

Intern Task

Report on Bank Loan Classification

Kabin Chongbang Limbu

Kabinchongbanglimbu2060@gmail.com

06 July, 2024

1. Introduction

1.1. Objective

The objective of this analysis is to build a machine learning model that can accurately classify whether a personal loan was accepted based on the customer information provided. The target variable is 'Personal Loan,' which indicates whether the loan was accepted (1) or not (0).

1.2. Dataset Description

The dataset contains information about bank customers, including 15 features and a target variable. The features include demographic details, financial information, and banking behavior. The target variable indicates whether the customer accepted a personal loan.

1.3. Feature Description

- **ID:** ID of the customer
- **Age:** Age of the customer
- **Gender:** Gender of the customer (M for Male, F for Female, O for Others)
- **Experience:** Amount of work experience in years
- **Income:** Annual income (in thousands)
- **HomeOwnership:** Home ownership status (Home Owner, Rent, Home Mortgage)
- **Zipcode:** Postal code in which the customer lives
- **Family:** Number of family members
- **CCAvg:** Average monthly spending with the credit card (in thousands)
- **Education:** Education level (1: bachelor's degree, 2: master's degree, 3: advanced/professional degree)
- **Mortgage:** Value of home mortgage (in thousands)
- **Securities Account:** Does the customer have a securities account with the bank?
- **CDAccount:** Does the customer have a certificate of deposit account with the bank?
- **Online:** Does the customer use internet banking facilities?
- **CreditCard:** Does the customer use a credit card issued by the bank?
- **Personal Loan:** Did this customer accept the personal loan offered in the last campaign? (Target Variable)

2. Data Preprocessing

2.1. Loading the Dataset

The dataset was loaded into a panda DataFrame for analysis.

2.2. Handling Missing Values

Missing values were handled by removing the value of the respective columns.

2.3. Encoding Categorical Variables

Categorical variables were encoded using Label Encoding.

2.4. Scaling Numerical Variables

Numerical variables were scaled using StandardScaler.

3. Exploratory Data Analysis

3.1. Distribution of Target Variable

The distribution of the target variable was analyzed to understand the balance of classes.

3.2. Scatter Plot and Box Plot

A Scatter plot was used to visualize relationships between Income and Credit Card Average.

A Box Plot was used to visualize Income by Personal Loan and Mortgage by Personal Loan.

3.3. Correlation Matric

A correlation matrix was created to identify relationships between features.

4. Model Building and Evaluation

4.1. Data Splitting

The dataset was split into training and testing sets. It is split in the ratio of 70:30.

4.2. Model Training

A Various models like Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine and Gradient Boosting is used on training.

4.3. Model Evaluation

The model was evaluated on the test set using accuracy, precision, recall and f1-score

4.4. Model Saving

The trained model was saved in “.pkl” format for future predictions.

5. Conclusion

5.1. Summary of Findings

The analysis successfully built a model to predict personal loan acceptance. The best model for prediction is Gradient Boosting which achieve an accuracy of 98.57% on the test set. Key factors influencing loan acceptance included income, average monthly credit card spending, and education level.

5.2. Recommendations and Future Work

Future work could explore additional features, such as customer transaction history, to improve model accuracy. Analyze historical data to detect shifts in customer behavior over time, which could influence model predictions.