|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| Data Science Project  DSCI-490-001 Submitted to Prof. Alan Rice, BBA, MA |
| |  |  |  | | --- | --- | --- | | Submitted by: | Sai Charan Guttikonda | Sai Nitin Bhadriraju | |

Table of Contents

[List of figures 1](#_Toc68855847)

[Introduction 1](#_Toc68855848)

[Background 1](#_Toc68855849)

[About the Company: 1](#_Toc68855850)

[Research objectives 2](#_Toc68855851)

[Methodology 2](#_Toc68855852)

[Population 2](#_Toc68855853)

[Sample 2](#_Toc68855854)

[Project Methodology 3](#_Toc68855855)

[Data exploration: 3](#_Toc68855856)

[Data wrangling: 3](#_Toc68855857)

[Data modelling: 3](#_Toc68855858)

[Data visualization: 3](#_Toc68855859)

[Methods used for data modeling 3](#_Toc68855860)

[Random forest: 3](#_Toc68855861)

[Kmeans clustering: 3](#_Toc68855862)

[Exploratory Data Analysis 4](#_Toc68855863)

[Demographics 4](#_Toc68855864)

[Age and Gender 4](#_Toc68855865)

[Cities 4](#_Toc68855866)

[Engagement 5](#_Toc68855867)

[Distribution of loan balance and deposit balance 7](#_Toc68855868)

[Results 8](#_Toc68855869)

[Research objective 1 8](#_Toc68855870)

[Research objective 2 8](#_Toc68855871)

[Research objective 3 10](#_Toc68855872)

[Data Modelling 10](#_Toc68855873)

[Optimizations of hyperparameters 10](#_Toc68855874)

[Wealth Management Model 10](#_Toc68855875)

[Mortgage Model 10](#_Toc68855876)

[Credit Card model 10](#_Toc68855877)

[Research objective 4 11](#_Toc68855878)

[Cluster 1: 11](#_Toc68855879)

[Cluster 2: 11](#_Toc68855880)

[Cluster 3: 12](#_Toc68855881)

[Cluster 4: 12](#_Toc68855882)

[Recommendations 12](#_Toc68855883)

[Limitations 13](#_Toc68855884)

[Annexure 14](#_Toc68855885)

[Technical 14](#_Toc68855886)

[Random forest model 16](#_Toc68855887)

[Kmeans model 18](#_Toc68855888)

[References 22](#_Toc68855889)

# List of figures

[Figure 1: Age and Gender 5](#_Toc68567346)

[Figure 2: City count 6](#_Toc68567347)

[Figure 3: Age group engagement 7](#_Toc68567348)

[Figure 4: Correlation matrix 9](#_Toc68567349)

[Figure 5: Deposit balance 10](about:blank#_Toc68567350)

[Figure 6: Loan balance 10](about:blank#_Toc68567351)

[Figure 7: Loan and deposit balance comparison 11](#_Toc68567352)

[Figure 8: Wealth Management and Mortgage 12](#_Toc68567353)

Figure 9: Credit card…..…………………………………………………………………………13

Figure10: Clusters………………………………………………………………………………. 14

# Introduction

# 

Background

The prominence of data analytics in the financial services market is growing at a rapid pace. According to industry reports, big data analytics in the financial sector is expected to grow with a CAGR of 22.97% for the period of 2021-2026. It is expected to reach approximately USD 62.1 billion by 2025 (mordorintelligence, 2020). One of the key factors contributing to the adaptation of analytics is the exponential growth of the digital footprint generated by the customers. The traditional business intelligence tools and reporting techniques are not efficient in analyzing larger volumes of data. Credit Unions (CU) collects and stores every transactional data related to members. CUs can leverage the technology to understand business trends, customer behavior and execute data-driven decisions. As of August 2019, the total number of CUs operating in BC is approximately 41 and manages around CAD 64 billion in assets (CPABC, 2019).

About the Company:

Prospera Credit Union (Prospera) is a British Columbia-based organization that offers various financial services, including personal banking, business banking, and wealth management. The company’s personal services include banking, investing, and borrowing. Through personal banking, Prospera offers daily and savings accounts, debit and credit cards, and additional services, including safety deposit boxes, foreign currency, and money and wire transfers. Its personal investment services include term deposits, mutual funds, online investments, registered saving plans, personal insurance, and advisory (Prospera, 2021).

Prospera’s business services include business borrowing, accounts, and cash management, business insurance and wealth management, and business tools and resources. Business borrowing solutions include the line of credit, term loans, commercial mortgages, overdraft protection, small business improvement loans, and credit cards. The accounts and cash management services include daily and savings accounts, payables, and receivables. The company’s various business solutions include personalized business plans, cash flow projections, and personal net worth statements (Prospera, 2021).

Besides, the company offers various online banking services including mobile banking, Interac e-transfer, prosperity tracker, and payment services via Apple pay, Google Pay, and Samsung pay. Prospera’s interest income in 2019 accounted for $132.4 million which increased by 8% from $122.5 million in 2018 (Consolidated Financial Statement, 2019).

