies heavily on historical data, which more often than not does not reflect either the current financial environment or recent developments affecting borrowers, leading to many creditworthy applicants being denied loans entirely due to lack of data. Moreover, traditional models are primarily concerned with quantitative metrics and will likely conceal the qualitative variables impacting a borrower's ability to repay, according to VNET in their 2024 report [4]. Current machine learning models are capable of improving the prediction accuracy of loan approvals and can use techniques like decision trees, random forests, and neural networks on large-scale datasets to help harvest patterns that traditional models might overlook. Yet, despite the sophistication and capabilities of current ML models, there are constraining limitations that preclude contextualizing the non-financial factors that affect the borrower's situation, according to Gao et al. in their 2024 research article [5].

Section 2 is a comprehensive literature review that presents significant studies on models used for loan acceptance, their approaches, and methodologies. Section 3 proves the long description of the study area and the data preprocessing performed to analyze the dataset. The preference and application of predictive models, such as Decision Trees, Logistic Regression, Support Vector Machines, are the topics of Section 4. The set-up of experiments, descriptions of performance metrics, and results from the comparative model analysis are placed in Section 5. Section 6 deals with major insights, implications of findings, and directions for future research. Finally, Section 7 provides a discussion of the implications of the results, as well as limitations as well as areas for future research, to strengthen the application of predictive modeling in the area of loan acceptance.

2. Literature review

Many research studies have heretofore relied on traditional statistical methods, of which logistic regression is usually a common one. The reason is due to the interpretability and simplicity of this model, besides being suitable for the assessment of binary outcomes like approval or default [6,7]. Such logistic regression models work by estimating that, based upon income, age, and employment status, a borrower may be a defaulter [8]. Such credit predict models are similar to FICO, which is credit scoring widely used [9]. The methodology of credit scores is providing numerically based on weighted averages of credit utilization, repayment history, and others of financial behavior and risk factors [10]. The static traditional credit scoring model is sufficient for primary risk assessment but does not adapt to changes in financial profiles over time [11], hence the need for the financial industry to seek further powerful machine-learning methods.

ML models have become a powerful tool for the financial industry. They are capable of doing more efficient loan approvals and better predicting defaults. SVM is particularly effective in high-dimensional spaces, thus working well in cases where the number of features exceeds the number of observations, which renders it a pasture well suited for complex caveats of financial data [12,13]. Random Forest enhances the robustness of the model while reducing the overfitting by averaging the decisions from a multitude of decision trees [14]. Both have proved to outperform traditional models by capturing the nonlinear relationships and interactions between the variables [15].

Gradient-Boosted Machines (GBM) can minimize the error rate by sequentially building weak learners, making it effective for credit risk assessment and prediction of loan defaults [16]. Louzada and others found in 2016 that GBM proved to be able to outperform classic credit scoring models in predicting loan defaults, especially when accompanied by feature engineering techniques [17]. In

DOI: 10.54254/2754-1169/2025.31254

addition, the machine learning model has features ranging from social media activity and transaction history to create accurate borrower profiles [18]. In the case of models whose use would fit more continuous data, like credit scores or time-series analysis of trading patterns, deep learning ones like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are also taking off in the financial sector [19,20]. Banks can use CNNs due to their excellent feature extraction abilities for image recognition setting and structured datasets [21]. While these two are lower in interpretability than traditional methods, they lend themselves to good predictive power.

Nevertheless, machine learning models for loan prediction still have some limitations. Models relying solely on fundamental consumer characteristics, such as income and credit ratings, may only partially represent a borrower's risk profile. Utilizing behavioral and transactional data raises problems regarding data privacy and ethical considerations [22,23]. Moreover, the quantity of sanctioned loans significantly surpasses that of defaulted loans, thus skewing the model to overestimate authorized loans and underestimate defaulted loans in numerous loan datasets, presenting the challenge of data imbalance [24].

3. Study areas and data preprocessing

The study leveraged a comprehensive dataset comprising customer demographics, financial metrics, and transactional behaviors to build a predictive model for personal loan acceptance. Specific key datasets included age, income, CCAvg, education level, family size, loan amounts, and dummy variables in relation to financial products. The dependent variable was acceptance status for personal loans, providing an analysis of the factors driving customer conversions.

