Identifying High-Potential Customers for Personal Loans: Predictive Insights for Banking

Haotian Zhang

March 2025

1 Abstract

In today's competitive banking environment, identifying high-potential customers for personal loans is essential for optimizing marketing strategies and improving profitability. This study develops and evaluates predictive models to estimate the likelihood of existing liability customers converting into personal loan holders. Using demographic, financial, and behavioral data from Thera Bank's customer base, we apply logistic regression and classification tree models to uncover the key factors influencing loan acceptance. Both models demonstrate strong predictive accuracy, with the classification tree achieving 98.02% accuracy. Key predictors include income, education level, family size, credit card spending, and certificate of deposit account ownership. The findings provide actionable insights for targeted marketing, enabling the bank to allocate promotional resources efficiently and tailor offers to customers with the highest probability of conversion. This research contributes to the growing application of interpretable machine learning techniques in financial marketing and customer segmentation.

2   Introduction

In today's rapidly evolving and highly competitive financial environment, banks face mounting pressure to innovate their customer engagement strategies, diversify revenue streams, and enhance operational efficiency (Krasnikov et al., 2009). A critical challenge for many financial institutions is optimizing the composition of their customer portfolios—specifically, striking an effective balance between liability customers, who maintain deposits, and asset customers, who generate interest income through lending products (Nikiel and Opiela, 2002). Thera Bank, a U.S.-based, growth-oriented retail bank, epitomizes this challenge. Currently, the majority of its customers are liability holders with varying levels of deposits. While these customers contribute to funding stability, their direct impact on revenue is relatively limited when compared to asset customers, who are more directly linked to income-generating loan products (Machauer and Morgner, 2001).

In response to this structural imbalance, Thera Bank has embarked on a strategic shift focused on expanding its asset customer base. Specifically, the bank is pursuing the conversion of existing liability customers into personal loan holders—an approach that not only leverages established customer relationships, thereby reducing acquisition costs, but also has the potential to significantly increase customer lifetime value (Moro et al., 2014). Central to this strategy is the need to determine which liability customers are most likely to respond positively to personal loan offers. This question lies at the heart of the present research, which aims to help the bank identify high-potential target customers in order to more effectively tailor its marketing strategies and optimize promotional resource allocation. Accurately predicting these conversion probabilities is essential not only for optimizing marketing spend and improving response rates, but also for increasing the return on promotional investments (Fenton, 2007).

Initial results from Thera Bank's targeted campaigns have been encouraging. Over the past year, the bank recorded a personal loan conversion rate exceeding 9%, a figure that

surpasses internal benchmarks and validates the promise of targeted marketing. Building on this momentum, the retail marketing team is now seeking more advanced methods to enhance campaign precision and scale outreach. The team contends that machine learning offers a robust and scalable framework for data-driven targeting—one capable of accurately predicting which customers are most likely to adopt a loan product (Ala'raj et al., 2021).

To this end, the current study develops a predictive classification model that estimates the likelihood of liability customers converting to personal loan users, based on variables such as demographics, financial behavior, and previous engagement with the bank. By uncovering patterns that distinguish likely converters from non-responders, the model is designed to support more efficient customer segmentation and tailored marketing interventions. In addition to enhancing campaign performance, this approach also advances personalization efforts by aligning offers with customer needs and preferences.

Broadly speaking, this research contributes to the expanding literature on predictive analytics in financial services by demonstrating how data-driven modeling can inform customer segmentation and marketing strategy in real-world banking contexts. The findings are relevant not only to practitioners seeking operational insights but also to scholars exploring the applications of machine learning in consumer financial behavior, marketing analytics, and strategic decision-making.

3 Literature Review

In an era of increasing data availability and technological advancement, retail banks are undergoing a paradigm shift toward data-driven marketing and decision-making (He et al., 2022). Predictive analytics has become central to understanding customer behavior, particularly in customer acquisition, product cross-selling, and loan adoption (Boustani et al., 2024). One critical challenge lies in identifying which liability customers, who contribute

deposit-based capital, are most likely to convert into asset customers through products such as personal loans (Chang et al., 2024). Accurate prediction of such conversions can significantly enhance marketing efficiency, reduce acquisition costs, and increase customer lifetime value (Ulug and Akyüz, 2025).

3.1 Predictive Modeling in Financial Services

The application of predictive analytics in retail banking has been widely studied in areas such as credit scoring, churn prediction, and loan default forecasting (Singh et al., 2024). However, research focusing on loan product adoption from existing customers—especially liability to asset transition—is relatively nascent (Văduva et al., 2024). Traditional marketing practices, such as rule-based segmentation and demographic targeting, are increasingly giving way to machine learning techniques that leverage complex, non-linear customer behaviors (de Waal et al., 2024).

In the context of personal loan marketing, predictive modeling allows banks to create highly targeted campaigns by identifying customers with the highest likelihood of responding to offers (Rahman et al., 2024). As noted by Bharathi S et al. (2022), the integration of customer-level behavioral features (e.g., transaction patterns, credit card usage) has dramatically improved classification accuracy in banking environments. These developments underline the importance of choosing the right algorithm—not only for predictive performance, but also for interpretability, scalability, and deployability (Al-Quraishi et al., 2025).

3.2 Logistic Regression

Logistic regression is among the most frequently used models in both marketing and credit risk prediction due to its interpretability and statistical foundation(Chen et al., 2023). As a generalized linear model, logistic regression is ideal for binary classification tasks, such as predicting whether a customer will accept a personal loan (de Waal et al., 2024). It estimates

the log-odds of the dependent variable as a linear combination of independent features such as income, education, or CD account ownership.

