**CONTENTS**

[INTRODUCTION 1](#_Toc180775026)

[TASK 1: EXPLANATORY ANALYSIS 1](#_Toc180775027)

[1. RFM score computing 2](#_Toc180775028)

[2. RFM analysis and insights 3](#_Toc180775029)

[2.1. Attritted customers 3](#_Toc180775030)

[2.2. Existing customers 4](#_Toc180775031)

[2.3. High value customers 5](#_Toc180775032)

[3. Interval variables distribution and insights 5](#_Toc180775033)

[4. Nominal variable distribution and insights 8](#_Toc180775034)

[5. Conclusion 10](#_Toc180775035)

[TASK 2: BUILDING PREDICTIVE MODEL 10](#_Toc180775036)

[1. Data preparation 10](#_Toc180775037)

[2. Data partition 11](#_Toc180775038)

[3. Dealing with missing values 11](#_Toc180775039)

[4. Skewness 11](#_Toc180775040)

[5. Dimension reduction 12](#_Toc180775041)

[6. Building models 12](#_Toc180775042)

[7. Result and insights 13](#_Toc180775043)

[7.1. Most important variable 14](#_Toc180775044)

[7.2. Most-likely-to-churn group of customers 15](#_Toc180775045)

[7.3. The most difficult groups of customers to predict their churn outcomes 15](#_Toc180775046)

[7.4. Performance of model compared to random guess 18](#_Toc180775047)

[8. Performance comparison 18](#_Toc180775048)

[8.1. Lift 19](#_Toc180775049)

[8.2. ROC curves: 19](#_Toc180775050)

[8.3. Misclassification rate 20](#_Toc180775051)

[8.4. Precision 20](#_Toc180775052)

[8.5. F1 score 20](#_Toc180775053)

[8.6. Choosing the optimal model 20](#_Toc180775054)

[PART C: RETENTION PROGRAM 21](#_Toc180775055)

[1. Establish detailed and constructed complaint handling procedures: 21](#_Toc180775056)

[2. Customer feedback and survey system 22](#_Toc180775057)

[3. Rewards and Loyalty program 22](#_Toc180775058)

[4. Alternative product recommendation system 23](#_Toc180775059)

[5. Retention -centric marketing 23](#_Toc180775060)

[CONCLUSION 23](#_Toc180775061)

[References 24](#_Toc180775062)

[APPENDIX 25](#_Toc180775063)

**List of figures**

[Figure 1: Top 5 RFM score mostly seen for Attritted group (1) 3](#_Toc180775064)

[Figure 2: Top 5 RFM score mostly seen for Attritted group (2) 4](#_Toc180775065)

[Figure 3: Top 5 RFM score mostly seen for Attritted group (3) 4](#_Toc180775066)

[Figure 4: Top 5 RFM score mostly seen for Existing group (1) 4](#_Toc180775067)

[Figure 5: Top 5 RFM score mostly seen for Existing group (2) 4](#_Toc180775068)

[Figure 6: Top 5 RFM score mostly seen for High value group (1) 5](#_Toc180775069)

[Figure 7: Top 5 RFM score mostly seen for High value group (2) 5](#_Toc180775070)

[Figure 8: Distribution of Satisfaction score 6](#_Toc180775071)

[Figure 9: Distribution of Hours spend on App 6](#_Toc180775072)

[Figure 10: Distribution of Hike in order amount 6](#_Toc180775073)

[Figure 11: Distribution of distance from warehouse to home 7](#_Toc180775074)

[Figure 12: Distribution of average tenure 7](#_Toc180775075)

[Figure 13: Distribution of coupon used 8](#_Toc180775076)

[Figure 14: Distribution of complains 8](#_Toc180775077)

[Figure 15: P-value for Chi-squared test on nominal variables 8](#_Toc180775078)

[Figure 16: City tiers distribution 9](#_Toc180775079)

[Figure 17: Preferred device distribution 9](#_Toc180775080)

[Figure 18: Marital status distribution 9](#_Toc180775081)

[Figure 19: Preferred ordered category distribution 9](#_Toc180775082)

[Figure 20: Preferred payment method distribution 10](#_Toc180775083)

[Figure 21: Data partition node - parameter setup 11](#_Toc180775084)

[Figure 22: Variable transformation node - Function and distribution after transformation 12](#_Toc180775085)

[Figure 23: Cluster node - parameter setup 12](#_Toc180775086)

[Figure 24: Decision tree model - Parameter setup for both models 13](#_Toc180775087)

[Figure 25: Regression model - parameter setup 13](#_Toc180775088)

[Figure 26: Variable significance - DT\_ASE 14](#_Toc180775089)

[Figure 27: Variable significance - DT\_Miscalssification 14](#_Toc180775090)

[Figure 28: Variable significance - Regression model 14](#_Toc180775091)

[Figure 29: Result of first decision tree (DT\_ASE) 15](#_Toc180775092)

[Figure 30: Nodes with highest misclassification rate 16](#_Toc180775093)

[Figure 31: Node with low confidence and high variability 16](#_Toc180775094)

[Figure 32: Nodes contains outliers and be potential rare segments 17](#_Toc180775095)

[Figure 33: Metrics determine model performance compared with random guess 18](#_Toc180775096)

[Figure 34: Model comparisons metric - Lift 19](#_Toc180775097)

[Figure 35: Model comparisons metric - ROC curves 20](#_Toc180775098)

[Figure 36: Model comparisons metric - Misclassification rate 20](#_Toc180775099)

[Figure 37: Model comparisons metric - Precision 20](#_Toc180775100)

[Figure 38: Model comparisons metric - F1 score 20](#_Toc180775101)

[Figure 39: Model comparisons metrics 21](#_Toc180775102)

**List of tables**

[Table 1: Summary of purchasing patterns for three customer segments 5](#_Toc180726809)

[Table 2: RFM score system 6](#_Toc180726810)

[Table 3: Important metrics for three predictive models 17](#_Toc180726811)

[Table 4: Summary table of group of customers that be difficult to predict churn probability 21](#_Toc180726812)

INTRODUCTION

The report below focuses on developing predictive models for churn analysis, aimed at informing necessary retention strategies for a global retail e-commerce company. The dataset encompasses a variety of information, ranging from demographic variables to behavioral characteristics.