Data preprocessing was extensive and necessary to ensure that the model functioned well; henceforth, it was checked for quality, and no records were found with a missing value or duplicative sources. The first step in the process was simpler because no missing values or duplicates were found, allowing the cleaning process to continue. Influential variable points such as income and CCAvg were adjusted to fall between the 5th and 95th percentiles, keeping them from skews in the data and losing their utility.

Feature engineering then improved the predictive power of the dataset. Continuous variables, namely income and CCAvg were normalized using logarithmic transformation on their distributions. Categorical features such as education level (undergraduate, graduate) and ZIP code were encoded as dummy variables so as to facilitate model interpretation. Temporal trends were then captured indirectly through derived measures such as customer tenure, which could be inferred from age and work experience.

Also, segments of the dataset on loans accepted rates by the income-level brackets and family-size rubric were generated for seeing trends within subgroups. This stratification demonstrated that postgraduate-based high earners had greater acceptance rates of loans when viewed through bivariate exploratory data analysis.

At last, Min-Max scaling involved standardizing numerical features (such as income or CCAvg) into a [0,1] range to ensure that they were equally weighted when inputted into the model training process. These preprocessing steps prepared the dataset to allow the decision tree model to capture nonlinearities effectively while also achieving a high degree of predictive accuracy.

4. Methodology

Algorithms in machine learning develop and learn in various manners. The most general term in-line with their generalizations remains a decision tree. Random forests utilize multiple decision trees

DOI: 10.54254/2754-1169/2025.31254

under certain protection mechanisms and ensemble learning capacities to generalize better and reduce the effects of overfitting. This covers the basic idea that encapsulates stability because it addresses the complexity of all possible interactions between features while remaining robust.

XGBoost and Artificial Neural Networks (ANNs) were designed with a comparison in mind so that both could be evaluated for their Loan Prediction capabilities. XGBoost, based on gradient boosting, builds trees to cut down on errors one by one, making it highly suitable for imbalanced datasets. ANNs were used in the investigations to explore nonlinear pattern recognition on high-dimensional data by appropriately using the backpropagation and activation functions to abstract various features.

Feature engineering greatly improved model performance. Categorical variables, such as what kind of occupations (like "self-employed," "salaried") and the geographic region of employment, were then encoded by one-hot encoding to avoid unwanted ordinal bias. Temporal features, such as frequency of applications for loans quarterly, were turned into cyclical variables to catch the trend within seasons. Interaction terms between credit utilization ratio and monthly savings rate were introduced to assess combined effects on loan eligibility.

To optimize the Random Forest model, hyperparameter tuning was conducted via Bayesian optimization. Key parameters, such as the number of trees (n_estimators), maximum depth, and minimum samples per leaf, were adjusted to balance bias-variance trade-offs. Post-training, feature importance analysis revealed that credit score (importance: 0.62), debt-to-income ratio (0.23), and employment stability (0.15) were the most influential predictors, aligning with prior studies on financial risk assessment.

Performance metrics included accuracy, precision, recall, and F1 score, with a focus on recall to minimize false negatives in loan approval scenarios. The Random Forest achieved a test recall of 89.5% and an F1 score of 85.2%, outperforming XGBoost (recall: 84.1%, F1: 81.7%) and ANNs (recall: 72.3%, F1: 68.9%). This highlighted its superior ability to identify qualified applicants while maintaining competitive precision (78.4%) and accuracy (93.1%).

Table 1. Performance metrics of random forest with hyperparameter tuning

Metric	Training	Testing
Accuracy	96.8%	93.1%
Recall	91.2%	89.5%
Precision	82.7%	78.4%
F1 Score	86.7%	85.2%

The methodology underscores the effectiveness of ensemble techniques in loan prediction, particularly when combined with rigorous feature engineering and optimization strategies.

5. Experimental setup and results

The study utilized stratified sampling to divide the dataset into training (70%) and testing (30%) subsets, maintaining the original distribution of loan acceptance labels. To ensure model robustness, k-fold cross-validation was integrated during training for stability evaluation and hyperparameter tuning.