Numerous studies have demonstrated the effectiveness of logistic regression in financial decision-making (Alonso Robisco and Carbó Martínez, 2022). Valluri et al. (2022) applied logistic regression to model churn among auto loan customers and reported robust results with clear coefficients useful for business decisions. Similarly, de Waal et al. (2024) applied logistic regression to P2P lending data and noted its strong performance on structured datasets with well-labeled inputs. Despite its linear assumptions, logistic regression provides valuable insights into feature significance, allowing marketers and analysts to interpret the influence of specific attributes like income level or family size on loan acceptance likelihood.

3.3   Tree Models

Classification trees are a form of decision tree model designed for supervised learning tasks where the response variable is categorical—making them particularly suited for binary classification problems such as predicting personal loan adoption (Momparler et al., 2016). These models divide the data space into disjoint regions using a tree-like structure that recursively applies decision rules on input variables, such as income, education, or account ownership (Safarkhani and Moro, 2021). In the context of retail banking, classification trees not only provide predictions but also yield interpretable insights, enabling marketers to understand customer segments at a glance. For example, a tree might reveal rules such as: "If income ¿ $70,000 and CD account = Yes, then likelihood of loan acceptance is high."

Their interpretability makes classification trees highly attractive in banking applications. As De Caigny et al. (2024) show, classification trees can reveal both main effects and feature interactions without extensive preprocessing or transformation of variables, making them suitable for marketing departments with limited data science capacity. Recent studies have reaffirmed the practical power of tree-based classifiers. For instance, Văduva et al.

(2024) demonstrated that decision trees could effectively detect churn risk in banks with high accuracy and minimal calibration. Likewise, Joung and Kim (2023) emphasized that classification trees serve as both predictive and descriptive tools for segmenting customers based on behavioral patterns, empowering personalized marketing and retention strategies.

3.4 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm that predicts outcomes by comparing new data points with the most similar cases in the training set (Joung and Kim, 2023). In marketing applications, KNN can identify customer profiles similar to past converters and infer likely behaviors (Zaki et al., 2024). However, KNN suffers from scalability issues and a lack of model transparency, which limits its practical utility in strategic marketing (Murugan and T, 2023). Alborzi and Khanbabaei (2016) observed that while KNN achieved moderate accuracy in predicting credit scoring outcomes, it was computationally expensive on larger datasets and difficult to explain to non-technical audiences. In a more recent study, Kruthika et al. (2024) compared KNN with decision tree and logistic regression models in the context of bank loan prediction and found that while KNN performed adequately in smaller datasets, it was consistently outperformed by decision trees and logistic regression in terms of accuracy and business deployment suitability. Another drawback of KNN is its sensitivity to feature scaling, which can distort distance-based calculations (Hussein Sayed et al., 2024). This requires preprocessing steps such as normalization or standardization, which may introduce additional complexity in real-time systems (Ahsan et al., 2021).

3.5 Model Selection

In banking environments—especially those governed by compliance and regulatory standards—interpretability is not optional. Stakeholders from legal, compliance, and executive teams require transparent models that can justify why a customer was targeted or denied

an offer(Rudin, 2019) . In this regard, logistic regression and decision trees offer significant advantages over models like KNN, neural networks, or support vector machines, which are typically considered black-box approaches.

As Bücker et al. (2022) emphasize, machine learning models deployed in regulated financial environments must strike a balance between predictive accuracy and explainability to foster stakeholder trust and ensure compliance with regulatory standards. A model that accurately predicts conversions but cannot explain why certain customers were targeted fails to support business decision-making and long-term marketing strategy (Bücker et al., 2022) . Moreover, model explainability facilitates continuous refinement through feature testing and campaign feedback, enabling marketing teams to optimize strategies over time (Melsom et al., 2022) .

4   Data

4.1   Data Exploration

The variables included in the dataset are listed in Table 1, and the correlation heat map of the independent variables is shown in Figure 2.1. This heat map illustrates the strength of pairwise correlations among all variables, where darker shades represent stronger correlations and lighter shades indicate weaker ones. Notably, the plot reveals a very high correlation between 'Age' and 'Experience', highlighted by the darkest cell in the matrix. This strong relationship likely reflects redundancy, as individuals with more experience also tend to be older.

To prevent multicollinearity and ensure more stable model estimation, 'Experience' is excluded from further analysis. The remaining variables exhibit only weak to moderate correlations, as indicated by the lighter shades in the heat map, suggesting that multicollinearity is not a major concern among them.

4.2 Visualization

Figure 2 illustrates the distribution of the dependent variable—Personal Loan—across the sample of 5,000 existing bank customers. The panel on the left displays the absolute count of customers who have accepted (value = 1) versus those who have not accepted (value = 0) a personal loan. The panel on the right expresses this distribution as proportions of the total sample.

From the charts, it is evident that a significant majority—approximately 4,000 customers—have not opted for a personal loan, whereas only around 400 customers, or roughly 10% of the total population, have done so. This pronounced imbalance in class distribution highlights a key challenge in predicting loan uptake: customer adoption of personal loans is relatively rare.

This insight underscores the importance of building a predictive model that can accurately identify the minority segment of likely loan adopters. From a marketing strategy perspective, such a model would be highly valuable for targeted campaigns, allowing the bank to direct promotional efforts and resources toward customers who exhibit the highest probability of conversion. Consequently, the outcomes of our research have direct implications for enhancing loan acquisition strategies and optimizing customer outreach, potentially improving loan portfolio growth while minimizing wasted marketing expenditure.

To better understand customers' likelihood of accepting personal loans, we first categorize them by income level, as income is a core demographic factor influencing financial decision-making and risk assessment. As shown in Table 2, we adopt the classification standards used by the U.S. Census Bureau to divide annual income into three tiers: High Income (greater than $156,000), Medium Income (between $52,000 and $156,000), and Low Income (less than $52,000).

