ST3189 MACHINE LEARNING

Tee Hwei Xin
220453998

Table of Contents

1. Introduction

In the banking and finance industry, banks must be able to acquire and retain customers to stay competitive. To accomplish this, banks need to develop marketing strategies to promote their products and services to the right customers. Understanding their existing and potential customer behaviours aids in cross-selling their products and services by offering better terms and conditions that meet their needs.

This research uses a dataset from a United States bank called Thera Bank. We aim to predict credit card spending so that we can discover the factors that drive it and identify the best model that can predict credit card spending accurately. Thera Bank would be able to offer credit limit increases or decreases based on spending predictions to meet customer needs as well as to expand its promotions and rewards deals, such as additional interest with minimum spending, to increase customer satisfaction and loyalty.

In addition, we aim to determine the attributes that increase the likelihood of a customer taking up a personal loan and develop the best models to predict if the customer will take up a personal loan based on their demographics. This enables banks to create target customer profiles and use personal loans as a gateway product for upselling or cross-selling other financial services like credit cards and mortgage loans. Lastly, we intend to use unsupervised learning techniques such as K-means clustering to identify customer segments, allowing Thera to design suitable marketing strategies for them while discovering possible fraud detection through unusual credit card spending.

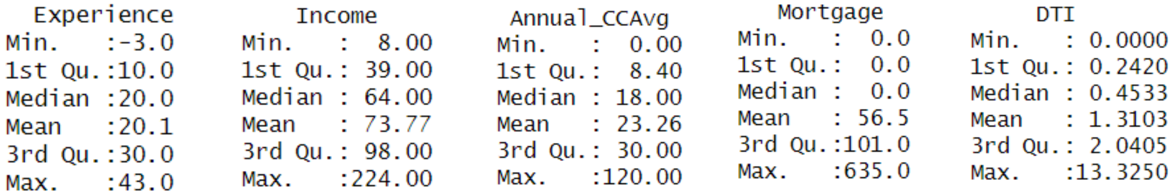
This dataset consists of 5,000 customers, and the variables, like Debt-to-Income ratio and Annual Credit Card Spending, were calculated and included. The final dataset contains the following variables:

ID: A unique identifier for each customer.
Age: Age of the customer (in years).
Experience: Years of professional experience.
Income: Annual income of the customer (in $1000s).
ZIP Code: ZIP code of the customer's residential area.
Family: Number of family members.
Education: Level of education (1: Undergraduate, 2: Graduate, 3: Advanced/Professional).
Mortgage: Value of the mortgage on the customer's house (in $1000s).
Personal Loan: Indicates whether the customer has accepted a personal loan offered by the bank (1: Yes, 0: No).
Securities Account: Indicates if the customer has a securities account with the bank (1: Yes, 0: No).
CD Account: Indicates if the customer has a certificate of deposit (CD) account (1: Yes, 0: No).
Online: Indicates if the customer uses online banking services (1: Yes, 0: No).
CreditCard: Indicates if the customer has a credit card issued by the bank (1: Yes, 0: No).
Annual_CCAvg: Average annual spending on credit cards (in $1000s).
Debt-to-income ratio(DTI): Amount of debt as compared to income (in percentage point)

2. Findings

2.1 Exploratory Data Analysis (EDA)

The Experience column contained negative values, so any values less than zero were removed. The debt-to-income ratio was shown as “DTI”, which could be useful for assessing the likelihood of customers accepting personal loans and predicting credit card usage. The summary statistic table shows that some variables have extremely high maximum values, indicating the presence of outliers. However, we do not remove the data point as it is not a result of data error and represents actual customer profiles in which such circumstances can happen in the real world.

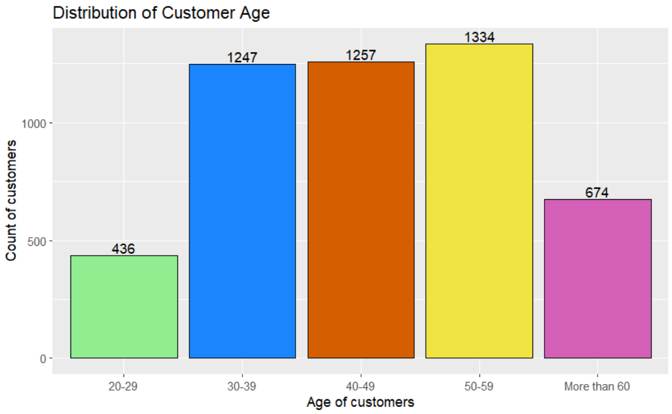


```
   Experience          Income          Annual_CCAvg          Mortgage              DTI
 Min.    :-3.0    Min.   :  8.00    Min.   :  0.00    Min.   :  0.0    Min.   : 0.0000
 1st Qu.:10.0    1st Qu.: 39.00    1st Qu.:  8.40    1st Qu.:  0.0    1st Qu.: 0.2420
 Median :20.0    Median : 64.00    Median : 18.00    Median :  0.0    Median : 0.4533
 Mean   :20.1    Mean   : 73.77    Mean   : 23.26    Mean   : 56.5    Mean   : 1.3103
 3rd Qu.:30.0    3rd Qu.: 98.00    3rd Qu.: 30.00    3rd Qu.:101.0    3rd Qu.: 2.0405
 Max.   :43.0    Max.   :224.00    Max.   :120.00    Max.   :635.0    Max.   :13.3250
```

