

Analysis of Neo Bank Personal Loan Campaign

jiayi xiao^{1*}, jinghao gao², junlong yan³

¹*School of Data Science, Capital University of Economics and Business, Beijing, China*

²*Macquarie Business School, Macquarie University, Sydney, Australia*

³*Smeal College, Pennsylvania State University, State College, USA*

**Corresponding Author. Email: xiaojiayi202109@163.com*

Abstract. This article focuses on the key customer attributes influencing personal loan purchases, which is a key area of interest for financial institutions seeking to expand their asset base. This study specifically examined the transition of debt customers to borrowers at Neo Bank, a US bank. This article identifies and quantifies the key determinants that influence customers' personal loan decisions, providing a reference for customer behavior and preferences. In this work, we developed a predictive model that uses customer data to predict the likelihood of indebted customers getting personal loans. The innovation of this model lies in its ability to identify customer groups with higher conversion potential, enabling targeted marketing strategies. The findings highlight the importance of customer analytics in improving loan acquisition rates and highlight the effectiveness of precision marketing. The conclusions drawn from this study are multifaceted, and by identifying the most important customer attributes, this research helps to develop more effective sales strategies and products to meet the needs of high-potential customer groups. The significance of this study is that it highlights the value of data analytics in driving business growth and customer satisfaction.

Keywords: Neo Bank, Segment of Customers, Personal Loan Campaign, Decision Tree

1. Introduction

Neo Bank, a financial institution in the United States, is experiencing a surge in its customer base, predominantly consisting of liability customers with a spectrum of deposit sizes. The proportion of customers who are also asset customers remains minimal. The bank is keen on swiftly augmenting this segment to bolster its loan portfolio and, consequently, to enhance earnings through interest accruals. The management is particularly interested in strategies that can convert existing depositors into personal loan customers without jeopardizing their status as depositors.

According to Yuzhen Wang and Jingqiao Qin [1], Product Advisor Duration, Product Advisor Uncontacted Days, Previous Interactions with Customer, the customer's account balance, loan status, and default records are factors that have a significant impact on customer purchasing behavior. Guo, Weidong, Zhou, and Zhizhong [2] used a tree-based feature selection method and classifier in a comparative study to predict personal loan defaults. After referring to and comparing diverse experimental methods, we chose a decision tree model to conduct our research.

This study intends to utilize data gathered from Neo Bank's clientele to scrutinize the determinants of customer loyalty within the banking sector, particularly focusing on the transition from liability customers to borrowers. In this paper, the authors construct a predictive model designed to assist the marketing department in pinpointing potential customers with a higher propensity to procure loans. The overarching aim of this project is to dissect the pivotal factors that influence a liability customer's decision to take out personal loans and to comprehend the customer attributes that are most influential in driving these decisions.

2. Methodology

2.1. Variables

The data set used for this study represents a survey of 5,000 Neo Bank customers across multiple dimensions. The survey included the customer's age, experience, income, ZIPCode, family, average spending on credit cards per month, education, value of house mortgage, whether to accept personal loan, whether to have a Securities Account, whether to have a CD Account, whether to use internet banking facilities, And whether to use credit cards from other banks. During the study, we set "whether to accept personal loan" as a dependency factor. At the same time, consider selecting Age, Experience, Income, ZIPCode, Family, CCAvg, Education, Mortgage, Securities_Account, CD_Account, Online and CreditCard are tested as independent factors. The codes in ZIPCode column contains five numbers, but we only takes the first two digits of the code for statistical analysis. The Education section has three categories: Education Level, Undergrad, Graduate, Advanced/Professional, which represent the customer's education level. For Securities_Account, CD_Account, Online(whether to use internet banking facilities), CreditCard(whether to use a credit card issued by any other Bank columns), We divide whether the state is owned or whether the action is used into 0 and 1. 0 indicates none and 1 indicates yes.

We divided whether the customer accepts a personal loan into 0 and 1. 0 indicates that the customer does not accept the personal loan, and 1 indicates that the customer accepts a personal loan.