This segmentation allows us to examine loan acceptance behavior across distinct income

brackets, which is essential for deriving actionable insights for customer targeting. From a marketing strategy standpoint, such classification is not only useful for refining predictive models but also for tailoring messaging and offers according to income-based personas. By aligning our analysis with widely recognized benchmarks, our findings gain both interpretive clarity and external validity, which enhances the practical relevance of our recommendations for the bank's campaign design.

Figure 3 visualizes the distribution of income classes among customers who have and have not accepted personal loans. The graph clearly indicates that customers in the Low Income category (in red) are entirely absent from the group that has accepted personal loans (right bar). Instead, the loan-accepting segment is composed exclusively of customers from the Medium (green) and High (blue) income brackets. This pattern suggests that income level plays a critical role in a customer's likelihood of qualifying for or choosing to utilize personal loan products.

From a marketing and financial segmentation perspective, this insight highlights two key implications. First, customers in the Medium and High income classes are likely perceived as more creditworthy due to their stable income and greater financial capacity. This enhances their eligibility and appeal for personal loan offers, aligning with standard risk-based pricing and approval strategies used in consumer banking. Second, these customer segments may also exhibit stronger loan-driven consumption and investment behavior, driven by needs such as property acquisition, automobile purchases, or business ventures. As such, they represent a more receptive and actionable target audience for personalized marketing campaigns promoting loan products.

From a bank perspective, by focusing efforts on customers within the Medium-to-High income ranges, banks can more efficiently allocate marketing resources and improve the conversion rate of loan offers. Simultaneously, the exclusion of low-income customers in actual loan uptake suggests a need for either alternative financial products or tailored financial education

to address their barriers to loan access.

Figure 4 examines customer acceptance of personal loans segmented by both education level and income class, offering a multidimensional perspective on financial decision-making. The faceted bar charts reveal a consistent trend across educational categories—Undergraduate, Graduate, and Advanced/Professional—wherein customers with medium and high incomes (green and blue bars) are more likely to take personal loans than those with low income (red).

Crucially, this figure provides additional nuance to our prior findings: it demonstrates that educational attainment amplifies the influence of income on personal loan acceptance. Among customers with Advanced/Professional degrees, a significantly larger proportion of personal loan users belong to the high-income class, compared to those with only undergraduate degrees.

From a marketing strategy standpoint, this intersection of education and income offers a powerful basis for target segmentation and message personalization. Financial institutions can tailor product offerings and communications for highly educated, high-income individuals by emphasizing sophisticated financial planning, investment leverage, or lifestyle enhancement. Conversely, for lower education or income segments, marketing efforts may need to address perceived barriers such as eligibility concerns or financial insecurity.

5 Methods

5.1 Model Selection

Logistic regression and classification tree models are particularly well-suited for marketing and credit risk prediction due to their interpretability, regulatory compliance, and business relevance. Logistic regression, as a generalized linear model, enables clear interpretation of

feature effects on binary outcomes, such as loan acceptance, and is widely adopted for its transparency and ease of deployment on structured data.

Classification trees, likewise, offer intuitive, rule-based segmentation that helps marketers identify actionable patterns (e.g., "If income greater than $70,000 and CD account is Yes, then high likelihood of acceptance") without extensive preprocessing. These models also capture feature interactions and support strategic decision-making in non-technical environments.

In contrast, K-Nearest Neighbors (KNN) poses notable drawbacks. It scales poorly with large datasets, requires intensive preprocessing (e.g., normalization), and lacks explainability— making it ill-suited for high-stakes financial applications where decisions must be justified to legal and compliance stakeholders. While useful in small-scale settings, KNN consistently underperforms relative to logistic regression and tree-based models in both accuracy and business applicability.

Therefore, logistic regression and classification trees are preferred for contexts demanding both predictive power and model transparency.

5.2   Model Building

5.2.1   Logistic Regression

To predict the likelihood of a customer accepting a personal loan, we developed a logistic regression model using R. The modeling process consists of several key steps. First, we specify a logistic regression model and set the L1 regularization penalty (lasso) to a small value of 0.01, indicating minimal regularization to approximate ordinary logistic regression. The mixing parameter is set to 1 to enforce pure lasso regularization, although the lower penalty actually makes the model similar to a standard logistic model.

The logistic regression model estimates the probability that a customer accepts a personal

loan using the following formula (1):

$$\text{log-odds(Personal Loan} = 1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{1}$$

where log-odds(Personal Loan = 1) represents the logarithm of the odds of accepting a personal loan, β_0 is the intercept, and $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients associated with the predictor variables $X_1, X_2, \ldots, X_p$, all estimated from the training data.

We then formalize the logistic regression formula that relates the dependent variable Personal.Loan to a set of independent variables, including Age, Income, Family, CCAvg, Education, Mortgage, Securities.Account, CD.Account, Online, and CreditCard. The model is trained using a training subset of the data, which is prepared by an initial random split into training and test sets.

5.2.2 Tree Model

We built a classification tree model using R to predict the likelihood of a customer accepting a personal loan. In developing the model, we considered three important hyperparameters: min_n, tree_depth, and cost_complexity. The min_n parameter controls the minimum number of observations required for a node before attempting a split. The tree_depth parameter limits the maximum number of layers, or splits, in the tree. Setting an appropriate tree depth helps prevent the tree from becoming too complex and memorizing noise in the data, thereby avoiding overfitting. Finally, the cost_complexity parameter controls the overall complexity of the tree by penalizing additional splits. Lower cost_complexity values allow the tree to grow larger and deeper by encouraging more splits, while higher values encourage earlier pruning of branches.

During our modeling process, we paid special attention to tuning the cost_complexity parameter to balance model complexity and prediction accuracy. After evaluating the cross-

validation error behavior, we set cost_complexity to 0.001, which allowed the tree to grow deep enough to capture important patterns without causing overfitting. The min_n and tree_depth parameters were retained at their default settings in the rpart package but may be adjusted in future extensions of the model.

