

**BC2407 Analytics II: Advanced Predictive Techniques**

**Semester 2 Academic Year 22/23**

**Semester Project Report**

|  |  |  |
| --- | --- | --- |
| Members | Elliot Yew Yi Le | U1911991E |
| Abhishekh Pandey | U2120840L |
| Ng Jie En Dominick | U2120310K |
| Yu Yaqi | U2110350J |
| Seminar Number | 6 | |
| Group Number | 2 | |
| Instructor | Neumann Chew C.H. | |
| Submission Date | 2 April 2023 | |

**Content Page**

[**Executive Summary** 4](#_Toc131369010)

[**1. Introduction** 5](#_Toc131369011)

[**1.1 Background of Telecommunications Industry in Singapore** 5](#_Toc131369012)

[**1.2 Business Problem of Customer Churn** 5](#_Toc131369013)

[**1.3 Key objectives** 6](#_Toc131369014)

[**2. Exploratory Data Analysis** 7](#_Toc131369015)

[**2.1 Data Description and Methodology** 7](#_Toc131369016)

[**2.1.1 Variables Easily Acquired** 7](#_Toc131369017)

[**2.1.2 Variable Relevance** 7](#_Toc131369018)

[**2.2 Data Preparation and Cleaning** 7](#_Toc131369019)

[**2.3 Data Visualization** 7](#_Toc131369020)

[**2.2.1 Data Correlation** 8](#_Toc131369021)

[**2.2.2 Insights on Relationship between Churn.Value and Predictor Values** 8](#_Toc131369022)

[**2.3 Feature Selection** 11](#_Toc131369023)

[**2.4 Train-Test Split** 12](#_Toc131369024)

[**2.5 Train set Balancing** 12](#_Toc131369025)

[**3. Significant Variable Analysis** 13](#_Toc131369026)

[**3.1 Data Modelling** 13](#_Toc131369027)

[**3.1.1 Logistic Regression** 13](#_Toc131369028)

[**3.1.2 MARS** 13](#_Toc131369029)

[**3.1.3 Random Forest** 13](#_Toc131369030)

[**3.2 Variable Importance** 14](#_Toc131369031)

[**3.3 Recommendations** 14](#_Toc131369032)

[**3.3.1 Total Customer Service Request** 14](#_Toc131369033)

[**3.3.2 Contract** 15](#_Toc131369034)

[**3.3.3 Number of Referrals** 16](#_Toc131369035)

[**3.3.4 Unlimited Data** 17](#_Toc131369036)

[**4. Prediction Model Analysis** 17](#_Toc131369037)

[**4.1 Methodology** 17](#_Toc131369038)

[**4.2 Logistic Regression** 18](#_Toc131369039)

[**4.3 Multivariate Adaptive Regression Splines (MARS)** 19](#_Toc131369040)

[**4.4 Random Forest** 20](#_Toc131369041)

[**4.5 Model Evaluation** 20](#_Toc131369042)

[**4.6 Recommendation** 21](#_Toc131369043)

[**4.6.1 Proactive Customer Retention Department** 21](#_Toc131369044)

[**5. Limitations and Future Work** 22](#_Toc131369045)

[**5.1 Imbalanced Initial Dataset** 22](#_Toc131369046)

[**5.2 Representativeness of the Data** 22](#_Toc131369047)

[**5.3 Insufficient Variable Details** 22](#_Toc131369048)

[**5.4 Future Work** 23](#_Toc131369049)

[**6. Conclusion** 23](#_Toc131369050)

[**7. References** 24](#_Toc131369051)

[**8. Appendix** 26](#_Toc131369052)

# **Executive Summary**

As the telecommunications market in Singapore gets increasingly competitive, customers now have a wide array of companies to choose from. As competitors continue to offer newer and more attractive services and products, customers are ultimately spoilt for choice. While good for customers, this also means that the threat of customer churn has never been higher, which telecommunication industry players like M1 must address in order to protect their market share, as well as their dominance in the market.

This report aims to address the business problem of customer churn, as well as recommend strategies that M1 could adopt to mitigate it. From an analytical perspective, our approach to this issue is two-pronged. Firstly, we sought to identify the significant factors and variables that affect the likelihood of customer churn, using analytical models like Logistic Regression, MARS, and Random Forest, which helped to filter out the key variables from the insignificant ones. The team derived crucial insights as to the important factors that may cause customer churn. Further research was then done on the significant variables identified for a better holistic understanding, and thus craft targeted recommendations which might help M1 improve its business strategic processes and reduce the likelihood of customer churn. Significant variables identified include *‘Total Customer Service Request’, ‘Contract’, ‘Number of Referrals’ and ‘Unlimited Data*.

The second part of our approach required the use of an accurate predictive model, which would be used to identify customers who are at high risk of churning. After comparing the output of several analytical models, the team concluded that the MARS model had the highest accuracy of 88.5%, and hence might be most suitable for this use-case. As a comparison, without the use of a predictive model, the chance of identifying a customer about to churn is only 26%. Thus, this huge improvement in predictive capabilities gives us the confidence to propose a proactive customer retention strategy, to engage with unsatisfied customers early to regain their trust and keep their business.

The report then concludes with suggestions for future research, as well as some limitations that have inhibited our analysis in one way or another. Regardless, our models and proof of concepts ultimately aim to offer insights and demonstrate how an analytics approach can help M1 in mitigating the significant business problem of Customer Churn.

# **1. Introduction**

## **1.1 Background of Telecommunications Industry in Singapore**

For close to 2 decades, the 3 main telecommunications companies in Singapore (Singtel, Starhub & M1) enjoyed oligopolistic competition in the industry, shielded from international players and almost exclusively closed to foreign competition. Rather than focusing on marketing and customer retention, the optimal strategy was to coordinate and keep prices high, while locking customers into contracts in order to grow the average revenue per user over time. So, in 2016 when the authorities wanted to facilitate the entry of a fourth telco to drive greater competition and innovation in the sector, the move was met with protests from the incumbents (Ng, 2021).

As authorities went ahead with deregulation to increase competition regardless, TPG joined the market as Singapore’s 4th telco. Later on, mobile virtual network operators (MVNOs) such as Circles. Life, Zero1 and RedOne would enter the playing field as well. With competition in the industry having never been higher, the 3 original players found themselves quickly losing customer and thus market share to the new entrants – a process known in the industry as churn.

## **1.2 Business Problem of Customer Churn**

Customer churn, also known as customer attrition, is when someone chooses to stop using your products or services. In effect, it is when a customer ceases to be a customer.

In general, telecommunications is a highly saturated sector, composed largely of commoditized services with limited areas for differentiation. The result of this combination is a highly competitive industry with exceptionally high customer churn (HEAVY.AI, 2022). In addition, the changing perceptions of customer choice, as well as more information than ever due to the digital accessibility, means that customers are less brand loyal than ever. In fact, studies have shown that 77% of all consumers retract their loyalty more quickly than they did three years ago. This means customer retention in the telecom industry is harder to achieve than ever (Axigen, 2022).

The financial implications of customer churn are well documented – it is estimated that it is 5 to 25 times more expensive to acquire a new customer compared to retaining an existing one (Gallo, 2014). Additional time and resources need to be spent in terms of marketing, onboarding processes, service fees etc. Even then, securing a new customer is far from guaranteed. In comparison, reducing churn is considerably cheaper. Research done by Frederick Reichheld of Bain & Company shows increasing customer retention rates by 5% would increase profits by 25% up to 95%.

Another implication of high customer churn is that it is likely to affect the brand image of the company. Above and beyond the loss of revenue from the churned customer, there could be potential damage to the company’s branding or reputation through negative word of mouth and bad reviews. This would undoubtedly have a serious impact on the company’s bottom line as well as future growth prospects (Qualtrics, 2023).