2.2. Exploratory data analysis

We use univariate analysis, bivariate analysis, and correlation analysis to visualize the data. Univariate analysis mainly focuses on the description and statistical inference of a single variable, which can describe the central or discrete trends of sample data. Bivariate analysis is used to study the relationship between two variables. Correlation analysis mainly uses heatmaps to display the correlation coefficients between multiple variables, used to quantify the degree of linear correlation between two variables, with values ranging from -1 to 1. The closer the correlation coefficient is to 1 or -1, the stronger the linear relationship between variables; Approaching 0 indicates a weak or non-existent linear relationship.

2.3. Modeling

The modeling method we have chosen is the decision tree. It simulates the human decision-making process by constructing a tree model, where each internal node represents a judgment on an attribute (i.e. a feature selection), each branch represents an output of a judgment result, and each leaf node corresponds to a final decision result (i.e. a category label). Finally, the performance of the decision tree model is evaluated using accuracy, precision, recall, and F1 score.

3. Methodology

3.1. Summary

According to survey data, the age of the bank's customers is centered around 45 years old, and the customer's experience is centered around 20 years and relatively evenly distributed between 0-40 years.

3.1.1. Customs' income

Most of customers have an annual income between 0-50 thousand US dollars, while the number of customers with an annual income of over 50 thousand dollars gradually decreases. Figure 1 shows statistics on the incomes of different people.

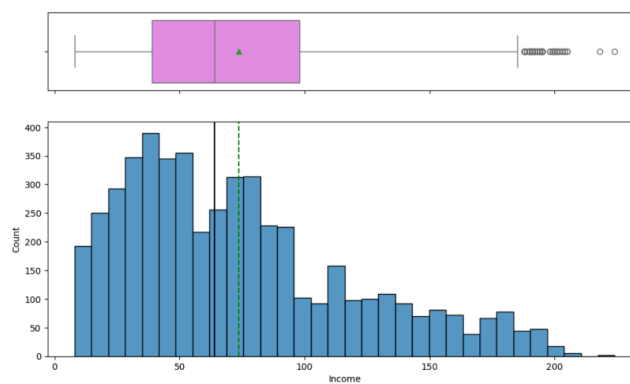


Figure 1. Income histogram

3.1.2. Average spending on credit cards per month

The average spending on credit cards per month of customers is generally decreases from 0 to 10 thousand. Figure 2 shows average monthly credit card spending by different people.

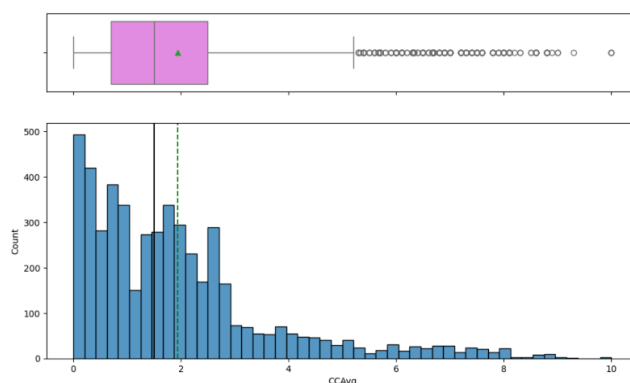


Figure 2. CCavg histogram

3.1.3. Observations on mortgage

Most of the customers have a mortgage value of 0. Figure 3 shows statistics of mortgages for different people.

