Predictive Analysis for Personal Loans by Using Machine Learning

Hui-I. Huang[1], Chou-Wen Wang[1], and Chin-Wen Wu[2(✉)]

[1] National Sun Yat-Sen University, Kaohsiung, Taiwan
[2] Nanhua University, Chiayi County, Taiwan
cwwu@nhu.edu.tw

Abstract. This study adopts five common machine learning algorithms for predicting consumer personal loan uptake, including Logistic Regression, Support Vector Machine, Multilayer Perceptron, Gradient Boosting Decision Trees Catboost, and Xgboost. The research utilizes data from Thera Bank available in the public database Kaggle, featuring fields like age, work experience, income, family size, average credit card expenditure, education level, home loans, securities account, deposit account, and internet banking usage. The study addresses the issue of imbalanced data using the SMOTE (Synthetic Minority Over-sampling Technique) method and compares the accuracy and stability of predictions using the five models with three different sampling rates to identify the optimal model and key factors. Empirical results show that the Gradient Boosting Catboost model and the Support Vector Machine model perform with stability and precision across different sampling ratios, making them the best models. Moreover, through the Gradient Boosting Xgboost model, the study identifies key features such as educational factors, income, family size, the existence of a deposit account, and annual credit card spending. The findings of this research can provide crucial factors for financial institutions when formulating marketing strategies for personal loans.

Keywords: Bank · Machine Learning · Personal Loans · Support Vector Machine · Gradient Boosting Model

1 Introduction

This research delves into the application of machine learning models in the domain of personal consumer loans and their impact on the decision-making processes of financial institutions. In the context of rapidly evolving financial technology, traditional evaluation methods face increasing challenges, not only from the rapidly changing market environment but also from the need to process large volumes of complex financial data and make precise predictions about market strategies. Therefore, the integration of machine learning models to enhance prediction efficiency and analyze the crucial factors affecting loan applications has become a priority.

The core purpose of this study is to use machine learning models to predict consumers' acceptance of personal consumer loan activities and explore which models are

F. F.-H. Nah and K. L. Siau (Eds.): HCII 2024, LNCS 14720, pp. 187–199, 2024.
https://doi.org/10.1007/978-3-031-61315-9_13

most effective in this field. Additionally, we investigate which factors decisively influence consumers' acceptance of loan activities from financial institutions, aiding these institutions in optimizing their marketing activities. The main objectives of the study are:

1. Which machine learning model can most effectively predict customer applications for personal loans?
2. What key factors are most likely to influence customer applications for personal credit loans?

The personal loan case study in this research is derived from the public data in Kaggle's database, provided around 2019 by Walke in the Thera Bank case. The data, covering 5,000 records with 14 variables, include demographic information (such as age, income), the customer's relationship with the bank (e.g., mortgage loans, securities accounts), and the customer's response to the last personal loan activity (whether they accepted the personal loan).

Through various machine learning models, the study explores which models can effectively predict whether customers will accept the bank's marketing activities, enabling financial institutions to understand customer needs better and adjust their financial service strategies accordingly.

In summary, the study found that machine learning models, particularly the Catboost and SVM models, excel in predicting whether customers will apply for credit loans. The SVM model showed higher accuracy in predicting customers who have already applied for loans. However, after a comprehensive comparison, the Catboost model outperformed in overall performance (including predictions for both types 0 and 1), indicating that financial institutions can leverage these efficient models to adjust marketing strategies, target different customer groups, and enhance the accuracy of predictions and loan application rates. For institutions with larger data volumes, the Catboost model is recommended; otherwise, the SVM model should be the primary choice.

Further analysis of feature importance from the Xgboost model revealed that education, income, and family situation are key factors affecting loan decisions. This suggests that financial institutions should pay more attention to these factors in customer credit assessments, assigning them higher weights. Additionally, institutions can delve deeper into the nuances of these features to better understand customers' credit risk and develop more targeted loan products.