To determine the most effective model-building method, a comprehensive explanatory analysis of the variables will be conducted. The primary focus will be on understanding the significance and distribution of each variable in relation to the target variable (churned vs. non-churned customers). Additionally, a new segment will be manually created based on transactional data, using RFM approach. This group is referred to as the High-Value Segment. This group consists of customers who are particularly valuable to the company due to their high profitability; thus, their behaviors may differ from those of other customer segments. Understanding the patterns of this high-value customer segment will contribute to the ultimate goal of the project: customizing retention strategies for high-value customers at risk of churning.

The second part of the task will involve building and evaluating three different predictive models designed to effectively identify customers with a high potential for churn. The optimal model will be determined after examining combination of different critical metrics.

The results from both the initial analysis and the model evaluations will be integrated to develop comprehensive retention strategies for the online retail e-commerce organization. This approach will enable the company to reduce churn risk and enhance customer engagement proactively, even before a churning event occurs.

TASK 1: EXPLANATORY ANALYSIS

Before commencing the detailed analysis and deriving insights, a brief explanation of how the RFM score is computed will be provided. Based on this score, a threshold will be established to identify high-value customers.

The insights will be categorized into three parts: RFM Insights, Interval Variable Insights, and Nominal Variable Insights. This division is necessary due to the different methods used for computing and visualizing the insights across these categories.

Below is a summary table of critial take-away insights concluded from Task 1. Detailed explanations of how each task was executed will be presented individually in the subsequent sections.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Domain** | | **Churned** | **Non-churned** | **High value** |
| **RMF score**  **insights** | Description | High Monetary (4 and 5) scores, indicating high spending before churn. customers ceased purchasing shortly before churning. 70% submitted complaints before attrition | High Recency (5 and 4) and Frequency (4 and 5) scores, indicating frequent purchases in the past but not recent activity. Some customers show overlap with the attritted group (score 142), potentially indicating at-risk customers | High scores in both Frequency and Monetary, indicating consistent spending and purchases. Not many belong to the top 20% for Recency, but they remain engaged. More likely to remain loyal after progressing past the 4-year mark |
| **Interval Variable Insights** | Tenure | Typically churn after 5 years of service | Remain loyal for up to 15 years. | Long-term customers, some with the company for over ~15 years |
| Coupons | Less frequent use compared to high-value customers | Used less frequently than high-value customers but more than attrited customers | Most frequent use of coupons, indicating more purchases and higher spending |
| Complaints | More likely to have filed complaints before leaving, suggesting unresolved dissatisfaction. | Fewer complaints compared to attritted customers, generally satisfied with resolutions | Less likely to file complaints, and if they do, they are typically satisfied with the solutions. |
| **Nominal variables insights** | City Tiers | More evenly distributed across tiers (53% from Tier 1, 48% from Tier 3) | Over 70% located in Tier 1 cities | Over 70% located in Tier 1 cities |
| Preferred Device | No strong preference for a specific device | No strong preference for a specific device | 60% prefer using mobile phones |
| Marital Status | Higher percentage of divorced customers | Mostly married, with fewer divorced and single individuals | Mostly married, with fewer divorced and single individuals |
| Preferred Order Categories | More inclined to purchase mobile phones | Favor Laptops and Accessories | Most inclined to purchase mobile phones and Laptops & Accessories |
| Preferred Payment Methods | Wide range of payment methods used from traditional to digital method | Mostly use credit or debit cards | Mostly use credit or debit cards |

Table 1: Summary of purchasing patterns for three customer segments

1. RFM score computing

RFM analysis focuses on measuring Recency, Frequency, and Monetary Value for each customer, assigning specific scores to determine customer value for customized CRM strategies (Hughes, 2005).

Recency refers to the time elapsed since a customer’s last order. In our dataset, the variable DaySinceLastOrder effectively represents this dimension.

Frequency indicates the total number of purchases made within a specified period. In the dataset, this concept is best captured by the variable OrderCount.

Monetary Value represents the total amount spent by customers over a defined time frame. While the current dataset lacks a variable that directly measures total spending, we can use the cashback variable as a proxy. Cashback rewards are given to customers based on their spending, so higher cashback amounts indicate higher total spending. Thus, for the Monetary Value dimension, the cashback variable will serve as the representative. However, obtaining direct data on the total spending amount would enhance the specificity and accuracy of this measurement.

The scores for each domain in the RFM analysis will be calculated based on their values within specified percentiles. The scoring system is designed as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Percentiles (Top down) | Recency score | Frequency score | Monetary score |
| **>20%** | 1 | 5 | 5 |
| **>40%** | 2 | 4 | 4 |
| **>60%** | 3 | 3 | 3 |
| **>80%** | 4 | 2 | 2 |
| **Remaining** | 5 | 1 | 1 |

Table 2: RFM score system

The top 20% of values for Frequency and Monetary Value will be assigned the highest score of 5. For Recency, the top 20% will receive a score of 1 (indicating a more recent interaction). After each 20% percentile, the scores will decrease by one point.

The final RFM score will be a 3 digits number taken from each domain.

Generally, high value customers are those that usually bring great profit to the company. Combining with RFM concep, we can define high value customers are those that still purchase until recently, with medium to high frequency and spendings above the average spendings from customer base. Assuming that the definitions align with the company’s criteria for identifying high-value customers, the RFM scores that indicate high-value customers will include the following combinations:

* Recency: 1-3 – Customers that recently make an order or purchase, above the average
* Frequency: 3-5 – Customers that make purchases frequently, above the average
* Monetary: 3-5 – Customers spend significantly, above the average
* Churning status: Not churn (0) – Existing customers

Customers within the above combination will be labelled as High value customers.

Please review the full computation from excel file that is attached together with the report.