In summary, customer churn is not simply a matter of trying to replace lost customers. It is a serious problem that needs to be thoroughly considered and mitigated, particularly for hyper competitive industries like telecommunications in Singapore. This is especially true in today’s market as the regulatory, economic, and social environment has become less favourable for customer retention, as more competitors enter the market and consumers become more educated and discerning. To not take the issue of customer churn seriously would come with a hefty price indeed.

## **1.3 Key objectives**

Our team seeks to aid M1 in reducing the rate of customer churn through the use of data analytics. To achieve this, our team proposes a 2-pronged approach involving an analysis of significant variables of customer churn, followed by the development of a prediction model that predicts whether a customer is likely to churn or otherwise.

Prior to our 2-step approach, in Section 2, our team conducted adequate data cleaning to ensure that the accuracy of the models developed will be reliable and relevant to our context. Similarly, variables which were irrelevant for analysis and data anomalies were further removed or changed, with data with missing values removed accordingly. As more insights were discovered during the analysis, data cleaning was conducted iteratively. In Section 2.2 data exploration was performed to acquire preliminary understanding of the cleaned dataset. In Section 2.3, further feature selection methods were used, removing variables that may have high multicollinearity with other variables and further narrowing down our variables through the use of Logistic Regression. Our team utilized a variety of visualization techniques, including stacked bar charts, box plots and heatmaps.

In Section 3, to fully explore the significant variables of customer churn, our team used predictive models, such as Logistic Regression, Multivariate Adaptive Regression Spline (MARS), and Random Forest to identify the most significant variables that affects customer churn. This offered critical insights into the key attributes that result in customer churn while bringing to light possible variables that have been overlooked in the past. With the key factors identified, our team proceeded to recommend solutions to specifically address these significant variables identified, to help lower rates of customer churn. This ultimately answers our key questions as to which variables are vital in concluding whether a customer churn will occur. By performing significant variable analysis, we provided holistic recommendations that address these key factors, and therefore, reduce the rates of customer churn.

Next, in Section 4, our team has developed a prediction model to predict which customers are at a higher risk of churning. The outcome variable “Churn” is a binary categorical variable based on past customer data. Our team initially made use of multiple prediction models for analysis including regression models such as Logistic Regression and MARS as well as other advanced predictive models with higher predictive accuracy such as Random Forest. The results acquired from each model will then be compared based on a variety of metrics to evaluate and determine which model is the most suited in predicting the likelihood of a customer churning.

Finally, using our chosen model, our team will be able to predict whether a customer is at risk of churning, and implement corrective measures to more effectively tackle this issue.

In summary, these 2 facets of the analytics solution would be synergistic as the “prevention & cure” of the significant business problem of customer churn. With a better understanding of the significant variables resulting in Customer Churn, as well as a predictive model to mitigate customers who are likely to churn, we would expect a significant reduction in Customer Churn and thus an improvement in M1’s bottom line.

# **2. Exploratory Data Analysis**

## **2.1 Data Description and Methodology**

To demonstrate our solution, we have chosen the “JB Link Telco Customer Churn” dataset which is a customised version of the widely known IBM Telco Customer Churn dataset. The dataset was chosen due to the following reasons:

### **2.1.1 Variables Easily Acquired**

To ensure the practicality of our solution, the data used must be easily collectable. Therefore, many of the variables in our dataset, such as ‘Phone.Service’, ‘Internet.Type’ and ‘Unlimited.Data’, are information already available in the company’s database. Some other variables may not be as readily available such as ‘Married’, ‘Dependents’ and ‘Customer.Satisfaction’, but can be easily obtained through surveys or part of the registration form when customers join the company.

### **2.1.2 Variable Relevance**

The dataset used is based on the widely known IBM Telco Customer Churn dataset that has been used in numerous machine learning studies and research. Therefore, the dataset is reliable and useful to build our prediction models upon.

## **2.2 Data Preparation and Cleaning**

Before data exploration, we performed an initial cleaning to ensure that the raw data is free from any missing values and has been formatted with the appropriate data types. By examining the raw dataset, it is evident that certain variables, such as ‘Gender’, ‘Offer’ and ‘Contract’ are categorical variables. Therefore, it is necessary to convert them into the factor datatype. The corrected structure of the dataset and a detailed overview of the modifications can be found in Appendix 1.

In addition, we conducted a thorough investigation to identify erroneous values and outliers. None were found. However, we noted that ‘Customer.Satisfaction’ had a large number of missing values. Nevertheless, we decided to retain the variable for analysis since it may contain valuable information relating to our target variable, ‘Churn.Value’.

We also decided to drop the ‘Zip.Code’, ‘Latitude’, ‘Longitude’, ‘Population’ and ‘City’ columns since they are not relevant variables in the context our case study. We also dropped the redundant ‘Customer.ID’ column as it is simply an identifier column.

Lastly, we added a new variable titled ‘Number.of.Subscriptions’ which is a count of the total number of services a customer is subscribed to. We hypothesise that the number of subscriptions a customer has may be related to Churn.Value hence we added the variable for further analysis.

## **2.3 Data Visualization**

To gain a deeper understanding of the relationships between the variables in our dataset, identify anomalies, and uncover meaningful insights, we conduct data exploration on the cleaned dataset.

### **2.2.1 Data Correlation**

Since our dataset contains a large number of variables, we utilise methods like the Chi-Square test of Independence (to determine the association between two categorical variables) and the Pearson correlation coefficient (to determine the association between 2 continuous variables) both of which can be found in Appendix 2A. This aids us in feature selection as a high correlation coefficient can be an indication of multicollinearity, negatively impacting our model’s accuracy. Our analysis via the Chi-square test indicated that all categorical variables are likely to be related to Churn.Value, while the Pearson Correlation matrix suggested that Number.of.Subscriptions has a strong and very strong positive linear correlation to Total.Regular.Charges and Monthly.Charge respectively, and that Total.Regular.Charges has a very strong positive linear correlation to Tenure.In.Months. This hypothesis is supported by plotting the variables against each other in a scatterplot and they can be found in Appendix 2A.

### **2.2.2 Insights on Relationship between Churn.Value and Predictor Values**

The analysis presented below was obtained through a set of visual representations. Because the findings were consistent across the entire dataset, we can conclude that the dataset is reliable and can be used for modelling purposes.

1. **Relationship between Contract and Churn.Value**

Customers with 2-year contracts were the least likely to churn whereas those with month-to-month contracts were the most likely. This may be due to numerous factors such as early termination charges of the contracts or other penalties imposed by the telco (*StarHub: Frequently Asked Questions, 2023*). However, more information on the policies of the telco is needed for any meaningful conclusions.

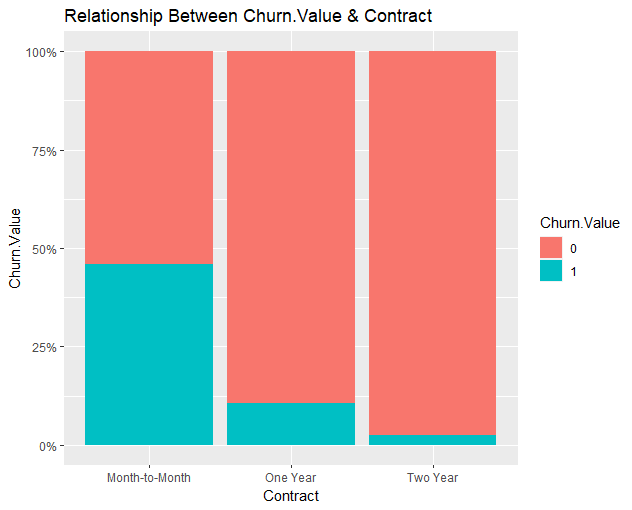


Figure 1. Relationship between Churn.Value and Contract

1. **Relationship between Number.of.Subscriptions and Churn.Value**

We see an increasing trend in the proportion of customers churning as the number of subscriptions of the customer increases to 3. Beyond 3 subscriptions, we see a decreasing trend in the proportion of customers churning. This might be due to the fact that customers satisfied with the telco are more likely to take up more subscriptions and also less likely to churn.