Figure 1: Summary Statistics Table of Continuous Variable

Most of the customers were aged between 30 and 60, and we noticed that undergraduates have the highest mean income when compared to other education levels. This demonstrates that a higher education level does not necessarily result in higher income. At least 70% of the customers do not own a credit card issued by the bank, and 90% of the customers do not have personal loans. This shows that customers have low loyalty towards Thera, which could be because Thera is new, and customers have little trust in the bank. Hence, it is crucial for Thera Bank to come up with strategies to acquire and sustain customer loyalty for its growth in this competitive industry.

The correlation matrix shows that Credit Card Spending increases with higher Income. In addition, variables like Work Experience, education level and number of family members have no obvious linear relationship with credit card spending. Furthermore, people who take up personal loans have higher Credit Card Spending and Income than people who do not. There were more CD account holders than non-CD account holders accepting a personal loan. In contrast, there is no obvious relationship between Personal Loan and Mortgage, nor between Personal Loan and DTI.



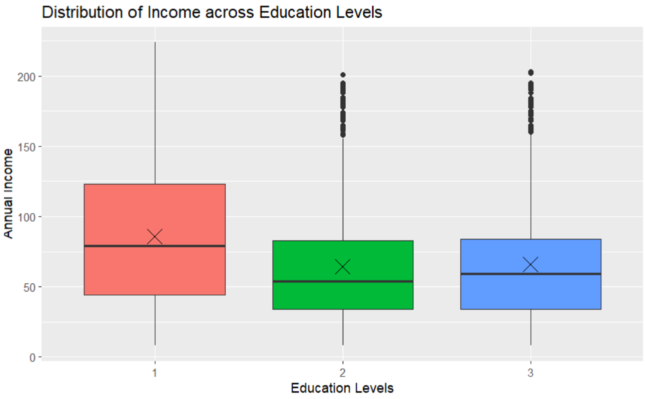


Figure 2: Age Distribution of Customers

Figure 3: Income Distribution across Education Levels

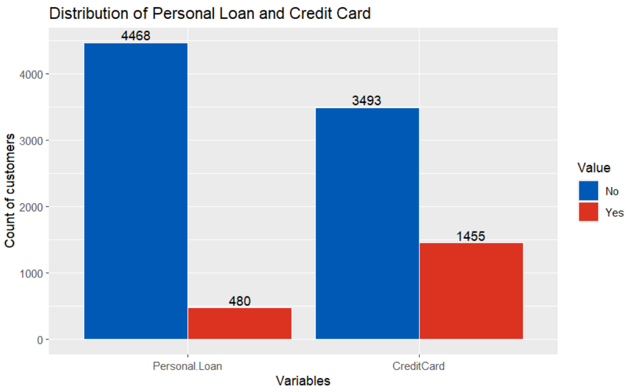


Figure 4: Bar plot of Personal Loan and Credit Card Ownership

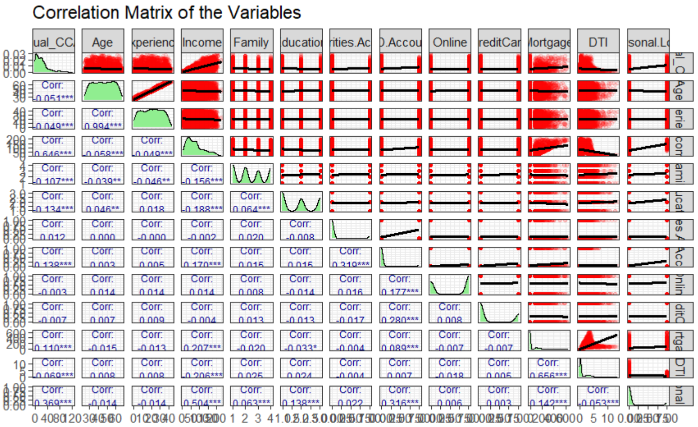
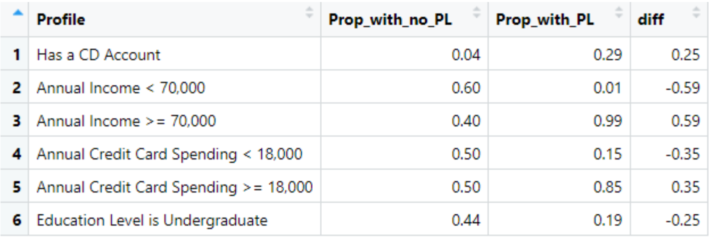


