Final Project

BT4211

Data-driven Marketing

AY 2018/2019 Semester 2

Group 4

Lee Ying Yang (A0170208N)

Tan Zheng Wei Nicholas (A0173054L)

Wang Zejia (A0142641N)

Yu Sheng Jie (A0142534M)

Table of Contents

1. Introduction

The key marketing objectives of any firm pertaining to their customers are acquisition and retention, both being key components of a firm's customer equity. The subject of this study is a telecommunications (telco) firm, and the aim is to examine several marketing issues pertaining to customer retention and customer acquisition, and propose methods that could further these marketing objectives. Since telco firms typically have a substantial user base, retention will be the greater focus over acquisition.

The first marketing problem will be churn prediction. Logistic regression will be used to model customer churn decisions and in the process, variables contributing towards notable increase or decrease in customer defection probabilities will be analysed.

The second marketing problem will be customer retention duration. Survival analysis will be utilised for modelling customer churn decisions in a continuous time setting. Similar to churn predictions, the goal is to determine the factors that contribute towards accelerating or decelerating customers' potential churn decision and thus affecting their customer lifetime duration.

The third marketing problem will be the contract choice of customers. It is common industry practice to offer customers multiple plan choices of varying lock-in durations which they cannot terminate without penalty. Longer duration contracts are more desired for telco firms, as customers are highly deterred from churning within their lock-in period, ensuring consistent long-term revenue. Multinomial logit will be used for modelling the contract choice of customers, and variables leading to higher probabilities of longer term contracts will be analysed.

The last marketing problem will be market basket analysis. Bundling strategies have long been the go-to pricing strategy for telco firms. However, key challenges are often faced in meeting the changing needs of consumers and modifying bundles to exactly meet the needs of customers whilst maintaining and improving profit margins (ResearchAndMarkets.com, 2018). Using market basket analysis through association rule mining, common co-purchases and strong purchasing associations are examined, to derive more attractive service offerings towards prospective customers.

In terms of findings, the first two analyses produced consistent results. Longer-term contracts and automatic electronic payments had positive impact on retention, and thus longer contracts and automatic payments should be encouraged by making them the default for customers. Customers with partners or dependents were less likely to churn and have longer retention rates, and thus family plans should be introduced to better cater to and entice these customers. Customers subscribing to any home Internet plan had higher churn probabilities and lower retention rates, likely reflective of dissatisfaction with home Internet services provided.

From the contract choice analysis, it was found that customers that prefer short versus long-term contracts tend to have different profiles. Senior citizens, customers without partners or dependents, and customers who required fewer number of services were more likely to select short-term monthly contracts. On the contrary, customers who have dependents or partners and those who require greater number of services had a higher probability of signing up for long-term contracts. Market basket analysis also found that customers who bought at least two additional online services

had high probabilities of needing tech support. The recommendation is thus to offer free tech support to customers who use multiple online services if they were to upgrade their contract to a two-year contract. Streaming movies and TV with device protection was also found to be frequently co-purchased, and could be bundled.

2. Literature Review

Many empirical studies have been done in the past to predict the probability of customer churning in different industries. In particular, Huang, Bingquan, et al. (2012) did a similar study in the context of a telecommunications industry. Using features such as line information, payment and account information, demographic profiles and complain information, the authors explored several prediction techniques, including logistic regression, decision trees and support vector machine, and developed an ensemble model for the prediction of churn. In our context, we used similar set of features and adopted prediction techniques such as logistic regression for the modelling. Furthermore, evaluation of models was based on classification accuracy, sensitivity and specificity, as outlined in the work by Nie, Guangli, et al. (2011).

In the telecommunications sector particularly, customer churn has also been studied in a continuous time setting, i.e. survival analysis. Wong (2011) used the Cox Proportional Hazards (PH) regression to model customers' time to churn in the Canadian wireless telecommunications industry. On the other hand, Portela & Menezes (2011) estimated a Cox PH model but found that the PH assumption based on Schoenfeld residuals did not hold. They proceeded to estimate Accelerated Failure Time (AFT) models postulating different data distributions, including exponential, Weibull, log-normal, and log-logistic. The model selected based on AIC was the log-logistic model, and the number of off-peak calls, customer spending on phone and internet plans, and number of overdue bills were among the predictors found significant in affecting survival time.

In identifying factors contributing to customers' contract choice, a multinomial choice model is used. El-Habil (2012) had used Multinomial Logit Choice model in modelling physical violence choices against children where there were multiple choice outcomes. The theoretical framework adopted was to start with exploratory data analysis for selecting a set of independent variables, based on intuition and contextual knowledge. Variables with standard errors more than 2 were further filtered out. Final model selection was based off the following criteria: computing by chance accuracy, sample size requirements, pseudo R-square, accuracy rates and goodness of fit measures such as the likelihood ratio test.

For Market Basket Analysis, prior work has been done by Agrawal, Imielinski, & Swami (1993), in which association rule mining was carried out using the dataset from a large retail company. In order to improve the computational efficiency in terms of time and memory, estimation and apriori pruning techniques were introduced to eliminate expected infrequent itemsets from consideration, instead of employing a naive approach of enumerating all possible itemset combinations. Following this, Borgelt & Kruse (2002), then Borgelt (2003) further introduced the prefix tree implementation of the apriori algorithm where branches of infrequent itemsets are pruned and stopped from further growing, thus eliminating derivations from the search space. We will be adopting an implementation of Borgelt's apriori algorithm in R for mining frequent itemsets.

3. Data Description

3.1 Dataset Overview

In this study, a cross-sectional dataset from an anonymised telecommunications firm was used. The dataset consists of information about 7043 customers of the firm, with 21 attributes ranging from demographic information, account information, to the services that the customer has signed up for. Table 1 below summarises the variables included.

Variable	Values	Description
customerID	7043 unique values	Unique Customer ID
gender	Male, Female	Whether the customer is a male or a female
SeniorCitizen	1, 0	Whether the customer is a senior citizen or not
Partner	Yes, No	Whether the customer has a partner or not
Dependents	Yes, No	Whether the customer has dependents or not
tenure	Ranged from 0 to 72	Number of months the customer has stayed with the company
PhoneService	Yes, No	Whether the customer has a phone service or not
MultipleLines	Yes, No, No phone service	Whether the customer has multiple lines or not
InternetService	DSL, Fiber optic, No	Customer's internet service provider
OnlineSecurity	Yes, No, No internet service	Whether the customer has online security or not
OnlineBackup	Yes, No, No internet service	Whether the customer has online backup or not
DeviceProtection	Yes, No, No internet service	Whether the customer has device protection or not
TechSupport	Yes, No, No internet service	Whether the customer has tech support or not
StreamingTV	Yes, No, No internet service	Whether the customer has streaming TV or not
StreamingMovies	Yes, No, No internet service	Whether the customer has streaming movies or not
Contract	Month-to-month, One year, Two year	The contract term of the customer
PaperlessBilling	Yes, No	Whether the customer has paperless billing or not
PaymentMethod	Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)	The customer's payment method
MonthlyCharges	Ranged from 18.3 to 119	The amount charged to the customer monthly
TotalCharges	Ranged from 18.8 to 8680	The total amount charged to the customer
Churn	Yes, No	Whether the customer churned or not within the last month

Table 1. Description of variables in the dataset

The original dataset contains 11 missing values in the column *TotalCharges*. For simplicity, we removed rows with such missing values, which is 0.1% of total number of rows in the dataset. As a result, the dataset contains 7032 rows and 20 variables (excluding *customerID*) that can be used in subsequent analyses.