Chart, bar chart

Description automatically generated

Figure 2. Relationship between Churn.Value and Number of Subscriptions

1. **Relationship between Offer and Churn.Value**

Customers that took up Offer A were the least likely to churn, whereas customers that took up Offer E had a churn rate of more than 50%. This could potentially imply that Offer A was the best offer out of all the offers, however, again, more information about the telco is required to draw deeper insights. This data could also serve as a key performance indicator for offers, signalling the customer’s perception of the product’s value proposition.

Chart, bar chart, histogram

Description automatically generated

Figure 3.Relationship between Churn.Value and offer

1. **Relationship between Number.of.Referrals and Churn.Value**

Customers that churned typically had fewer referrals as compared to customers that did not churn. This may be due to the fact that unsatisfied customers are less likely to refer their friends and family to the company’s offerings when they themselves are unsatisfied.

Chart, box and whisker chart

Description automatically generated

Figure 4. Distribution of Number of Referrals by Churn.Value

1. **Relationship between Total.Customer.Svc.Requests and Churn.Value**

Customers that churned typically also had higher customer service requests than those that did not.

Chart, box and whisker chart

Description automatically generated

Figure 5. Distribution of Total Customer Service Requests by Churn.Value

1. **Remaining Predictor Variables**

The table below is a summary of the relationship between Churn.Value and the remaining predictor values. The corresponding set of visual representations can be found in Appendix 2B.

|  |  |
| --- | --- |
| 𝑿-variable | Relevance/Insights |
| Dependents | **Customers with dependents are less likely to churn** |
| Internet.Service | **Internet service subscribers less likely to churn** |
| Internet.Type | 1. **Fiber Optic Internet users are less like to churn** 2. **More than 50% of Cable Internet users churn** |
| Online.Security | **Online security subscribers are less likely to churn** |
| Paperless.Billing | **Customers that chose paperless billing are more likely to churn** |
| Payment.Method | **Customers that pay by credit card are least likely to churn** |
| Premium.Tech.Support | **Customers with premium tech support are less likely to churn** |
| Unlimited.Data | **Customers with unlimited data are less likely to churn** |
| Customer.Satisfaction | **The more satisfied the customer, the less likely they are to churn**  All customers that responded “Very Unsatisfied” and “Unsatisfied” churned, whereas all the customers that responded “Satisfied” and “Very Satisfied” did not churn |
| Monthly.Charge | **Customers that churned typically had higher monthly charges** |
| Tenure.in.Months | **Customers that churned typically had a shorter tenure period** |
| Product.Service.Issues.Reported | **Customer that churned reported more product service issues** |

## **2.3 Feature Selection**

In the initial phase of feature selection, we dropped Churn.Category and Churn.Reason as any value present in them imply that the customer has churned. We also dropped Customer.Satisfaction due to the large number of missing values. We also dropped the categorical variable Referred.a.Friend, Internet.Service, Under.30, Senior.Citizen and Number.of.Dependents as the information for the variables are already captured in Number.of.Referrals, Internet.Type, Age and Dependents respectively. Lastly, we dropped Total.Regular.Charges and Number.of.Subscriptions based on the Pearson Correlation coefficient to avoid multicollinearity.

In the next phase we used logistic regression to determine the most important variables from the remaining variables and selected them for our model development. This was done through the use of the backward elimination algorithm (Bursac et al, 2008). This algorithm iteratively removes the least significant features to minimize the model’s Akaike Information Criterion (AIC) value. This helps to optimize and balance the model’s complexity to its goodness of fit, reducing likelihood of our model overfitting. The table below is a summary of the final feature set.

|  |  |  |
| --- | --- | --- |
| Variable Name | Variable Structure | Data Type |
| Number.of.Referrals | Continuous | Integer |
| Tenure.in.Months | Continuous | Integer |
| Offer | Nominal Categorical | Factor |
| Phone.Service | Nominal Categorical | Factor |
| Avg.Monthly.Long.Distance.Charges | Continuous | Integer |
| Multiple.Lines | Nominal Categorical | Factor |
| Internet.Type | Nominal Categorical | Factor |
| Avg.Monthly.GB.Download | Continuous | Integer |
| Online.Security | Nominal Categorical | Factor |
| Online.Backup | Nominal Categorical | Factor |
| Device.Protection.Plan | Nominal Categorical | Factor |
| Premium.Tech.Support | Nominal Categorical | Factor |
| Streaming.TV | Nominal Categorical | Factor |
| Streaming.Movies | Nominal Categorical | Factor |
| Streaming.Music | Nominal Categorical | Factor |
| Unlimited.Data | Nominal Categorical | Factor |
| Contract | Nominal Categorical | Factor |
| Paperless.Billing | Nominal Categorical | Factor |
| Payment.Method | Nominal Categorical | Factor |
| Monthly.Charge | Continuous | Integer |
| Total.Refunds | Continuous | Integer |
| Total.Extra.Data.Charges | Continuous | Integer |
| Total.Long.Distance.Charges | Continuous | Integer |
| Gender | Nominal Categorical | Factor |
| Age | Continuous | Integer |
| Married | Nominal Categorical | Factor |
| Dependents | Nominal Categorical | Factor |
| CLTV | Continuous | Integer |
| Total.Customer.Svc.Requests | Continuous | Integer |
| Product.Service.Issues.Reported | Continuous | Integer |

## **2.4 Train-Test Split**

To properly evaluate the performance of the Logistic Regression, MARS, and Random Forest models, we first split the data into a 70-30 train test split. By using the same train-test split, we could ensure that any differences in performance between the models were not due to differences in the data used for training or testing. Additionally, this approach allowed our team to estimate how well each model would perform on new, independent data providing us with a more useful insight on their performance. However, it should be noted that Random Forest does not necessarily require a train-test split, as it has an out-of-bag (OOB) feature that can estimate the performance of the model on missing data.

## **2.5 Train set Balancing**

In response to the imbalanced distribution of the cleaned data set, with the majority class (Churn.Value==0) consisting of 5174 samples and the minority class (Churn.Value==1) consisting of 1869 samples, we employed a down-sampling technique on our train set to balance the dataset. Our team obtained a train set with an even split of 1308 data on both the majority and minority class.

# **3. Significant Variable A­nalysis**

The selection of 30 statistically significant variables in our model earlier in Section 2.3 provides a comprehensive understanding of the factors that influence the likelihood of customer churn. Although all 30 variables will be used for our analytics, due to the scarcity of resources, it may not be practical for M1 to address all 30 variables. Therefore, our team believe that we should also focus our efforts on identifying the most significant variable that has the greatest impact on customer churn. By prioritizing the key variables, we can optimize our resources and develop targeted strategies to address the main cause of churn. Additionally, we can further investigate the relationship between this variable and customer churn to gain a more concrete understanding of the underlying factors that cause customer churn.

## **3.1 Data Modelling**

### **3.1.1 Logistic Regression**

From our Logistic Regression model, we measure a variable’s significance by comparing its p-value. However, we should take note that we should not solely depend on the p-value to determine the significant variable as the p-value can be affected by the sample size and collinearity.