In January 2020, Prospera merged with Westminster Savings Credit Union, a full-service financial institution offering personal financial services products and services. The company operates a network of 29 branches, online and mobile banking, the exchange ATM network, and a contact center. The company serves approximately 120,000 members and manages around $9 billion worth of assets (futurestrong.ca, 2020).

Research objectives

**RO1:** Evaluate freely available external data sources (online) through secondary research, which could add more value to study the company’s customer base.

**RO2:** Analyze customers' behavior associated with various services offered by the company.

**RO3:** Explore the degree to which a customer could subscribe to one or more company’s services.

**RO4:** Segment prospective customers to upsell Prospera’s products and services.

# Methodology

Population

The population of interest is Prospera’s customer base.

Sample

Prospera data: The data set contains Prospera’s customer information with a sample size of 1000 and 35 variables. The data set is limited to British Columbia and dated to the year 2016.

Additional data: The additional data was downloaded from the Statistics Canada website. The data set contains city, postal code, and income information for the year 2016. The data set is limited to British Columbia.

Project Methodology

The data analysis is implemented in the following four phases using the tools, including R Studio and Excel:

### Data exploration:

Explore the data set provided by Prospera, and additional data gathered to find patterns and profile their customers.

### Data wrangling:

Cleaned the raw dataset, identified gaps in the data, and applied appropriate methods to resolve the issues. The cleaned dataset was used for the analysis.

### Data modeling:

Various statistical and nonstatistical techniques were implemented to understand different features and patterns in the dataset. Prediction and classification models were developed according to the research objectives.

### Data visualization:

Various charts, graphs, and tables were used to graphically represent the analysis. Important trends and patterns in the data were recognized and presented through the data visualization.

Methods used for data modeling

### Random forest:

To address the third research objective, we have performed Random forest classification. A random forest is an ensemble learning technique for classification and regression problems. The random forest consists of many individual decision trees that operate as an ensemble. Each tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction (Yiu, 2019).

### Kmeans clustering:

To address the fourth research objective, we have performed Kmeans clustering. Kmeans is an unsupervised learning algorithm, which tries to partition the dataset into ‘K’ pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. The algorithm tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different as possible. Kmeans assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid is at the minimum. The less variation we have within clusters, the more homogeneous the data points are within the same cluster (Dabbura, 2018).

# Exploratory Data Analysis

Demographics

### Age and Gender

Figure 1: Age and Gender

The age is categorized into four groups namely Young Adult (0-35), Adult (36-50), Senior (51-60), and Super Senior (60+). Seniors are the largest age group accounting for 37.3% of the data, followed by Adult (27.7%), Young Adult (18%), and Super Senior (17%). Male members in the data accounted for 51% and female accounted for 49%. The adult age category has a slightly more number of female customers when compared to the other categories. A chi-square test was performed to identify the difference in the proportion of gender in different age groups. According to the results, we have identified a statistically significant difference between the gender and the age groups.

### Cities

The data set contains a total of 28 cities. The maximum number of customers are from Kamloops and Kelowna accounting for 33% and 32%, respectively. West Kelowna, Lake Country, and Penticton have a customer base of 6%, 6%, and 5%, respectively. The remaining cities are categorized into Other, which holds 18% of the customer base. The below pie chart illustrates the city proportion.

Figure 2: City

Engagement

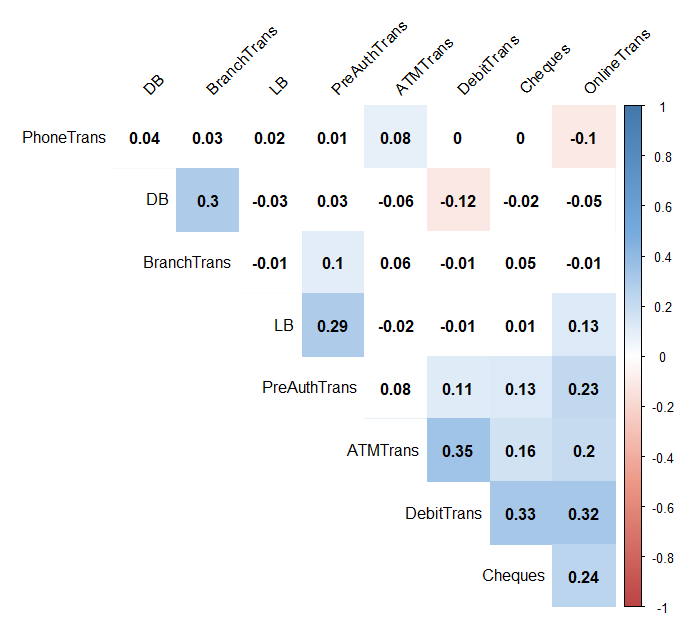
A new variable was created to assess the engagement of Prospera’s clients. The variables considered for engagement include the number of ATM transactions, branch transactions, cheques, online transactions, phone transactions, debit transactions, and preauthorized transactions. The most engaged age groups are seniors and adults with 33% and 31%, respectively of the total engagement. Young adults have an engagement rate of 23% and the least engaged group is super seniors with 14%. Figure 3 illustrates the engagement percentage for different age categories.

