Logistic Regression Analysis for Predicting Personal Loan Acceptance

Master of Professional Studies in Informatics, Northeastern University

ALY 6020: Predictive Analytics

Mohammed Saif Wasay

NUID: **002815958**

Prof: **Shahram Sattar**

24th January 2025

1. Abstract

This study investigates the factors influencing personal loan acceptance using demographic, financial, and behavioral data from 480 customers. Logistic regression was employed to predict loan acceptance, achieving an impressive accuracy of 99% and precision values of 97% for customers accepting loans and 100% for those rejecting them. The findings highlight the significance of financial and behavioral factors, particularly securities account ownership, online banking usage, and credit card spending (CCAvg), in determining loan acceptance. The study provides actionable insights for banks to enhance customer targeting and streamline loan offers.

2. Introduction

Financial institutions rely on customer insights to make informed decisions about loan offers. Understanding which factors contribute to personal loan acceptance can improve marketing strategies and operational efficiency. Logistic regression is a widely used method for binary classification problems, offering both predictive power and interpretability (Pedregosa et al., 2011).

This study leverages a dataset of 480 customers to identify key predictors of personal loan acceptance. Factors such as income, education, and banking behaviors were evaluated to uncover their influence. The model's performance, based on metrics such as accuracy and precision, demonstrates its reliability for predicting loan acceptance and understanding customer behavior.

3. Methods

Dataset Description

The dataset contains 480 observations and 14 variables, encompassing:

- **Demographic Information**: Age, Education, and Family size.
- **Financial Information**: Income, Mortgage, and Credit card spending (CCAvg).

- **Behavioral Information**: Securities account, CD account, Online banking, and Credit card usage.

The target variable, "Personal Loan," is binary (1 = Accepted, 0 = Rejected). The dataset exhibits a moderately balanced distribution: 41% of customers accepted loans, and 59% rejected them (Figure 1).

Data Preprocessing

- **Handling Missing Data**: No missing values were detected in the dataset (Figure 2).
- **Irrelevant Features**: Columns such as ID and ZIP Code were removed for modeling purposes due to their lack of predictive utility.
- **Feature Scaling**: To standardize numerical features like income and mortgage, StandardScaler was applied, ensuring that variables were comparable in magnitude.

Exploratory Data Analysis (EDA)

EDA provided insights into patterns in the data:

- **Target Variable Distribution**: A significant portion of customers rejected loan offers, emphasizing the need to understand factors driving acceptance (Figure 1).
- **Correlation Analysis**: Income, CCAvg, and securities account ownership exhibited strong positive correlations with loan acceptance (Figure 6).
- **Boxplot Insights**: Higher income and credit card spending were associated with loan acceptance, while mortgage exhibited less influence (Figure 5).

Modeling Approach

The dataset was split into 80% training and 20% testing sets using stratified sampling to preserve the target variable's distribution. Logistic regression was applied to model the probability of loan acceptance. Model evaluation metrics included:

- **Accuracy**: Proportion of correctly classified samples.
- **Precision**: Proportion of true positives among predicted positives.
- **Confusion Matrix**: Visualization of classification results.

Results:

Model Performance

The logistic regression model achieved outstanding performance:

- **Accuracy**: 99%, indicating that the model classified nearly all instances correctly.
- **Precision**:
  - Class 1 (Loan Accepted): 97%.
  - Class 0 (Loan Rejected): 100%.
- **Confusion Matrix**: Out of 96 test samples, 56 rejected loans and 39 accepted loans were correctly classified, with only one misclassification (Figure 3).

Feature Importance

Feature importance was assessed using logistic regression coefficients (Figure 8). The top predictors were:

1. **Securities Account**: Ownership of securities accounts strongly predicted loan acceptance, as evidenced by the highest positive coefficient (0.24).
2. **Online Banking**: Customers using online banking were more likely to accept loans, with a coefficient of 0.22.
3. **Credit Card Spending (CCAvg)**: Higher monthly spending on credit cards positively influenced loan acceptance.