2 Literature Review

2.1 Application of Machine Learning Models in the Finance

In recent years, the development of machine learning in the financial area has been rapid, with a wide range of applications. For instance, in financial credit scoring, machine learning can more finely evaluate the risk of borrowers; in fraud detection, it can effectively prevent fraudulent activities through the analysis of large amounts of data. Automated trading systems utilize machine learning to process and analyze market data, making real-time trading decisions. Furthermore, financial institutions leverage machine learning to optimize customer service, such as developing intelligent investment advisors and

personalized banking services, demonstrating the significance of machine learning in the innovation of financial products and services. These advancements not only revolutionize traditional banking and financial services but also bring a more data-driven, efficient, and customer-centered development direction to the finance industry.

Huang, J., Chai, J., & Cho, S. (2020) comprehensively evaluated the application of deep learning (DL) models in the financial and banking sectors, systematically assessing these models, including data preprocessing, input data, and model evaluation, and exploring the factors affecting the outcomes of financial deep learning models. The study provides observations on the latest applications of deep learning in the financial and banking sectors, highlighting the rapid development trend of deep learning in these industries, pointing out the shortcomings of existing literature, and directing future research in this field.

Zhu, Zhou, Xie, Wang, and Nguyen (2019) proposed a new hybrid ensemble machine learning method to improve the accuracy of predicting credit risk for small and medium-sized enterprises (SMEs) in supply chain finance. The study utilized random subspace and MultiBoosting techniques, analyzing data from 46 SMEs and 7 core companies listed on the Chinese stock market from 2014 to 2015, to verify the feasibility and effectiveness of the new method. The results showed that this new approach performs well in handling small sample data, providing practical insights for improving the financing capabilities of SMEs.

Tax, K. J., Dosoula, Smith, and Bernardi (2021) explored how to use supervised machine learning algorithms to improve fraud detection performance in anti-money laundering. The study tested four different machine learning algorithms: logistic regression, SVM, random forest, and artificial neural networks (ANN), using a simulated dataset for experimentation. The study found that, among these techniques, random forest performed best in terms of accuracy, while the accuracy of artificial neural networks was the lowest. This research provides new insights into the anti-money laundering field, showcasing the potential application of machine learning technologies in detecting money laundering activities.

Ma and Sun (2020) delved into the application of machine learning and artificial intelligence in market marketing, reviewing machine learning tasks and methods, and discussing AI-driven marketing trends and practices. The article proposes a unified conceptual framework and a multifaceted research agenda, exploring how machine learning methods can be integrated into market marketing research, highlighting the advantages of machine learning methods in handling unstructured data, making predictions, and extracting features, and discussing the potential applications of these methods in market marketing.

The aforementioned literature studies on various financial aspects illustrate that machine learning is ubiquitous. Through machine learning models, financial institutions can comprehensively review and strategize in marketing, trading, fraud, risk management, and other areas, simultaneously improving accuracy and efficiency. In the next section, we will explore the literature on personal loans combined with machine learning models.

2.2 Combining Machine Learning Models with Personal Loan Prediction

Arun, K., Ishan, and Sanmeet (2016) delved into machine learning's role in streamlining the personal loan approval process. By employing algorithms such as decision trees, random forests, and SVMs, their research aimed to mitigate risk in loan vetting, seeking the most accurate predictive model. Eletter and Yaseen (2017) applied Linear Discriminant Analysis, MLP, and CART to evaluate loan default probabilities, aiding credit decisions in Jordanian banks. Their comparison highlighted MLP's effectiveness in minimizing misclassification costs and identifying potential defaulters.

Ibrahim, Ridwan, Muhammed, Abdulaziz, and Saheed (2020) investigated the Catboost model's application in predicting loan approval and employee promotions, highlighting Catboost's superior performance over other classifiers. Their research points to the significance of feature engineering in enhancing model accuracy. Nosratabadi et al. (2020) reviewed various data science methodologies, including deep learning and hybrid models, across economic sectors. Their systematic review underscored the advanced performance and future potential of hybrid deep learning models in economic analyses. Zhang, Gong, Yu, and Wang (2020) combined factor analysis with Xgboost models to predict investor lending willingness in P2P platforms under incomplete information scenarios. Their method improved prediction accuracy, informing better investment decisions in the P2P sector.