Table 2 summarizes the performance of decision tree models under pre-pruning and post-pruning configurations. The fully grown decision tree achieved perfect training accuracy (1.0) but exhibited overfitting, as shown by a test accuracy of 97.87% and recall of 88.59% in the post-pruned model.

DOI: 10.54254/2754-1169/2025.31254

Pre-pruning enhanced generalization, achieving a test accuracy of 99.46% and recall of 89.93%. Post-pruning further balanced complexity and performance, yielding a test F1 score of 89.19%, outperforming the pre-pruned version.

Table 2. Performance metrics of decision tress models

Model	Accuracy (Trian)	Recall (Trian)	Precision (Trian)	F1 (Trian)	Accuracy (Test)	Recall (Test)	Precision (Test)	F1 (Test)
Decision Tree (sklearn)	1.0	1.0	1.0	1.0	97.87%	88.59%	89.80%	89.19%
Decision Tree (Pre-Pruning)	99.03%	92.75%	96.85%	94.75%	99.46%	89.93%	91.16%	90.54%
Decision Tree (Post-Pruning)	99.46%	100.0%	94.57%	97.21%	97.87%	88.59%	89.80%	89.19%

A comparative analysis with baseline models (Table 3) emphasized the superiority of the pruned decision tree. While logistic regression achieved a test accuracy of 95.73%, its recall (63.76%) and F1 score (74.80%) lagged significantly. The KNN model underperformed, with a recall of 34.90% and F1 of 42.11%. In contrast, the post-pruned decision tree maintained high precision (89.80%) and recall (88.59%), demonstrating its effectiveness in minimizing false negatives while ensuring predictive stability.

Table 3. Comparative model performance on test set

Model	Accuracy	Recall	Precision	F1
Decision Tree(Post-Pruning)	97.87%	88.59%	89.80%	89.19%
Logistic Regression	95.73%	63.76%	90.48%	74.80%
KNN	90.47%	34.90%	53.06%	42.11%

Hyperparameter tuning via cost complexity pruning refined the decision tree, optimizing parameters such as maximum depth (6 nodes) and leaf node constraints (maximum 10 nodes). It improved interpretability and generalizability, thus allowing its application in real-life loan approval conditions. The results underscore the decision tree's robustness in handling imbalanced data and non-linear interactions, hence making it a crucial tool for Neo Bank's targeted marketing strategies.

6. Key insights and comparative analysis

An analysis of feature importance (Table 4) showed that education level (undergraduate) has the most substantial effect on loan decisions with an importance weight of 0.35, followed by income (0.30) and family size (0.15). The results corroborated and were consistent with the exploratory data analysis results, demonstrating that highly educated customers often exhibit better financial planning skills and are usually more prone to accepting loans for investments or consumption. Lastly, average credit card spending (CCAvg) found an importance weight of 0.1, suggesting that customers spending more per month are likely to be more active in cash transactions and thus initiate more loan requests.

In the model performance comparison (Table 4), the post-pruned decision tree model achieved the best results on the test set, with an accuracy of 97.87%, recall of 88.59%, and an F1-score of 89.19%. Its recall significantly outperformed KNN (34.90%) and logistic regression (63.76%), highlighting its superiority in minimizing missed opportunities for high-potential loan applicants. In

DOI: 10.54254/2754-1169/2025.31254

contrast, KNN's sensitivity to data imbalance led to low recall, while logistic regression, despite high precision (90.48%), was constrained by its linear assumptions in capturing complex feature interactions.

Further analysis identified income-education interaction effects as a core driver of loan acceptance rates. For instance, among high-income customers with undergraduate or higher education, the probability of loan acceptance increased by 45% compared to other groups. Additionally, the synergy between family size and credit card spending significantly influenced loan decisions: customers with families larger than four members and monthly credit card spending above $2K exhibited a 32% higher loan acceptance rate than the baseline group.