3.2 Descriptive Analysis

In this section, the distribution of each variable in the dataset was examined to derive preliminary insights about customer behavior of the company. Additionally, the relationship of these variables with the target variable *Churn* was investigated to explore factors that could affect the possibility of customer churning.

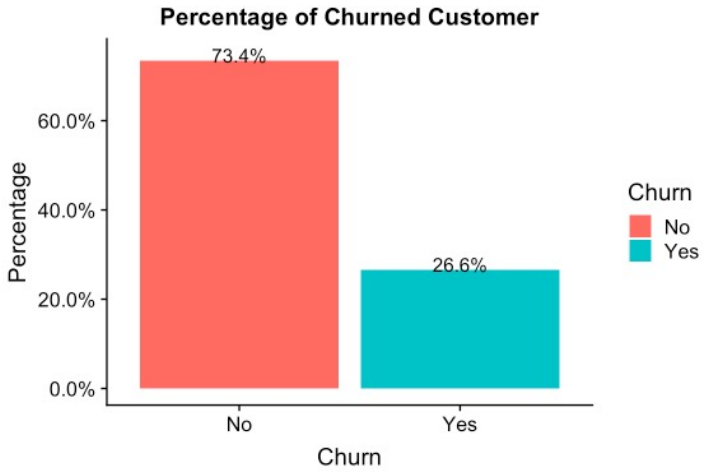


Figure 1. Percentage of churned customers

Firstly, out of 7032 customers, 26.6% of them churned in the last month (Figure 1). Moreover, churned customers tend to have a shorter tenure as compared to those who stayed with the company, as indicated by a mean tenure of 10 months and 38 months for the respective groups (Figure 2).

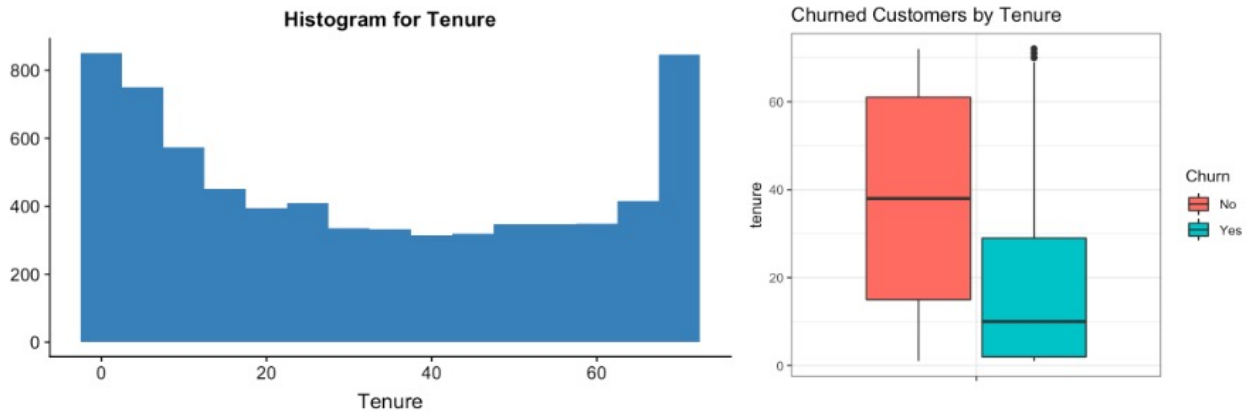


Figure 2. Distribution of tenure; Boxplot of churned customers by tenure

Monthly charges incurred by churned customers tend to be higher, possibly due to more types of services subscribed to (Figure 3). Nevertheless, the total amount charged is lower than retained customers as the tenure of churned customers is generally shorter (Figure 4). Notably, the distribution of both *MonthlyCharges* and *TotalCharges* are rightly skewed and further processing is done to discretize the variable into bins of equal interval, which will be elaborated in Section 3.3.

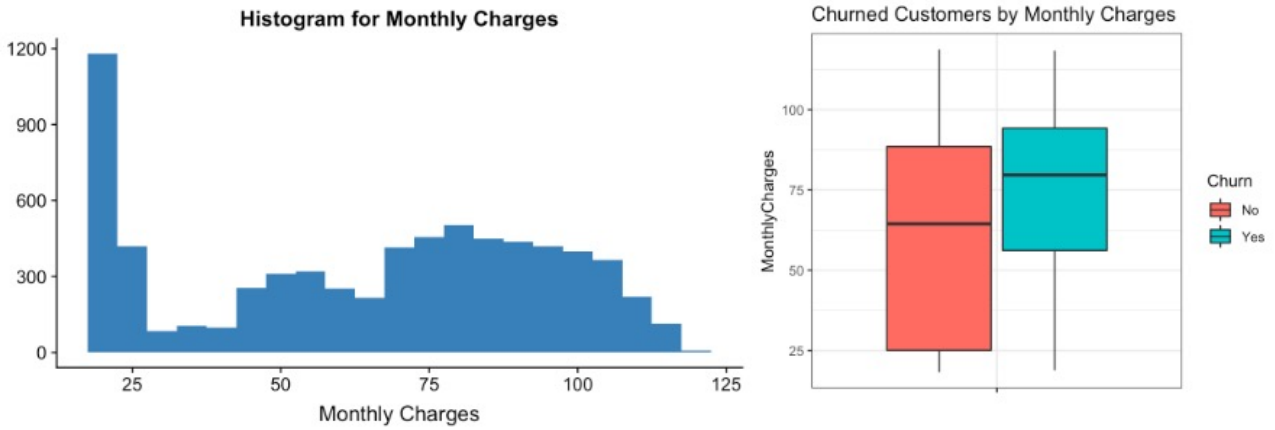


Figure 3. Distribution of monthly charges; Boxplot of churned customers by monthly charges

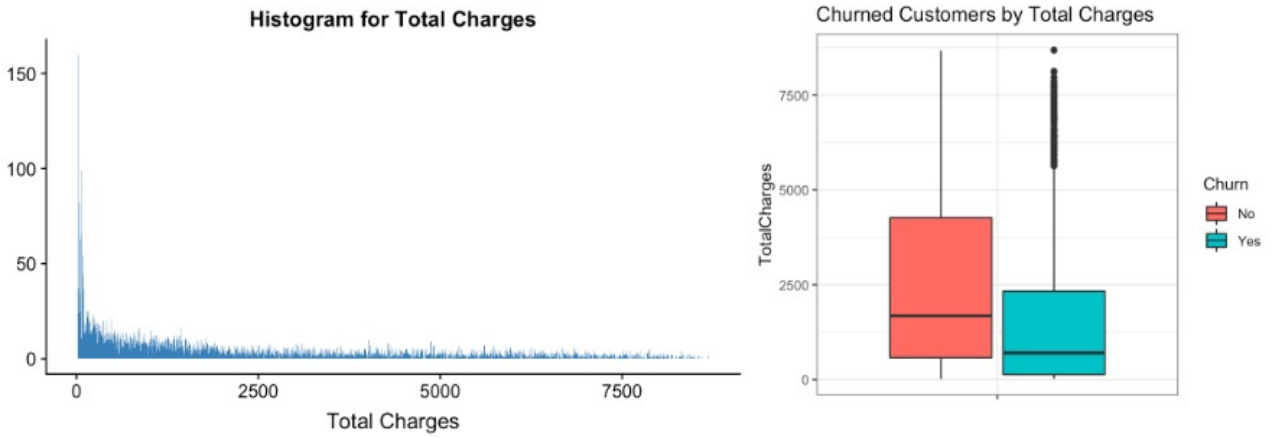


Figure 4. Distribution of total charges; Boxplot of churned customers by total charges

Figure 5 below shows the distribution of 16 categorical variables and their relationship with *Churn*. Some key insights are:

- Senior citizens tend to have higher churn rates.
- Customers without a partner or dependents tend to have higher churn rates.
- Whether the customer subscribes to phone service and/or multiple lines does not have significant impact on churn rate.
- Customers who subscribe to Internet services tend to have higher churn rates, especially those opt for fiber optic. However, among customers who have Internet services, those who purchased add-on services (e.g. online security, online backup) have lower churn rate.
- Customers under month-to-month contracts have higher churn rates than those under one-year or two-year contracts.
- Customers who opt for paperless billing or the payment method of electronic check tend to have higher churn rates than others.