6 Findings

In this section, we will present the output and performance of each model. It is primarily done through the confusion matrix to evaluate and compare their performance.

6.1 Logistic Regression

The results of the logistic regression analysis (Figure 5) reveal several important predictors of personal loan acceptance. Variables such as Income, Family, CCAvg, Education2, Education3, CD.Account1, and CreditCard1 were statistically significant at the 0.05 level or better. Higher income levels (Income) and higher average credit card spending (CCAvg) were associated with an increased likelihood of accepting a personal loan. Family size (Family) was also positively associated with loan acceptance. Educational attainment, particularly having a graduate degree (Education3), substantially increased the probability of accepting a loan compared to the baseline education group. Moreover, holding a certificate of deposit account (CD.Account1) strongly increased the odds of loan acceptance. Conversely, having an online account (Online1) or owning a credit card (CreditCard1) was negatively associated with loan acceptance, suggesting that these customers might already have access to alternative financial services and thus lower demand for personal loans.

The fitted logistic regression model demonstrated strong predictive performance. As shown in the confusion matrix (Table 3), the model correctly classified 1180 true negatives and 67 true positives, while making 9 false positive errors and 56 false negative errors. The relatively low number of false positives suggests that the model is conservative in predicting positive

13

cases (loan acceptance), which is desirable in financial contexts where over-approving loans may incur significant risk. Overall, the model achieved a high level of predictive accuracy (95.05%) and identified meaningful customer characteristics associated with personal loan acceptance.

6.2 Tree Module

The results of the classification tree analysis (Figure 6) reveal several important predictors of personal loan acceptance. Key variables such as Income, CCAvg, CD.Account, Education, and Family played critical roles in segmenting customers by their likelihood of accepting a personal loan. Higher income levels (Income) and higher average credit card spending (CCAvg) were associated with an increased likelihood of loan acceptance, as indicated by multiple early splits in the tree. Customers holding a certificate of deposit account (CD.Account) also showed a significantly higher propensity to accept a personal loan. Additionally, education level (Education), particularly having a graduate degree, was a strong differentiator among customer groups. Family size (Family) further contributed to distinguishing customers' likelihood to accept a loan, with smaller families associated with higher acceptance rates in certain branches of the tree. The combination of these variables allowed the model to capture complex interactions that influence loan acceptance decisions.

The fitted classification tree model demonstrated strong predictive performance. As shown in the confusion matrix (Table 4), the model correctly classified 1182 true negatives and 104 true positives, while making 7 false positive errors and 19 false negative errors. The relatively small number of misclassifications indicates that the tree model was effective in distinguishing between customers who were likely versus unlikely to accept a personal loan. In particular, the very low number of false positives suggests that the model is conservative in predicting loan acceptance, minimizing the financial risk associated with incorrectly approving unqualified customers. Overall, the classification tree achieved a high level of

accuracy (98.02%) and provided interpretable decision rules that can guide marketing and risk management strategies.

7 Conclusion

Based on the model results, we analyzed some management issues, the core purpose of which is to predict whether indebted customers will purchase personal loans, and use these models to help banks identify target customers and adjust marketing strategies. This not only helps banks better convert indebted customers into personal loan customers, but also optimizes the bank's resource allocation and market positioning.

Combining the positive correlation coefficient of logistic regression and the results of the classification tree, we can identify the key factors that affect customer decisions. We found that "income," "education," "family," "average credit card consumption amount (CCAvg)," and "certificate of deposit account" (CD_Account) are the five most important variables. To be more specific, people with higher income, larger family size, higher credit card consumption amount, higher education level, and holding a certificate of deposit account (CD) are more likely to apply for personal loans. This finding provides clear guidance for our marketing strategy for specific customer groups.

Therefore, in the next marketing campaign for indebted customers, banks should pay more attention to this group. This precise marketing strategy can not only improve marketing efficiency, but also meet the needs of specific customer groups more accurately, thereby increasing the conversion rate of personal loan products. Through such a strategy, banks can use resources more efficiently and provide customers with services that better meet their needs, while also creating greater profit margins for the banks themselves.

References

Ahsan, M., Mahmud, M., Saha, P., Gupta, K., and Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies*, 9(3):52.

Al-Quraishi, T., Albahri, O., Albahri, A., Alamoodi, A., and Sharaf, I. M. (2025). Bridging Predictive Insights and Retention Strategies: The Role of Account Balance in Banking Churn Prediction. *AI*, 6(4):73.

Ala'raj, M., Abbod, M. F., and Majdalawieh, M. (2021). Modelling customers credit card behaviour using bidirectional LSTM neural networks. *Journal of Big Data*, 8(1):69.

Alborzi, M. and Khanbabaei, M. (2016). Using data mining and neural networks techniques to propose a new hybrid customer behaviour analysis and credit scoring model in banking services based on a developed RFM analysis method. *International Journal of Business Information Systems*, 23(1):1.

Alonso Robisco, A. and Carbó Martínez, J. M. (2022). Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. *Financial Innovation*, 8(1):70.

Bharathi S, V., Pramod, D., and Raman, R. (2022). An Ensemble Model for Predicting Retail Banking Churn in the Youth Segment of Customers. *Data*, 7(5):61.

Boustani, N., Emrouznejad, A., Gholami, R., Despic, O., and Ioannou, A. (2024). Improving the predictive accuracy of the cross-selling of consumer loans using deep learning networks. *Annals of Operations Research*, 339(1-2):613–630.

Bücker, M., Szepannek, G., Gosiewska, A., and Biecek, P. (2022). Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, 73(1):70–90.