1. RFM analysis and insights
   1. Attritted customers

The scores that appear most frequently are 125,142,244,245, and 344 with the 125 score group being the most predominant. The majority of scores fall within the 100 and 200 ranges, suggesting that attritted customers likely ceased purchasing products for not long before officially churning

A graph with green and yellow lines

Description automatically generated

Figure 1: Top 5 RFM score mostly seen for Attritted group (1)

Furthermore, the scores frequently observed are predominantly in the 4 range, indicating that these customers were once frequent purchasers and that their churn is unexpected. Notably, their purchasing behavior was characterized by extremely high spending, as evidenced by their high monetary scores (4 and 5). This suggests that these customers had a strong engagement with the brand prior to their churn.

A grid of numbers and letters

Description automatically generated

Figure 2: Top 5 RFM score mostly seen for Attritted group (2)

However, up to 70% of this group most likely submitted complaints before their attrition. It would be beneficial to investigate the specific points of service that led to their dissatisfaction in order to improve customer engagement and retention

A table with numbers and text

Description automatically generated

Figure 3: Top 5 RFM score mostly seen for Attritted group (3)

* 1. Existing customers

The most commonly recorded RFM scores are 142, 225, 443, 543, and 551 with the highest frequency being the 225 and 551 score. Given that we have reclassified a portion of existing customers into high-value customers, this distribution is understandable, particularly with the significantly high scores for Recency (5 and 4).

A graph with green and yellow lines

Description automatically generated

Figure 4: Top 5 RFM score mostly seen for Existing group (1)

These scores indicate that these customers have not made a purchase in quite some time. However, prior to halting their purchasing behavior, it is likely that they were either frequent purchasers (frequency scores around 4 and 5) or spent substantial amounts of money (monetary score around 2-3). This insight suggests that while these customers may currently be inactive, they possess a history of engagement that could be leveraged to encourage reactivation through targeted marketing strategies.

A green and black grid with numbers

Description automatically generated

Figure 5: Top 5 RFM score mostly seen for Existing group (2)

Nevertheless, there are groups of customers that collide with attritted group: 142. It is an alarm that they might potentially become the next group of attritted. It might be worth studying the previous attritional behavior patterns and predict the potential of churning from this group of customers.

* 1. High value customers

The scores most frequently observed are 243, 244, 245, 343, and 344.

A graph with green and yellow bars

Description automatically generated

Figure 6: Top 5 RFM score mostly seen for High value group (1)

These customers have been pre-defined as those who still make purchases regularly, demonstrating a decent amount of spending and high frequency. However, the scores indicate that not many of these high-value customers belong to the top 20% of those who have purchased recently or the top 20% biggest spenders of the company.

A screenshot of a number table

Description automatically generated

Figure 7: Top 5 RFM score mostly seen for High value group (2)

In comparison to attritted profiles, it appears that customers in this segment tend to make their decisions about churning or continuing their engagement with the company relatively early in their customer journey. In fact, attritted customers tend to stay with the company for no more than 4 years. Once they progress past this initial stage, there is a higher likelihood that they will remain engaged and develop into high-value customers.

1. Interval variables distribution and insights

In general, we can identify several similarities in customer behaviors across the three segments:

The ***average satisfaction score*** is approximately 3 to 3.5, with the most frequently observed score being 3. However, attrited customers tend to score 5 on satisfaction more frequently than the other groups. In fact, this is the second highest frequency score for attrited customers, while the second highest score for the other groups is 1. This suggests that attrited customers appear to be significantly more satisfied with the services compared to the other segments.

A graph with red and green lines

Description automatically generated

Figure 8: Distribution of Satisfaction score

All customers spend a considerable amount of time on the shopping app, generally averaging around 2.6 to 3.1 hours. Most customers report spending up to 4 hours on a daily basis.

A graph with green lines and red circle

Description automatically generated

Figure 9: Distribution of Hours spend on App

There is a noticeable increase in purchase frequency, with a clear rise in the total amount of orders—approximately 15% on average—indicating a positive trend in business growth. The distribution of all segments appear to contain less to no outliers. Since all three distribution experience Mean > Median but only with slight difference, the distributions seem to be slightly right-skewed.

A graph showing a number of green squares

Description automatically generated with medium confidence

Figure 10: Distribution of Hike in order amount

The average distance between the warehouse and customers’ homes is around 15 km. However, the boxplot for the attrited group shows a longer range at the upper end than the other two groups, indicating higher variance in distances. In other words, attrited customers tend to live significantly further away from the average distance compared to the other segments. Additionally, Existing and High value customers distribution have significant amount of outliers in the upper range, suggesting some exceptional individuals from these two segments also live quite far away from the warehouse. Lastly, since the Mean slightly higher than the median, we can assume that the distribution shape is slightly right skewed as well.

A diagram of a distribution of a number of boxes

Description automatically generated with medium confidence

Figure 11: Distribution of distance from warehouse to home

The descriptive statistics suggests that satisfaction score, total spending time on app, distance or total order amount might not be the most significant variable in distinguishing the segments.

On the other hand, there are clear differences among the customer groups:

**Tenure**: Attritted group tends to leave the company after approximately **5 years** of service. Nonetheless, the group experience a great number of outliers in the upper range, making the mean located outside the IQR, suggeseting that there are individuals whose attritted after long-term relationship with the company. The distribution of the attritted group thus is signficantly right-skewed. In contrast, existing and high-value customers have demonstrated loyalty, with some remaining with the company for up to **15 years**. The general distribution of this group does not contain any outliers, indicating a consistent behaviors of engagement time for existing and high value group.

A diagram of a distribution of average

Description automatically generated

Figure 12: Distribution of average tenure

**Coupons:** All three segments exhibit a significantly high number of outliers in the upper range, indicating the presence of individuals who utilize more coupons than average. However, in comparison to the attrited and high-value segments, the existing group shows a wider distribution at the lower end, with the mean being slightly higher than the median. This distribution shape suggests that most existing customers do not utilize coupons as frequently as the average customer base.