Figure 5. Distribution of categorical variables; Percentage of churned customers

W3.3 Construction of New Variables

On top of the existing variables, we also constructed new variables to better capture the information required in subsequent analyses. Table 2 shows the list of new variables derived and their definition.

Variable	Description
TenureBin	Categorizes *tenure* into bins of 0-1 years, 1-2 years, 2-3 years, 3-4 years, 4-5 years, 5-6 years
MonthlyChargesBin	Categorizes *MonthlyCharges* into bins of 0-30, 30-60, 60-90, 90-120
AvgMonthlyCharges	*TotalCharges* divided by *tenure*
NumInternetAddons	Count of number of internet add-ons the customer subscribes to (incl. Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, and Streaming Movies)
PaymentMethodAuto	True if customer uses an automated payment method (Bank transfer, Credit Card), False if customer uses a manual payment method (Electronic check, Mailed check)

Table 2. List of new variables constructed

4. Churn Prediction

4.1 Model Specification

In order to predict whether a customer will churn in the following month, logistic regression is applied to model the relationship of the binary dependent variable *Churn* with a set of independent variables. These independent variables can be categorised into:

- Type of services: *PhoneService*, *MultipleLines*, *InternetService*, *OnlineSecurity*, *OnlineBackup*, *DeviceProtection*, *TechSupport*, *StreamingTV*, *StreamingMovies*
- Contract and payment information: *Contract*, *PaperlessBilling*, *PaymentMethod*
- Customer behavior: *tenure*, *MonthlyCharges*, *TotalCharges*, *AvgMonthlyCharges*, *TenureBin*, *MonthlyChargesBin*

In addition, control variables pertaining to customer demographics were included, which are *gender*, *SeniorCitizen*, *Partner* and *Dependents*.

Equation 1 below specifies, for a customer *i*, the log-odds of churn.

$$
\begin{aligned}
&LogOdds\ of\ churn, for\ customer\ i\ = \\
&\beta_0 + \beta_1 \times tenure_i + \beta_2 \times MonthlyCharges_i + \beta_3 \times TotalCharges_i + \beta_4 \times AvgMonthlyCharges_i \\
&+ \beta_5 \times gender(Male)_i + \beta_6 \times SeniorCitizen(Yes)_i + \beta_7 \times Partner(Yes)_i + \beta_8 \times Dependents(Yes)_i \\
&+ \beta_9 \times PhoneService(Yes)_i + \beta_{10} \times MultipleLines(Yes)_i + \beta_{11} \times InternetService(Fiber\ optic)_i \\
&+ \beta_{12} \times InternetService(No)_i + \beta_{13} \times OnlineSecurity(Yes)_i + \beta_{14} \times OnlineBackup(Yes)_i \\
&+ \beta_{15} \times DeviceProtection(Yes)_i + \beta_{16} \times TechSupport(Yes)_i + \beta_{17} \times StreamingTV(Yes)_i \\
&+ \beta_{18} \times StreamingMovies(Yes)_i + \beta_{19} \times Contract(One\ year)_i + \beta_{20} \times Contract(Two\ year)_i \\
&+ \beta_{21} \times PaperlessBilling(Yes)_i + \beta_{22} \times PaymentMethod(Credit\ card\ automatic)_i \\
&+ \beta_{23} \times PaymentMethod(Electronic\ check)_i + \beta_{24} \times PaymentMethod(Mailed\ check)_i \\
&+ \beta_{25} \times TenureBin(1-2\ years)_i + \beta_{26} \times TenureBin(2-3\ years)_i + \beta_{27} \times TenureBin(3-4\ years)_i \\
&+ \beta_{28} \times TenureBin(4-5\ years)_i + \beta_{29} \times TenureBin(5-6\ years)_i + \beta_{30} \times MonthlyChargesBin(30-60)_i \\
&+ \beta_{31} \times MonthlyChargesBin(60-90)_i + \beta_{32} \times MonthlyChargesBin(90-120)_i
\end{aligned}
$$

Subsequently, the probability of customer *i* churning can be computed as:

$$
Probability\ of\ churn, for\ customer\ i\ = \frac{1}{1+e^{-(RHS\ of\ log-odds\ of\ churn\ (equation\ 1))}}
$$

The dataset was then split into training and validation sets following a 70:30 ratio. Logistic regression was performed with all the independent variables included on the training set, using the *glm* function in R with the parameter *family="binomial*". Subsequently, we used stepAIC in the R MASS package for variable selection, which is a stepwise iterative process of adding or removing variables, in order to get a subset of variables that gives the best performing model. In addition, to control for multicollinearity in the model, variance inflation factor (VIF) was used to identify

variables that are highly correlated with others for removal from the model. Higher VIF values suggest higher correlation of the variable with respect to other variables in the model. In our case, variables with VIF more than 10 were removed. Moreover, variables with p-value more than 0.05 were discarded in the final model.

Table 3 below summarises the independent variables and their expected signs inferred from descriptive analysis. Variables not included in the table were expected to be insignificant as churn rates were relatively similar across the different categorical levels, as found in the descriptive analysis.

Variable	Expected sign (+: positive, -: negative)
tenure	(-) The longer the tenure, the lower the probability of churn
InternetService	Fiber optic: (+) Churn rate is higher relative to customers with DSL No: (-) Churn rate is lower relative to customers with DSL
OnlineSecurity	(-) Churn rate is lower for customers with online security
OnlineBackup	(-) Churn rate is lower for customers with online backup
DeviceProtection	(-) Churn rate is lower for customers with device protection
TechSupport	(-) Churn rate is lower for customers with tech support
Contract	One year: (-) Churn rate is lower relative to customers under month-to-month contract Two year: (-) Churn rate is lower relative to customers under month-to-month contract
PaperlessBilling	(+) Churn rate is higher for customers opt for paperless billing
PaymentMethod	Electronic check: (+) Churn rate is higher compared to bank transfer
MonthlyCharges	(-) The higher the monthly charges, the higher the probability of churn
TotalCharges	(+) The higher the total charges, the lower the probability of churn
TenureBin	(-) With respect to customers in bin 0-1 years, churn rate is lower for other bins
MonthlyChargesBin	(+) With respect to customers in bin 0-30, churn rate is higher for other bins

Table 3. Expected signs of independent variables and explanations

4.2 Model Estimation and Result

Following the methodology outlined in Section 4.1, a logistic regression model was derived for predicting a customer's churn decision. Table 4 below summarises the final model. Models in the intermediate steps of variable selection can be found in Appendix 1.