As seen on the figure in Appendix 3, the following variables seems to be the most significant for the Logistic Regression model in no ranking order:

Equation 1: Significant Variables for Logistic Regression

### **3.1.2 MARS**

From our MARS model, utilizing the ‘evimp’ function, we measured a variable’s significance (as shown in Appendix 4) by comparing the number of model subset that include the variable (nsubsets), the variable’s net decrease in Residual Sums of Squares (RSS) and its net decrease in Generalized Cross Validation (GCV) Score. Our team selected the top 8 most significant variables from these measurements. They are:

Equation 2: Significant Variables for MARS

### **3.1.3 Random Forest**

From our Random Forest model, our team measured a variable's significance (as shown in Appendix 5) by comparing its mean decrease accuracy. Our team selected the top 8 most significant variables from these measurements. They are:

Equation 3: Significant Variables for Random Forest

## **3.2 Variable Importance**

As each model has its own limitations and weaknesses, our team believe that there is a need to use multiple models to further identify the common significant features that influence the likelihood of a customer churning as compared to relying solely on any one model. This will also help to improve the robustness and stability of our results and reduce the likelihood of our variables being too overfitted to suit a single model.

By selecting the variables that are significant in multiple models in predicting the likelihood of customer churning, we obtained the following key factors:

Equation 4: Key Factors

Despite the significance of these 8 variables, our team has prioritized the variables that we deemed would be the most actionable for M1's business purposes. This would allow the M1 to address their customer churn issue efficiently. These variables will be further discussed in Section 3.5.

## **3.3 Recommendations**

### **3.3.1 Total Customer Service Request**

One of the most significant variables that contribute to customer churn is the variable ‘*Total Customer Service Request’,* which counts the number of times the customer contacted customer service in the past quarter. A parallel can be drawn to another significant variable ‘*Product or Service Issue Reported’,* which describes the number of times the customer reported an issue with a product or service in the past quarter. An interesting insight however, is that these 2 variables have little correlation to each other, which might be contradictory to what we might expect.

This suggests that the way customer service is handled plays an extremely significant role in determining customer churn, regardless of product or service quality. Putting it in another way, even customers that are generally satisfied and have few product/service issues, might become frustrated with poor customer service that fails to resolve their issues in a clear and timely manner. This might necessitate multiple customer service requests, and ultimately be a determining factor that results in customer churn. In recognition of this principle, some CEOs have even started emphasizing better customer service over product quality and price (Hyken, 2021).

***Insight***

It is not uncommon for companies to underestimate the value of an excellent customer service department and treat customer service merely as another area for potential cost cutting. This can have severe implications on performance metrics like customer waiting times for customer service, which would adversely impact customer satisfaction (Ecseley, 2015). In this regard, every minute counts when it comes towards customer service.

Another cost-cutting consideration is whether to outsource customer service. Undeniably, outsourcing can have significant cost savings, especially in the short run. However, the negative implications are important to consider as well. The customer service experience is likely to lose that personal touch with valued customers, and maintaining quality would become a challenge as well. Most importantly, studies have shown that outsourced customer services tend to result in drastic decreases in overall satisfaction (Dignan, 2008), due to factors like poor understanding of the business, as well as cultural and linguistic differences.

Lastly, an increasingly popular method of customer service is the use of automated customer service response bots, which require customers to press keypad numbers in response to scripted questions. It is debatable if such a system actually enhances the customer experience. While some customers might enjoy a smoother, automated experience, especially for simpler problems, it might irritate and enrage other customers who require a more complex solution to their problems (McGrath, 2021).

In aggregate, all these factors might result in unsatisfied customers which is reflected in an increase in *‘Total Customer Service Request’* which is highly likely to end up with customer churn.

***Recommendation***

In light of the above insights, we strongly recommend that customer service is one area where cost cutting should be minimized as far as possible. Cost cutting measures like reduction of customer service staff, outsourcing and automation of customer service would likely deteriorate the customer experience, and thus have a significant influence on the rates of customer attrition. While the short-term cost savings might be attractive, it might be penny wise and pound foolish in the long run.

### **3.3.2 Contract**

Another significant variable that contributes to customer churn is ‘*Contract’,* which describes the current length of contract. The contract can have 3 values, month to month, one year and two years. The data suggests that customers with longer contract durations are more likely to renew and thus not churn, whereas customers with shorter contracts are less likely to renew, which increases churn rate.

In general, longer telco contracts helps to build brand loyalty as customers are integrated into the ecosystem, while simultaneously building structural barriers to churn (Hopkinson, 2022). Brand loyalty is also built as contract customers enjoy special rates and participate in loyalty programmes as well (Kaur, 2021).

***Insight***

In today’s hyper competitive telecommunications arena, no-contract sim plans have become increasingly popular, both for customers, as well as for telcos in a bid to attract and retain customers. To illustrate the point, just recently in Feb 2023, M1 announced its latest SIM-only plan, Maxx SIMO at a promotional price of $15 a month. Such contracts were practically unheard of in the past. Prior to 2015, telcos were hesitant to offer such contracts which were bound to have high customer churn (Cheok, 2016).

Indeed, such no-contract options are likely to attract high churning customers by its very nature. While having more customers is a good thing, the issue is that these deals typically attract deal seekers that are quick to churn when they find a better deal with another company. Having such a fickle customer base can cause a multitude of problems for the company, and some have pointed to the issues of *Groupon* as a prime case example (Gallo, 2014).

The pandora’s box of “No-Contract Plans” have long been opened, and it is unlikely that such plans can ever be removed from the market. However, it should be noted that while such no-contract plans can be useful for short term customer acquisition and revenue injection, these customers are likewise easily lost and therefore unsustainable for long term growth.

***Recommendation***

In light of the above insights, we strongly recommend that while “no-contract sim plans” should continue to be offered to match competitors, it should remain only as an alternative to the traditional yearly contract plans. The business strategy and direction should remain focused on these contract plans, in order to build brand loyalty and long term customer relationships. As resources are limited, marketing and promotional resources should primarily go towards these plans, with the objective of providing great value to customers in order to encourage customers to renew these contracts. Overall, this would be more sustainable in minimizing customer churn and maximizing growth potential in the long run.

### **3.3.3 Number of Referrals**

‘Number of referrals’ was also identified as a significant variable that contributes to customer churn. The numeric value scaled from 0 (no referrals made), up to 11 referrals made in this dataset. The data suggests that customers who have made more referrals are significantly less likely to churn, compared to those that did not.

***Insight***

An attractive referral system is an essential component of any loyalty programme to make customers feel more valued and engaged, which would contribute to reducing their churn rate (Jakimovski, 2019). An additional benefit is that new customers acquired through referrals are significantly less likely to churn compared to other methods as well, as customers are usually more trusting of recommendations from friends and families (Galera, 2019), which would contribute to lowering the company’s overall churn rate.

In Singapore, the various telco companies have different strategies with respect to referral programmes. Singtel and M1 primarily rely on an extensive loyalty rewards programme, but crucially opts out of a referral programme for their customers. Starhub, as well as newer entrants to the market such as Circles.Life as well as MyRepublic have attractive referral programmes. Under the Starhub Friends Programme, referrers can be rewarded with up to $75 upon a successful referral. For Circles.Life, referrers and referees can enjoy up to an additional 30GB or data per month, and for MyRepublic, referrers receives a tiered monthly cashback income based on their number of referrals. (Appendix 6 &7 )

***Recommendation***

In light of the above insights, we strongly recommend that M1 implement a referral programme as well. Compared to its competitors, M1 could potentially be losing out as its customers are drawn away by attractive referral programmes. Overall, the data and empirical research suggests that this would help M1 to reduce customer churn by rewarding loyal customers, as well as acquire new, churn resistant customers.

### **3.3.4 Unlimited Data**

Another significant variable that contributes to customer churn brought to light from our analytics is ‘Unlimited Data’. A customer can subscribe to either a plan with unlimited data or limited data. Our analytics suggest that customers with an unlimited data plan are less likely to churn than those with a limited data plan. A possible reason for the churning of customers with limited data could be due to the lack of sufficient data, resulting in them switching to another internet provider that offers a more attractive package.

***Insight***

With the increase ease of access and affordability of data, consumers are consuming a larger amount of data than ever before. The global monthly average usage per smartphone is anticipated to be 19 GB in 2023 and is forecast to reach 46 GB by the end of 2028 (Ericsson, 2022). As more customers becomes accustomed to the usage of mobile data, they are likely to be more attracted to packages that provides larger data allowances or unlimited data as compared to the traditional packages that focuses on providing more call minutes and short messaging services (SMS).