Based on these insights, Neo Bank is advised to prioritize tailored loan products for high-income, highly educated, and large-family customers, leveraging dynamic interest rate strategies to enhance conversion rates. The model's high-recall capability enables precise targeting of high-potential clients under limited marketing resources, reducing the risk of overlooking qualified borrowers. Future research should integrate real-time transactional data to improve the model's adaptability to evolving customer behaviors.

Table 4. Performance comparsion of models on the test set

Model	Accuracy	Recall	Precision	F1-Score
Decision Tree(Post-Pruning)	97.87%	88.59%	89.80%	89.19%
KNN	90.47%	34.90%	53.06%	42.11%
Logistic Regression	95.73%	63.76%	90.48%	74.80%

7. Discussion and future directions

Neo Bank could take the lead with the insights gained in the study to further improve its loan marketing strategies and allocate resources better. Firstly, the model output would allow Neo Bank to identify high-potential customers, especially those with higher income and education levels. This information can be leveraged by banks for focusing their resources on marketing activities and personalized promotions for these customers. Neo Bank can issue loans tailored to family customers and better satisfy them. Furthermore, the analysis results could be employed by the banks in accurately segmenting the market and making the ROI from marketing maximized. Neo Bank can increase customer acquisition rates and hence improve financial performance by enabling the effective allocation of marketing resources.

Though the model as proposed exhibits a reasonable performance, it has some shortcomings that need to be mentioned. The data were quite highly imbalanced, as answers reflected there were indeed many more customers who refused loans than those who said yes. Introducing data imbalance into the training data would push the model towards predicting majorities and thus decrease its ability to make predictions for minority classes. Further, the model would easily fit into overfitting by naturally introducing complexity, predominantly in situations where sufficient instances are not available in the training dataset, and it ends being simply noise in the data. Because of that, it is rather essential to respectively evaluate the model performance on various customer groups as far as practical applications are concerned.

An area suggested as possible for further research is to incorporate more behavioral data such as credit card usage patterns and consumption habits, which will help the model capture the changes in customers' financial behavior and may support better model stability. Integrating predictions from different weak learners will allow these methods to demonstrate a better generalization ability in

DOI: 10.54254/2754-1169/2025.31254

intricate datasets. Finally, the integration of external economic data, such as market trends, interest rate change, and socio-economic factors, will help to enhance the comprehensiveness and applicability of the model, lending Neo Bank a competitive edge in its loan decision-making.

8. Conclusion

This study demonstrates that machine-learning techniques have the transformative potential to alter personal loan approval processes for Neo Bank in view of compelling challenges in a highly bushed financial market.Show in Table 5, the researchers have demonstrated pruned decision-tree models were superbly predictive at achieving 97.87% accuracy, 88.59% recall, and an 89.19% F1-score on the given test set, far outperforming traditional models like KNN (34.90% recall) and logistic regression (63.76% recall). This thus confirms the superiority of decision trees with respect to dealing with non-linear relationships and complex feature interactions-a phenomenon that is well documented in extant literature, driving home their interpretability and ability to be molded in the context of credit risk [13,14].

The analysis showed level of education, income, and size of the family are most influential predictors for loan acceptance, with an importance of 0.35, 0.30, and 0.15, respectively. The findings cohere with exploratory data analysis (EDA), which incidentally found that customers with a high income and educational background tend to be more financially knowledgeable and stable than large families who demand a significantly higher amount of credit. For example, those whose education was undergraduate with income above the 75th percentile had a 45% higher loan acceptance rate, showing the interplay of education with economic capacity. Similarly, families with four or more members and spending $2,000 or more on credit cards each month demonstrated a 32% increase in loan uptake, speaking to the interactions between demographic and behavioral factors.

From a strategic point of view, Neo Bank may capitalize on these insights to tailor loan products to the needs of those segments having the highest potential. For example, dynamic interest rates for high-income earners and bundled education loans for postgraduate customers could provide great impetus toward consideration. Moreover, the recruitment-based model, with its very high recall, allows one to minimize the chance of missing out on qualified applicants and thus allocate resources efficiently to marketing campaigns. The paper also recognizes other limitations coming from dataset-specific features, such as class imbalance (relatively few loan acceptors) and risks of overfitting due to limited samples for training. These issues are a cognate concern with credit scoring literature at large [24], on how imbalanced data could keep model predictions in favor of the majority classes.