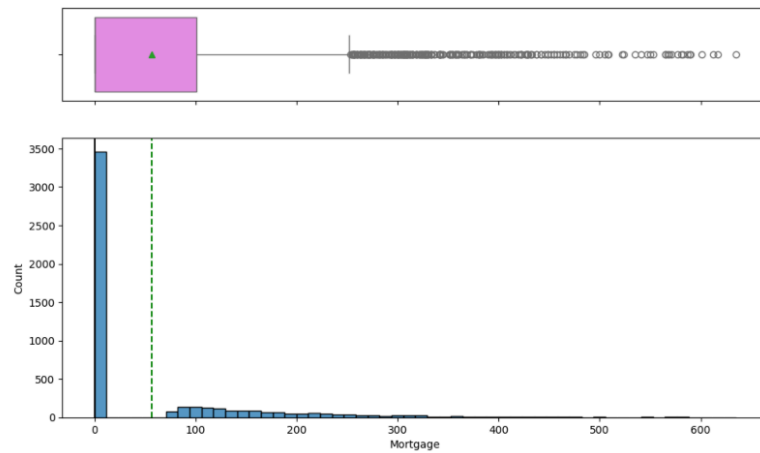


Figure 3. Mortgage histogram

3.1.4. Observations on family

Customers with a family size of 1 have the highest number, while customers with a family size of 3 have the lowest number, but the difference is not big. Figure 4 shows statistics of the number of family members in the client's household.

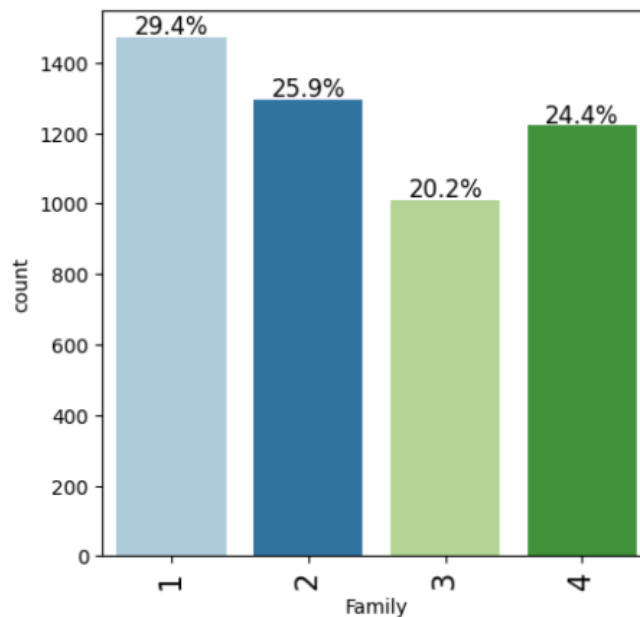


Figure 4. Family size histogram

3.1.5. Customs education

The highest number of customers have an undergraduate educational level, and the number of graduates and professionals is almost the same. Figure 5 shows statistics on customer education.

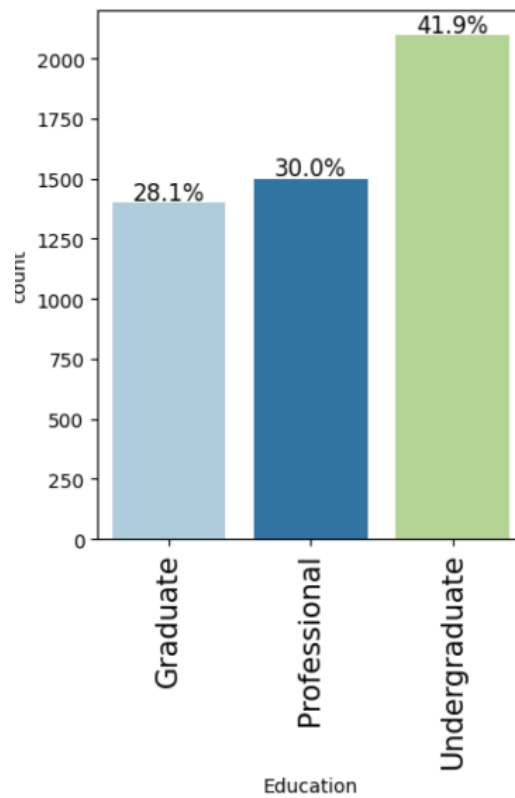


Figure 5. Educational level histogram

3.1.6. Bank CD account

The plot shows that most people don't have a CD-account. Figure 6 shows statistics on whether customers have CD accounts.