Figure 5: Correlation Matrix of Variables

Furthermore, the difference of 0.25 shows that customers with a CD account are more common among those with a personal loan. The table also shows that customers with an income of more than $70,000 are more likely to accept the personal loan, accounting for 99% of the loan acceptance. Moreover, higher annual credit card spending of $18,000 and more is associated with accepting a personal loan, as they account for 85% of the loan acceptance. Lastly, customers with an education level of undergraduate are less likely to accept the loan.



	Profile	Prop_with_no_PL	Prop_with_PL	diff
1	Has a CD Account	0.04	0.29	0.25
2	Annual Income < 70,000	0.60	0.01	-0.59
3	Annual Income >= 70,000	0.40	0.99	0.59
4	Annual Credit Card Spending < 18,000	0.50	0.15	-0.35
5	Annual Credit Card Spending >= 18,000	0.50	0.85	0.35
6	Education Level is Undergraduate	0.44	0.19	-0.25

Figure 6: Proportion of Customers with and without Loan based on Demographics

2.2 Regression

Linear Regression

The backward elimination method was applied to remove insignificant variables, as stepwise selection is more computationally efficient than the best subset selection method and considers the combined effect of all variables, unlike forward selection. In the GVIF detection, Age and Experience have extremely high GVIFs of 9.557 and 9.546 respectively, indicating severe multicollinearity. This makes it harder to separate their individual effects, and we cannot interpret the coefficients the usual way by comparing them to a unit change in credit card spending. We could still use these variables as they are significant to the model, as evident by their low P-value. Diagnostic plots were performed and there is an upward slope in the Scale-Location plot, which signifies that the errors are dependent on X variables and have no constant variance. The interpretation of the coefficients of X variables may not be that accurate due to heteroscedasticity issues. The final model is $\text{Annual CCAvg} = -17.922 + 0.715(\text{Age}) - 0.744(\text{Experience}) + 0.322(\text{Income}) - 0.971(\text{Education2}) - 1.055\ (\text{Education3}) - 0.033(\text{Mortgage}) + 3.460(\text{CD.Account1}) - 0.782(\text{Online1}) - 0.772(\text{CreditCard1}) + 2.177(\text{DTI})$. Income was found to be the most important variable in predicting credit card spending, followed by DTI and the remaining variables.

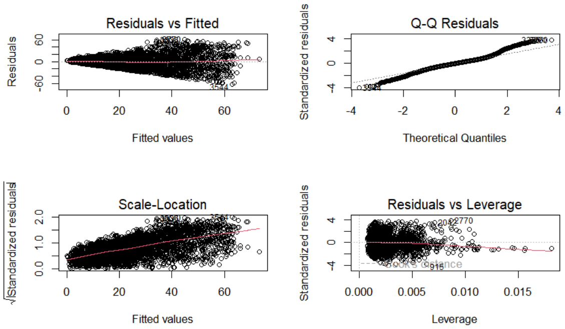


Figure 7: Diagnostic Plot of the Linear Model

Lasso Regression

Lasso regression was chosen as regularisation may help to deal with the multicollinearity and overfitting issues in linear regression. It is also easier to interpret and help with feature selection by shrinking the coefficient to 0, unlike Ridge Regression. The trainset was transformed with categorical variables encoded with 0 and 1, as the package does not handle categorical variables. A cross-validation was applied to find the optimal lambda, which was 0.003377. In the final model, not all variables, like 4 family members, are included as insignificant variables will have a higher penalty, which results in the coefficient possibly being zero. Hence, the final model is $\text{Annual CCAvg} = -12.914 + 0.520(\text{Age}) - 0.548(\text{Experience}) + 0.32(\text{Income}) + 0.339(\text{Family2}) - 0.789(\text{Family3}) - 0.76(\text{Education2}) - 0.858(\text{Education3}) - 0.0326(\text{Mortgage}) + 0.0642(\text{Securities.Account1}) + 3.452(\text{CD.Account1}) - 0.749(\text{Online1}) - 0.761(\text{CreditCard1}) + 2.132(\text{DTI})$. Unlike Linear Regression, CD Account is the most important variable, followed by DTI and the remaining variables.

CART

This model was chosen as it is more robust to the heteroscedasticity issue and handles multicollinearity better than linear regression. A cross-validation of different trees was conducted to find the benchmark for optimal cost penalty using Leo Breiman's one standard error approach, which was 0.24028. Thus, the maximum tree was pruned by a cost penalty of 0.0009086402, located between the 45th and 46th tree. Similar to linear regression, Income is the most important variable in predicting credit card spending, followed by DTI and other variables.