Figure 3: Engagement by age group

The most engaged cities are Kamloops and Kelowna with an engagement rate of 36% and 30%, respectively, followed by Lake Country with 6%, West Kelowna and Penticton with 4% each. The remaining 23 cities which were grouped into Other accounted for 21%. Table 1 illustrates the total engagement and the percentage by city.

Figure 4: Engagement by city

Figure 4 emphasizes the correlation between various types of transactions carried out by the customers. It provides an insight to identify the dependence between different types of transactions, including the number of online, debit cards, ATM, cheques, branch, and phone. It also includes the correlation between deposit and loan balance with other variables. There is no significant strong association between different types of transactions. However, there is a weak correlation between debit card and branch transactions with correlation coefficient of 0.30, followed by ; loan balance and preauthorized transactions (0.29); ATM and debit transactions (0.35); debit and cheque transactions (0.33); and online and debit transactions (0.32).



**Correlation coefficient**

**ranges from -1 to 1**

Figure 5: Correlation matrix

Distribution of loan balance and deposit balance

**Chart

Description automatically generated**

Figure 6: Deposit balance

Figure 7: Loan balance

**Chart

Description automatically generated**

Figure 7: Loan balance

The distribution of the deposit balance is right skewed. Over 57% of the customers maintain deposit balance of fewer than 5 thousand dollars. However, around 2% of the customers have over 188 thousand dollars as balance. As the data is extremely skewed, the median will be the optimal measure to represent the central tendency. The median deposit balance for the given sample dataset is 2,952 dollars. The distribution of loan balance is also right skewed with 16.7% of the customers having less than 5 thousand dollars as loan balance and around 2.9% of customers have over 300 thousand dollars. The median loan balance is 13,170 and it is 4.46 times the deposit balance.

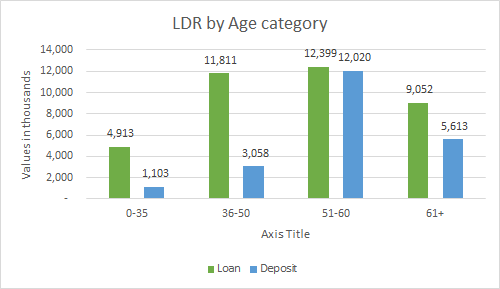


Figure 8: Loan and deposit balance comparison

The loan to deposit ratio (LDR) assesses the liquidity of the organization. LDR is calculated by the ratio of loans and deposits. There is a high LDR variance between different age groups. Young adults and adults have high LDR with 4.45 times and 3.86 times, followed by super seniors (1.61 times) and seniors (1.03 times). The overall loan balance is 1.75 times the deposit balance.

# Results

Research objective 1

The objective was to evaluate freely available external data sources (online) through secondary research, which could add more value to study the company’s customer base. Additional data has been collected to further understand the characteristics of demographics in the data. The additional data sources were evaluated based on four criteria including credibility, reliability, relevance, and ease of access (data format). Statistics Canada was selected as the optimal source for secondary data because it is managed by the Canadian government, the information is reliable and can be extracted in xlsx format.

Information was extracted from the 2016 Census based on the postal codes provided in the original data set. The information extracted from census data is based on cities and includes median income, average income, median age, average age, average population, and median population. Census information was cleaned, formatted, and merged with the primary dataset based on the postal code.

Research objective 2

The variables considered here are the customers who have mortgage, credit card, payroll, and wealth management. The percentage of the total number of customers who opted for these services includes payroll (50.1%), credit card (46.8%), wealth management (33.3%), and mortgage (20.4%). The most preferred services in the data set are payroll and credit card. The least preferred service is the mortgage.

Table 1 below, illustrates that seniors opted for the greatest percentage of services across all the service categories. Super seniors opted the least for payroll and mortgage services. Furthermore, young adults and adults are least engaged with wealth management and mortgage.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age Category | Payroll | Credit Card | WM | Mortgage |  |  |  |
| Young Adult | 60.00% | 75.60% | 23.30% | 16.70% |  |  | Low% |
| Adult | 56.30% | 41.20% | 28.50% | 21.30% |  |  | High% |
| Senior | 50.10% | 40.50% | 35.10% | 20.10% |  |  |  |
| Super Senior | 29.40% | 40.00% | 47.60% | 23.50% |  |  |  |

Table 1: Proportion within the age groups for various products

Even though senior customers have the highest number of subscriptions across the services, proportionately they took fewer products compared to young adult and adult. Only 40.5% of seniors have credit card and about 20% of them opted for the mortgage. In contrast, 75.6% of the young adults have credit card and 60% has payroll. Besides, 56.3% of the adults have payroll and 41.2% has credit card. Most of the super seniors opted for wealth management services with 47.6%.

The proportion of customers having wealth management is largely from Kelowna (34%) and Kamloops (29%), followed by West Kelowna (6%), Penticton (6%), Osoyoos (5%), and Other (21%). The number of customers having mortgage is majorly from Kelowna (37%) and Kamloops (24%), followed by Lake Country (10%), West Kelowna (6%), Vernon (6%), and Other (17%). The below-given figure 9 depicts the customer count for the top 5 cities and others.