	Dependent variable:
	Churn
	Estimated coefficient (p-value)
SeniorCitizen	0.311^{***} (0.099)
Dependents	-0.204^{**} (0.097)
MultipleLines	0.352^{***} (0.095)
InternetService.xFiber.optic	1.302^{***} (0.177)

InternetService.xNo	-1.060*** (0.160)
OnlineSecurity	-0.322*** (0.101)
StreamingTV	0.392*** (0.107)
StreamingMovies	0.367*** (0.106)
Contract.xOne.year	-0.738*** (0.127)
Contract.xTwo.year	-1.654*** (0.214)
PaperlessBilling	0.333*** (0.088)
PaymentMethod.xElectronic.check	0.266*** (0.083)
TenureBin.x1.2.years	-0.879*** (0.114)
TenureBin.x2.3.years	-1.245*** (0.135)
TenureBin.x3.4.years	-1.253*** (0.148)
TenureBin.x4.5.years	-1.672*** (0.165)
TenureBin.x5.6.years	-1.868*** (0.196)
MonthlyChargesBin.x60.90	-0.646*** (0.181)
MonthlyChargesBin.x90.120	-0.972*** (0.267)
Constant	-0.603*** (0.112)
Observations	4,922
Log Likelihood	-2,055.272
Akaike Inf. Crit.	4,150.544
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4. Model summary for churn prediction

The estimated coefficients of each variable imply the unit increase in the log-odds probability of churn for a customer, given one unit increase in the focal variable while holding other variables constant. Positive coefficients indicates that the probability of churn increases as the value of the variable increases and vice versa. Most of the results are within our expectation except the fact that subscription to streaming TV or movies has a positive effect on the probability of churn. Moreover, *MonthlyChargesBin* shows negative sign, which is the opposite direction as what we have expected. This suggests that with other variables holding constant, increase in monthly charges could actually reduce the churn rate, possibly due to greater exposure of the range of services provided and the customer is locked in.

To evaluate the model fit, we compared the accuracy of prediction from the model against the baseline accuracy, which can be computed by assigning all cases to the majority in the dataset. In this case, if all cases are assigned to *Churn* = No, the baseline accuracy obtained is 73.5% in the validation dataset.

As the logistic regression model outputs the probability of churn for each customer in the validation set, we would have to choose an optimal cutoff threshold to convert the probability into a binary term of *Churn*. The optimal cutoff value was selected based on maximising accuracy, sensitivity and specificity, which is the intersection point as shown in Figure 6. Using the optimal cutoff of 0.31, we predict that a customer will churn if the probability is more than or equal to 0.31, else he/she will stay with the company.

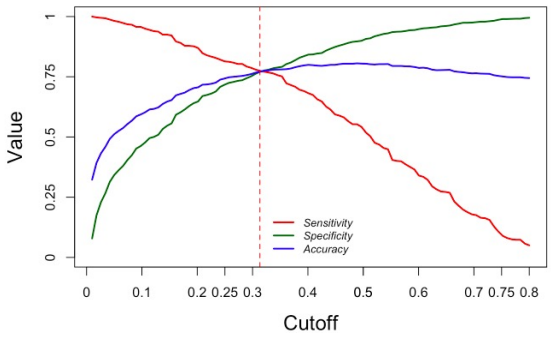


Figure 6. Selection of optimal cutoff threshold

	Actual	
Prediction	No	Yes
No	1195	126
Yes	356	433

Table 5. Confusion matrix of churn prediction

Table 5 presents the confusion matrix of the model prediction with the positive class being *Churn* = Yes. The accuracy of the model is 77.2%, which is 3.7% higher than the baseline accuracy. This suggests 77.2% of the time, the model produces accurate prediction. Moreover, sensitivity of the model is 77.5%, which implies that if a customer will actually churn, the model is able to correctly predict 77.5% of the time. Specificity of the model is 77.1% and this implies that if a customer will not churn, there is 77.1% chance that the model will predict correctly.

4.3 Discussion

Based on the logistic regression model, the company can identify customers with high risk of churning and customize marketing push or promotional activities to retain these users.

In addition, we found that senior citizens have a higher chance of churning. The company can thus target other age groups to expand its client base while organizing marketing campaigns such as offline sales events for elderly.

Moreover, customers under month-to-month contracts are more likely to churn as compared to one-year or two-year plans. Even though a month-to-month contract is more flexible and may be more effective when acquiring customers, the company should consider and develop subsequent sales plan to convert existing customers to long-term contracts.

On top of that, subscription to Internet service is found to have negative impact on customer's decision to churn, especially the subscription to Fiber Optic. This highlights the potential unsatisfactory experience customers have with regards to the Internet service provided by the company and it is worth looking into how the service can be improved.

5. Survival Analysis

In this section, survival analysis using time duration models (i.e. hazard models) are used to examine probabilities of customers churning while taking into account their length of subscription so far.

5.1 Model Specification

From the dataset, the total length of subscription for each churned customer is known. However for customers who are still subscribed, it is unknown how much longer they will stay — effectively meaning that the dataset is right-censored.

A linear regression model may be the simplest method to derive a relationship between customers' characteristics (independent variables) and subscription duration (dependent variable). However, right-censored observations have to be eliminated from the model, leading to potential censoring bias.

Hazard models have an advantage since they are designed to analyze duration data. Both semi-parametric and parametric models will be examined.

Proportional Hazards

The Proportional Hazards (PH) model is a type of semi-parametric model. It assumes that a change in the covariates scales the hazard by a proportionate amount at all times. Under the Cox PH model, the conditional hazard rate is described by the equation

$$h(t \mid X) = h_0(t)\, exp(\beta' X)$$

where h_0 is the baseline hazard and is a function of time alone. exp(β) is interpreted as the hazard ratio, where a hazard ratio above 1 implies that the covariate increases the risk of churning.

Accelerated Failure Time

The Accelerated Failure Time (AFT) model is a parametric model. Under AFT models, we measure the direct effect of the covariates on survival time instead of hazard, as in the PH model. The conditional hazard rate is described by the equation

$$h(t \mid X) = h_0(t\, exp(-\beta' X))\, exp(\beta' X)$$

The corresponding log-linear form of the model is given by

$$ln\, T_i = \mu + \beta' X_i + \sigma\varepsilon_i$$

where μ is the intercept, σ is the scale parameter, and ε_i is a random variable assumed to have a particular distribution. The usual corresponding distributions of T include Weibull, log-normal, and log-logistic. The estimates are produced using the method of maximum likelihood estimation.

$\exp(\beta)$ is interpreted as the time ratio, where a time ratio above 1 implies that the covariate prolongs the time to churn. We would expect, for instance, the choice of a two-year contract over month-to-month renewal to produce a time ratio above 1, since it is more prohibitive for customers under contracts to churn, due to factors such as cancellation fees.

5.2 Model Estimation and Result

Prior to fitting the model, several new features were created to better test our hypotheses.

The firm offers six internet add-ons for customers who subscribe to an internet plan, which include Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, and Streaming Movies. For the purpose of this analysis, the feature *NumInternetAddons* representing the number of add-ons purchased by the customer was created. Customers who purchased more add-ons are expected to stay longer with the company, as they could be more content with the services provided.

Customers currently pay their phone bills via four different methods, either manually by electronic or mailed check, or automatically by bank transfer or credit card. A boolean feature *PaymentMethodAuto* was created to differentiate automated payments and manual payments.

A Kaplan-Meier survival curve was first fitted to the data to examine the overall survival rate over time. It is observed that the probability of survival (not churning) after 20 months is around 80%.