Random Forest

Similar to using CART, we use Random Forest to determine if using multiple CART trees performs better. A hyperparameter tuning was conducted to find that the optimal number of trees is 500. The number of features is 3 as int(M/3) is calculated, which is the default setting in the randomForest package. Similar to other models, Income is the most important variable in predicting credit card spending, followed by DTI and the remaining variables.

Models Evaluation

Linear Regression, CART and Random Forest have a similar conclusion, which is that Income is the most important variable, followed by DTI. However, CD Account is the most important variable, followed by DTI for Lasso Regression. What was common is that DTI has been ranked at least the second most important variable in predicting credit card spending. Thus, this variable should be included in the model prediction. In addition, there is no difference in prediction error between Linear Regression and Lasso Regression as the Root Mean Square Error(RMSE) is similar. Among all the 4 models, Random Forest is the best as it has the lowest error, while the other two models have more than 40% of its error.

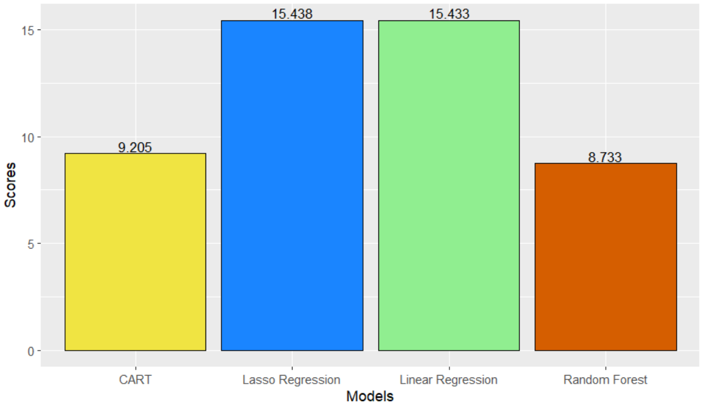


Figure 8: Bar Chart of RMSE across Models

2.3 Classification

Logistic Regression

The best combination of X variables was selected based on P-value. As a result, age, Experience, ZIP Code, Mortgage and DTI were removed. The GVIF values of all X variables were lower than 2, showing the absence of multicollinearity. When looking at the odds ratios, we saw that customers who have an education level that is Graduate and Advanced/Professional are more likely to accept a personal loan by 50.502 and 60.304 times respectively, as compared to customers who are undergraduates. Moreover, customers who own a CD Account are likely to accept a personal loan by 41.642 times than those who do not. The final model is $\widehat{Personal\ Loan} = -12.700 + 0.063(Income) - 0.182(Family2) + 1.972(Family3) + 1.642(Family4) + 3.922(Education2) + 4.099(Education3) - 0.865(Securities.Account1) + 3.729(CD.Account1) - 0.712(Online1) - 0.959(CreditCard1) + 0.0149(Annual\_CCAvg)$. When fitting the model using the train set data, the P-value for variables like Family2 is much higher due to data sacrificed for the test set, leading to less confidence in results. After we generate the probability of each customer in the test set, we set the threshold to 0.097 instead of 0.5, as the proportion of customers accepting a personal loan is 0.097 based on the whole dataset, and false negatives are more costly than false positives. We found that the Income and Education level of Advanced/Professional are important variables in determining if a customer will accept a personal loan.

K-Nearest Neighbours (KNN)

This model was used to see if the performance is better with non-linear boundaries and distance-based learning. The trainset was transformed with categorical variables encoded with 0 and 1 as KNN calculate the distance to determine the majority class. The model included weighted KNN as there is an imbalance in classes. A 10-fold cross-validation and grid search were used to find the optimal k value. It was found that the optimal value is 4. As the model identifies the neighbours using Euclidean distance, scaling of the data is required before we train the model to ensure all variables are on the same scale. Also, we found that Income, followed by Annual Credit Card Spending, are important variables in identifying customers who accept a personal loan.

CART

This model was used to see if the performance is better with rule-based learning, unlike KNN, which is distance-based. Unlike logistic regression, it was also robust to outliers, which are common in this data set. Heavier weight was assigned to the minority class as there was a 9:1 class imbalance. There was cross-validation of different trees to find the benchmark for optimal cost penalty using the 1 standard error method by Leo Breiman, which helps to prevent overfitting. The optimal cost penalty is 0.01424447, located between the 3rd and 4th tree. Similarly to KNN, Income and Annual Credit Card Spending were the top two most important variables. Among the customers with an annual income of $93,000 and more, customers with an education level of undergraduate and more than 2 family members have a 95% chance of accepting the loan, while family members of 1 to 2 have only 7%. This shows that the number of family members decides whether a customer will accept the loan among undergraduates with annual income of $93,000 and more. Lastly, customers with an annual income of less than $93,000 and credit card spending of $35,000 will have no chance of accepting the loan.