Figure 9: Wealth Management and Mortgage

The number of customers having credit card largely came from Kelowna (33%) and Kamloops (28%), followed by West Kelowna (8%), Lake Country (6%), Penticton (5%), and Other (19%). In our exploratory data analysis, we have observed that the most engaged cities are Kamloops (36%) and Kelowna (30%). The below figure illustrates the number of customers having credit cards from the top 5 cities and others.

Figure 10: Credit card

Research objective 3

### Data Modelling

Random forest (RF) algorithms were implemented to classify whether a customer will opt for a mortgage, credit card, and wealth management services. RF models train multiple decision trees in parallel and provide the classification based on majority rule. The dataset is divided into training and test datasets with the 70-30 rule, that is 70% of the data are used for training the model and 30% of the data is used to evaluate the model. The metrics selected to evaluate the model include sensitivity, specificity, accuracy, precision, and f1score.

### Optimizations of hyperparameters

The K-Fold cross-validation technique was used to increase the efficiency of the model by tunning the parameters including the number of variables considered at each split and the number of trees. If the number of trees was more than the optimum number, then the model could overfit the data and performs poorly when new data is introduced.

### Wealth Management Model

The optimized model to predict if a customer will opt for wealth management service has an accuracy of 82%. The precision and recall values of the model are 0.652 and 0.8681. According to the model, the five most important variables to classify were: number of products, deposit balance, number of online transactions, and tenure group. As per the model, the customers opting for wealth management have an average deposit balance of 47,405 CAD and an income of approximately 31,000 CAD. On average, clients subscribed to four products offered by Prospera.

### Mortgage Model

The stratified bootstrap sampling technique was used to design the model to classify whether a customer would subscribe to mortgage services. The dataset is imbalanced because of the extreme disproportion in the number of data points for each class in the predictor, that is 79.6% of the customers do not have mortgage and only 20.4% have a mortgage. An oversampling technique was implemented to balance the dataset. This technique resamples the smaller proportion class and approximates it to the number of the majority’s class. The dataset used for training the modeling consists of 50.4% of positive data points (has mortgage) and 49.6% of negative (no mortgage). The optimized model consists of 19 variables and has an accuracy of 78%. The recall and precision values of the model 0.28 and 0.38. The five most important variables to classify the customers are the number of products subscribed by a customer, deposit balance, annual income after tax, number of cheques, and online transactions. The clients opting for mortgage have an average annual income of 31,112 CAD and a mean deposit balance of 18,722 CAD. Besides, the mean number of products opted by the customers is five.

### Credit Card model

The final random forest model to predict if a customer would opt-in for a credit card has an accuracy of 73.67%. A total of 19 variables are included in the model and have 0.73 precision and 0.71 recall. The top five important variables to access whether a customer will subscribe to the credit card service or not are number of products, Age category, number of ATM transactions, online banking user, and statement type.

Research objective 4

The fourth research objective is to segment prospective customers to upsell Prospera’s products and services. Kmeans, an unsupervised clustering technique was used to segment customers based on the variables, including deposit balance, loan balance, mortgage, credit card, payroll, wealth management, number of products, and tenure group. The data were normalized to obtain better performing clusters, and four such clusters were recognized.

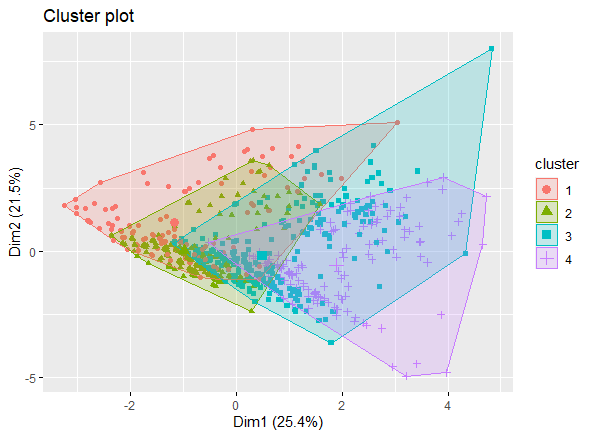


Figure 11: Clusters

### Cluster 1:

The cluster’s traits are:

* The cluster size i.e., the number of customers in this cluster is 138. The median deposit balance is 1,265 dollars, which is the lowest among the clusters. The median loan balance is 25,877 dollars, which is the second-highest among the clusters.
* The cluster has opted for very few services and has minimal engagement with Prospera.
* Adults are the dominating age category in this cluster.

### Cluster 2:

The cluster’s traits are:

* The cluster size i.e., the number of customers in this cluster is 316. The median deposit balance is 1,451 dollars, which is the second-lowest among the clusters. The median loan balance is 5,878 dollars, which is the lowest among the clusters.
* The cluster’s most opted services are wealth management and payroll. The engagement rate is second highest among the clusters with 25% of the overall engagement coming from this cluster.
* Seniors are the dominating age category in this cluster.