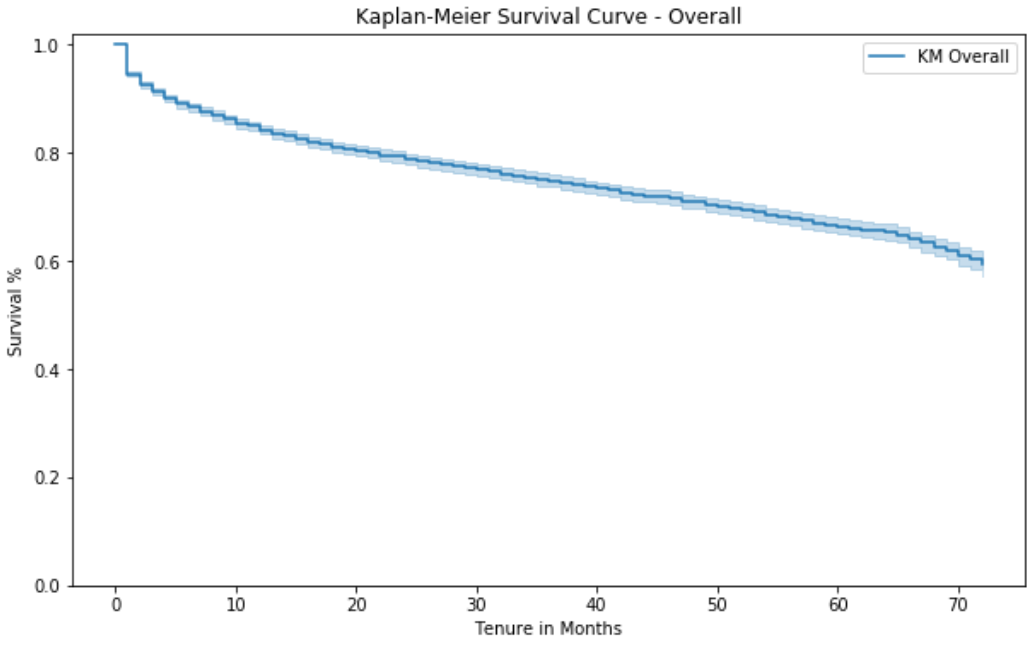


Figure 7. Kaplan-Meier survival curve

In order to account for the effect of predictor variables, we performed survival analysis using the semi-parametric and parametric models discussed earlier. The predictor variables used in our analysis include:

- Contract and payment information: *Contract*, *PaymentMethodAuto*
- Customer demographics: *Partner*, *Dependent*, *SeniorCitizen*
- Services: *MultipleLines*, *InternetService*, *NumInternetAddons*

Four models were fitted in total: the Cox PH model, Log-Logistic AFT model, Log-Normal AFT model, and Weibull AFT model.

Model	AIC
Cox Proportional Hazard	27938.603298
Accelerated Failure Time (LogLogistic)	17894.735429
Accelerated Failure Time (LogNormal)	17787.959853
Accelerated Failure Time (Weibull)	17990.617936

Table 6. Comparison of AIC across models

The Log-Normal AFT model was chosen as the best model considering it produced the lowest AIC. From the model coefficients, we calculated the time ratio by taking $\exp(\beta)$, and then subtracting 1 to obtain the relative change.

	Coefficient	p-value	Time Ratio	Relative Change (%)
SeniorCitizen	0.040	0.557	1.041	4.105
Partner	0.556	0.000	1.743	74.320
Dependents	0.119	0.109	1.126	12.605
Num_Internet_Addons	0.344	0.000	1.411	41.089
PaymentMethod_Auto	0.742	0.000	2.100	110.048
MultipleLines_No	-0.006	0.952	0.994	-0.639
MultipleLines_Yes	0.531	0.000	1.701	70.139
InternetService_DSL	-1.027	0.000	0.358	-64.181
InternetService_Fiber optic	-1.589	0.000	0.204	-79.587
Contract_One year	1.752	0.000	5.767	476.710
Contract_Two year	2.655	0.000	14.232	1323.155

Table 7. Summary of Log-Normal AFT model

Focusing on the variables that were found to be statistically significant, namely *Partner*, *Dependents*, *NumInternetAddons*, *PaymentMethodAuto*, *InternetService*, and *Contract*, we plotted the Kaplan-Meier survival curves (Figure 8) in order to compare survival probabilities over time, segmented by each of these predictor variables.

5.3 Discussion

Contract and Payment Type
Compared to customers under a month-to-month contract,

- time to churn for customers under a one-year contract is 476.7% longer
- time to churn for customers under a two-year contract is 1323.2% longer

Compared to customers who pay manually (via electronic or mailed check),

- time to churn for customers who make automatic payments (via bank transfer or credit card) is 110.0% longer

The firm should always encourage customers to sign up for longer period contracts, which will make them slower to churn. Secondly, it is recommended to make the default payment method automatic, to reduce the effort of staying subscribed - customers who set up an automatic payment method are slower to churn.

Customer Type

- time to churn for customers with a partner is 74.3% longer than customers without a partner
- time to churn for customers with dependents is 12.6% longer than customers without dependents

Customers with partners or dependents (i.e. with family members) are slower to churn, possibly due to less time to consider options. The firm should entice these customers to sign up by offering family plans.

Services Subscribed
Compared to customers who does not subscribe to an Internet service,

- time to churn for customers who subscribe to a DSL plan is 64.2% shorter
- time to churn for customers who subscribe to a Fiber Optic plan is 79.6% shorter

This indicates that customers are unsatisfied with the Internet services offered by the firm, since customers who subscribed to an Internet plan churn *faster* than those who did not.

On average, an additional Internet add-on subscription leads to 41.1% increase in time to churn. The firm should encourage internet subscribers to purchase add-ons, for example by offering product bundles.