Chang, V., Hahm, N., Xu, Q. A., Vijayakumar, P., and Liu, L. (2024). Towards data and analytics driven B2B-banking for green finance: A cross-selling use case study. *Technological Forecasting and Social Change*, 206:123542.

Chen, D., Ye, J., and Ye, W. (2023). Interpretable selective learning in credit risk. *Research in International Business and Finance*, 65:101940.

De Caigny, A., De Bock, K. W., and Verboven, S. (2024). Hybrid black-box classification for customer churn prediction with segmented interpretability analysis. *Decision Support Systems*, 181:114217.

de Waal, H., Nyawa, S., and Wamba, S. F. (2024). Consumers' Financial Distress: Prediction and Prescription Using Interpretable Machine Learning. *Information Systems Frontiers*.

Fenton, E. M. (2007). Visualising Strategic Change:. *European Management Journal*, 25(2):104–117.

He, W., Hung, J.-L., and Liu, L. (2022). Impact of big data analytics on banking: a case study. *Journal of Enterprise Information Management*.

Hussein Sayed, E., Alabrah, A., Hussein Rahouma, K., Zohaib, M., and Badry, R. M. (2024). Machine Learning and Deep Learning for Loan Prediction in Banking: Exploring Ensemble Methods and Data Balancing. *IEEE Access*, 12:193997–194019.

Joung, J. and Kim, H. (2023). Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. *International Journal of Information Management*, 70:102641.

Krasnikov, A., Jayachandran, S., and Kumar, V. (2009). The Impact of Customer Relationship Management Implementation on Cost and Profit Efficiencies: Evidence from the U.S. Commercial Banking Industry. *Journal of Marketing*, 73(6):61–76.

Kruthika, M., Rajadevi, R., Senthilvadivu, K., Lavanya, J., Deepak, D., and Lavanya, G. (2024). Comparative study of Machine Learning Approaches for Loan Approval Prediction in financial sector. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

Machauer, A. and Morgner, S. (2001). Segmentation of bank customers by expected benefits and attitudes. *International Journal of Bank Marketing*, 19(1):6–18.

Melsom, B., Bakke Vennerød, C., de Lange, P. E., Hjelkrem, L. O., and Westgaard, S. (2022). Explainable artificial intelligence for credit scoring in banking. *Journal of Risk*.

Momparler, A., Carmona, P., and Climent, F. (2016). Banking failure prediction: a boosting classification tree approach. *Spanish Journal of Finance and Accounting / Revista Española de Financiación y Contabilidad*, 45(1):63–91.

Moro, S., Cortez, P., and Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31.

Murugan, M. S. and T, S. K. (2023). Large-scale data-driven financial risk management & analysis using machine learning strategies. *Measurement: Sensors*, 27:100756.

Nikiel, E. M. and Opiela, T. P. (2002). CUSTOMER TYPE AND BANK EFFICIENCY IN POLAND: IMPLICATIONS FOR EMERGING MARKET BANKING. *Contemporary Economic Policy*, 20(3):255–271.

Rahman, M. H., Chandra Das, A., Shak, M. S., Uddin, M. K., Alam, M. I., Anjum, N., Bony, M. N. V. A., and Alam, M. (2024). TRANSFORMING CUSTOMER RETENTION IN FINTECH INDUSTRY THROUGH PREDICTIVE ANALYTICS AND MACHINE LEARNING. *The American Journal of Engineering and Technology*, 06(10):150–163.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Safarkhani, F. and Moro, S. (2021). Improving the Accuracy of Predicting Bank Depositor's Behavior Using a Decision Tree. *Applied Sciences*, 11(19):9016.

Singh, P. P., Anik, F. I., Senapati, R., Sinha, A., Sakib, N., and Hossain, E. (2024). Investigating customer churn in banking: a machine learning approach and visualization app for data science and management. *Data Science and Management*, 7(1):7–16.

Ulug, B. and Akyüz, S. (2025). Optimized churn prediction using ensemble-based feature selection via second-order cone programming. *Annals of Operations Research*.

Văduva, A.-G., Oprea, S.-V., Niculae, A.-M., Bâra, A., and Andreescu, A.-I. (2024). Improving Churn Detection in the Banking Sector: A Machine Learning Approach with Probability Calibration Techniques. *Electronics*, 13(22):4527.

Valluri, C., Raju, S., and Patil, V. H. (2022). Customer determinants of used auto loan churn: comparing predictive performance using machine learning techniques. *Journal of Marketing Analytics*, 10(3):279–296.

Zaki, A. M., Khodadadi, N., Lim, W. H., and Towfek, S. K. (2024). Predictive Analytics and Machine Learning in Direct Marketing for Anticipating Bank Term Deposit Subscriptions. *American Journal of Business and Operations Research*, 11(1):79–88.

Variable Name	Variable Definition
ID	Customer ID.
Age	Customer's age in completed years
Experience	Number of years of professional experience
Income	Annual income of the customer (in thousand dollars)
ZIP Code	Home Address ZIP code
Family	Family size of the customer
CCAvg	Average spending on credit cards per month (in thousand dollars)
Mortgage	Value of house mortgage if any (in thousand dollars)
Education	Education Level 1: Undergrad Education Level 2: Graduate Education Level 3: Advanced/Professional
Personal_Loan	Did this customer accept the personal loan offered in the last campaign?
Securities_Account	Does the customer have securities account with the bank?
CD_Account	Does the customer have a certificate of deposit (CD) account with the bank?
Online	Do customers use internet banking facilities?
CreditCard	Does the customer use a credit card issued by any other Bank (excluding All life Bank)?