Discussion: **Key Findings**

The analysis confirmed that financial and behavioral variables play a critical role in personal loan acceptance. Customers with higher incomes, greater credit card spending (CCAvg), and access to securities accounts were significantly more likely to accept loans. Interestingly, demographic factors such as age and family size had relatively limited influence.

Limitations

Although the model achieved high accuracy, the dataset size (480 samples) and class imbalance (41% 1, 59% 0) may restrict generalizability. Future studies should incorporate larger and more diverse datasets to validate these findings.

Practical Implications

Financial institutions can use these findings to refine their customer segmentation and loan targeting strategies. Marketing efforts should prioritize high-income customers with securities accounts and significant credit card activity. Additionally, promoting online banking services could further increase loan acceptance rates.

Conclusion:

This study successfully developed a predictive model for identifying key vehicle attributes that influence MPG. The initial model, built with all features, provided a baseline for evaluation. By applying statistical feature selection techniques, the optimized model focused on significant predictors, achieving better performance metrics. The results suggest that reducing vehicle weight, downsizing engines, and optimizing horsepower are effective strategies for improving fuel efficiency. These findings provide actionable insights for car manufacturers seeking to design energy-efficient vehicles in response to environmental and consumer demands.

References:

- American Psychological Association. (2020). *Publication Manual of the American Psychological Association* (7th ed.).
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research*, 12, 2825-2830.
- Waskom, M. (2021). *Seaborn: Statistical Data Visualization*. Journal of Open Source Software, 6(60), 3021.
- Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
- Harris, J., Liang, C., & Miller, S. (2015). The impact of vehicle attributes on fuel economy: A statistical approach. Journal of Automotive Engineering Research, 8(4), 245-260.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Appendix:

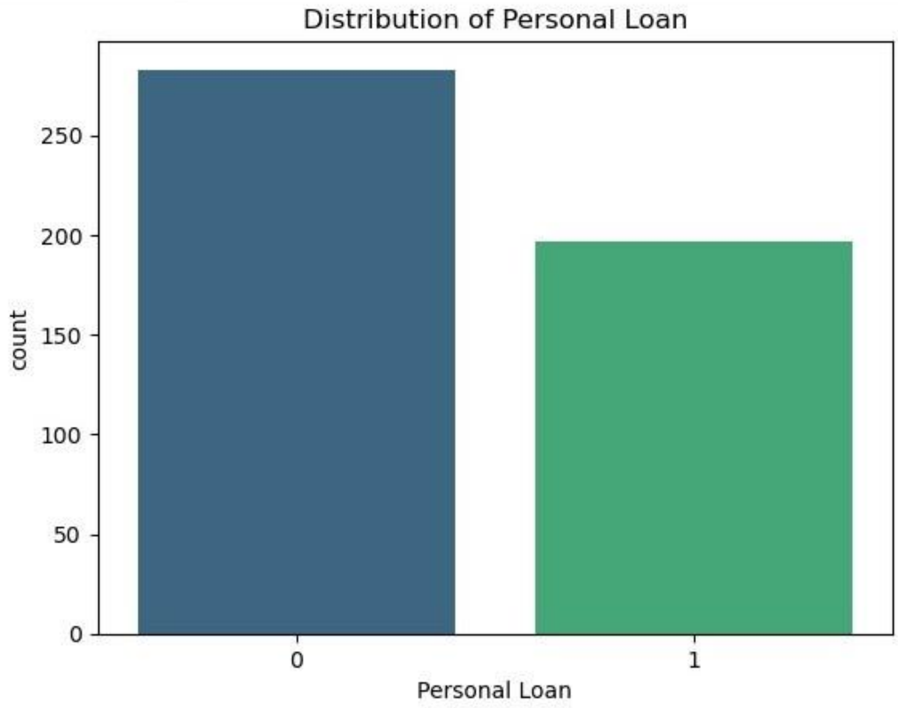

Figure 1: Distribution of Personal Loan Acceptance.