Sreesouthry, S., Ayubkhan, Rizwan, Lokesh, and Ra (2021) explored logistic regression within machine learning to forecast loan approvals. Their work, leveraging Kaggle's public dataset, emphasized machine learning's capability to refine credit scoring and risk assessments, showcasing the predictive strength of logistic regression in banking. Akça and Sevli (2022) demonstrated the use of Support Vector Machines (SVM) with various kernel functions to predict bank loan application outcomes, finding the polynomial kernel to be most effective with an impressive 97.2% accuracy rate. This study underlines the importance of choosing the right kernel function and the impact of data imbalance on model outcomes.

Anand, Velu, and Whig (2022) employed a variety of machine learning models to predict personal loan behavior, contributing to safer credit lending practices. Their comprehensive approach, assessing models from logistic regression to Gaussian Naive Bayes and SVMs, demonstrated the effectiveness of machine learning in identifying loan default risk. Li, Zhang, Qiu, Cai, and Ma (2022) focused on predicting loan defaults in P2P lending using models like logistic regression, random forest, and Catboost, enhanced by blending techniques. Their findings suggest that such integrated approaches can significantly reduce credit risk for online lending platforms.

Collectively, these studies not only showcase the broad applicability and potential of machine learning models in financial services, particularly in personal loan approval and risk assessment, but also offer insights into the methodological advancements and data strategies that enhance predictive accuracy and reduce risk.

3 Research Methodology

3.1 Data Source

The personal loan case comes from the public data repository on Kaggle, provided by Walke around 2019, featuring a real-case scenario of Thera bank. However, the data release description does not record the period of data acquisition; it is inferred that the data published in 2019 were obtained from the 2018 database. Therefore, it is understood that the data represent 5,000 customers who applied for business due to personal loan marketing activities in the previous year. The original dataset contains a total of 5,000 entries and covers 14 variables, including demographic information of the customers (such as age, income, etc.), the relationship between customers and the bank (e.g., mortgage loans, securities accounts, etc.), and the customers' response to the last personal loan activity (whether they accepted a personal loan or not). Through different machine learning models, we will delve into which model can effectively predict whether customers will accept the bank's marketing activities, thereby allowing financial institutions to deeply understand customer needs and adjust corresponding financial service strategies accordingly.

3.2 Feature Preliminary Analysis

In the case study, we examine a growing customer base of T Bank, which primarily comprises depositors with varying balance sizes and a limited proportion of loan customers. T Bank aims to increase its loan customer ratio by leveraging its existing depositor base to introduce loan services, aiming to generate more interest income. Analysis of the previous year's data on depositors converting to loan customers showed a successful conversion rate of over 9%, prompting the personal loan department to devise more targeted activities to increase this ratio under minimal budget constraints. The bank provided data on 5,000 customers, of which only 480 (9.6%) accepted the personal loans offered in the previous activity.

Table 1 presents the data features of T Bank's case, including mean, standard deviation, maximum and minimum values, and types. The data, free from missing or duplicate values and devoid of character-type variables, had 56 outliers removed across variables like age, experience, income, family size, average credit card spending, education level, mortgage, securities account, CD account, online banking, and credit card ownership, ensuring no significant bias in assessment results. Credit card average spending (CCAvg) was annualized to align with income measurement standards.

Data Source: Compiled for this Study. This customer data table offers an in-depth analysis of variables including age, work experience, income, family status, education level, and other financial-related variables. These insights provide objective references for banks or financial institutions to understand their customer base, develop more effective marketing strategies, and offer suitable financial products (Table 2).