Table 1: Variable Descriptions Used in the Dataset

Income Class	Income (in thousand dollars)
High Income	Greater than 156
Medium Income	[52, 156]
Low Income	Less than 52

Table 2: Income Classification by Income Level

	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	1180 (TN)	9 (FP)
$Y = 1$	56 (FN)	67 (TP)

Table 3: Confusion Matrix for Logistic Regression Model

	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	1182 (TN)	19 (FP)
$Y = 1$	7 (FN)	104 (TP)

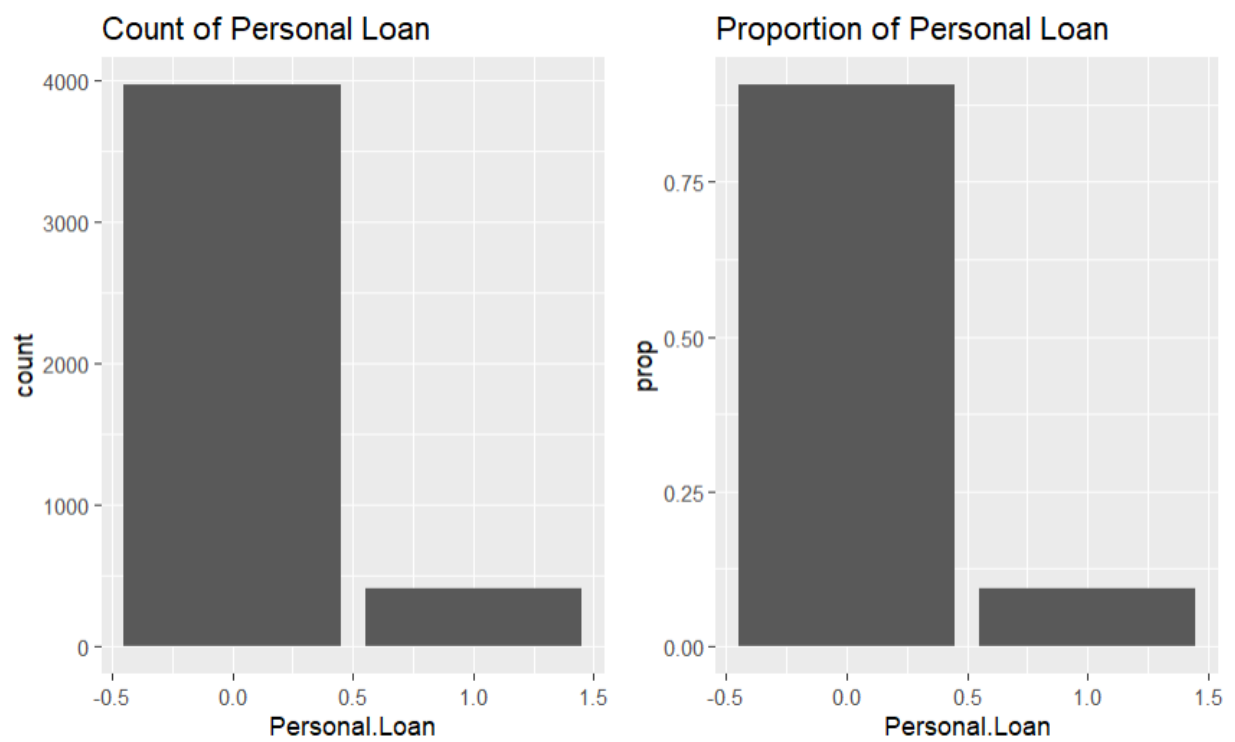Table 4: Confusion Matrix for Classification Tree Model
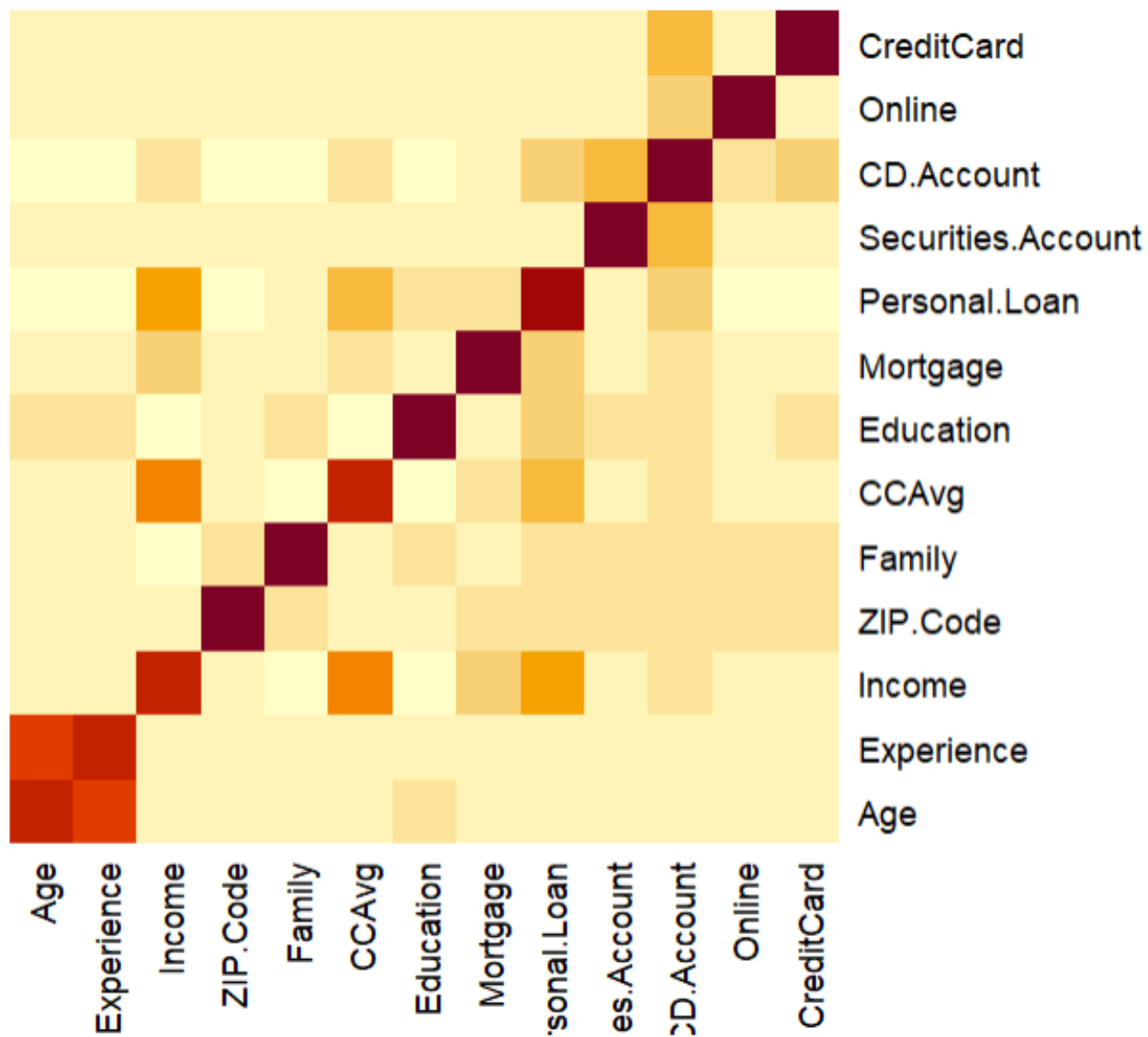
Figure 1: Correlation Matrix