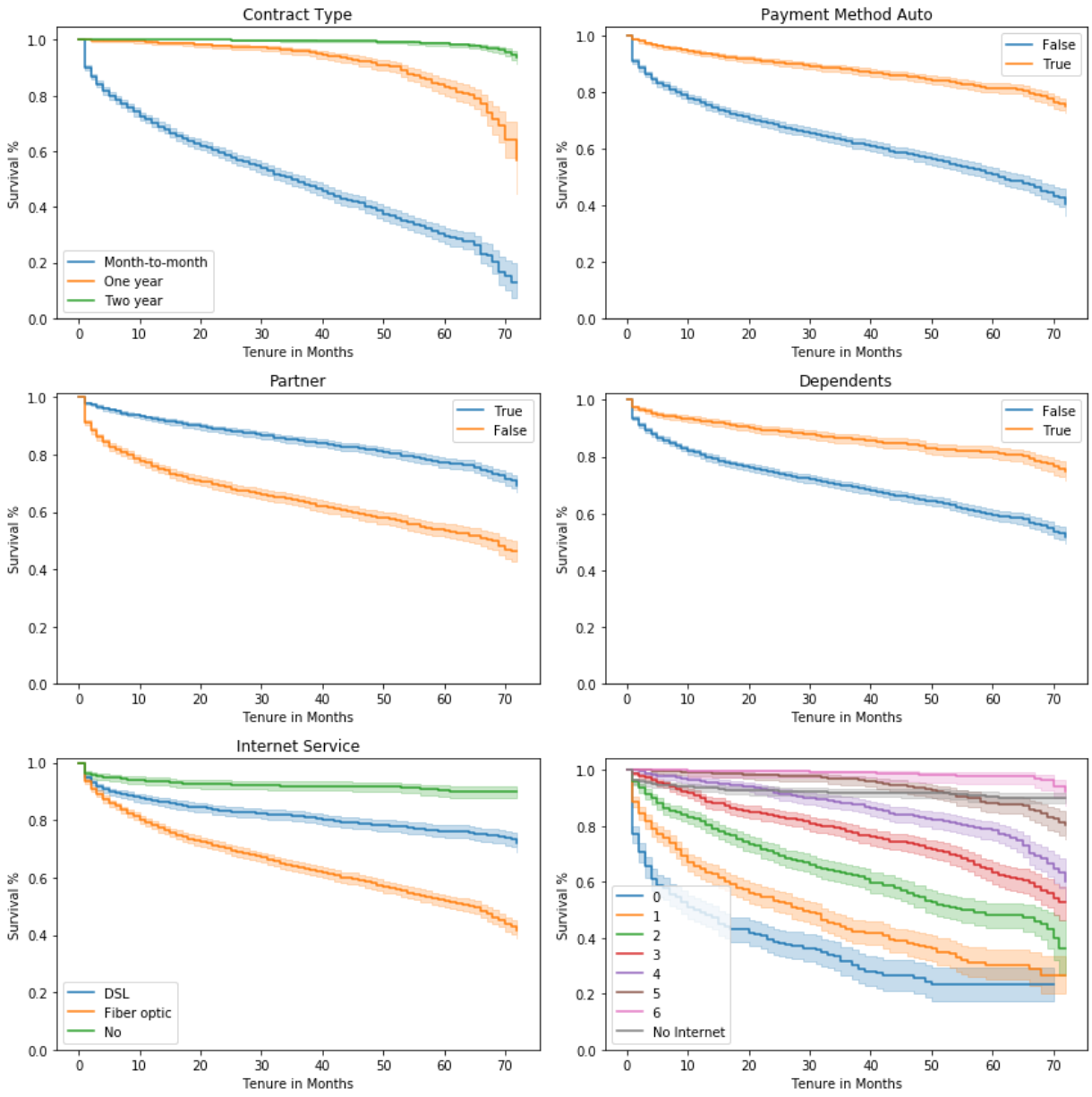


Figure 8. Kaplan-Meier survival curve, segmented by predictor variables

6. Choice of Contract

6.1 Model Specification

Baseline Category Logit Model
For the Multinomial Logit Response model, the parameter estimates can be identified relative to a baseline category.

At a fixed setting **x** for explanatory variables, let:

$$\Pi_j(x) = p(Y = j \mid x) \quad , \sum \Pi_j(x) = 1$$

At each of the J categories of Y is a multinomial with probabilities, i.e. $\{\Pi_1(x), ..., \Pi_j(x)\}$, and each response category is paired with the baseline category, with the model:

$$log\ \frac{\Pi_j(x)}{\Pi_J(x)} = \alpha_j + \beta_j' x \text{ , where j = 1,…,(J-1)}$$

The parameters of the other pairs of response and baseline categories will be determined by the other (J-1) equations. For this model, the baseline category chosen for contract choice will be the "Month-to-month" response.

Model Variable Properties
The response variables at hand here are the contract types: "Month-to-month", "One year", and "Two year". The control variable identified is *gender*.

The exogenous independent variables are as follows, along with the expected sign of their respective coefficients:
SeniorCitizens (-), *Partner* (+), *Dependents* (+), *PhoneService* (+), *InternetService* (-), *MultipleLines* (+), *OnlineSecurity* (+), *TechSupport* (+), *StreamingTV* (+), *StreamingMovies* (+), *PaperlessBilling* (+), *PaymentMethod* (+).

Similarly, the endogenous independent variables identified are:
MonthlyCharges (+), *Tenure* (+), *agg_services* (+)

Demographically, senior citizens are more likely to subscribe to month-to-month contracts, while customers with partners and dependents are more likely to subscribe to two-year contracts due to the prevalent use of bundling between multiple lines. Intuitively, customers under month-to-month contracts usually subscribe to fewer services to cut costs, and opt for paperless billing.

Churn and *TotalCharges* were dropped. While intuitively, *Churn* may be correlated to contract choice, it is a spurious correlation that defies a causal relationship due to the wrong time-order - a customer has to first choose a contract before they can possibly "Churn" from that said contract. *TotalCharges* will be a redundant variable since it can be estimated from *MonthlyCharges* multipled by *tenure*, both of which are found in the dataset.

Feature Engineering
On top of the existing variables within the dataset, we had engineered a new feature *agg_services*, that aggregates the number of contract services chosen, to investigate its impact on contract choice. The possible services include: Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies.

6.2 Model Estimation and Result

Model Estimation
Three multinomial logit models were estimated, whereby the full model includes all the variables listed in the previous section. The second model was estimated using only the control variable, which is *gender*. The third model is derived from ANOVA test (Appendix 4), by dropping insignificant variables. It includes the following variables: *SeniorCitizen, Partner, Dependents, InternetService, PaperlessBilling,* and *PaymentMethod.*

The following results (Table 8) were from our first model, which we found to be the best model. Variables that were found to be statistically insignificant were removed from the table. Add-on services with similar coefficients to that of *agg_services* were also omitted from the table as they reflect similar sentiments.

	Dependent variable:	
	Model 1 (Full Model)	
	One year	Two year
SeniorCitizen1	-0.293***	-0.571***
	(0.106)	(0.137)
PartnerYes	0.573***	1.132***
	(0.080)	(0.094)
DependentsYes	0.233***	0.304***
	(0.087)	(0.096)
PhoneServiceYes	-0.642***	-1.620***
	(0.070)	(0.092)
InternetServiceFiber optic	-0.737***	-1.052***
	(0.102)	(0.126)
PaperlessBillingYes	-0.350***	-0.508***
	(0.077)	(0.088)
PaymentMethodElectronic check	-0.892***	-1.760***
	(0.100)	(0.127)
PaymentMethodMailed check	-0.573***	-0.887***
	(0.107)	(0.115)
MonthlyCharges	-0.891***	-1.801***
	(0.045)	(0.065)
agg_services	0.436***	0.851***
	(0.018)	(0.026)
Constant	-0.891***	-1.801***
	(0.045)	(0.065)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 8. Statistically significant coefficient values for Model 1

Model Evaluation

The three models were evaluated based on model accuracy and AIC values.

Model 1's accuracy obtained is 68.69% as compared to Model 2 (55%) and 3 (62.72%) Confusion matrices for the other models can be found in Appendix 3.

	Month-to-month	One year	Two year
Specificity	0.8907	0.1088	0.7224
Sensitivity	0.6417	0.96407	0.8372

Table 9. Confusion matrix of Model 1

Model accuracy of 68.69% is higher than proportional by chance accuracy (50.5%), which was calculated with the percentage of composition of each response category, hence it fulfils this criterion (El-Habil, 2012). Insufficient data points possibly led to poor specificity for the response category "One year", which might have in turn, adversely affected the specificity of "Two year".

AIC score of Model 1 was the lowest (9892) as compared to Models 2 (14076) and 3 (12006). Hence we chose Model 1 as our final model, as it performed better on both accounts (AIC and accuracy).

Multicollinearity can affect the reliability of estimates (Garson, 2009), causing standard errors to become inflated. As such, multicollinearity between variables can be inferred from the standard errors, and variables found with standard error more than 2 will be removed (El-Habil, 2012), as a first layer of check.