1. Age and Experience: The wide age distribution, with an average age of 45 years, indicates the bank's appeal to various age groups. The average work experience of

Table 1. Feature Description

ITEM	VARIABLE	QUANTITY	MEAN	STANDARD DEVIATION	MINIMUM	MAXIMUM
1	Age	4944	45.35	11.46	23	67
2	Experience	4944	20.35	11.25	0	43
3	Income	4944	72.85	45.27	8	218
4	Family	4944	2.40	1.15	1	4
5	Education	4944	1.88	0.84	1	3
6	Mortgage	4944	54.13	97.68	0	617
7	Personal Loan	4944	0.09	0.29	0	1
8	Securities Account	4944	0.10	0.30	0	1
9	CD Account	4944	0.05	0.22	0	1
10	Online	4944	0.59	0.49	0	1
11	Credit Card	4944	0.29	0.45	0	1
12	ann_CV	4944	22.80	20.45	0	120

Table 2. Comparison of Model Results at Different Sampling Ratios

Model	Accuracy		
Training and Testing Sets	10%	20%	30%
Logistic Regression	89.09%	89.56%	88.85%
SVM	98.11%	98.17%	97.44%
Xgboost	98.99%	98.79%	98.52%
Catboost	99.19%	98.58%	98.58%
MLP	98.38%	97.44%	97.44%
All Zeros Model	90.11%	90.90%	90.11%

Data Source: Compiled for this study.

20.35 years suggests a customer base with considerable professional tenure, informing age-related financial products and services like retirement plans or long-term investments.

2. Income and Family Status: An average income of 72.85% shows stable annual earnings among customers, though a high standard deviation indicates significant income disparity. An average family size of 2.40 suggests most customers have smaller families.

3. Education Level: Customers' education levels range from below college (1) to master's (2) and doctorate (3) degrees, with an average level of 1.88, indicating a higher education level among most customers. This may affect their financial understanding and risk tolerance, necessitating consideration in financial education and communication.
4. Mortgage: The average mortgage amount and its high standard deviation indicate a segment of customers with substantial mortgage needs, suggesting targeted financial planning and services for this group.
5. Personal Loan: The low average uptake of personal loans (0.09) from the previous activity indicates relatively low demand, necessitating a deeper understanding of customers' attitudes and needs regarding loan products.
6. Financial Products (Securities Account, CD Account, Online Banking, Credit Card): These binary variables reveal customer participation in these financial products or services, suggesting the need for better promotion or the development of new financial services to attract customers.
7. Credit Card Spending (ann_CV): Variability in credit card usage among customers, indicated by the average and standard deviation of credit card spending, suggests the potential for targeted rewards programs and promotions for specific customer segments.

Based on these analyses, banks can tailor financial products like retirement plans and family loans to customers' diverse age, income, and family situations. For customers with higher education levels, more in-depth and professional financial education can enhance their understanding of financial products. Banks could offer customized financial and loan solutions for those with significant mortgages and strengthen the promotion of financial products like securities accounts, online banking, etc., to attract more participation. For frequent credit card users, launching corresponding rewards programs could increase loyalty and potentially encourage personal loan applications.

3.3 Brief Review of Machine Learning Models

This study divides the dataset into training and testing sets with ratios of 90%/10%, 80%/20%, and 70%/30%, respectively. We utilize five machine learning models for training: Logistic Regression, Support Vector Machine (SVM), Xgboost, Catboost, and Multilayer Perceptron (MLP). Here's a concise overview of each model:

1. Logistic Regression (LR): A well-known classification algorithm, particularly effective for binary classification problems. Despite its name, LR is a classifier that predicts the probability of an instance belonging to one of two classes using a logistic function. It assumes a linear relationship between input features and the log odds of the output.
2. Support Vector Machine (SVM): Developed by Vapnik, Boser and Guyon in 1992, SVM works well for both linear and nonlinear problems, particularly in high-dimensional spaces. It aims to find a hyperplane that best divides a dataset into two classes, maximizing the margin between the closest points of the classes, which are known as support vectors.