Figure 2: Personal Loan Distribution
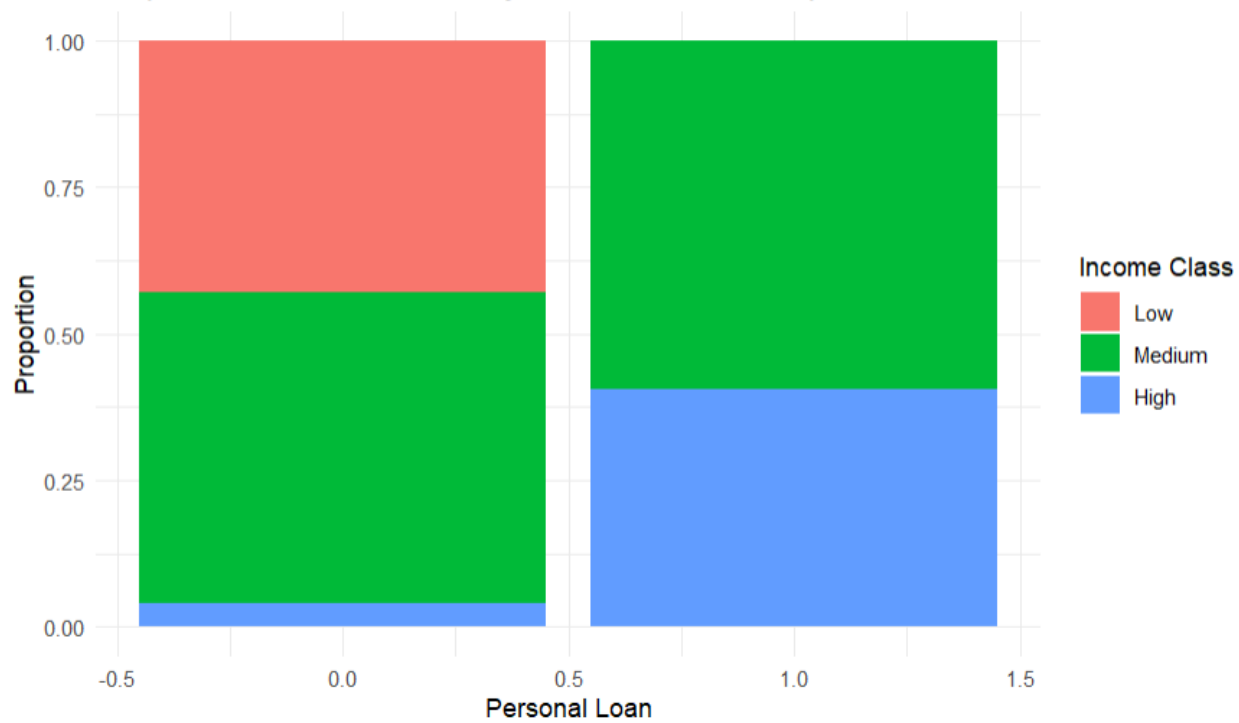
Figure 3: Proportion of Income Class by Personal Loan Acceptance
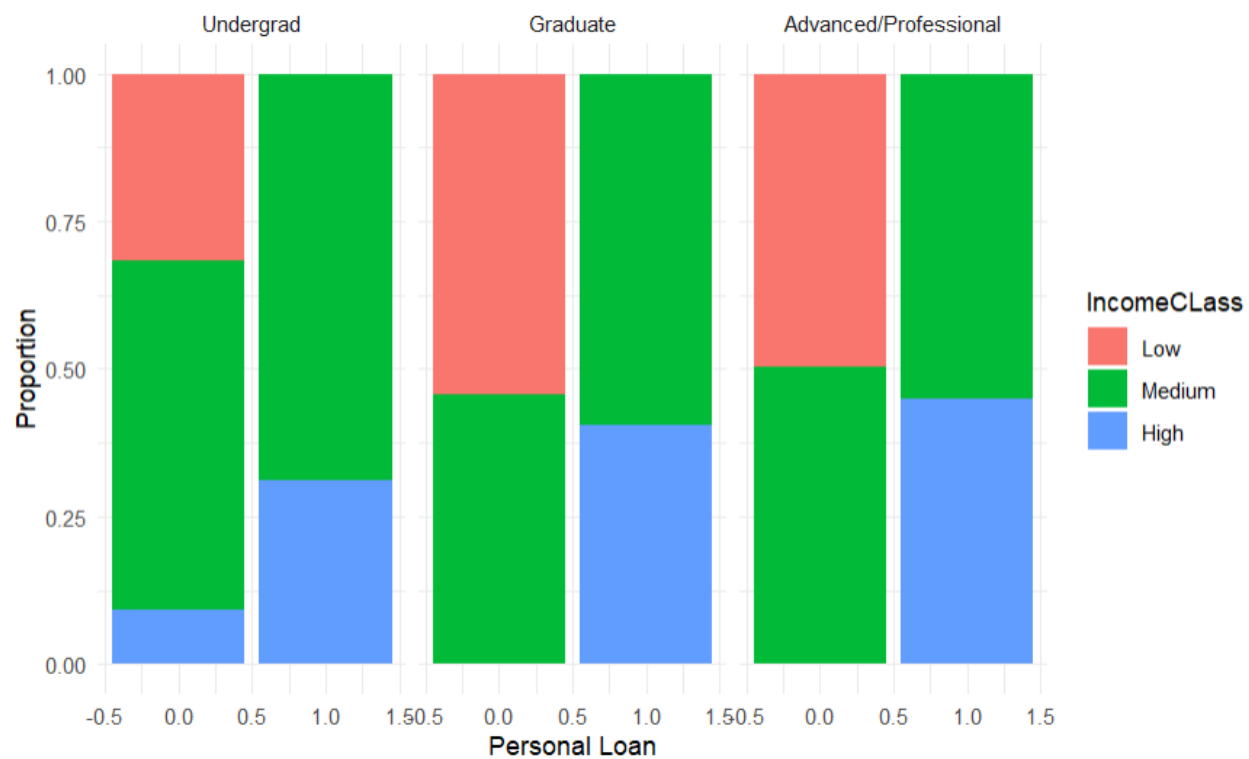
Figure 4: Personal Loan by Education and Income