### Cluster 3:

The cluster’s traits are:

* The cluster size i.e., the number of customers in this cluster is 400. The median deposit balance and the median loan balance are 3,530 dollars and 11,268 dollars, which is the second-highest among the clusters.
* Proportionately, the cluster has opted for the most number of services and products among the clusters. The engagement is highest among the clusters with 46% of the overall engagement coming from this cluster.
* Seniors are the dominating age category in this cluster.

### Cluster 4:

The cluster’s traits are:

* The cluster size i.e., the number of customers in this cluster is 145. The median deposit balance and the median loan balance are 15,060 dollars and 41,760 dollars, which is the highest among the clusters.
* The cluster’s most opted services are wealth management, mortgage, and credit card. The cluster has opted for the second-highest number of products across clusters with 24%. The engagement of this cluster with Prospera is moderate with 19%.
* Seniors are the dominating age category in this cluster.

# Recommendations

* The wealth management model identified 58 potential customers. Narrowed to 9 customers after consideration of model error. 4 seniors, 2 adults, and 3 young adults from Kelowna, Armstrong, and Chase.

The mortgage model identified 138 potential customers and narrowed it to 6 clients based on key characteristics. They belong to the senior age category and are from Kelowna and Lake Country. 39 possible clients for credit card and narrowed down to 11. The majority fall under young adult and seniors.

* In cluster 1, the loan amount is 20 times more than the deposit balance, which is the highest compared to other clusters. It is important to keep an eye on this cluster for any defaults or delinquencies.
* Cluster 2 has the second-highest percentage of customers who signed up for payroll services, which means there is more number of working-class customers. These customers can be targeted for the credit card as the percentage of customers who opted for the credit card is comparatively less (15%).
* Cluster 3 is the most loyal customer base among the clusters. Customers in this cluster opted for the maximum number of services and the highest percentage of products. These customers can be targeted for wealth management as it the least opted service in the cluster, as well as the deposit balance, is fairly affluent.
* Cluster 4 has affluent customers with the highest deposit balance. Customers in this cluster can be targeted for wealth management.
* According to industry estimates, the Okanagan Valley real estate market is set to increase by 5% in 2021 (Mortgage Sandbox, 2021). Besides, the average sales price in Kelowna for the year 2020 increased 5.6% to 553,175 dollars from 523,832 dollars in 2019 (Jackson, 2020). Target clusters 2 & 3 for mortgage services as the majority of customers from those clusters are from Kelowna and other Okanagan cities.

# 

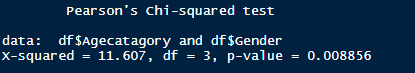
# Limitations

* Due to imbalances in some variables, we tried oversampling and under-sampling techniques. These techniques further increased imbalances in the data. To overcome this, collecting more data points can be helpful to further optimize the models.
* An unsupervised learning algorithm is implemented to cluster the customer into different groups. The number of clusters and the size of each cluster might vary when new data points are introduced into the model.
* The predictive models are trained and tested based on the provided 1000 data points. The dataset was small to split into training, test, and validation sets for model selection. The selected models might underfit or overfit while predicting the classifications.
* The census information extracted through secondary research was from 2016 and it is collected every five years. If Prospera wants to use those variables, it might need to purchase or utilize estimated data.
* The company’s database management system has data inconsistency issues with date format and storage of addresses as a single entry. The inconsistency could lead to staff productivity losses and an inability to analyze the data. Prospera should implement a central semantic store approach that focuses on accurately logging and storing the rules used by the database integration process in a single centralized repository.

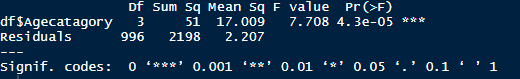
# Annexure

Technical analysis

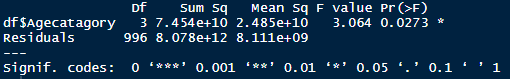
Pearson's Chi-squared test between age category and gender

`

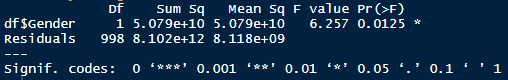
ANOVA test between the number of products and age category.



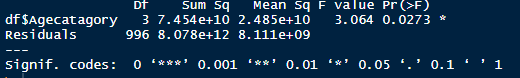
ANOVA test between loan balance and age category.



ANOVA test between the loan balance and gender.



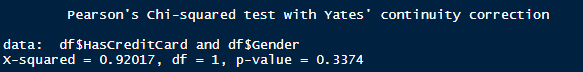
ANOVA test between deposit balance and age category.



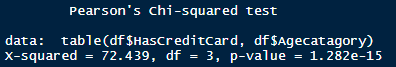
ANOVA test between deposit balance and gender



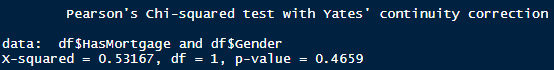
Pearson's Chi-squared test between has a credit card and gender.



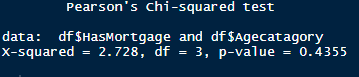
Pearson's Chi-squared test between has credit card and age category.