A graph with green and black lines

Description automatically generated

Figure 13: Distribution of coupon used

**Complaints and Dissatisfaction**: Compared to existing customers (both high-value and non-high-value), the attritted group is more likely to have previously lodged complaints. This may indicate that they are attempting to express their dissatisfaction, and their decision to leave the company could stem from unresolved issues. In contrast, existing customers generally do not experience discomfort with the company’s services or find the solutions to their complaints to be reasonable.

A graph showing a number of green and gray bars

Description automatically generated

Figure 14: Distribution of complains

1. Nominal variable distribution and insights

To identify which nominal variables is significance in distinguishing the segments, chi-square test have been applied on all nominal variables (including Preferred login device, Preferred ordered category, Gender, City tiers and Preferred payment method) against the target variables Segments. The p-value[[1]](#footnote-1) obtained from Chi-squared test indicates that, apart from Gender, all other variables contribute significantly into defining the features of each segments.

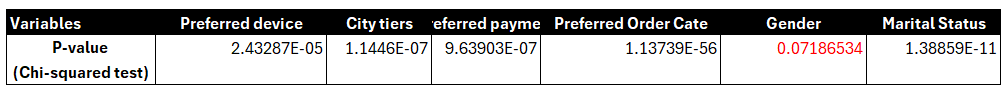


Figure 15: P-value for Chi-squared test on nominal variables

**City Tiers:** existing and high-value customers are predominantly located in Tier 1 cities (over 70% of the population), whereas are attritted customers more evenly distributed, with approximately 53% originating from Tier 1 and 48% from Tier 3 cities.

A green circle with a white triangle

Description automatically generated

Figure 16: City tiers distribution

**Preferred Device:** Both existing and attritted customers show equal interest in all three types of devices. However, 60% of high-value customers typically prefer to use mobile phones.

A green and grey pie chart

Description automatically generated

Figure 17: Preferred device distribution

**Marital Status:** Generally, the customer base consists mostly of married individuals. However, attritted customers have a higher percentage of divorce rates compared to high-value and existing customers, who demonstrate lower divorce percentages.

A green pie chart with a white text

Description automatically generated

Figure 18: Marital status distribution

**Preferred Order Categories:** While Laptops and Accessories are generally favored by most customers, the attritted group shows a greater inclination to purchase mobile phones. Mobile phones are also the most popular product among high-value customers.

A table with numbers and text

Description automatically generated

Figure 19: Preferred ordered category distribution

**Preferred Payment Methods:** Attritted customers exhibit a wide range of choices for payment methods. In contrast, existing and high-value customers predominantly use credit or debit cards, with other payment methods being utilized infrequently, except in a few exceptional cases.

A table with numbers and text

Description automatically generated

Figure 20: Preferred payment method distribution

1. Conclusion

In conclusion, the high-value and existing customer groups exhibit minimal differences in their purchasing behaviors. However, the initial observations of variable distributions between existing and attrited customers show significant variation.

To simplify the modeling process, we will construct a predictive model using a binary target variable that distinguishes between churned and non-churned customers. The characteristics of high-value customers will be leveraged when suggesting retention strategies in Task 3.

Additionally, while building the model, several variables exhibiting skewness will require transformation, such as Distance from Warehouse to Home and Tenure. The rationale for these transformations and the specific processes involved will be detailed in the following sections.

TASK 2: BUILDING PREDICTIVE MODEL

The main objective of building the predictive model, according to the assigned business question, is to predict churn probability. We will proceed to build three different predictive models. The first model we will develop is the **Decision Tree**, which is the compulsory model for this analysis. However, constructing a decision tree requires establishing a suitable assessment method to determine the termination points for the tree branches.

In SAS Enterprise Miner, several built-in assessment methods are available by default, with two primary options being **Average Squared Error** and **Misclassification Rate**. Hence, we will built **two different decision trees** that incorporate two different assessment method accordingly. The selelections will assist to determine the most appropriate construction for decision tree model accordingly. The third model to be build will be **logistic regression model**.

1. Data preparation

All three models are supervised learning model, hence, target variable is required to be selected. Churn indicator as binary variable (1 and 0) will be assigned as target variable in this report to answer the question of predicting the churning probability.

Since the dataset is input directly after performing RFM analysis from task 1, several variables are computed such as Segment column. This column is directly correlated with Churn indicator due to its computation method and nature, which will create data leakage should the variable is set as explanatory input. Hence, to ensure the integrity and avoid overfitting issue, Segment variable will be rejected from the data.

Another rejected column will be RowID which does not contribute meaningful insight into the model. The unique identify of each observations can be determined using CustomerID instead. The variable, as a consequence, will be assign as ID at measurement level. The remaining indenpendence variable will be treated as input variable.

1. Data partition

The data will be divided into two different set: Training and Validation. While the training dataset is used for building the model, validation set will be used to validate and assess the performance of build model on unseen data. In this study, the proportion of training and validation is 7:3. The same will be setup in parameter of *Data partition* node

A screenshot of a computer

Description automatically generated

Figure 21: Data partition node - parameter setup

1. Dealing with missing values

Preliminary data exploration from task 1 advises that the dataset currently does not have any missing values. Hence, we will not proceed with solution to deal with missing values.

1. Skewness

Among the interval variables, four exhibit an extreme right-skewed distribution as studied in Task 1. To enhance model performance and stabilize estimation, it is essential to address outliers carefully, as the presence of skewness often indicates the existence of outliers. Furthermore, all three models—particularly the logistic regression model—require the initial assumption of normality, meaning that the data should ideally follow a normal distribution.

To address the issue of skewness, the *Transform Variable* node has been added to apply appropriate transformations. However, of the four variables, only one—Tenure—showed significant improvement in its distribution following transformation. The distributions of the remaining variables did not change substantially after applying transformations.

Consequently, to avoid unnecessary complications, only Tenure will be transformed using the square root method, while the other variables will remain unchanged. This approach allows for a more effective modeling process while maintaining simplicity in the analysis.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Figure 22: Variable transformation node - Function and distribution after transformation

1. Dimension reduction

To avoid the multicolinearity issue and decrease the complexity level of the model, it is essential to proceed with dimension reduction during data preparation stage. The process will look into the correlation among the explanatory variables and proceed to eliminate those that have high correlation accordingly.