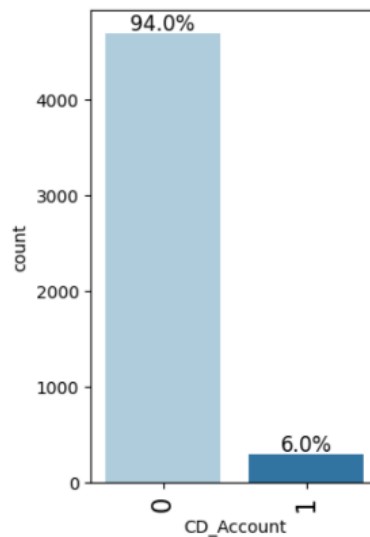


Figure 6. Bank cd account histogram

3.1.7. Online banking using

More people also like using Internet banking facilities. Figure 7 shows statistics on whether customers use banking facilities.

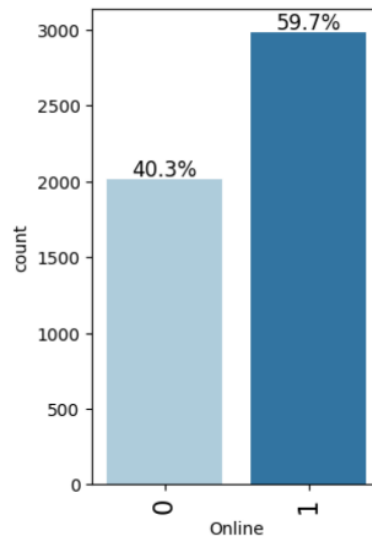


Figure 7. Internet banking facilities using histogram

3.1.8. Using of credit card

A majority of customers don't use a credit card issued by other banks. Figure 8 shows statistics on whether customers have credit cards.

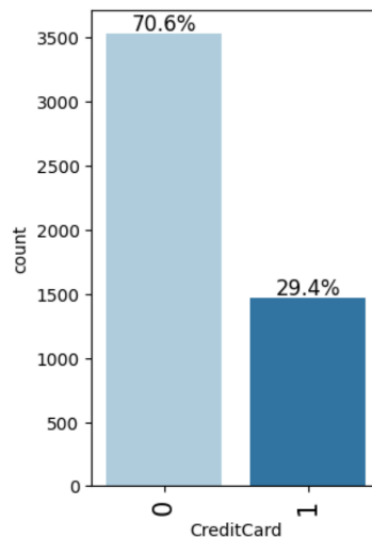


Figure 8. Credit card issued histogram

3.1.9. Total influencing factors

There are many other factors that influence the results, but the relationships are very small. (zip code, online, securities account, credit card)

All in all, for the feature importance, Table 1,2 shows the proportion of impacts. (High to Low)

Table 1. Importance of each factor (1)

variable quantity	Importance
Education_Undergraduate	0.403732
Income	0.304761
Family	0.161717
Average spending on credit cards per month	0.053107
Age	0.036035
CD_Account	0.025711
Mortgage	0.005557

Table 2. Importance of each factor (2)

Education_Professional	0.005144
ZIPCode_92	0.003080
ZIPCode_93	0.000594
Online	0.000561
Securities_Account	0.000000
ZIPCode_91,94,95,96	0.000000
Credit Card	0.000000

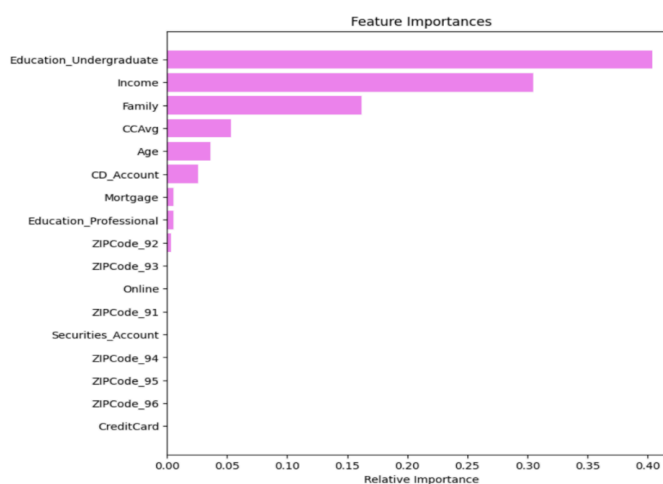


Figure 9. Feature importance histogram

Figure 9 shows statistics of other minor total impacts. Educational, income and family have a highly correlated with Neo Bank Personal Loan Campaign.