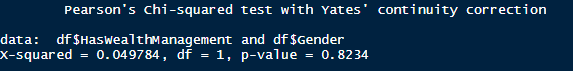
Pearson's Chi-squared test between has mortgage and gender.



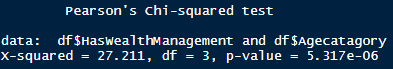
Pearson's Chi-squared test between has mortgage and age category.



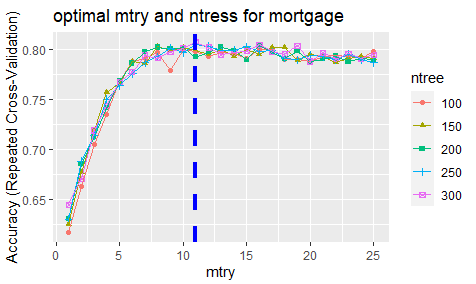
Pearson's Chi-squared test between has wealth management and gender.



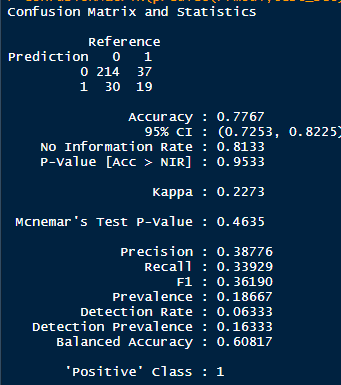
Pearson's Chi-squared test between has wealth management and age category.



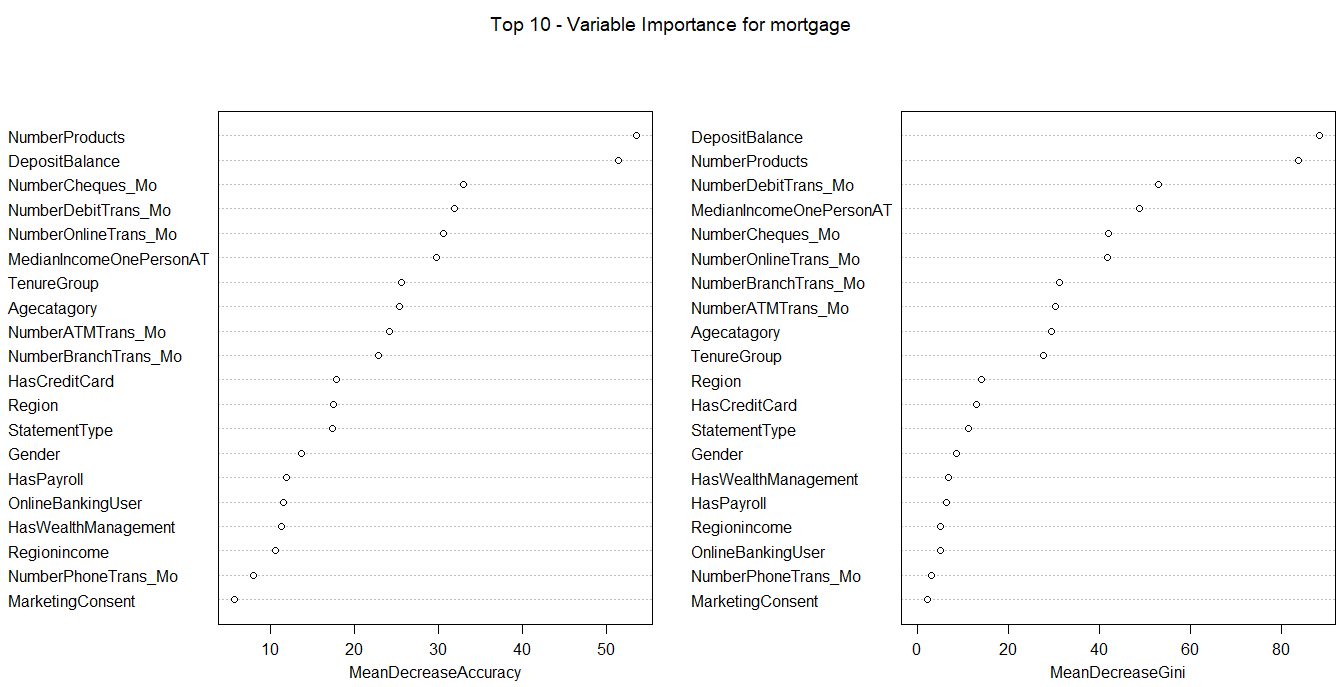
Random forest model

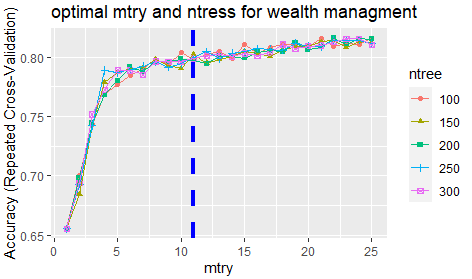


The output of random forest model for mortgage.