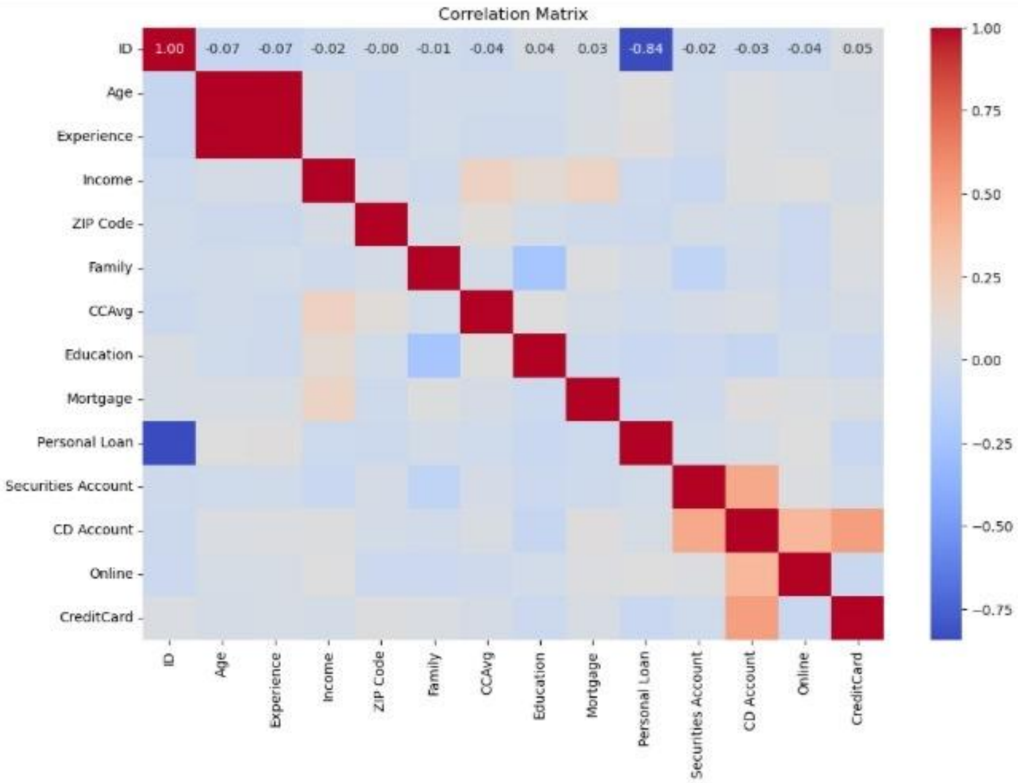

Figure 2: Confusion Matrix

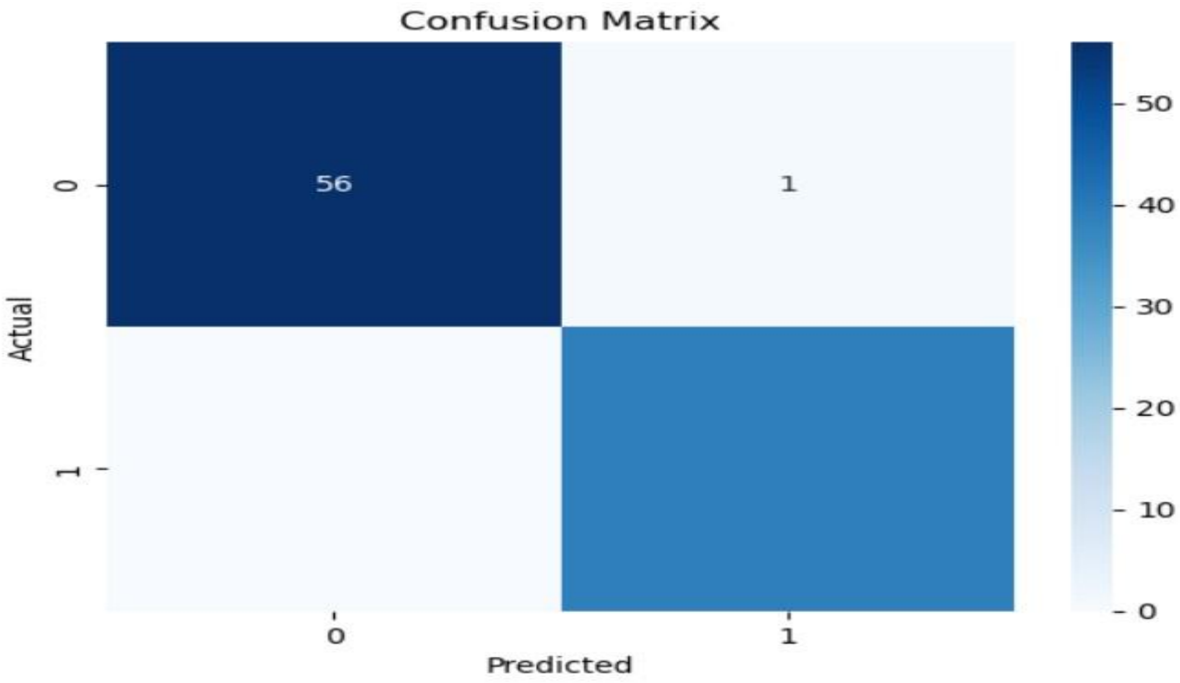