3.2. Correlation analysis

The correlation test, There is a strong positive correlation between age and experience, and the correlation is as high as 0.99, which indicates that there may be multi-collinearity in the data set. In addition, the correlation between Income and CCAvg is also high, at 0.65. Figure 10 shows a correlation analysis of different features.



Figure 10. Correlation analysis of feature importance

3.3. Fitting result

Figure 11 shows the results of the data accuracy check.

[93]:

	Accuracy	Recall	Precision	F1
0	0.981333	0.899329	0.911565	0.905405

Figure 11. Fitting result

Our model has extremely high accuracy (including accuracy, recall, precision, and f1), which means that models are better able to predict and classify data with higher reliability.

4. Discussion

4.1. Key findings

4.1.1. Performance in training and testing data

The model performs very well on the training data, with all metrics (accuracy recall precision F1 score) being 1.0. This indicates that the model has no misclassifications on the training set. This usually indicates that the model may be overfitting, that is, the model is too complex to perform very well on the training data. To prevent overfitting, we can consider regularizing the decision tree model such as setting parameters such as the maximum tree depth and the minimum number of sample splits. We can try adding or removing some features, retraining the model, and observing the impact on performance as well. The model's accuracy on the test data is 0.981, indicating that the model's overall performance on the test data is also very good. The recall rate (0.899) and precision (0.912) are also high, and the F1 score is 0.905, indicating that the model performs well in identifying customers who accept loans (positive class), but there is still a small amount of error. From the confusion matrix, we can see that the model is relatively accurate in predicting the negative class (not accepting loans) (1338 correct predictions and only 13 wrong predictions), but

there are a few misclassifications in the positive class (accepting loans) (15 wrong predictions and 134 correct predictions).

4.1.2. Output analysis

As can be seen from the output of the decision tree model, the model predicts whether a customer will accept a personal loan by segmenting the data layer by layer. Income and average monthly credit card spending (CCAvg) are the main segmentation criteria, indicating that these two variables have a significant impact on customers' loan decisions. In addition, family size, age, and education level are also used as segmentation criteria at multiple nodes, showing their certain importance in customer decision-making. Other factors, such as the first two digits of ZIPCode, whether the customer holds a time deposit account (CD_Account), etc., will also become segmentation criteria in some cases. In particular, customers with incomes above 116.50 and education levels below a bachelor's degree are more inclined to accept loans (222 customers). Overall, income is the most critical factor, and groups with higher incomes are more likely to accept loans. At the same time, the average monthly credit card consumption is also an important indicator. Customers with higher consumption levels are more inclined to accept loans. Factors such as family size, age, education level, etc. have an impact on customers' loan decisions in different situations.

In the decision tree model, the importance of a feature is calculated by the total impairment of the feature to the decision criterion (such as the Gini coefficient) [3]. Through feature importance analysis, it is concluded that education level (especially those below a bachelor's degree) and income are the most critical factors affecting customers' loan decisions. Household size and average credit card spending are also important factors. Other characteristics such as age, term deposit account, and mortgage amount have some, but relatively small, impact on lending decisions. The effects of geographic and behavioral characteristics (e.g. postal code, use of online banking) are relatively weak. Securities accounts, specific postal codes, and whether credit cards from other banks were used had no significant impact on lending decisions.

From the data analysis, it can be seen that education is the most important variable in converting savers into borrowers. We believe there may be multiple reasons behind this. First of all, college students have more entrepreneurial and financial needs than other groups. They hope to achieve more at this stage, so they have greater borrowing needs [4]. Secondly, compared to researchers or professionals, undergraduates receive relatively less support, especially financial support, so they are more likely to borrow bank loans to achieve self-development. Income is the second important factor. It is not difficult to understand that people with lower incomes tend to borrow from banks to improve their quality of life and meet their basic needs [5]. Third, family size is also a key factor. People with a family size of 1 are usually looking for a partner or are in a period of career advancement, so their borrowing needs are higher. Relatively speaking, a family of three usually includes two adults and one child. Such family units have relatively little pursuit of quality of life and have a smaller financial burden of raising children, so their loan needs are relatively low.