In SAS miner, the dimension reduction can be performed by applying *Variable Clustering* node. To make sure the node successfully clustering variable with high correlation together, two critical setup is required:

* Clustering source: Correlation
* Includes class variables: Yes:
* Variable selection: Best variables

A screenshot of a computer

Description automatically generated

Figure 23: Cluster node - parameter setup

1. Building models

As stated previously, three different models will be built accordingly

* Decision tree using Average Squared Error as Assessment method
* Decision tree using Misclassification rate as Assessment method
* Logistic Regression

For the first two decision tree, parameter setup will be similar, except for assessment method setup:

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Figure 24: Decision tree model - Parameter setup for both models

For Logistic Regression, parameter adjustment includes choosing Stepwise as selection model. This will allow the system to automate the process of dining the best subset of predictors by simultaneously adding and removing the explanatory variables until no variables meet the criteria for entering or leaving the model

A screenshot of a computer

Description automatically generated

Figure 25: Regression model - parameter setup

1. Result and insights

The assignment question requires dividing the results and insights into two parts, beginning with an analysis of the first decision tree as a foundational comparison standard before evaluating the remaining two models.

To enhance the effectiveness of comparing the performance of all three models and to ensure clarity and cohesiveness in the evaluation, the following section will integrate the performance metrics for all three models simultaneously. While addressing the required questions for the first decision tree, the comparison process will be conducted in relevant areas. This approach enables a more robust understanding of how each model performs relative to the others.

Importance metrics of three models can be viewed from the summary table as below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | DT\_ASE | DT\_Misclassification | | Logistic Regression | |
| Most important variable | Complain  CashbackAmount  PreferedOrderCat\_  Laptop&Accessory  OrderAmountHikeFromLastYear  MaritalStatusSingle  SatisfactionScore | Complain  CashbackAmount  PreferedOrderCat\_  Laptop&Accessory | Complain  Preferred PaymentMethodMode\_EWallet  MaritalStatusSingle  PreferedOrderCat\_  Laptop&Accessory  GenderMale  SatisfactionScore  OrderCount  CashbackAmount | |
| Lift | 3.12 | 3.36 | 3.21 | |
| ROC curves | 0.76 | 0.708 | 0.792 | |
| Misclassification rate | 0.1531 | 0.1531 | 0.1611 | |
| Average Squared Error | 0.1131 | 0.1192 | 0.1150 | |
| Precision | 0.4031 | 0.4031 | 0.2326 | |
| Recall | 0.5778 | 0.5778 | 0.5769 | |
| F1 score | 0.4749 | 0.4749 | 0.3315 | |

Table 3: Important metrics for three predictive models

* 1. Most important variable

Both decision tree models identify similar important variables, including Complaint, Cashback Amount, and the Preferred Category (Laptop and Accessory). However, the second decision tree, which utilizes the Misclassification Rate as its assessment method, has a smaller number of leaves and terminates at an earlier stage. Consequently, the total number of identified important variables in this model is significantly less than that in the first decision tree.

A close-up of a computer screen

Description automatically generated

Figure 26: Variable significance - DT\_ASE

A close-up of a computer

Description automatically generated

Figure 27: Variable significance - DT\_Miscalssification

In contrast, the regression model retains most of the important variables identified in the first decision tree. Additionally, it successfully identifies two more variables: Gender (Male) and Payment Method (E-wallet). The small p-values (<.005) associated with these variables indicate their significance in predicting churn probability. Furthermore, the high coefficients suggest a meaningful linear contribution to the likelihood of churning, which may not be as evident in the decision tree model.

A screenshot of a computer

Description automatically generated

Figure 28: Variable significance - Regression model

While there are slight differences between the regression and decision tree models, this does not imply inconsistency between them. The decision tree model focuses on variables that produce strong splits early in the tree, while the regression model captures the linear relationships between explanatory and response variables by assigning values to them. Thus, even though Gender and Payment Method may not lead to splits in the early stages of the decision tree, they have cumulative effects that allow them to contribute meaningfully to the logistic regression.

Ultimately, the fact that the regression model captures the importance of all variables identified by the decision tree confirms their significant contribution to distinguishing between churned and non-churned customers, particularly in the context of the current business case.

* 1. Most-likely-to-churn group of customers

Identifying the group with the highest probability of attritted customers using decision trees involves analyzing the statistics of each leaf. For every split rule, the percentages of churned and non-churned customers are explicitly defined (highlighted in the attached figure). The leaf or branch with the highest churn probability indicates the group most likely to churn. In our case study, the churned customers are represented by an indicator of 0, allowing us to focus on the leaf that has the highest percentage associated with this value.

A computer screen shot of a diagram

Description automatically generated

Figure 29: Result of first decision tree (DT\_ASE)

From the first decision tree, we can conclude that customers who have reported complaints are more likely to churn. Among these customers, those with a cashback amount lower than AUD 163,000 and who typically order **Laptop & Accessory** items are particularly at risk of churning. The same has been indicated in the second decision tree model.

Moreover, the first decision tree highlights the fact that an increase in order amounts compared to the previous year—specifically, a rise greater than **15.5%**—also correlates with a higher probability of churn. For this group of customers, the likelihood of churning increases as the hike in order amount grows. Notably, if the increase exceeds **21.5%**, the probability of churning escalates to over **80%**. This is not being captured in the first decision tree due to early termination in split rule.

This insight highlights the importance of addressing customer complaints and adjusting cashback policies while closely monitoring the order trends to proactively address potential churn risk.

* 1. The most difficult groups of customers to predict their churn outcomes

The most difficult group to accurately predict the churn outcomes can be identified via few criteria:

1. **The High Misclassification Rate Group**:

The misclassification rate reflects the probability of incorrectly predicting true positives and true negatives in the outcomes. Therefore, a higher misclassification rate indicates a greater likelihood that the model will capture incorrect results. Each leaf in the decision tree has its own misclassification rate, which serves as an assessment method for evaluating model performance.