3. Catboost (Categorical Boosting): A gradient boosting decision tree algorithm optimized for categorical data without the need for extensive data preprocessing like one-hot encoding. Introduced by Yandex in 2017, Catboost excels in handling categorical features directly, improving efficiency and performance in large datasets.
4. Xgboost (eXtreme Gradient Boosting): An efficient and scalable implementation of gradient boosting framework by Tianqi Chen in 2014. Xgboost is known for its speed and performance, especially in structured data competitions. It uses advanced regularization (L1 and L2), which helps in reducing overfitting and improving model performance.
5. Multilayer Perceptron (MLP): A type of artificial neural network with multiple layers, including input, hidden, and output layers, where each layer is fully connected to the next. MLP can model complex nonlinear relationships through its network structure and is particularly useful for deep learning tasks.

Each model offers unique advantages in handling specific types of data and learning tasks. Logistic Regression is straightforward and interpretable, making it suitable for simpler binary classification problems. SVM is powerful for datasets with a clear margin of separation, even in high-dimensional spaces. Catboost and Xgboost are advanced tree-based algorithms that perform well on structured data, with Catboost being particularly adept at handling categorical features directly. MLP, representing deep learning approaches, is versatile in modeling complex patterns but requires substantial data and computational resources.

In summary, the selection of a machine learning model depends on the specific characteristics of the dataset, including the linearity of relationships, presence of categorical features, and the size of the data. This study aims to compare these models' performance in predicting customer responses to bank loan marketing activities, providing insights into the effectiveness of each model in financial services applications.

3.4 Assessing Model Predictive Metrics

Following the introduction of the five models in the previous section, this section explains how we use evaluation metrics such as the confusion matrix, accuracy, precision, recall, F1 score, and ROC curve to thoroughly analyze and assess the performance of the models. These metrics will help reveal the strengths and limitations of the models in various aspects, providing a more compelling interpretation and discussion of the research findings.

Confusion Matrix is a table used to evaluate the performance of classification models, particularly in binary classification. The confusion matrix categorizes predictions into True Positives (TP) and True Negatives (TN) for correct predictions, and False Positives (FP, Type I error) and False Negatives (FN, Type II error) for incorrect predictions, based on the actual observation being positive (P) or negative (N).

Using the confusion matrix, we can calculate various evaluation metrics, including accuracy, precision, recall, and F1 score, offering a comprehensive assessment of the model's performance in different aspects. Common evaluation metrics include Accuracy, which assesses overall model performance by the proportion of correct predictions; Precision, indicating the accuracy of positive predictions; Recall (Sensitivity), measuring the

model's ability to identify all positive cases; and F1 Score, a balance between precision and recall, with values closer to 1 denoting higher model accuracy and effectiveness.

4 Research Methodology

This chapter delves into the predictive factors of personal loan uptake using statistical analyses and machine learning models. Initially, a correlation heatmap revealed associations between variables like age, work experience, and the positive correlation between income and personal loan applications. Scatter plot analyses showed that higher-income individuals are more likely to apply for loans, while age has a lesser impact; however, higher annual credit card spending and certain levels of work experience influence loan applications, with mid-experience individuals less likely to apply.

In terms of machine learning model outcomes, the study addressed data imbalance using the SMOTE technique and employed various models including Logistic Regression, SVM, Xgboost, Catboost, and MLP for prediction. These models demonstrated varying degrees of accuracy and stability across different sampling ratios, with Catboost and Xgboost showing superior performance in accuracy, precision, and recall, particularly in handling class imbalance. The SVM model excelled in predicting Type 1 outcomes. Furthermore, an importance analysis of factors in the Xgboost model highlighted education, income, and family size as key determinants of loan uptake, with education's importance slightly increasing with higher sampling ratios, underscoring its pivotal role in prediction. Overall, this chapter thoroughly examines various factors affecting personal loan uptake and identifies the most effective predictive models, offering strategic insights for financial institutions in formulating personal loan strategies.