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.500e+01  9.448e-01 -15.873  < 2e-16 ***
Age                     1.920e-02  9.160e-03   2.097 0.036030 *
Income                  6.513e-02  4.155e-03  15.674  < 2e-16 ***
Family                  7.095e-01  1.049e-01   6.762 1.36e-11 ***
CCAvg                   1.517e-01  6.068e-02   2.500 0.012406 *
Education2              4.289e+00  3.716e-01  11.543  < 2e-16 ***
Education3              4.211e+00  3.670e-01  11.473  < 2e-16 ***
Mortgage                8.625e-04  7.711e-04   1.119 0.263344
Securities.Account1    -1.160e+00  4.422e-01  -2.623 0.008724 **
CD.Account1             3.769e+00  4.706e-01   8.010 1.15e-15 ***
Online1                -5.736e-01  2.195e-01  -2.613 0.008985 **
CreditCard1            -1.054e+00  2.859e-01  -3.685 0.000228 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

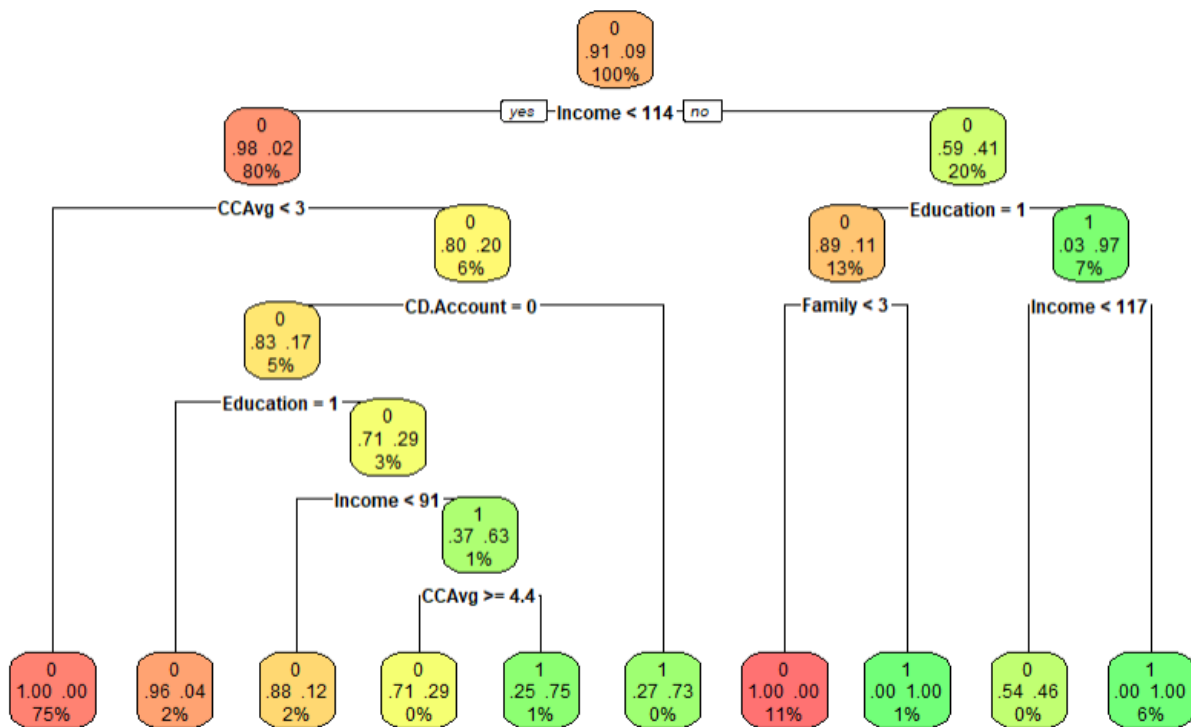Figure 5: The Results of the Logistic Regression Analysis

Figure 6: Tree Module