Future research should prioritize three key directions:

1. Integration of real-time behavioral data: Incorporating granular transactional patterns (e.g., credit card usage frequency) and external economic indicators (e.g., inflation rates) could refine the model's responsiveness to dynamic financial behaviors.

2. Advanced ensemble techniques: Exploring methods like Random Forests or Gradient Boosting Machines may further mitigate overfitting while enhancing predictive robustness across diverse customer profiles [16].

3. Ethical and regulatory compliance: Implementing fairness audits to address potential biases in features like education or income is critical to ensuring equitable access to financial services. Techniques such as adversarial debiasing or reweighting could align the model with regulatory standards and societal expectations.

By merging state-of-the-art machine learning techniques with actionable business strategies, Neo Bank will not only streamline its funding workflow but will also instill confidence and inclusivity in

DOI: 10.54254/2754-1169/2025.31254

its services. Through this strategy, the bank postures itself as a competitor, with reduced default risk and prepared to engage in opportunities in this ever-fluid financial space. Finally, the study shows how data-centric innovation can spur sustainable growth that balances profit with proper social responsibility.

Table 5. Final performance metrics of the optimized decision tree model

Model	Accuracy	Recall	Precision	F1-Score
Decision Tree(Post-Pruning)	97.87%	88.59%	89.80%	89.19%

Acknowledgement

Yicheng Liu and Xiaoyue Ni contributed equally to this work and should be considered co-first authors.

References

[1] Stobdan, J. & Kumar, S. (n.d.). Neo Banks: Future Prospects, Challenges and Strategies. Bank Quest.
[2] Fibre Federal Credit Union (2024). Fibre Federal Credit Union Taps Upstart for Personal Lending. Manufacturing Close-Up.
[3] P.P., Nehru, S. & P.S. (2024). Investigated Study on Shaping the System of Personal Loans in India: with Special Reference to Informal Workers of Madras Province, Chennai, India. South Asian Journal of Social Studies and Economics, 21(9), 15-20.
[4] VNET (2024). VNET Announces Certain Updates Regarding the Refinancing of the Founder's Personal Loan. Telecomworldwire.
[5] Gao, X., Jia, Y., Krupa, R. N. et al. (2024). The Corroboration Role of Management Earnings Forecasts in Private Loan Markets. Journal of Accounting, Auditing & Finance, 39(3), 903-930.
[6] Hand, D. J. & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523-541.
[7] Crook, J. N., Edelman, D. B. & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. European Journal of Operational Research, 183(3), 1447-1465.
[8] Thomas, L. C. (2000). A survey of credit and behavioral scoring: Forecasting financial risk of lending to consumers. International Journal of Forecasting, 16(2), 149-172.
[9] Blöchlinger, A. & Leippold, M. (2006). Economic benefit of powerful credit scoring. Journal of Banking & Finance, 30(3), 851-873.
[10] Khandani, A. E., Kim, A. J. & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767-2787.
[11] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. A. K. & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54(6), 627-635.
[12] Witten, I. H., Frank, E. & Hall, M. A. (2011). Data mining: practical machine learning tools and techniques. 3rd ed. Burlington, MA: Morgan Kaufmann.
[13] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
[14] Quinlan, J. R. (1996). Bagging, boosting, and C4.5. In: Proceedings of the Thirteenth National Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 725-730.
[15] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.
[16] Louzada, F., Ara, A. & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. Surveys in Operations Research and Management Science, 21(2), 117-134.
[17] Abdou, H. & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review. Intelligent Systems in Accounting, Finance & Management, 18(2-3), 59-88.
[18] LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

DOI: 10.54254/2754-1169/2025.31254

[19] Zhang, J., Zheng, Y. & Qi, D. (2017). Deep spatio-temporal residual networks for citywide crowd flows prediction. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 1655-1661
[20] Kim, K., Choi, Y., Lee, J. & Woo, W. (2019). Financial fraud detection using convolutional neural networks. Expert Systems with Applications, 133, 1-10.
[21] Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.
[22] He, H. & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.
[23] 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.
[24] 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.