Random Forest

This model was used to see if ensemble learning could improve the prediction performance with the influence of more trees instead of a single decision tree. A hyperparameter tuning was conducted. The optimal number of trees was 200, and the number of features was 3 as calculated using int(sqrt(M)). Income is the most important variable, followed by Undergraduates in Education and the rest.

Models Evaluation

Among all the models, CART has the lowest false negative rate. If Thera Bank is low in budget and needs to identify as many actual customers as possible who would actually accept the loan, CART would be the best model. However, Random Forest is the best model in terms of overall performance as it has the lowest false positive rate and highest F1, accuracy as well as precision scores. By identifying actual customers who would not accept personal loan, Thera bank can immediately eliminate them as part of the marketing strategy and better utilise the money for attracting the actual customers who would take up the loan. Thera bank can consider using this model if they have a higher budget to hire bank representatives to call customers and send personalised offers through emails to compensate for the difference in false negative rate between CART and Random Forest.

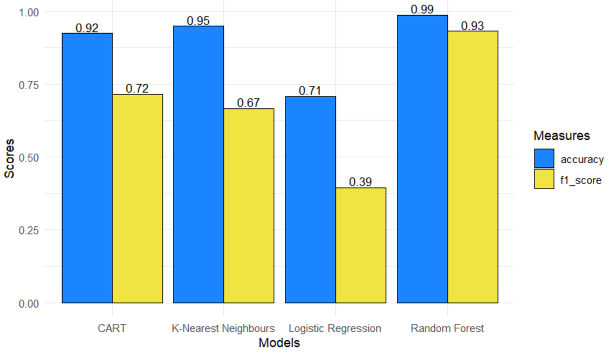


Figure 9: Accuracy and F1-Score of Models

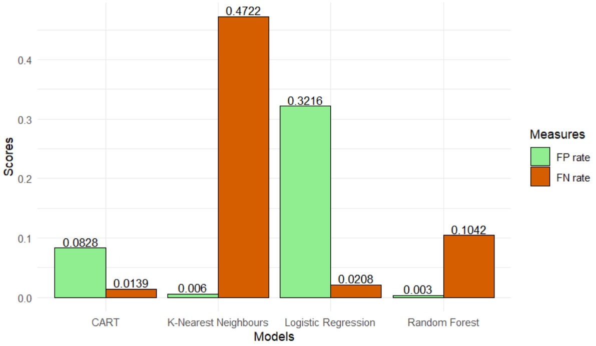


Figure 10: Recall and Precision of Models

2.4 Unsupervised learning

K-means clustering was deployed to find the customer segments. The optimal number of clusters, k, can be found using the elbow and silhouette methods in which one identify by the lowest decrease rate of within-cluster sum of square and another identify by how well the data points fit into its clusters. The optimal k was 4 and sizes of each cluster were 1636, 1598, 1054 and 660 respectively.

Interpretation of Clusters

Age, Income, Working Experience, Education Level

In the case of a variable having many outliers, the median will be used to measure the comparison instead of the mean, as it is not influenced by outliers. Cluster 1 appears to be customers aged more than 40 years old, cluster 2 is aged 20 to 49 years old, and cluster 3 and 4 are more likely to be middle-aged of 30-59. As compared to other clusters, cluster 4 tends to be high monthly income earners with a median of at least 12,000, whereas the other have a similar monthly income of 5,000 range. Cluster 1 has the most working experience with an average of 31 years, which is triple the average years of experience for Cluster 2. Clusters 3 and 4 have similar working experience of at least 18 years. The customers in Cluster 4 are likely to be undergraduates, while other clusters have a mix of education levels.