To capitalize on this opportunity, M1’s competitors such as Singtel and Circles.Life has been aggressively marketing their data packages with a higher limit to attract new subscribers, possibly “stealing” away customers from M1.

***Recommendation***

Considering the above insights, we strongly recommend that M1 should revise their limited mobile data packages by offering more attractive mobile plans with a higher data limit or unlimited data. By providing a compelling service through analysing customer’s changing requirements and preferences, M1 would be able to retain their existing customers through providing them with a better mobile data-focused package as compared to their competitors, reducing churn.

# **4. Prediction Model Analysis**

## **4.1 Methodology**

To evaluate our models, we will be using certain metrics which are explained below.

**Accuracy:** Accuracy is the measure of how often a classification model correctly predicts the class label of an instance in the dataset. It is a commonly used evaluation metric for classification problems and is expressed as a ratio of the number of correct predictions to the total number of predictions made by the model. The formula for accuracy is as follows.

**False Negative Rate:** False negative rate (FNR) is the measure of the number of instances belonging to a positive class (customers that churn) that are incorrectly predicted as negative by a classification model. FNR is a ratio of the number of false negative predictions to the total number of instances belonging to the positive class. Given the context of our case study, we wish to minimise the FNR in order to in order to maximise customer retention. The formula for FNR is as follows.

**False Positive Rate:** False positive rate (FPR) is the measure of the number of instances belonging to the negative class (customers that stay) that are incorrectly predicted as positive by a classification model. FPR is a ratio of the number of false positive predictions to the total number of instances belonging to the negative class. The formula for FPR is as follows.

**Recall:** Recall is the measure of the ability of a classification model to correctly identify instances belonging to the positive class (customers that churn). It is a ratio of the number of true positive predictions to the total number of instances belonging to the positive class. In the context of our study, we wish to maximise recall as it provides information on the model's ability to correctly identify customers that are likely to churn. The formula for recall is as follows.

**Precision:** Precision is a measure of the ability of a classification model to correctly identify instances belonging to the positive class among all instances predicted as positive. It is the ratio of the number of true positive predictions to the total number of instances predicted as positive.

Given the context of our case study we have decided not to include specificity as it is not a relevant metric in our model evaluation.

## **4.2 Logistic Regression**

Logistic Regression can be used to predict the outcome of binary categorical data. Due to the relatively simplistic nature of Logistic Regression, it allows for better interpretability of the relationship between the input variables and the target variable (Jain et al., 2020). To achieve our goal of identifying key variables as well as to accurately predict the likelihood of a customer churning, our team decided to use Logistic Regression as our first model.

As discussed earlier in section 2.3, the remaining features were used in our Logistic Regression model as they are the features that will optimize the model’s complexity and goodness of fit. The model is then trained using the balanced train set obtained in section 2.4.2. A threshold value of 0.5 was set for the prediction of whether a customer will churn. The model was then used to predict the outcome for the train set and test set.

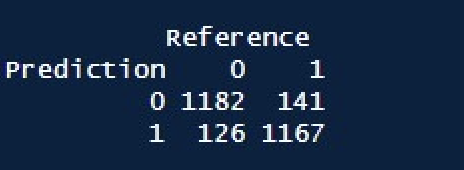
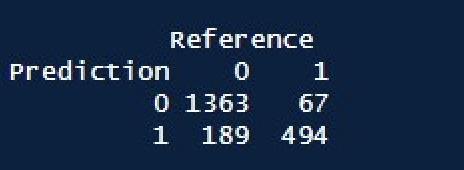
To evaluate the model, a confusion matrix was created to visualize the model accuracy on both the train set and test set.

Figure 7. Confusion Matrix of Test Set

Figure 6. Confusion Matrix of Train Set

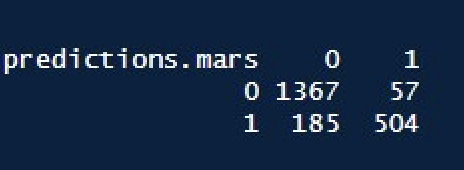
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | False Negative Rate | False Positive Rate | Recall | Precision |
| Train set | 0.8979 | 0.1078 | 0.0963 | 0.8922 | 0.9026 |
| Test set | 0.8788 | 0.1194 | 0.1218 | 0.8806 | 0.7233 |

From the train set confusion matrix, the model has an accuracy of 0.8979 with a false negative rate of 0.1078, false positive rate of 0.0963, recall of 0.8922 and precision of 0.9026. While for the test set confusion matrix, the model has an accuracy of 0.8788 with a false negative rate of 0.1194, false positive rate of 0.1218 recall of 0.8806 and precision of 0.7233. As the accuracy between both the models are relatively high and like each other, the model is not overfitted. However, the precision of the model is significantly lower than that of the train set.

## **4.3 Multivariate Adaptive Regression Splines (MARS)**

MARS is a non-parametric regression model that is suitable for predicting both continuous and categorical outcomes. The model utilizes a linear combination of hinge functions found from the data, without the need for the linearity assumption required by Logistic Regression (Nisbet et al., 2018). This makes the model suitable in predicting the outcome variable in the case where the relationship between the predictor variable and target variable is non-linear. Due to MARS adaptability and flexibility, our team chose to use it as our second model.

The MARS model was trained using the same features selected in section 2.3. To tune our model to improve its goodness of fit, we set the maximum degree of interaction of the model to 1.0. A similar procedure to that of Logistic Regression was then used to predict the outcome of both the train and test set to test the accuracy of the MARS model.

To evaluate the model, a confusion matrix was created to visualize the model accuracy on both the train set and test set.

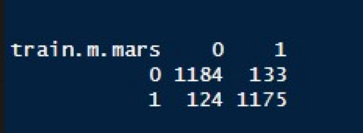


Figure 9. Confusion Matrix of the Test Set

Figure 8. Confusion Matrix of the Train Set

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | False Negative Rate | False Positive Rate | Recall | Precision |
| Train set | 0.9018 | 0.1017 | 0.0948 | 0.8983 | 0.9045 |
| Test set | 0.8855 | 0.1016 | 0.1192 | 0.8984 | 0.7315 |

From the train set confusion matrix, the model has an accuracy of 0.9018 with a false negative rate of 0.1017, false positive rate of 0.0948, recall of 0.8983 and precision of 0.9045. Compared to the test set confusion matrix with an accuracy of 0.8855 with a false negative rate of 0.1016, false positive rate of 0.1192, recall of 0.8984 and precision of 0.7315, the model is slightly less accurate on the test set, with the test set having a lower significance and lower precision as well.

## **4.4 Random Forest**

Random Forest works well with both continuous and categorical variables. Random Forest is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). The accuracy of Random Forest is high, and it can provide an estimate of variables importance. And unlike other model, Random Forest does not overfit, since the data for each tree is selected using a method called bagging which selects a random set of data points from the data set for each tree (Team, G. L., 2022).

Like the previous two models, Random Forest was trained by using the same features selected in section 2.3. To optimize the hyperparameters of the model, we found the optimal mtry value to be 5, with the lowest Out of Bag Error. The Random Forest model's accuracy was evaluated by using a similar procedure as Logistic Regression and Mars to predict the outcome of the train and test sets. To assess the model's accuracy, a confusion matrix was created to visualize how well the model performed on both the train and test sets. This allowed for a clear understanding of the model's ability to make accurate predictions.