Marginal Effects and Interpretation

With reference to Table 10, after dropping variables that were statistically insignificant, the marginal effects of each variable on the probabilities of the different contract choices were computed.

	"Month-to-month"	"One year"	"Two year"
SeniorCitizen1	0.0557	-0.00816	-0.0476
PartnerYes	-0.11	0.0149	0.0948
DependentsYes	-0.0364	0.0159	0.0205
InternetServiceFiber optic	0.12	-0.0448	-0.0751
InternetServiceNo	-0.008	-0.00959	0.0176
OnlineSecurityNo internet service	-0.008	-0.00959	0.0176
OnlineSecurityYes	-0.0856	0.0247	0.061
OnlineBackupNo internet service	-0.008	-0.00959	0.0176
OnlineBackupYes	-0.0435	0.0218	0.0217
DeviceProtectionNo internet service	-0.008	-0.00959	0.0176
DeviceProtectionYes	-0.0766	0.0229	0.0537
TechSupportYes	-0.104	0.00185	0.102
PaperlessBillingYes	0.0574	-0.0208	-0.0366
PaymentMethodElectronic check	0.171	-0.0232	-0.147
PaymentMethodMailed check	0.0969	-0.0305	-0.0664
MonthlyCharges	0.173	-0.0206	-0.152
agg_services	-0.0829	0.0119	0.071

Table 10. Marginal effects based on Model 1

With reference to table 10, key indicators that lead to a higher probability of selecting each contract type were identified.

Senior citizens have a higher probability of selecting “Month-to-month” contracts, with the coefficient value of 0.0557. Subscribers of this type of contract also tend to not have dependents (-0.0364) nor partners (-0.11) and are likely subscribed to Fiber Optic internet services, while not being subscribed to the other services such as *OnlineSecurity*, *DeviceProtection*, and *TechSupport*. This tallies with the negative coefficient of *agg_services* (-0.0829).

It was harder to identify clear variables that were closely related to customers opting for “One year” contracts. In general, the coefficients share the same sign as that of the “Two year” contracts, while having a smaller magnitude.

As for “Two year” contracts, subscribers tend to exhibit profiles that were the opposite of the “Month-to-month” contract subscribers. They have dependents and partners, are not subscribed to Internet service, and are usually subscribed to other services such as *OnlineSecurity*, *DeviceProtection*, and *TechSupport*. A positive coefficient was obtained for *agg_services* (0.071).

6.3 Discussion

Each group is demographically different, with the “Month-to-month” customers being more price-sensitive, especially since they tend to be senior citizens and/or are without partners or dependents. From the lower number of services subscribed, they are likely to seek flexibility in their contracts. The “Two year” group are customers who have dependents and partners, who are most likely working class, that prefer more add-on services. However they tend to not subscribe to the home internet services, as they might already have internet on their cellphone plans.

Recommendations will seek to encourage customers under a “One year” contract to opt for a “Two year” contract, as well as to identify optimal product bundles. To avoid cannibalisation between broadband service and cellphone internet services, the firm can limit the data usage of cellphone internet services. This opens up another viable option for bundling whereby a broadband internet service can be bundled with base plans, to encourage customers to take up the internet service as part of the two-year plan.

Since “Month to month” contract subscribers are more likely to have a fiber optic internet plan, the firm could offer an Internet Bundle (Internet Service, Online Security, Online Backup, Device Protection) for customers subscribed to this type of contract. This would be beneficial especially for senior citizens who are less tech-savvy, and would be less informed about internet security.

7. Market Basket Analysis

7.1 Model Specification

In market basket analysis, association rule mining is utilised. The first variable of interest is support of a rule, representing the probability that both an antecedent itemset combination and the consequent item of interest appears together in all transactions. The second variable of interest is the confidence of an association rule, which refers to the probability of an item of interest is purchased given a specific preceding itemset.

$$Support = P(E \cap I)$$
$$Confidence = P(E \mid I) = \frac{P(E \cap I)}{P(I)}$$

In the equations, I denotes an antecedent itemset and E denotes the consequent item purchase of interest. Although having high confidence in a rule is important, the lift ratio must still be considered. A lift ratio greater than 1 indicates that a customer buys the consequent item of interest more frequently given that the customer is also purchasing the antecedent itemset.

$$Lift\ Ratio = \frac{P(E \mid I)}{P(E)}$$

In computation however, checking over all possible antecedent itemset combinations in order to find frequent itemsets is highly inefficient due to the sheer number of possible combinations (Appendix 5) that grows greatly with each additional item being introduced.

Therefore, the Apriori algorithm implementation by Borgelt (2003) is used in computation, in which a prefix tree is used to represent possible itemset combinations (Appendix 6), along with the heuristic that all supersets containing an infrequent smaller itemset are all also consequently infrequent. All derivations of infrequent itemsets below a certain support threshold are thus pruned from the tree, removed from the search space to improve computational efficiency.

In our analysis, all itemsets that have support below 0.05, which corresponds to less than 5% of all transactions, are pruned. Association rules with confidence below 0.1 are also not considered. This is to ensure discovery of association rules that have substantial transaction volume to be relevant in marketing efforts. The maximum size of the antecedent itemset is also restricted to five items, in order to help with the interpretability of association rule results.

7.2 Model Estimation and Result

The resulting top 10 rules in terms of lift-ratio from association rule mining were extracted and analysed. The following table shows the antecedent itemsets, consequent items, support, confidence, lift ratio, and transaction counts of the association rules.

Antecedent	Consequent	Support	Confidence	Lift Ratio	Count
InternetService=DSL, OnlineSecurity=Yes, StreamingTV=Yes	TechSupport=Yes	0.05310237	0.7290448	2.512066	374
PhoneService=Yes, OnlineSecurity=Yes, DeviceProtection=Yes, StreamingTV=Yes, StreamingMovies=Yes	TechSupport=Yes	0.05253443	0.7182836	2.474986	385
InternetService=DSL, OnlineSecurity=Yes, StreamingMovies=Yes	TechSupport=Yes	0.05466421	0.7167530	2.469712	415
PhoneService=Yes, OnlineSecurity=Yes, DeviceProtection=Yes, StreamingMovies=Yes	TechSupport=Yes	0.06460315	0.7032457	2.423170	
InternetService=DSL, OnlineSecurity=Yes, DeviceProtection=Yes	TechSupport=Yes	0.06346727	0.7028302	2.421738	447
InternetService=DSL, DeviceProtection=Yes, StreamingTV=Yes	TechSupport=Yes	0.06005963	0.7026578	2.421144	423
OnlineSecurity=Yes, DeviceProtection=Yes, StreamingMovies=Yes	TechSupport=Yes	0.07340622	0.6986486	2.407330	517
InternetService=DSL, OnlineBackup=Yes, StreamingTV=Yes	TechSupport=Yes	0.05026267	0.6954813	2.396416	354
OnlineSecurity=Yes, DeviceProtection=Yes, StreamingTV=Yes	TechSupport=Yes	0.07056652	0.6951049	2.395119	497
PhoneService=Yes, OnlineSecurity=Yes, DeviceProtection=Yes, StreamingTV=Yes	TechSupport=Yes	0.06233139	0.6935229	2.389668	439

Table 11. Top 10 association rules in terms of lift ratio, at 0.05 minimum support

There are overlaps between the above association rules, however it is immediately apparent that Tech Support is an item that is very frequently bought with other online services such as Streaming TV, Streaming Movie, Online Security, Online Backup, and Device Protection with a very high confidence of >0.69 and high lift ratio. It is also observable within these 10 instances, that antecedent itemsets leading to Tech Support purchase include at least two or more online services.