In the decision tree model, the initial nodes of the decision tree tend to experience higher misclassification rates compared to the later nodes. This pattern suggests that the earlier splits are less effective at accurately predicting churn outcomes, making them more prone to misclassifications

A computer screen shot of a diagram

Description automatically generatedA graph with a line and a line

Description automatically generated with medium confidence

Figure 30: Nodes with highest misclassification rate

In our current case, the first 3 leaves experience the highest misclassification rate, approximately 0.1718, higher than the remaining 8 leaves which are 0.1531. This suggests that leaves 1,2 and 3 (node 8, 18 and 19 respectively) have almost 3% chance of incorrectly predicting the churn outcomes compared to the remaining leaves.

1. **Low Confidence Predictions with High Variability within Branches group:**

This group is represented by a leaf with statistics indicating a high probability for both churned and non-churned customers, with the numbers closely matched at around 50%. Such balance in the outcomes indicates uncertainty, making the model lacks confidence in predicting churn for these customers. As a result, it might be necessary to include more factors to better understand the impact on churn probability for these particular groups. In our business study, this is the case of node 21 and node 22

A screenshot of a computer

Description automatically generated

Figure 31: Node with low confidence and high variability

1. **Outliers and Rare Segments:**

Due to the limited number of observations in these categories, the model lacks sufficient predictive power when encountering unseen data that falls within them as it has not yet learned the underlying patterns associated with these outliers. This is the case of node 23, in which only 8/751 observations for churned group and 33/1748 observations non-churned group. The churn probability in this case is currently predicted at **100%**, which indicates a potential overfitting issue. This overfitting suggests that the model is overly reliant on the limited data available for this node, leading to significant challenges when predicting future observations.

A screenshot of a computer

Description automatically generated

Figure 32: Nodes contains outliers and be potential rare segments

In conclusion, the table below will summarize group of customers that might be difficult to predict churning probability

|  |  |  |  |
| --- | --- | --- | --- |
| **Node** | **Interpretation** | **Reason** | **Details** |
| 8 | Low possibility of reporting complains, might not be single and do not prefer to order Laptop & accessory | The High Misclassification Rate | Misclassification rate: 0.1718 |
| 18 | Low possibility of reporting complains, might not be single, prefer Laptop & Accessory and satisfaction score is lower than 4.5 |
| 19 | Low possibility of reporting complains, might not be single, prefer Laptop & Accessory and satisfaction score is higher than 4.5 |
| 21 | Low possibility of reporting complains, Single status, satisfaction score higher than 4.5 and the order amount increase more than 15.5% compared to last year | Low Confidence Predictions with High Variability within Branches group | Churned:50%  Non-churned: 50% |
| 22 | Already submitting complains, Cashback amount higher than AUD 163, favorite category is not laptop & accessory and total order amount increase lower than 21.5% compared to last year | Churned: 46.34%  Non-churned: 53.66% |
| 23 | Already submitting complains, Cashback amount higher than AUD 163, favorite category is not laptop & accessory and total order amount increase higher than 21.5% compared to last year | Outliers and Rare Segments | Churned: 8/751  Non-churned: 33/1748 |

Table 4: Summary table of group of customers that be difficult to predict churn probability

* 1. Performance of model compared to random guess

To determine the performance of model compared to random guess, two critical metrics can be measured, including ROC curves and Lift.

1. **ROC curves:**

ROC curves plot the true positive against the false positive rate at various threshold levels, using AUC (area under the curves) as the metrics for model’s performance. The ROC curve allows for comparison between the model's performance and that of a random guess, with the random guess represented by an AUC of **0.5 (Baseline)**. A model that has AUC greater than 0.5 indicates the better-than-random performance.

1. **Lift:**

This metric represents the ratio between actual results and the outcomes expected from a random guess. By plotting the lift across different percentiles of the population, we can evaluate how much better a predictive model performs compared to random guessing. If the lift value exceeds **1**, it indicates that the model has better-than-random predictive power.

*From the summarized table provided in the early stage of this part, ROC of all three models is larger than 0.5(approximately 0.7), the lift at top 10% is all higher than 3 and lift at top 20% is still higher than 1. Hence, it is logical to assume that all three models perform better than random guess.*

A graph with a line

Description automatically generatedA graph with green bars

Description automatically generated

Figure 33: Metrics determine model performance compared with random guess

1. Performance comparison

For classification predictive model, the best metrics for showcasing the performance would be ROC curves, Lift, Precision, F1 score and Misclassification. The definition and its contribution for defining the predictive power of the model of ROC curves, Lift and Misclassification rates have been clarified in the previous parts. The Precision and F1 score have different approach:

1. **Precision:**

Measure the proportion between true positive predictions out of all positive predictions. In our business case, precision can assist to answer the question Among all customers predicted to churn, how many of them are actually churn? Hence, the higher precision rate, the better the model is at predictive churn customers

1. **F1 score:**

The F1 score is particularly important for the current dataset due to the significant imbalance between churned and non-churned observations. Since the F1 score balances Precision (which minimizes false positives) and Recall (which minimizes false negatives), it effectively addresses the bias that may arise from the dominance of the majority class over the minority class.

Calculating the F1 score allows for a measurement that emphasizes the performance of the model on the minority class, ensuring that both Precision and Recall are considered[[2]](#footnote-2). This provides viewers with a comprehensive overview of the model's effectiveness by highlighting its ability to accurately identify churned customers while also considering the importance of avoiding false positives. Consequently, the F1 score serves as a valuable metric for evaluating model performance in situations with class imbalance

* 1. Lift

The second model exhibits the highest lift measurement, approximately 3.36 at 10% depth. This value is significantly greater than that of the first decision tree and the logistic regression model. This metric suggests that, in terms of predicting churned customers, the second decision tree is more effective at accurately identifying the top 10% of customers most likely to attrit.