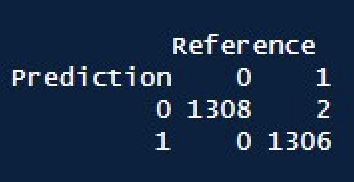
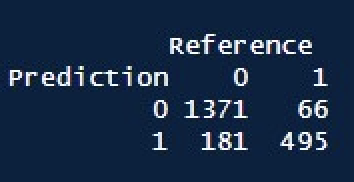


Figure 11. Confusion Matrix of the Test Set

Figure 10. Confusion Matrix of the Train Set

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | False Negative Rate | False Positive Rate | Recall | Precision |
| Train set | 0.9992 | 0.0015 | 0.0000 | 0.9985 | 1.0000 |
| Test set | 0.8831 | 0.1176 | 0.1166 | 0.8824 | 0.7322 |

Looking at the confusion matrix for the train set, it's clear that the model has a high accuracy rate of 0.9992. Additionally, the false negative rate is low at 0.0015 and the false positive rate is 0.0000. The recall and precision values are also quite high at 0.9985 and 1.0000, respectively.

However, the test set confusion matrix shows that the model's accuracy is slightly lower at 0.8831. The false negative rate is higher at 0.1176 and the false positive rate is 0.1166. The recall value is 0.8824, and the precision value is significantly lower at 0.7322. Therefore, the model is slightly less accurate on the test set, with a significantly lower precision value.

## **4.5 Model Evaluation**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | False Negative Rate | False Positive Rate | Recall | Precision |
| Logistic Regression | 0.8788 | 0.1194 | 0.1218 | 0.8806 | 0.7233 |
| MARS | 0.8855 | 0.1016 | 0.1192 | 0.8984 | 0.7315 |
| Random Forest | 0.8831 | 0.1176 | 0.1166 | 0.8824 | 0.7322 |

With reference to the table above, a comparison between the test set accuracy of Logistic Regression, MARS and Random Forest was performed. Although model accuracy is a commonly used metrics to evaluate the performance of machine learning models, our team believes that it should not be the sole deciding factor to select the model best suited for predicting customer churn. Due to the nature of the telecommunication industry, where the cost of attracting new customers outweighs the cost of maintaining existing ones, the recall would be a better measure of model performance.

Referring to the table above, the MARS model has the highest recall value out of all three models, meaning that it was the model which most accurately predicted the actual cases of churn occurrence, resulting in the lowest false negative rate. As such, our team deemed the MARS model to be the most suitable model in predicting the likelihood of a customer churning.

Compared to relying on random sampling alone, by utilizing our MARS predictive model, M1 will be able to achieve a significant improvement in their churn prediction accuracy, with an increase of more than 2 times, from 0.2654 to 0.8855. This enables them to more accurately provide the necessary corrective measures to those customers with higher risk of churning.

## **4.6 Recommendation**

On average, the monthly churn rate is about 1.9% and 2.1%, and it could increase to as high as between 10% and 67% annually for telecommunications companies (Arthur, n.d.), while the cost of acquiring a new customer is up to 25 times higher than that of retaining a customer (Heavy.AI, 2021). Hence, customer retention should be prioritized as it undoubtedly plays a key role in increasing a telecommunications company's profitability.

Yet, studies have shown that it is exceedingly difficult to change a customer’s mind once they have decided to cancel their contracts, and such efforts are usually unsuccessful, which ultimately leads to customer churn (McGrath, 2021).

Therefore, it is imperative that M1 takes the initiative and be proactive in addressing customer churn before it happens, in order to maximize the success rate of customer retention. This can be achieved by leveraging on the predictive capabilities of our analytics model, in order to pre-emptively target customers at the highest risk of churning.

### **4.6.1 Proactive Customer Retention Department**

To implement a strategy of pre-emptive customer retention as described earlier, we recommend that M1 should integrate a Proactive Customer Retention Department (PCRD) that uses models to conduct churn analytics.

As with all business strategies, there would be a cost trade-off involved with customer retention strategies. Wide scale discounts and promotions, while effective in keeping customers hooked on the service, would undoubtedly and perhaps even unnecessarily eat into M1’s revenue. Furthermore, there might even be the risk of customers experiencing promotional fatigue (Bearne, 2020), weakening marketing efforts in the long run.

Our analytics model can assist the PCRD in identifying the specific customers who are likely to churn with a high degree of accuracy, to achieve the aim of pre-emptive customer retention. At the same time, it can mitigate the costs and disadvantages associated with customer retention policies earlier described due to its specific and targeted nature, ensuring that resources and efforts are focused on customers that most need them as compared to a mass customer retention strategy in general.

To be specific, the churn analytics model would analyse customers’ metadata stored in the company’s database and provide a list of customers who are likely to leave, in a similar process as demonstrated in our proof-of-concept model. Upon identification, M1’s PCRD specialists would reach out to these customers who are at high risk of churning, to receive their feedback and understanding, as well as provide attractive and personalized offers in an effort to retain their valuable patronage.

A proactive approach would also help the company improve its products or services before customers decide to switch to another provider. Often, customers may be reluctant to share the reasons for their dissatisfaction and would express their unhappiness with the service by simply cancelling. By proactively engaging these customers and thus identifying these factors that drive customers to leave, such as slow or unstable network speed, the PCRD team can communicate with the relevant departments to address these issues promptly.

To further enhance the effectiveness of our churn analytics model, we further recommend that M1 can integrate and leverage on the synergistic benefits between a predictive model with Sentimental analysis. M1 can use social media monitoring tools to collect and analyse the data collected from a variety of digital channels, such as Instagram, Facebook, Twitter, etc. The company can then utilize the data to track customers' emotions, identify common complaints, as well as monitor trends of customer behaviour. This can also be done by analysing the survey feedback given by customers.

The aim here is to gain insights into customers’ needs and preferences. The proactive customer retention team then can use the result of sentimental analysis and be able to develop a more effective communication strategy when reaching out to the customers. By understanding their preferences and pain points, we can tailor our messaging to better resonate with them and increase the likelihood of retaining their business.

# **5. Limitations and Future Work**

## **5.1 Imbalanced Initial Dataset**

Initial dataset is extremely imbalanced before down sampling, with ratio of 5174 to 1869. Down sampling the dataset might affect the accuracy of the model performance as the dataset has been modified and less realistic.

## **5.2 Representativeness of the Data**

As mentioned earlier, the data we're working with was gathered from IBM, a U.S. company. While we've removed some columns that we deemed irrelevant to our analysis, it's important to note that the features we've selected as being most correlated with customer churn may not necessarily apply in the context of Singapore.

## **5.3 Insufficient Variable Details**

With regards to the dataset, variables such as “Offer” did not contain information on the differentiating factor between each offer. Therefore, despite “Offer” being somewhat significant in predicting the likelihood of a customer churning, we are unable to determine precisely what aspect of the offer is driving the association with customer churn. To improve on this, M1 should keep track of the detailed information of each offer to improve our ability to accurately identify actionable strategies to reduce churn.

## **5.4 Future Work**

Further research and improvements would be possible by leveraging on a more specific and targeted dataset to the Singapore context, and we would expect greater insights and accuracy in that case. The scope of machine learning in the given context is vast. For example, association rules can be used to discover patterns and relationships in customer behaviour and usage. Association rules can help M1 identify which services or products are often used together, such as Internet Service Subscription with TV and music streaming services. This could help M1 develop more competitive offers and packages to acquire new customers and at the same time retain existing ones. Lastly, data analytics can also help measure the performance of offers and campaigns and identify specific factors that affect their success or failures. This would aid in decision making and ensure the competitiveness and profitability of M1 in the increasingly competitive telco market.

# **6. Conclusion**

In summary, this report has demonstrated the use of analytical techniques in a 2-pronged approach to identify the significant variables that might lead to customer churning, as well as development of a predictive model to identify customers most at risk of churning. Based on the insights gathered, the team has also proposed 5 possible recommendations that M1 could consider, which is likely to mitigate the issue of customer churn.