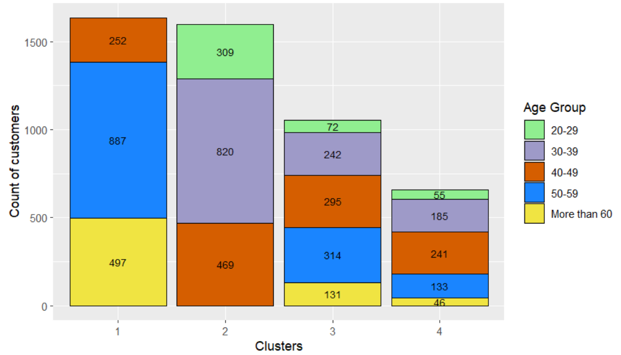


Figure 11: Age Distribution across Clusters

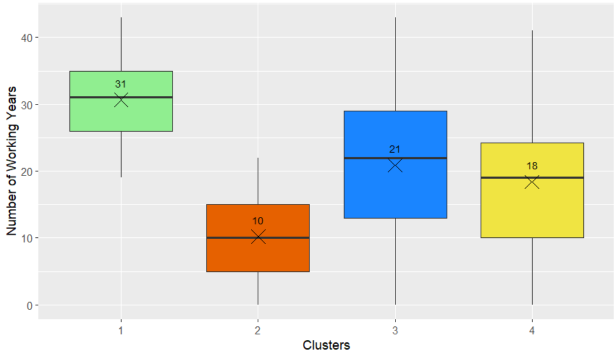


Figure 12: Years of Work Experience across Clusters

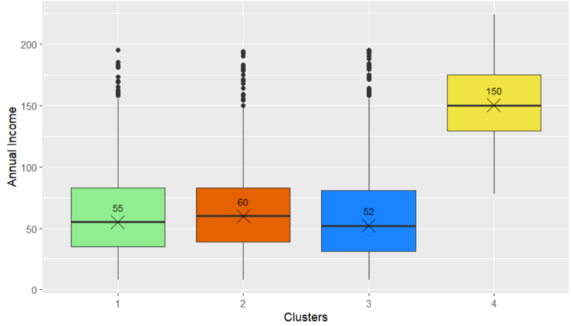


Figure 13: Income Distribution across Clusters

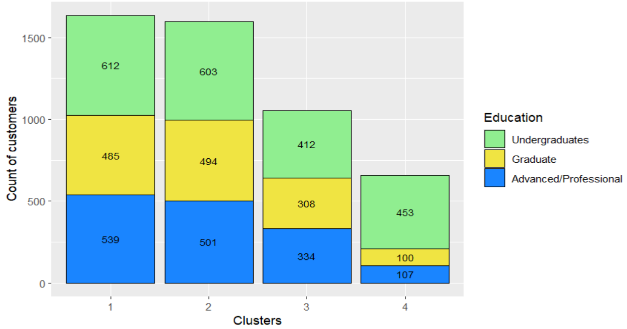


Figure 14: Education Level across Clusters

Family Size, Mortgage, Personal Loan and Credit Card Ownership

Cluster 4 seems to have a smaller family size of not more than 2, while other clusters have a balanced amount of family size groups. In addition, cluster 3 has an extremely high median mortgage of 166,000, whereas other clusters mostly have zero mortgages. Furthermore, cluster 4 has a higher number of credit card ownerships as compared to other clusters.