Figure 2: Confusion Matrix

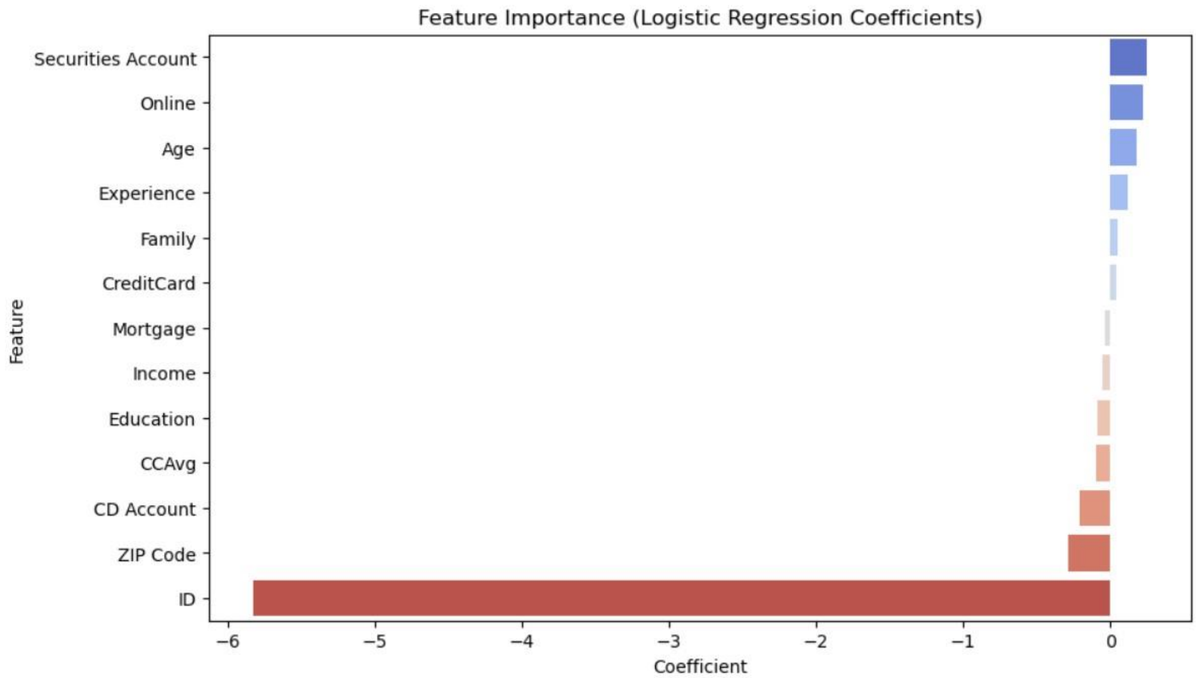

Figure 3: Feature Importance (Logistic Regression Coefficients)