4.1 Model Comparison

In this study, six models were employed to tackle a binary classification problem, including a simplistic model that predicts all zeros, Logistic Regression, SVM, Xgboost, Catboost, and MLP. These models exhibited distinct performances and characteristics when applied to the data. Here is a concise comparison of their results for training and testing sets with ratios of 90%/10%.

1. All Zeros Model: Predicts all instances as class 0, achieving an accuracy rate of approximately 89%, indicating a bias in handling imbalanced datasets with poor recognition of class 1.
2. Logistic Regression: Exhibits high computational efficiency with training time around 4.4 s and almost instant prediction, reaching an accuracy of 90.2%. The model shows balanced performance across classes with precision, recall, and F1 score close to 0.9 for both classes, validated by a cross-validation score of about 89.8%.
3. SVM: Requires longer training time around 2 min but demonstrates exceptional predictive capability with an accuracy of roughly 98.1%. It shows almost perfect precision and recall for both classes, slightly lagging in cross-validation scores compared to Catboost and Xgboost.

4. Xgboost: Features quick training (10 s) and prediction times, achieving an impressive accuracy of approximately 98.99%. It shows high accuracy for both primary and secondary classes with a cross-validation score of about 98.58%, indicating robustness and efficiency.
5. Catboost: Stands out for its training efficiency and remarkable accuracy of around 99.19%, demonstrating superior performance in correctly predicting both classes and handling data imbalance. It slightly leads in performance metrics, especially in recall rates for less prevalent classes, with a cross-validation score of about 98.70%.
6. MLP: Despite longer training times, it excels in prediction accuracy (approximately 98.38%) and showcases high precision, recall, and F1 scores for both classes, supported by a cross-validation score of around 97.68%, proving its consistency and robustness.

For Comparative Summary, to clearly compare the models, Catboost and Xgboost outshine others in balance across accuracy, precision, recall, and stability in cross-validation, showing superior generalization capabilities and effectiveness in managing class imbalance. SVM excels in predicting class 1 but slightly falls short in cross-validation performance compared to Catboost and Xgboost. Catboost marginally leads across most performance indicators, demonstrating its strong predictive accuracy and reliability in identifying less common classes, an essential factor in model performance evaluation.

4.2 Robustness Check

For datasets split into 80% training and 20% testing, a comparative analysis shows that the Xgboost model excels with the highest accuracy at 98.79%, followed by Catboost at 98.58%, and SVM at 98.17%. These models demonstrate high consistency and reliability across metrics for class 0, with Xgboost and Catboost also showing superior cross-validation scores, indicating their robustness and generalization ability. In contrast, the MLP and Logistic Regression models have slightly lower overall accuracy but still perform well above 97% and 89%, respectively. Notably, the all zeros strategy, while achieving a 90.90% accuracy rate due to class imbalance, fails entirely to identify the minority class, highlighting its ineffectiveness in imbalanced datasets.

For datasets split into 70% training and 30% testing, all five models display high accuracy rates in personal loan prediction tasks, with Xgboost and Catboost slightly outperforming others, achieving over 98.5% accuracy. These models particularly excel in predicting class 0, with near-perfect precision and recall. SVM and Xgboost show higher precision for class 1, while Catboost's lower recall indicates a more cautious approach in identifying class 1. The MLP model reveals some weaknesses in recall for class 1 but still maintains commendable overall performance. Compared to the all zeros strategy, which reaches 91% accuracy but fails to identify class 1, resulting in significantly lower macro-average metrics, SVM shows standout performance in class 1 predictions but slightly lags behind Catboost and Xgboost in cross-validation scores. Overall, Xgboost and Catboost provide the most balanced and reliable performance for this task.

For class 0 predictions, nearly all models achieve or approach perfect scores, likely due to the larger sample size of class 0. In predicting class 1, accuracy and recall vary

across models, but Catboost and Xgboost generally provide higher values. Considering the average cross-validation scores, an important indicator of model generalization ability, both Catboost and Xgboost excel, particularly Catboost, which reaches 98.70% at a 10% sampling ratio.