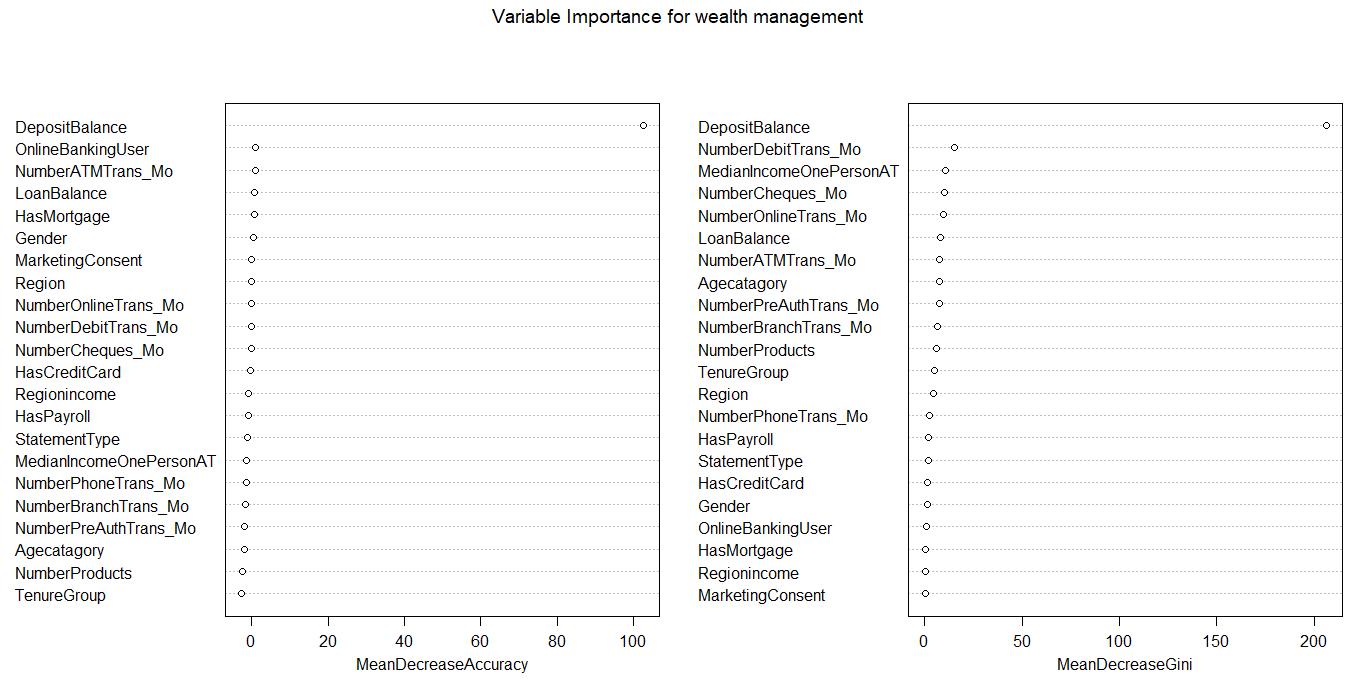


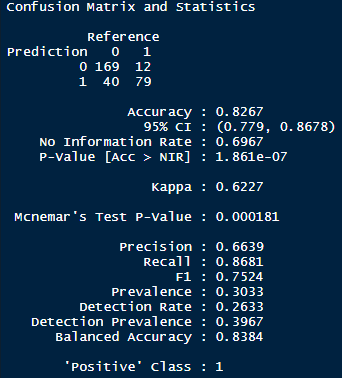
Important variables for mortgage predictive model.





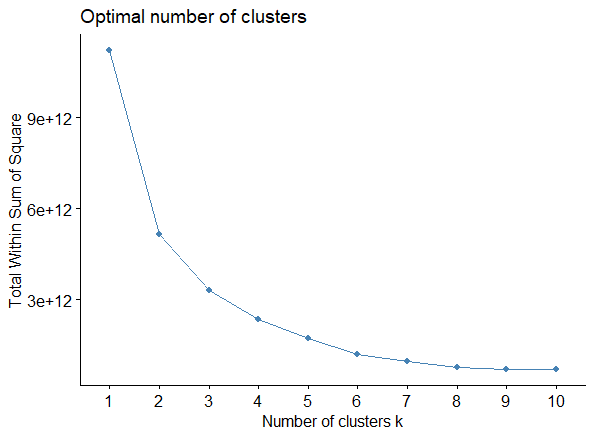
Important variables for mortgage predictive model.



the output of random forest model for wealth management.

Kmeans model

Selecting optimal k using the elbow method with the raw data.



Using the elbow method, we can observe that the optimal k is between 3 and 5. We have performed K-means clustering with 3, 4, and 5 as optimal k. We have obtained higher efficiency with 4 clusters. The below-given picture illustrates the output for K-means with 4 clusters.