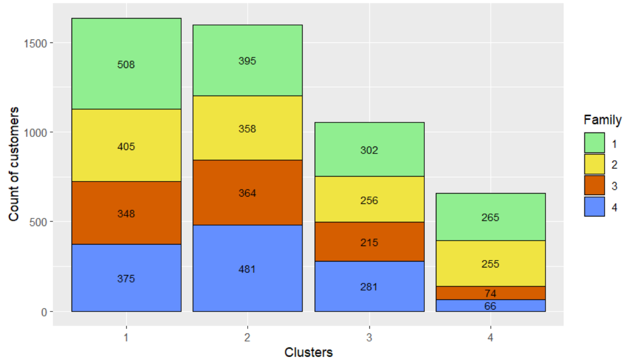


Figure 15: Age Distribution across Clusters

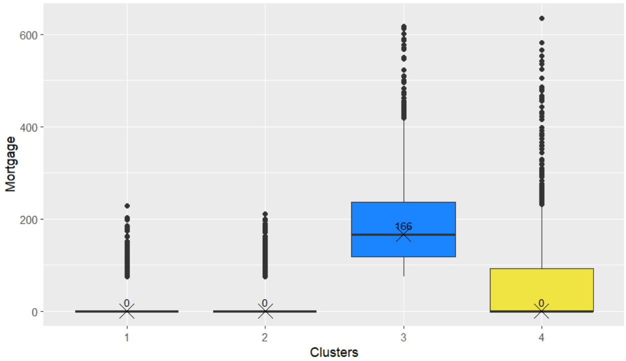


Figure 16: Mortgage across Clusters

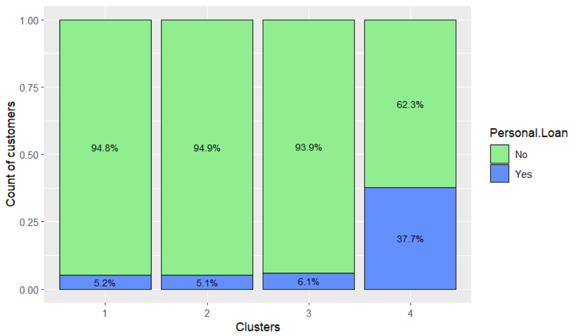


Figure 17: Personal loan across Clusters

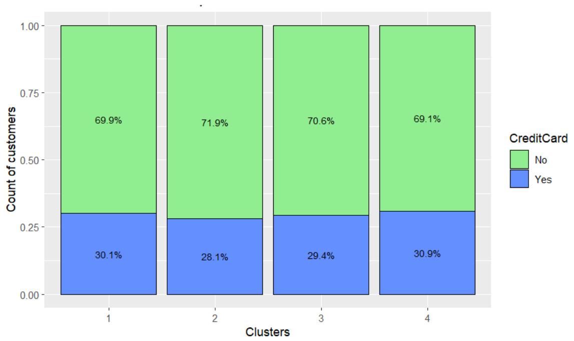


Figure 18: Credit Card across Clusters

Card Spending, DTI, CD Account and Securities Account

It was found that less than half of the customers had Thera bank credit cards across all clusters, and cluster 4 has the highest mean credit card spending of roughly 5,000 as compared to other clusters of 1,300 to 1,400. Moreover, cluster 3 has the highest mean DTI of 4.14 as compared to other clusters, and more than 50% of the customers in each cluster do not own a securities account. For CD accounts, cluster 4 has the highest number of CD Account holders, yet less than half of them do not possess one.

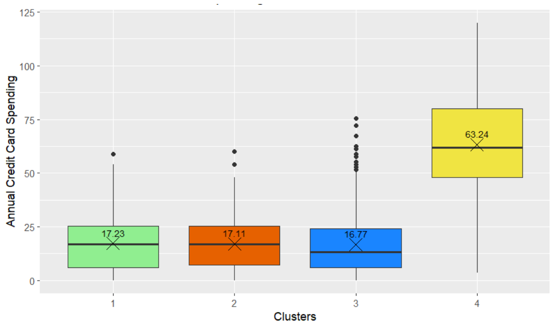


Figure 19: Card Spending across Clusters

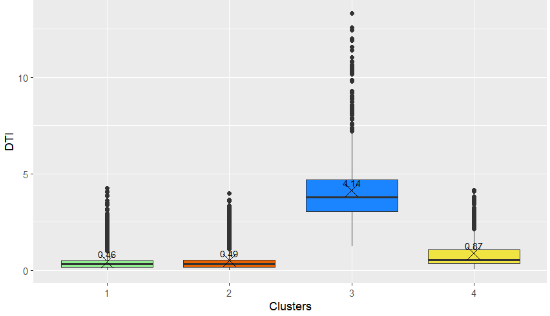


Figure 20: DTI across Clusters

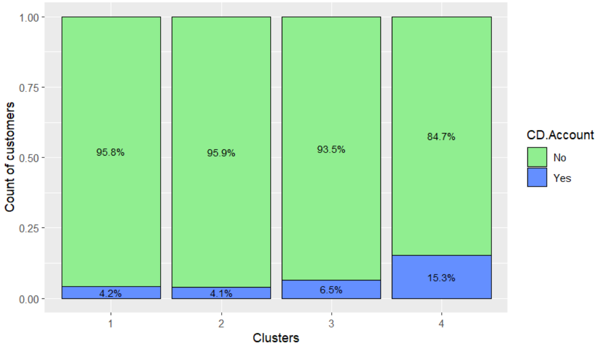


Figure 21: CD Account across Clusters

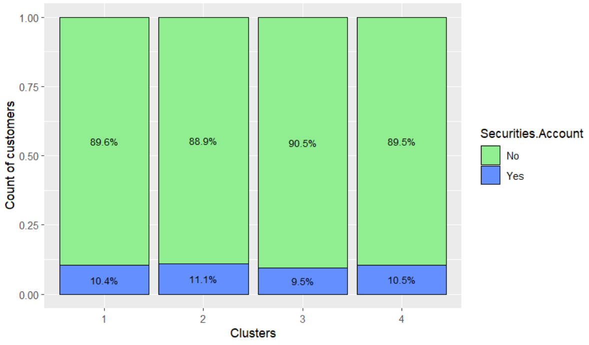


Figure 22: Securities Account across Clusters

Customer Segment in each Cluster

Cluster 1 customers tend to be older than 40 years old with much working experience and a stable income of $5,000. Cluster 2 customers tend to be young adults as they have the least working experience of 10 years. Cluster 3 is most likely to be customers who have high debts due to high DTI. Cluster 4 tends to be higher-income earners who are aged between 30 to 49 years old, and their highest education level is undergraduate. They have a small family size of not more than 2 family members, and they are more likely to take up personal loan than other clusters.