In essence, the work done in this report aims to offer insights on how the business problem of customer churn can be investigated and mitigated from an analytical perspective. The team acknowledges that the analytics were done on a generic dataset, but the general insights and key takeaways remain applicable to the telecommunications industry. Further research and improvements would be possible by leveraging on a more specific and targeted dataset to the Singapore context, and we would expect greater insights and accuracy in that case.

# **7. References**

Arthur, (n.d.). *Churn reduction in the telecom industry*. Database Marketing Institute. Retrieved April 2, 2023, from [http://www.dbmarketing.com/2010/03/churn-reduction-in-the-telecom-industry/](http://www.dbmarketing.com/2010/03/churn-reduction-in-the-telecom-industry/ )

Axigen. (2022). *How to improve customer retention in the telecom industry*. Axigen. Retrieved April 2, 2023, from <https://www.axigen.com/articles/improve-customer-retention-telecom-industry_94.html>

Bearne, S. (2020, June 7). Shoppers suffer 'promotional fatigue'. Drapers. Retrieved April 2, 2023, from https://www.drapersonline.com/news/shoppers-suffer-promotional-fatigue

Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine*, *3*(1). [https://doi.org/10.1186/1751-0473-3-17](https://doi.org/10.1186/1751-0473-3-17 )

Cheok, J. (2016). Contractless mobile plans yet to catch on here. Retrieved April 2, 2023, from <https://www.businesstimes.com.sg/companies-markets/telcos-media-tech/contractless-mobile-plans-yet-catch-here>

Dignan, L. (2008). *Study: Offshore outsourcing dings customer satisfaction; taking back office offshore OK*. ZDNET. Retrieved April 2, 2023, from <https://www.zdnet.com/article/study-offshore-outsourcing-dings-customer-satisfaction-taking-back-office-offshore-ok/>

Ecseley. (2015, September 27). *Waiting on hold a minute too long*. Uniphore. Retrieved April 2, 2023, from <https://www.uniphore.com/blog/waiting-on-hold-a-minute-too-long/>

Gallo, A. (2014, November 5). *The value of keeping the right customers*. Harvard Business Review. Retrieved April 2, 2023, from <https://hbr.org/2014/10/the-value-of-keeping-the-right-customers>

Galera, R. (2019, June 29). *Hussling: 4 surefire ways to reduce customer churn*. ReferralCandy. Retrieved April 2, 2023, from <https://www.referralcandy.com/blog/reduce-customer-churn>

HEAVY.AI.(2021). Strategies for reducing churn rate in the telecom industry. Retrieved April 2, 2023, from [https://www.heavy.ai/blog/strategies-for-reducing-churn-rate-in-the-telecom-industry#:~:text=Customer%20churn%20in%20the%20telecom%20industry%20poses%20one,as%20high%20as%2067%25%20annually%20for%20prepaid%20services%2A.](https://www.heavy.ai/blog/strategies-for-reducing-churn-rate-in-the-telecom-industry#:~:text=Customer%20churn%20in%20the%20telecom%20industry%20poses%20one,as%20high%20as%2067%25%20annually%20for%20prepaid%20services%2A. )

HEAVY.AI. (2022). *Top 7 ways to optimize customer retention in the Telecom Industry*. RSS. Retrieved April 2, 2023, from <https://www.heavy.ai/blog/top-7-ways-to-optimize-customer-retention-in-the-telecom-industry#:~:text=A%20highly%20effective%20customer%20retention,%2Dsize%2Dfits%2Dall>

Hopkinson, J. (2022). *Telcos: The importance of branding in building customer loyalty*. Retrieved April 1, 2023, from <https://blog.huawei.com/2022/02/10/telcos-importance-branding-customer-loyalty/>

Hyken, S. (2021, December 10). *Businesses lose $75 billion due to poor customer service*. Forbes. Retrieved April 2, 2023, from <https://www.forbes.com/sites/shephyken/2018/05/17/businesses-lose-75-billion-due-to-poor-customer-service/?sh=a49838116f92>

Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn prediction in telecommunication using logistic regression and logit boost. *Procedia Computer Science*, *167*, 101–112. [https://doi.org/10.1016/j.procs.2020.03.187](https://doi.org/10.1016/j.procs.2020.03.187 )

Jakimovski, K. (2019, September 4). *5 proven strategies to reduce customer churn rate*. Alite International. Retrieved April 2, 2023, from <https://alite-international.com/5-proven-strategies-to-reduce-customer-churn-rate/>

Kaur, I. (2021). *Winning with customer loyalty - A practical guide*. Retrieved April 2, 2023, from <https://twimbit.com/telecoms/winning-with-customer-loyalty>

McGrath, R. (2021, September 28). *The Better Your Automated Customer Service Options, the worse your customer service experience*. Rita McGrath. Retrieved April 2, 2023, from <https://www.ritamcgrath.com/sparks/2021/04/the-better-your-automated-customer-service-options-the-worse-your-customer-service-experience/>

Mort, A. (2020, March 9). *Andrew Mort*. CustomerThink. Retrieved April 2, 2023, from <https://customerthink.com/customer-retention-in-the-telecom-industry-new-thinking/>

N.d. (November 2022).*Ericsson Mobility Report November 2022*. Retrieved April 2, 2023, from <https://www.ericsson.com/en/reports-and-papers/mobility-report/reports/november-2022>

Nisbet, R., Miner, G., & Yale, K. (2018). Advanced algorithms for Data Mining. *Handbook of Statistical Analysis and Data Mining Applications*, 149–167. [https://doi.org/10.1016/b978-0-12-416632-5.00008-6](https://doi.org/10.1016/b978-0-12-416632-5.00008-6 )

Ng, J. S. (2021). *The big read: Consumers smiling as prices tumble but is that all there is to greater telco competition?* CNA. Retrieved April 2, 2023, from <https://www.channelnewsasia.com/singapore/telcos-competition-mobile-phones-5g-singtel-tpg-m1-starhub-209121>

Qualtrics. (2023, January 19). *What is customer churn? learn how to prevent it: Qualtrics*. Qualtrics AU. Retrieved April 2, 2023, from <https://www.qualtrics.com/au/experience-management/customer/customer-churn/>

Starhub: Frequently asked questions. Frequently Asked Questions. (n.d.). Retrieved April 2, 2023, from <https://www.starhub.com/personal/support/article.html?id=824nY6LB9OAH9t2yASxIU6>

Team, G. L. (2022, October 25). *Random Forest algorithm in Machine Learning: An overview*. Great Learning Blog: Free Resources what Matters to shape your Career! Retrieved April 2, 2023, from <https://www.mygreatlearning.com/blog/random-forest-algorithm/>