K-means clustering with 4 clusters of sizes 114, 801, 42, 42

Cluster means:

DepositBalance LoanBalance HasWealthManagement HasMortgage HasCreditCard HasPayroll NumberProducts

1 13123.47 163978.071 0.3333333 0.84210526 0.5263158 0.5350877 4.649123

2 12515.99 4849.029 0.3046192 0.07615481 0.4581773 0.5056180 3.612984

3 11886.30 354302.220 0.2142857 1.00000000 0.4761905 0.3571429 4.595238

4 232685.82 16925.374 1.0000000 0.11904762 0.5000000 0.4523810 5.309524

TenureGroup

1 6.377193

2 6.347066

3 6.095238

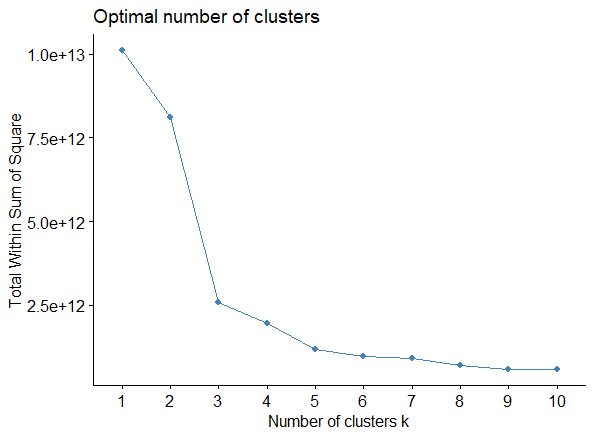
4 6.714286

Within cluster sum of squares by cluster:

[1] 309679946808 531174994103 754790038800 750911623971

(between\_SS / total\_SS = 79.1 %)

We also have observed an outlier. After removing the outlier, the model efficiency increased. The below-given figure was used to select optimal k without the outlier.



Using elbow method, 3,4, and 5 can be optimal k. We have observed a high percentage of efficiency with 4 clusters. The below-given picture illustrates the K-means output with 4 clusters without an outlier.

K-means clustering with 4 clusters of sizes 108, 48, 804, 37

Cluster means:

DepositBalance LoanBalance HasWealthManagement HasMortgage HasCreditCard HasPayroll NumberProducts

1 13705.98 157822.073 0.3333333 0.83333333 0.5277778 0.5277778 4.657407

2 11032.41 323369.007 0.2500000 1.00000000 0.5000000 0.4166667 4.604167

3 13214.98 4717.261 0.3072139 0.07462687 0.4589552 0.5037313 3.613184

4 247080.28 19212.587 1.0000000 0.13513514 0.4594595 0.4594595 5.378378

TenureGroup

1 6.453704

2 5.958333

3 6.349502

4 6.702703

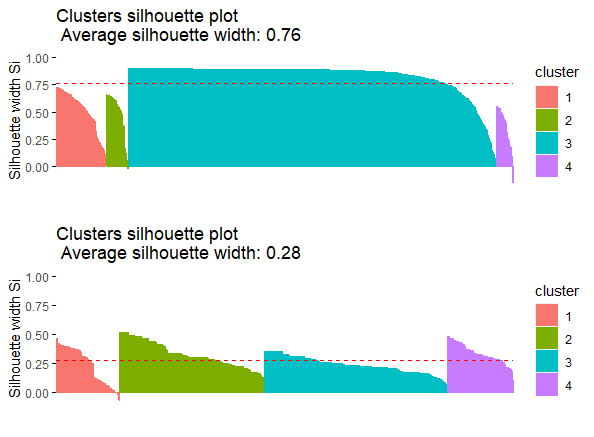
Within cluster sum of squares by cluster:

[1] 262940802388 248835429223 589459261518 684882408719

(between\_SS / total\_SS = 82.3 %)

The model without outlier gave us an efficiency of 82.3%. We further checked the model with scaled data as the magnitude of the variables deposit balance and loan balance is more than the other binary variables. Normalization reduces the variables' weight between 0 and 1, which is aligned with our binary variables between 0 and 1.

After normalizing the data, we have again checked for the optimal k. The number of clusters used in the model is 4 and we have obtained an efficiency of 62.1%. The below-given picture explains the difference between K-means without regularization and with regularization. The model without regularization has negative spillage and we can observe high bias in cluster 3. Whereas the cluster with regularization has minor negative spillage and the clusters are almost distributed normally.



Further oversampling and undersampling has been performed to reduce imbalances in some binary variables. Further normalization techniques are creating more imbalances in the data. Hence, the model with normalized data and 4 clusters has been finalized for the analysis of results.

# References

Dabbura, I. (2018, September 17). *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*. Retrieved from towardsdatascience: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Jackson, K. (2020, December 1st). *Kelowna Housing Market Outlook (2021)*. Retrieved from remax: https://blog.remax.ca/kelowna-housing-market-outlook/

Mortgage Sandbox. (2021, March). *Okanagan real estate forecast*. Retrieved from mortgagesandbox: https://www.mortgagesandbox.com/okanagan-real-estate-forecast#:~:text=Re%2FMAX%20forecasts%20Okanagan%20prices,in%20their%20previous%202021%20forecast).

Yiu, T. (2019, Jun 12). *Understanding Random Forest*. Retrieved from towardsdatascience: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