However, at 20% depth, the performance of the second decision tree shows a noticeable decline compared to the other models. Despite this drop in performance, the overall stability of the model improves as it delves deeper into the population. This indicates that the second model has a higher likelihood of correctly predicting churning customers compared to the other models. Thus, while the model may experience fluctuations at greater depths, it remains a strong contender for accurately identifying at-risk customers.

A graph with different colored lines

Description automatically generated

Figure 34: Model comparisons metric - Lift

* 1. ROC curves:

The model most capable of distinguishing between churned and non-churned customers is the regression model, which has an AUC of approximately 0.79. In comparison, the second decision tree demonstrates the lowest AUC at 0.708. This indicates that, while the second tree can still classify the two types of customers, its accuracy in correctly identifying new customers is slightly lower than that of the other two models.

A graph with green rectangles and white text

Description automatically generated

Figure 35: Model comparisons metric - ROC curves

* 1. Misclassification rate

Logistic regression reports the highest misclassification rate among the three models, while the two decision trees demonstrate identical performance in this metric. This statistic indicates that the logistic regression model is more likely to produce false predictions compared to the other two models.

A graph with green squares

Description automatically generated

Figure 36: Model comparisons metric - Misclassification rate

* 1. Precision

Both decision trees have the same precision measurement, approximately **0.4031**, suggesting that when predicting churn probability, they achieve correct predictions about **40.31%** of the time. In contrast, the regression analysis shows much weaker performance, producing correct outcomes only about **23%** of the time.

A graph with green bars

Description automatically generated

Figure 37: Model comparisons metric - Precision

* 1. F1 score

With a slightly higher F1 score, both decision trees are more effective at correctly predicting churned customers while maintaining low false positives outcomes. Conversely, the regression model has a slightly lower F1 score, indicating that it tends to make more errors in its predictions.

A green bar graph with black text

Description automatically generated

Figure 38: Model comparisons metric - F1 score

* 1. Choosing the optimal model

Observing all the five measurements, adding in our business questions which is identifying the top attritted customers to sketch the retention program and improve customer engagements, the second decision tree (DT\_Misclassification) will be the optimal model. The model is not only well balanced the correct and incorrect churning probability but also produce the highest number of correct prediction (hightest lift).

A screenshot of a computer

Description automatically generated

Figure 39: Model comparisons metrics

While the second decision tree demonstrates the lowest ROC curves (AUC value), its difference compared to the regression (best model interms of AUC) is not that much (0.0702 and 0.792 respectively), suggesting that its predictive power in distinguishing the churned and non-churned customers is still comparable.

Furthermore, rather than focus on distinguishing power power of the model, the business question requires to correctly predict churning customers more. Minimizing the the predicting errors would be prioritized. Hence, lift is slightly more importance in this case than ROC curves.

Given this analysis, ***the optimal model*** for the business case is the second decision tree, referred to as ***DT\_Misclassification***. By following the split rules that produce medium to high churning probabilities, we can effectively identify the target group of potential candidates for the retention program.Based on the behaviors of these customers, we can customize retention strategies aimed at proactively addressing the risk of churn

PART C: RETENTION PROGRAM

According to the chosen model, the group of customers most likely to churn exhibits the following characteristics:

* They have generally made complaints.
* They have a low cashback amount (lower than 162.5).
* Their preferred order category is not Laptop & Accessory.

The retention strategy should address these issue carefully to reducing the risk of churn and improve the customer satisfaction. Hence, several suggestion can be made as follows:

1. Establish detailed and constructed complaint handling procedures:

While complaints may often be viewed as negative outcomes, a company that knows how to effectively embrace and manage these challenges can turn the situation to its advantage (Jannach et al., 2014). Given that the model emphasizes the importance of complaints in customer engagement, it is essential to study and develop (or enhance, if not already established) suitable ***complaint handling procedures as SOPs*** *(Gruber et al., 2009)*. These procedures should align with the company's vision and goals.

Moreover, it is crucial to create a concise and straightforward version of these procedures that employees and customer service agents can easily understand and implement without spending excessive time on training. This underscores the need for ***periodic training sessions****(How to Handle Customer Complaints The L.E.A.R.N. Technique, 2011)* to reinforce proper complaint handling practices.

Importantly, ***all complaints should be documented*** *(How to Handle Customer Complaints The L.E.A.R.N. Technique, 2011)*, including the date, time, and category, and followed up accordingly. This systematic recording will serve as valuable data for analysis, helping to identify service segments that require further innovation or enhancement. Additionally, it will allow the company to assess the effectiveness of different complaint handling methods on customer satisfaction.

Depending on the size and budget of the company, integrating a ***complaint handling system or software*** (such as Xero or Zoho Desk) can simplify the process (*How to Handle Customer Complaints The L.E.A.R.N. Technique*, 2011). Although implementing such solutions may involve an investment, they facilitate timely complaint management and record essential details for further analysis without the need for extensive training or outsourcing agents.

1. Customer feedback and survey system

Given the significant role complaints play in differentiating between existing and attrited customers, the company should consider enhancing its survey and customer feedback systems (Casas‐Arce et al., 2017). These statistics can serve a dual purpose: on one hand, they act as indicators of potentially churning customers, and on the other, they provide constructive feedback that can be used to improve key service areas.

By systematically gathering and analyzing customer feedback, the company can identify specific pain points and areas for enhancement. This proactive approach not only increases customer satisfaction but also helps to reduce churn risk by addressing issues before they escalate (Casas‐Arce et al., 2017). Implementing a robust feedback mechanism enables the company to stay attuned to customer needs and expectations, ultimately fostering stronger relationships and loyalty (Jannach et al., 2014).

1. Rewards and Loyalty program

The significant decrease in customer engagement after the threshold of 162.5 suggests that customers do not perceive sufficient benefits in continuing the services or products. This indicates that the current cashback system is failing to engage customers effectively, highlighting the importance of implementing the targeted marketing campaigns.

To effectively implementing the program, the customers segmentation must be applied to group customers based on their value. Then the approach can work in two different ways

1. **High value customers:**

Offer additional incentives and exclusively customized service to retain the loyalty (Evans, 2002). Based on the data analysis from the first part, it is evident that many attrited customers with potentially offering high value tend to leave the company after approximately 4-6 years of membership. It is important to recognize this timeframe and consider proactively communicating about exclusive incentives such as annual loyalty e-voucher on their preffered categories which are mobile phone.

