Predictive Analysis of Personal Loan Approval Using Logistic Regression

Daniel Makala Mackaemba

Master of Professional Studies in Data Analytics

Northeastern University

January 25, 2026

Introduction

Financial institutions must balance efficiency, accuracy, and risk management when evaluating loan applications. Traditional manual review processes are resource-intensive and may introduce inconsistencies in decision-making. Predictive analytics offers a scalable solution by leveraging historical data to estimate loan acceptance probability.

This study applies logistic regression to predict personal loan acceptance based on customer attributes. Logistic regression was selected due to its interpretability, probabilistic outputs, and suitability for binary classification problems. The objectives of this analysis are to (1) identify the most significant variables influencing loan acceptance, (2) determine which factors negatively affect acceptance, and (3) assess the overall accuracy and precision of the predictive model.

Methodology

Data Preparation

The dataset consists of 5,000 customer records with 14 variables capturing demographic characteristics (e.g., age, education, family size), financial indicators (income, mortgage balance, average credit card spending), and banking relationships (CD account, securities account, online banking usage). Initial data inspection revealed no missing values. However, inconsistencies were identified in the professional experience variable, including negative values and cases where experience exceeded age. These were corrected by converting negative values to absolute values and ensuring experience did not exceed age minus 18 years.

Outliers were identified using the Interquartile Range (IQR) method in income, mortgage, and credit card spending variables. These values were retained, as they represent valid customer profiles rather than data entry errors. Additional engineered features were created to

improve predictive capability, including income per family member, mortgage-to-income ratio, and binary indicators for mortgage presence and high income.

Model Development

The dataset was split into training (70%) and testing (30%) sets using stratified sampling to preserve the loan acceptance rate. All numeric features were standardized using z-score normalization. A logistic regression model was trained on the training set, and performance was evaluated on the test set using accuracy, precision, recall, F1-score, and ROC-AUC metrics.

Results

Significant Predictors of Loan Acceptance

The three most significant variables influencing personal loan acceptance were Income, CD Account ownership, and Family Size, based on coefficient magnitude and statistical significance.

Income was the strongest predictor, with higher-income customers substantially more likely to accept personal loans. CD account ownership also showed a strong positive association, suggesting that customers with established long-term savings relationships are more receptive to additional banking products. Family size demonstrated a positive relationship with loan acceptance, likely reflecting increased financial needs among larger households.

Negative Influence on Loan Acceptance

Among all predictors, Age exhibited the most substantial negative influence on loan acceptance. Older customers were less likely to accept personal loans, potentially due to reduced reliance on credit, accumulated assets, or conservative financial preferences later in life. While age showed a negative association, it reflects behavioral patterns rather than discriminatory intent and remains a legally permissible variable in credit modeling when properly monitored.

Model Performance

The logistic regression model demonstrated excellent predictive performance. Overall accuracy reached 98.73%, correctly classifying 1,481 out of 1,500 test cases. Precision was 93.50%, indicating that the majority of customers predicted to accept loans did so in reality. The model also achieved 91.85% recall, an F1-score of 92.67%, and an ROC-AUC of 0.9924, reflecting near-perfect class separation. Only 19 misclassifications were observed, underscoring the model's robustness.

Business Implications

The results support immediate operational use of the model in loan approval workflows. Automated pre-screening can prioritize high-probability applicants for review, improving efficiency and reducing processing time. Targeted marketing efforts can focus on customers with higher income levels, CD accounts, and larger families, improving campaign effectiveness and resource allocation.

To ensure long-term effectiveness, the model should be periodically retrained using recent data and monitored for fairness across demographic groups. Classification thresholds may also be adjusted depending on business priorities, such as maximizing acceptance rates or minimizing false positives.

Conclusion

This analysis demonstrates that logistic regression is a highly effective and interpretable method for predicting personal loan acceptance. Income, CD account ownership, and family size emerged as the most significant predictors, while age showed the strongest negative influence. With 98.73% accuracy and 93.50% precision, the model provides a reliable decision-support tool for automating loan approval processes. By adopting this data-driven approach, the bank can enhance operational efficiency, improve targeting strategies, and maintain rigorous risk management standards.

References

Agresti, A. (2018). An introduction to categorical data analysis (3rd ed.). Wiley.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning with applications in R (2nd ed.). Springer.

Scikit-learn Development Team. (2024). Scikit-learn: Machine learning in Python.