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

Haotian Zhang

University of Oklahoma

March 2025

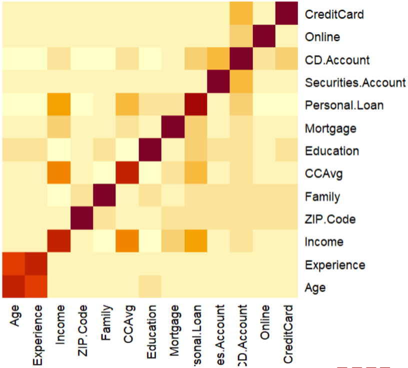
Introduction

- Thera Bank aims to convert deposit customers into personal loan customers.
- Personal loans improve profitability compared to deposits.
- Predictive modeling can optimize customer targeting and marketing.

Data Collection and Description

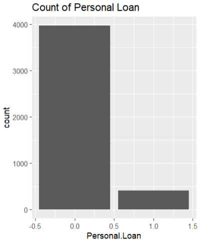
- Dataset: 5,000 Thera Bank customers (4373 after clean).
- Variables: Age, Income, Family, Credit Card Average Spending (CCAvg), Education, CD Account, etc.
- Removed Experience due to high correlation with Age.

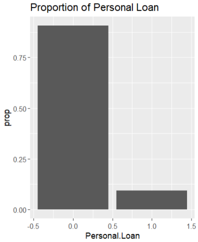
Variable Name	Variable Definition
ID	Customer ID.
Age	Customer's age in completed years
Experience	Number of years of professional experience
Income	Annual income of the customer (in thousand dollars)
ZIP Code	Home Address ZIP code
Family	Family size of the customer
CCAvg	Average spending on credit cards 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)?

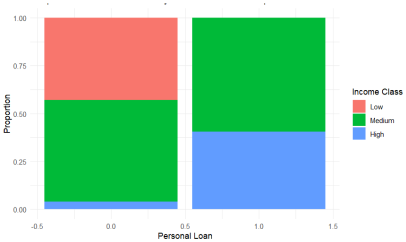


Visualization

- Personal Loan Distribution. (Only 10% of customers accepted a loan.)
- Income-based segmentation: **No low-income** customers accepted loans.







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

Methodology

- **Model Selection:**
  - **Logistic Regression**: Chosen for transparency, interpretability, and strong performance in structured financial datasets.
  - **Tree Model**: Selected for rule-based segmentation and easy interpretation of customer groups.
  - **K-Nearest Neighbors (KNN)**: Rejected due to scalability issues, preprocessing needs, and lack of explainability.
- **Model Building:**
  - Logistic regression: Specified a logistic regression model with a small L1 penalty (0.01) to approximate ordinary logistic regression, setting the mixing parameter to 1 for pure lasso regularization.
  - Tree model: Set the cost_complexity to 0.001 to balance depth and overfitting, while retaining the default settings for min_n and tree_depth for potential future adjustment.

Findings: Logistic Regression

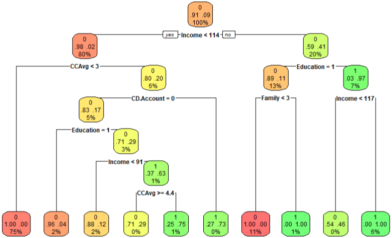
- Significant predictors: Income, Family size, CCAvg, Education, CD Account.
- Higher income and graduate education increase likelihood.
- Online banking and Credit Card ownership negatively associated.
- Logistic Regression Accuracy: 95.05%

```
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
```

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

Findings: Tree Model

- Key splits: Income, CCAvg, CD Account, Education.
- Clear rule-based segmentation.
- Tree Model Accuracy: 98.02%



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

Conclusion and Implications

- Predictive models help the bank identify target customers and adjust marketing strategies.
- People with the following features are more likely to buy personal loans:
  - Higher income
  - Larger family size
  - Higher spending on credit cards
  - Higher level of education (graduate or advanced degree)
  - Certificate of deposit (CD) account ownership
- Improve personal loan conversion rates, optimize resource allocation, and increase bank profitability.

How Did AI Contribute to This Research?

This research utilized AI, specifically ChatGPT 4o, for:

- Code development and troubleshooting in RStudio and LaTeX.
- Internet searches for academic research[1].
- Advice on methodologies and argument structures.
- Assisting in basic mathematics and data cleaning tasks.

The use of AI is equated to any other research tool, ensuring the paper's academic integrity. Critical intellectual processes were conducted by the human research team.

[1] All AI-found research was verified for accuracy by the researcher.