1. **Low value customers:**

Offer tier incentives to encourage their increase in loyalty tiers (Evans, 2002). Adopting insights from task 1, careful calculations must ensure that the rewards—specifically in terms of cashback—exceed AUD 163 for higher-tier customers. This will help create a more compelling incentive for customers to remain engaged with the company.

1. Alternative product recommendation system

The identification of unfavorable sentiments towards **Laptops & Accessories**—the company’s most popular products—highlights the diverse range of demands for products, particularly within a global e-commerce context (Rosário & Raimundo, 2021). Attritted customers may be seeking more than just laptops; they might either lack the knowledge about other available products or perceive that the customer service for these alternative offerings is not as strong as that for the primary products(Renjith, 2015).

Hence, it is recommended to analyze the segmentation clusters to identify the preferred categories for each group of customers and to customize rewards, recommendations, or selling programs accordingly. This approach aims to broaden the focus beyond just Laptops & Accessories and to cater to a wider range of options available, especially for global retail e-commerce as the current company.

In particular, promotional campaigns and reward designs could be strategically shifted towards Mobile Phones. Descriptive data indicates that attrited customers tend to be significant users of mobile phones, in contrast to existing customers who favor laptops and accessories. This targeted strategy will allow the business to meet the specific needs of its diverse customer base more effectively.

1. Retention -centric marketing

All of the above suggestions can be categorized under a general strategy known as a retention-centric marketing strategy. By segmenting the group of attrited customers to identify their specific interests and preferences, the company can select appropriate retention strategies, as outlined previously, and craft customized messages to win back customers as soon as they exhibit early signs of churn (as indicated by the decision trees).

For example, if customer segmentation reveals that cashback rewards significantly influence customer decisions, it becomes essential to consider their loyalty tier and provide suitable cashback rewards or discounts and promotions on relevant categories beyond just Laptops and Accessories.

This tailored approach not only addresses the unique needs of different customer segments but also enhances the likelihood of re-engaging customers, ultimately fostering loyalty and reducing churn.

CONCLUSION

In conclusion, the primary objective of this report is to successfully develop the most optimal predictive model to identify customers with a high probability of churning. To achieve this desired outcome, a preliminary explanatory analysis was conducted. This initial stage provided an overall understanding of the distribution of each input variable in relation to the target variable, as well as identifying potential issues that could violate assumptions when building predictive models, such as outliers (skewness), multicollinearity, and missing values.

Additionally, the first step included implementing the RFM (Recency, Frequency, Monetary) approach to identify high-value customers and analyze their behaviors alongside those of churned and non-churned customers. This process facilitated an in-depth understanding of the exceptional behaviors associated with valuable customers, which will support the development of customized retention strategies.

Following this analysis, three different predictive models were constructed, including two decision trees and one logistic regression model. After evaluating five key performance metrics (Lift, ROC curves, Misclassification Rate, Precision, and F1 Score), the optimal model chosen was the decision tree that utilized the misclassification rate as its assessment method.

Utilizing the results obtained from the model to identify common behaviors among the group of customers with the highest churning probability, and integrating insights from earlier tasks, five retention strategies are recommended:

1. Establish Detailed and Structured Complaint Handling Procedures: Create effective processes to address customer complaints, ensuring resolution and enhancing customer satisfaction.
2. Improve Customer Feedback and Survey Systems: Implement robust mechanisms to gather customer feedback, enabling the company to identify and address areas for improvement.
3. Enhance Rewards and Loyalty Programs: Revise the existing rewards system to provide more compelling incentives for high-value customers, encouraging their continued engagement.
4. Adopt Alternative Product Recommendation Systems: Expand product offerings and tailor recommendations to better meet the diverse needs of customers, especially those who are not primarily interested in Laptops & Accessories.
5. Sketch a Retention-Centric Marketing Strategy: Combine all the above recommendations into a cohesive marketing strategy focused on retaining at-risk customers.

By implementing these strategies, the company can effectively reduce churn risk and strengthen customer engagement, ultimately enhancing overall business performance.

References

Casas‐Arce, P., Lourenço, S. M., & Martínez‐Jerez, F. A. (2017). The performance effect of feedback frequency and detail: Evidence from a field experiment in customer satisfaction. *Journal of Accounting Research*, *55*(5), 1051-1088.

Evans, G. (2002). Measuring and managing customer value. *Work study*, *51*(3), 134-139.

Gruber, T., Szmigin, I., & Voss, R. (2009). Handling customer complaints effectively: A comparison of the value maps of female and male complainants. *Managing Service Quality: An International Journal*, *19*(6), 636-656.

*How to Handle Customer Complaints The L.E.A.R.N. Technique*. (2011). Retrieved October from <https://www.softexsw.com/en/How-to-Handle-Customer-Complaints-The-L.E.A.R.N.-Technique-_639_1-viewed>

Hughes, A. M. (2005). *Strategic database marketing*. McGraw-Hill Pub. Co.

Jannach, D., Zanker, M., & Fuchs, M. (2014). Leveraging multi-criteria customer feedback for satisfaction analysis and improved recommendations. *Information Technology & Tourism*, *14*, 119-149.

Renjith, S. (2015). An integrated framework to recommend personalized retention actions to control B2C E-commerce customer churn. *arXiv preprint arXiv:1511.06975*.

Rosário, A., & Raimundo, R. (2021). Consumer marketing strategy and E-commerce in the last decade: a literature review. *Journal of theoretical and applied electronic commerce research*, *16*(7), 3003-3024.

APPENDIX

**Task A -4: Chi-squared computation**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Task C -8: Precision and F1 score**

A screenshot of a table

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Please view the the detailed computation in appendix part 1-4: Chi-squared computation [↑](#footnote-ref-1)
2. Please view the detailed computation in Appendix part C-8: Precision and F1 score computation [↑](#footnote-ref-2)