Comparing all metrics across sampling ratios, Xgboost and Catboost exhibit superior and consistent performance. However, given Catboost's slightly higher accuracy at a 10% sampling ratio and its relatively better performance in predicting class 1, Catboost emerges slightly ahead among these models. Nonetheless, selecting the best model also depends on the data characteristics in practical applications, including the uniformity of data distribution and class balance. If class 1 prediction is equally critical, a model with a higher F1 score for class 1 might be preferred. Additionally, training and prediction times are practical considerations in model selection. While the SVM model leads in class 1 metrics, it slightly lags behind Catboost in class 0 predictions, impacting its average cross-validation score. Overall, considering stability and performance across different sampling ratios, the Catboost model is identified as one of the best models.

4.3 Key Factors for Applying Personal Credit Loans

The analysis of key factors influencing customer applications for personal credit loans reveals that education, income, and family size are paramount. Education stands out as the most significant predictor across various sampling ratios, suggesting that individuals' educational backgrounds play a crucial role in their likelihood of applying for credit loans. This importance slightly increases with larger sample sizes, highlighting the growing relevance of education in the predictive models as more data is considered.

Income follows as the second most influential factor, indicating that higher earnings are associated with a greater propensity to apply for loans. This relationship is consistent across all sampling ratios, underscoring the importance of financial stability in loan application decisions.

Family size ranks third in importance, maintaining a stable position across different sampling ratios but showing minor fluctuations compared to education and income. This suggests that while family size has a steady impact on loan application likelihood, its influence is somewhat less pronounced than that of education and income.

Additionally, the feature importance of having a deposit account shows more variation across sampling ratios but does not emerge as a top predictor, indicating its lesser, though still present, contribution to predicting loan applications. Overall, education, income, and family size are key to understanding and predicting customer behavior regarding personal credit loan applications, with education being the most consistent and significant factor.

5 Conclusion

From the findings and analyses of this study, several conclusions can be drawn. Firstly, machine learning models, especially Catboost and SVM, have shown excellent performance in predicting whether customers will apply for credit loans, with the SVM model being particularly accurate for customers who have already applied for loans.

This aligns with Akça and Sevli's recommendation for financial institutions to use the SVM model for predictions. However, after a comprehensive comparison, the Catboost model performed better overall, indicating that financial institutions can utilize these efficient models to refine marketing strategies and tailor plans for different customer segments to enhance prediction accuracy and loan application rates. For large datasets, the Catboost model is recommended; otherwise, the SVM model may be preferable.

Secondly, the feature importance analysis from the Xgboost model further reveals education, income, and family situation as key factors influencing loan decisions, suggesting financial institutions should pay more attention to these factors and assign them higher weights in customer credit evaluations. Additionally, institutions could delve deeper into these features' nuances to better understand customer credit risk and develop more targeted loan products.

Furthermore, financial institutions should enhance their data analysis of these features to optimize loan products and services, including designing customized loan plans for different customer groups and introducing more flexible loan products to adapt to changing market demands.

Lastly, this study suggests financial institutions explore using machine learning models for deeper customer behavior analysis, monitoring credit histories and behavior patterns, and considering external environmental factors. This will help institutions better understand customer credit needs and risks and more accurately tailor marketing strategies and risk management measures.

In summary, the insights from this study aim to improve loan marketing strategies and services, enhance the likelihood of individual loan applications, leverage machine learning models, focus on important features, and deepen customer behavior analysis. This approach can help financial institutions maintain a competitive edge in the financial market, offer better loan products and services, reduce the risk of bad loans, and contribute to sustainable growth and customer satisfaction.

The original data obtained for this study did not include the loan amounts approved by financial institutions for customers who consented to personal loans, preventing an analysis of whether customers with good educational backgrounds and income levels are more likely to receive bank loans or whether higher-educated and higher-income customers prefer applying for personal loans. Additionally, if financial institutions can obtain sufficient information internally, they could explore loan applications, credit records, and approved amounts by age group, for instance, to tailor marketing strategies for target customer groups, thereby making marketing more efficient and precise. Building on this study's foundation, it aims to assist financial institutions in constructing effective machine learning models and implementing precise marketing strategies.