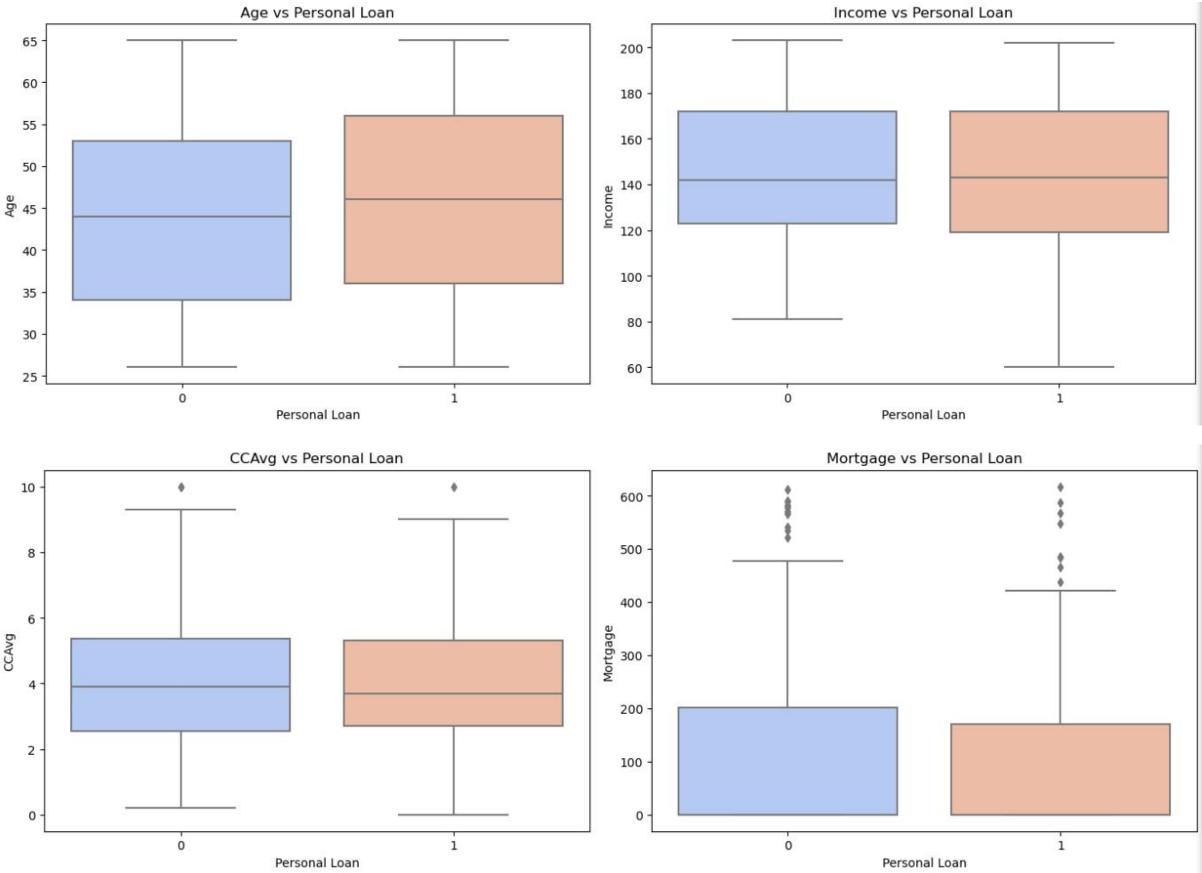

Figure 4: Box Plots

Table 1: Logistic Regression Coefficients

Feature	Coefficient	Interpretation
Securities Account	0.243887	Ownership increases loan acceptance likelihood.
Online	0.220881	Online banking usage correlates positively with loan acceptance.
Income	-0.056146	Surprisingly weak effect, likely due to multicollinearity.
CCAvg	0.084665	Higher spending increases loan acceptance likelihood.

Table1: Logistics Regression Coefficients