3. Literature Review

A study by Wang and Xie (2023) found that customers with higher incomes and undergraduate qualifications are more likely to take up personal loans. They also suggest that higher mortgage debt would increase the likelihood of personal loan uptake (Wang & Xie, 2023). My findings agreed to a certain extent that higher income would increase the probability of personal loan acceptance, as seen in the Logistic Regression model, where every $1000 increase in income can increase the likelihood by 0.063, holding other variables constant.

However, my findings disagreed with the statement that higher mortgage increases the likelihood of personal loan acceptance, as Mortgage was an insignificant variable in the Logistic Regression. In addition, Education is indeed a key factor in explaining the change in personal loan acceptance. However, it may be higher education, like Graduate and Advanced/Professional, that increases the likelihood of personal loan acceptance.

In terms of regression, a study found that customers who are 61 and older would have higher credit card spending as compared to those who are younger. In addition, higher income and marital status affect spending as higher income provides higher financial freedom and married couples spend more than customers who are single (Teoh et al., 2013). Credit card spending was also found to be positively correlated to any loans (Lin et al., 2019).

My findings agreed with these findings that older age and higher income generate higher credit card spending, as shown in the Linear Regression model. Moreover, all models conclude that DTI is an important variable in credit card spending prediction. This agrees with the study that any loans can increase credit card spending, as higher loans will incur higher DTI. In contrast, research found that older age does not have a higher credit card spending (Lin et al., 2019). The unsupervised learning agreed with this statement, as cluster 1 tends to be older than 40 years old and yet they are not the highest credit card spenders like cluster 4, which has a younger age of less than 40. Thus, more data is needed to validate these two statements.

4. Limitations and Conclusion

In conclusion, Random Forest is best for regression as it has the lowest RMSE. For classification, Random Forest and CART can be the best models, depending on Thera Bank’s budget, as there is a trade-off between the two models. According to our findings, we should include DTI and income in the future prediction model for credit card spending because they are important variables, and three out of four models produce the same conclusions. Moreover, we can consider including Income and Annual Credit Card Spending for the future personal loan prediction model, as all models agree that Income is the most important variable, and Annual Credit Card Spending was ranked second across the two models. Since Education has no common conclusion on which level is most important, the future model can include Education and use more data to validate the significance of each level.

Our regression and classification predictions are not as accurate as they should be because they may be missing some other important features, such as the bank interest rate, household income, financial literacy score and region of residence, all of which can affect the acceptance of a personal loan and credit card spending. Moreover, the DTI can include other debts, like a car loan, to better reflect the true DTI value. With more accurate and relevant data, we can improve our prediction, allowing the bank to easily and precisely discover new customers. This captures the opportunity for Thera Bank to increase profitability and customer bases by implementing appropriate targeted marketing efforts.

In unsupervised learning, we identified that cluster 4 is Thera Bank's valuable customer as since cluster 4 has the highest average income and number of personal loans, with less than 50% of them taking up personal loans. Thus, Thera Bank can create marketing strategies to promote personal loans to cluster 4 customers. With cluster 4 having the highest number of CD account holders, it may suggest that they are risk-averse. Hence, Thera Bank can offer guaranteed terms for personal loans, such as providing an option to choose loan repayment periods to ease their uncertainties in taking up personal loans, as well as attractive monthly interest for minimum spending at Thera Bank’s partner stores.

Furthermore, cluster 4 has the highest monthly credit card spending of at least 5,000 and yet less than half of them have credit cards issued by the bank. These suggest that Thera Bank can create a bundle package in which minimum credit card spending allows the personal loan to have lower bank interest charges, encouraging Cluster 4 customers to sign up for credit cards and use them for their daily activities. Finally, it is worthwhile investigating the attitude of customers towards Thera Bank, as the low rate of securities account sign-ups with the bank may signify that customers do not have much confidence in the bank to manage their assets. This could help the bank to develop a plan to improve its brand image and strategic position as a small player in the banking industry.

5. References

Azam, R., Muhammad, D., & Akbar, S. S. (2012, November 2). The significance of socioeconomic factors on personal loan decision a study of consumer banking local private banks in Pakistan. *RePEc: Research Papers in Economics*.

Lin, L.Q., Revindo, M. D. R., Gan, C., & Cohen, D. A. (2019). Determinants of credit card spending and debt of Chinese consumers. *International Journal of Bank Marketing*, 37(2), 545–564. doi: 10.1108/IJBM-01-2018-0010

Teoh, W. M. Y., Chong, S. S., & Yong, S. M. (2013). Exploring the factors influencing credit card spending behavior among Malaysians. *International Journal of Bank Marketing*, 31(6), 481–500. doi: 10.1108/IJBM-04-2013-0037

Wang, H. Y., & Xie, Z. Y. (2023). Analyze the influence of various factors on personal loans. *Advances in Economics, Management and Political Sciences*, 6(1), 448–459. doi: 10.54254/2754-1169/6/20220186