References

1. Akça, M.F., Sevli, O.: Predicting acceptance of the bank loan offers by using support vector machines. Int. Adv. Res. Eng. J. **6**(2), 142–147 (2022)
2. Agarwal, K., Jain, M., & Kumawat, A.: Comparing classification algorithms on predicting loans. In: Information Systems and Management Science: Conference Proceedings of 3rd International Conference on Information Systems and Management Science (ISMS) 2020

(pp. 240–249). Springer International Publishing (2022). https://doi.org/10.1007/978-3-030-86223-7_21
3. Amari, S.: A theory of adaptive pattern classifiers. IEEE Trans. Electron. Comput. **3**, 299–307 (1967)
4. Anand, M., Velu, A., Whig, P.: Prediction of loan behaviour with machine learning models for secure banking. J. Comput. Sci. Eng. (JCSE) **3**(1), 1–13 (2022)
5. Arun, K., Ishan, G., Sanmeet, K.: Loan approval prediction based on machine learning approach. IOSR J. Comput. Eng **18**(3), 18–21 (2016)
6. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144–152 (1992)
7. Cox, D.R.: The regression analysis of binary sequences. J. R. Stat. Soc. Ser. B Stat Methodol. **20**(2), 215–232 (1958)
8. Cramer, J.S.: The origins of logistic regression (2002)
9. Eletter, S.F., Yaseen, S.G.: Loan decision models for the Jordanian commercial banks. Global Bus. Econ. Rev. **19**(3), 323–338 (2017)
10. Huang, J., Chai, J., Cho, S.: Deep learning in finance and banking: a literature review and classification. Front. Bus. Res. China **14**(1), 1–24 (2020)
11. Ibrahim, A.A., Ridwan, R.L., Muhammed, M.M., Abdulaziz, R.O., Saheed, G.A.: Comparison of the CatBoost classifier with other machine learning methods. Int. J. Adv. Comput. Sci. Appl. **11**(11) (2020)
12. Li, X., Ergu, D., Zhang, D., Qiu, D., Cai, Y., Ma, B.: Prediction of loan default based on multi-model fusion. Procedia Comput. Sci. **199**, 757–764 (2022)
13. Ma, L., Sun, B.: Machine learning and AI in marketing–Connecting computing power to human insights. Int. J. Res. Mark. **37**(3), 481–504 (2020)
14. Nosratabadi, S., et al.: Data science in economics: comprehensive review of advanced machine learning and deep learning methods. Mathematics **8**(10), 1799 (2020)
15. Prasad, K.G.S., Chidvilas, P.V.S., Kumar, V.V.: Customer loan approval classification by supervised learning model. Int. J. Recent Technol. Eng. **8**(4), 9898–9901 (2019)
16. Sreesouthry, S., Ayubkhan, A., Rizwan, M.M., Lokesh, D., Raj, K.P.: Loan prediction using logistic regression in machine learning. Ann. Romanian Soc. Cell Biol. **25**(4), 2790–2794 (2021)
17. Tax, N., et al.: Machine learning for fraud detection in e-commerce: a research agenda. In: Deployable Machine Learning for Security Defense: Second International Workshop, MLHat 2021, Virtual Event, August 15, 2021, Proceedings 2 (pp. 30–54). Springer International Publishing (2021). https://doi.org/10.1007/978-3-030-87839-9_2
18. Zhang, D., Gong, Y., Yu, L., Wang, X.: P2P online loan willingness prediction and influencing factors analysis based on factor analysis and XGBoost. J. Phys.: Conf. Ser. **1624**(4), 042039 (2020). IOP Publishing
19. Zhu, Y., Zhou, L., Xie, C., Wang, G.J., Nguyen, T.V.: Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. Int. J. Prod. Econ. **211**, 22–33 (2019)