# **8. Appendix**

Appendix 1

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Variable Structure | Default Data Type | Correct Data Type |
| Customer.ID | Column Dropped | | |
| Referred.a.Friend | Nominal Categorical | Character | Factor |
| Number.of.Referals | Continuous | Integer | Integer |
| Tenure.in.Months | Continuous | Integer | Integer |
| Offer | Nominal Categorical | Character | Factor |
| Phone.Service | Nominal Categorical | Character | Factor |
| Avg.Monthly.Long.Distance.Charges | Continuous | Integer | Integer |
| Multiple.Lines | Nominal Categorical | Character | Factor |
| Internet.Service | Nominal Categorical | Character | Factor |
| Internet.Type | Nominal Categorical | Character | Factor |
| Avg.Monthly.GB.Download | Continuous | Integer | Integer |
| Online.Security | Nominal Categorical | Character | Factor |
| Online.Backup | Nominal Categorical | Character | Factor |
| Device.Protection.Plan | Nominal Categorical | Character | Factor |
| Premium.Tech.Support | Nominal Categorical | Character | Factor |
| Streaming.TV | Nominal Categorical | Character | Factor |
| Streaming.Movies | Nominal Categorical | Character | Factor |
| Streaming.Music | Nominal Categorical | Character | Factor |
| Unlimited.Data | Nominal Categorical | Character | Factor |
| Contract | Nominal Categorical | Character | Factor |
| Paperless.Billing | Nominal Categorical | Character | Factor |
| Payment.Method | Nominal Categorical | Character | Factor |
| Monthly.Charge | Continuous | Integer | Integer |
| Total.Regular.Charges | Continuous | Integer | Integer |
| Total.Refunds | Continuous | Integer | Integer |
| Total.Extra.Data.Charges | Continuous | Integer | Integer |
| Total.Long.Distance.Charges | Continuous | Integer | Integer |
| Gender | Nominal Categorical | Character | Factor |
| Age | Continuous | Integer | Integer |
| Under.30 | Nominal Categorical | Character | Factor |
| Senior.Citizen | Nominal Categorical | Character | Factor |
| Married | Nominal Categorical | Character | Factor |
| Dependents | Nominal Categorical | Character | Factor |
| Number.of.Dependents | Continuous | Integer | Integer |
| City | Column Dropped | | |
| Zip.Code | Column Dropped | | |
| Latitude | Column Dropped | | |
| Longitude | Column Dropped | | |
| Population | Column Dropped | | |
| Churn.Value | Nominal Categorical | Integer | Factor |
| CLTV | Continuous | Integer | Integer |
| Churn.Category | Nominal Categorical | Character | Factor |
| Churn.Reason | Column Dropped | | |
| Total.Customer.Svc.Requests | Continuous | Integer | Integer |
| Product.Service.Issues.Reported | Continuous | Integer | Integer |
| Customer.Satisfaction | Ordinal Categorical | Integer | Ordinal Factor |

Appendix 2A

Chart, scatter chart

Description automatically generated

Appendix 2A 1: Chi-Square Test of Independence Matrix

Graphical user interface

Description automatically generated

Appendix 2A 2: Unfiltered Pearson Correlation Matrix

Chart

Description automatically generated

Appendix 2A 3: Filtered Pearson Correlation Matrix to show absolute coefficient values greater than 0.7

Chart

Description automatically generated

Appendix 2A 4: Number of Subscriptions against Monthly Charge

Chart, line chart

Description automatically generated

Appendix 2A 5: Number of Subscriptions against Total Regular Charges

Chart, histogram

Description automatically generated

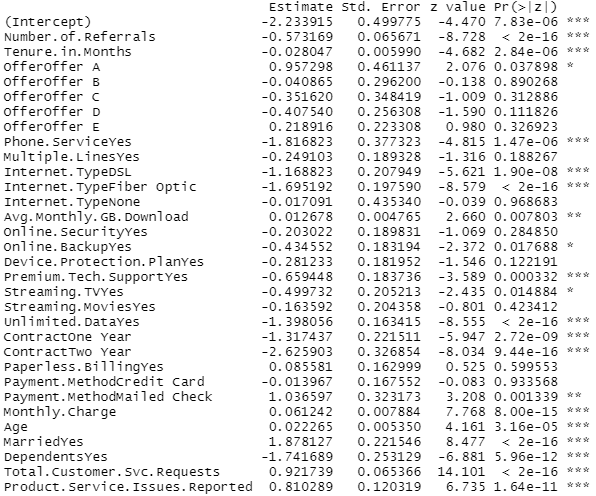
Appendix 2A 6: Tenure in months against Total Regular Charges

Appendix 2B

|  |  |
| --- | --- |
| 𝑿-variable | Relevance/Insights |
| Dependents | **Chart, bar chart  Description automatically generated** |
| Internet.Service | **Chart, bar chart  Description automatically generated** |
| Internet.Type | **Chart, bar chart  Description automatically generated** |
| Online.Security | **Chart, bar chart  Description automatically generated** |

|  |  |
| --- | --- |
| **𝑿-variable** | **Relevance/Insights** |
| Paperless.Billing | Chart, bar chart  Description automatically generated |
| Payment.Method | Chart, bar chart  Description automatically generated |
| Premium.Tech.Support | Chart, bar chart  Description automatically generated |
| Unlimited.Data | Chart, bar chart  Description automatically generated |

|  |  |
| --- | --- |
| **𝑿-variable** | **Relevance/Insights** |
| Customer.Satisfaction | Chart, bar chart  Description automatically generated |
| Monthly.Charge | Chart, box and whisker chart  Description automatically generated |
| Tenure.in.Months | Chart, box and whisker chart  Description automatically generated |
| Product.Service.Issues.Reported | Chart, box and whisker chart  Description automatically generated |

Appendix 3

**A picture containing table

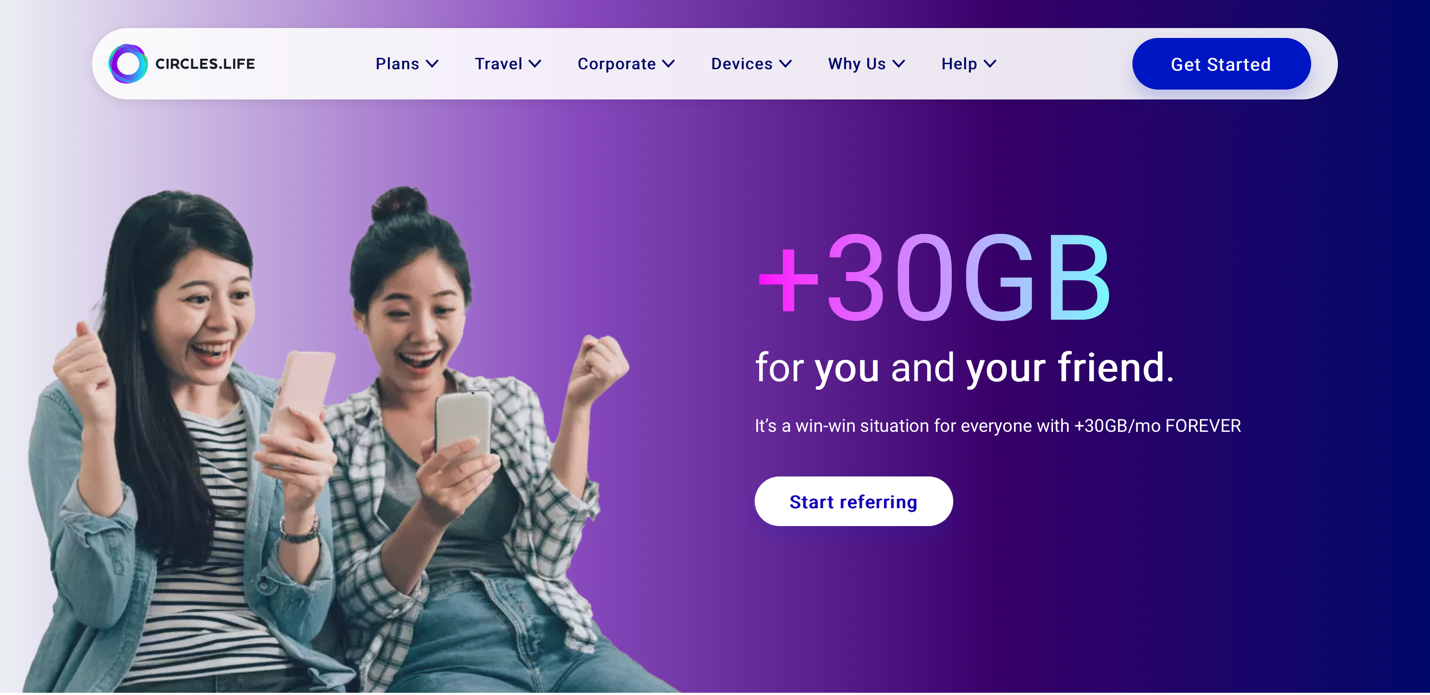
Description automatically generated**Appendix 4

Appendix 5

Table

Description automatically generated

Appendix 6



Appendix 7

Graphical user interface, application, Teams

Description automatically generated