It is recommended that banks pay more attention to college students and low- and middle-income people when promoting loan programs, because these groups are more likely to successfully apply for loans and bring greater profits to the banks.

4.2. Limitations and recommendations

The decision tree model provides valuable insights for the marketing team, helping them identify customer characteristics that are highly correlated with loan applications, allowing marketing

campaigns to more accurately target potential loan customers. At the same time, the model also reveals areas that may require further research or strategy adjustments, such as optimizing loan product design for low-income or customers with specific educational backgrounds. To further improve the predictive performance of the model, it is possible to consider introducing ensemble learning methods such as random forests, gradient-boosted trees (GBM), XGBoost, etc. [6]. These methods generally perform better and are more robust than single decision tree models. In addition, it is possible to explore ways to combine multiple models, such as combining decision trees with other algorithms (such as logistic regression and support vector machines) and improving prediction accuracy through voting or stacking methods. Based on the results of the model analysis, the new bank can design more targeted marketing strategies. For example, customized loan products or preferential measures can be provided for customers with higher incomes but who have not yet accepted loans. In addition, special credit products or more attractive loan terms can be developed for young customers and customers with higher education [3]. At the management level, it is recommended that customer service and relationship management be strengthened, especially for potential loan customers, and a deeper understanding of their needs and expectations should be gained through regular communication and customer satisfaction surveys. At the same time, data-driven methods are used to regularly assess customer loyalty and credit risk to optimize pricing strategies and terms of loan products [4,7].

5. Conclusion

This study aimed to identify the key customer attributes influencing the decision to accept personal loans from Neo Bank. By constructing a predictive decision tree model based on data from 5000 customers, we were able to pinpoint the most significant factors that drive this decision. Our analysis revealed that education level, income, and family size are the primary determinants influencing a customer's likelihood of taking out a personal loan.

The model's high accuracy on both training and test data suggests its robustness, although the perfect scores on training data indicate a potential overfitting issue. Adjustments such as regularization or using ensemble methods could enhance the model's generalizability.

From a practical standpoint, Neo Bank can leverage these insights to refine its marketing strategies. Targeting undergraduates and low-to-middle-income customers with tailored loan products could increase loan uptake. Moreover, enhancing customer relationship management and conducting regular satisfaction surveys can help in understanding and meeting customer needs better.

Overall, this study provides a valuable framework for Neo Bank to identify and convert depositors into personal loan customers effectively, thereby boosting its loan portfolio and earnings. Further optimization of the model and strategic adjustments based on continuous data evaluation will be crucial for sustained success.

Acknowledgement

Jiayi Xiao, Jinghao Gao, and Junlong Yan contributed equally to this work and should be considered co-first authors.

References

- [1] Wang, Yuzhen; Qin, Jingqiao, Analysis of financial product purchases based on logistic regression, Journal of physics. Conference series, 2021-04, Vol.1848 (1), p.12164

- [2] Guo, Weidong; Zhou, Zach Zhizhong, A comparative study of combining tree-based feature selection methods and classifiers in personal loan default prediction, *Journal of forecasting*, 2022-09, Vol.41 (6), p.1248-1313
- [3] Yang S B, Chen T L. Uncertain decision tree for bank marketing classification [J]. *Journal of Computational and Applied Mathematics*, 2020, 371: 112710.
- [4] Safarkhani F, Moro S. Improving the accuracy of predicting bank depositor's behavior using a decision tree [J]. *Applied Sciences*, 2021, 11(19): 9016.
- [5] Amin R K, Sibaroni Y. Implementation of decision tree using C4. 5 algorithm in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region) [C]//2015 3rd international conference on information and communication technology (ICoICT). IEEE, 2015: 75-80.
- [6] Islam M R, Habib M A. A data mining approach to predict prospective business sectors for lending in retail banking using decision tree [J]. *arXiv preprint arXiv: 1504.02018*, 2015.
- [7] Sivasree M S, Sunny T R. Loan credibility prediction system based on decision tree algorithm [J]. *Int. J. Eng. Res. Technol*, 2015, 4(9): 825.