Tech Support is also currently the second least purchased item for the firm (Appendix 7), and therefore it is possible that the high lift ratio is partially due to the low initial base purchase probability of the item itself.

It is observed that that most of the high lift ratio association rules have support nearing the 0.05 limit. In order to discover more insights, another round of mining was run with the minimum support threshold raised to 0.10 to find association rules that may have lower lift but higher transaction frequencies. The following are the top 10 rules.

Antecedent	Consequent	Support	Confidence	Lift Ratio	Count
OnlineSecurity=Yes, DeviceProtection=Yes	TechSupport=Yes	0.1012353	0.6417642	2.211323	713
TechSupport=Yes, StreamingTV=Yes, StreamingMovies=Yes	DeviceProtection=Yes	0.1000994	0.7468220	2.171704	705
PhoneService=Yes, MultipleLines=Yes, DeviceProtection=Yes, StreamingMovies=Yes	StreamingTV=Yes	0.1171376	0.8291457	2.157249	825
PhoneService=Yes, InternetService=Fiber optic, DeviceProtection=Yes, StreamingMovies=Yes	StreamingTV=Yes	0.1137299	0.8283351	2.155140	801
DeviceProtection=Yes, TechSupport=Yes, StreamingTV=Yes	StreamingMovies=Yes	0.1000994	0.8353081	2.153395	705
PhoneService=Yes, MultipleLines=Yes, DeviceProtection=Yes, StreamingTV=Yes	StreamingMovies=Yes	0.1171376	0.8341759	2.150476	825
InternetService=DSL, TechSupport=Yes	OnlineSecurity=Yes	0.1029391	0.6154499	2.146911	725
PhoneService=Yes, InternetService=Fiber optic, DeviceProtection=Yes, StreamingTV=Yes	StreamingMovies=Yes	0.1137299	0.8300518	2.139844	801
DeviceProtection=Yes, TechSupport=Yes, StreamingMovies=Yes	StreamingTV=Yes	0.1000994	0.8150289	2.120520	705
InternetService=DSL, OnlineSecurity=Yes	TechSupport=Yes	0.1029391	0.6144068	2.117058	725

Table 12. Top 10 association rules in terms of lift ratio, at 0.1 minimum support

Consistent with the results of the previous analysis, the first rule with the highest lift ratio has Tech Support as a consequent of two other online services.

From the rest of the association rules generated, a common pattern observed is that Device Protection, Streaming TV and Streaming Movies are frequently purchased together and antecedent itemsets that contain two of the above had the third as the consequent item. 7 out of 10 of the association rules held this pattern.

7.3 Discussion

It is likely that customers have a greater preference in purchasing Tech Support when they are also purchasing multiple online services. This could be due to customers feeling the need for extra help with the variety of online services that will potentially need setup and troubleshooting.

Streaming TV and Streaming Movies being regularly bought together makes logical sense due to the complementary nature of TV series and movies. Device Protection being regularly bought with these other two services could be due to customers frequently purchasing these streaming services to be used on their mobiles devices for entertainment consumption on the move.

With these insights, several recommendations can be made:

Firstly, if customers sign up for a minimum of two or three online services, Tech Support could be offered for free or heavily discounted if customers are willing to sign up to a two-year contract. This could be used to encourage more customers to opt for longer contracts and help the firm better lock in customers and improve retention.

The second recommendation will be to provide bundling for Streaming TV, Streaming Movie and Device Protection services. Furthermore, this bundle can be more aggressively pushed towards customers who have multiple lines subscribed, since it likely means that they own multiple mobile devices. The services proposed could be utilized and shared across all of their devices.

8. Summary of Findings and Implications

Having identified four marketing problems (churn prediction, survival analysis, contract choice, market basket analysis) to investigate the different aspects of consumer behaviour, we consolidated the insights obtained and came up with the following recommendations.

Key, resonant insights are that:

Customer demographics are important in determining a customer's behaviour. Consumers with dependents and partners are more likely to opt for "Two year" plans, which in turn, results in lower probability to churn as well as taking a longer time to churn. These customers are less likely to have Internet Service, while being more likely to subscribe to add-ons, which suggests that internet services are being shared across family members. Among the add-ons, Tech Support is most likely purchased with Internet-related options, while Device Protection is mostly likely purchased with Streaming TV and Streaming Movies.

In general, an internet plan subscription, especially the fiber optic option, led to higher probability of churn as well as shorter time to churn. This suggests that customers are dissatisfied with the internet services provided. Internet plan subscription is also associated with a month-to-month contract choice, although these customers are typically less likely to purchase internet add-on services.

In response, the firm can adopt the following recommendations:

Firstly, it is imperative to improve the internet services provided to increase customer satisfaction. In doing so, there is potential in upselling customers on internet add-on services, as well as converting month-to-month contract subscribers into longer-term subscribers. The purchase of add-on services and longer term contracts are associated with lower churn probabilities, which translates to better customer retention for the firm.

The firm can consider offering an Internet Services Plan that bundles Internet Service with various add-ons to improve the attractiveness of their offerings. Customers who opt for a two-year contracts can be given a special price, to encourage more subscriptions for longer term contracts. Based on our market basket analysis, potential product bundles are as follows:

Internet Service (compulsory), with the option of
A: DeviceProtection, StreamingTV, StreamingMovies, and/or
B: TechSupport, OnlineSecurity, OnlineBackup

To encourage customers to opt for the bundled services, the firm can increase the price of individual add-ons.

Branding strategies for the two-year contracts can be more generic, seeking to connect with audience on universally relatable points and focusing on value (as positive coefficient of *MonthlyCharges* in churn prediction suggests that consumers are price sensitive). This can help to attract different groups, such as the senior citizens, encouraging them to opt for the two-year contract. Monthly contracts can be positioned to be flexible and low commitment products instead, with limited services. Additionally, the firm can set up family plans to acquire more customers with partners or dependents, who are found to retain better in the long run.

9. Conclusion

The four marketing problems identified in our study, namely churn prediction, retention duration, contract selection, and market basket analysis were explored and analyzed using separate modeling techniques. However, in reality, customer purchase and churn behaviors are more intricate, and the error terms could be correlated across several of the models estimated. Further studies could explore the use of simultaneous equation models, such as a three-stage least squares (3SLS) estimation to achieve a more holistic analysis of customer behavior.

Moreover, the dataset that was used in our study is a cross-sectional one. While the dataset was sufficient to produce insights based on time-invariant customer characteristics such as demographics, it lacks additional information which could have been obtained in a panel dataset, such as monthly billings and services subscribed which are subject to changes over the course of a customer's subscription period. Analyzing the problems identified using panel data would further substantiate the results obtained, as it can control for unobserved customer behavioural characteristics.

In summary, the analyses performed in our study provide valuable insight on customer behavior within the telecommunications industry, and is helpful to the firm when evaluating their marketing and customer retention strategies.

10. References

Agrawal R., Imielinski T. , & Swami A. (1993), Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 207–216.

Borgelt C. (2003), Efficient Implementations of Apriori and Eclat. *Workshop of Frequent Item Set Mining Implementations*.

Borgelt C. & Kruse R. (2002), Induction of Association Rules: Apriori Implementation. *15th Conference on Computational Statistics*.

El-Habil, A. M. (2012), An Application on Multinomial Logistic Regression. *Pakistan Journal of Statistics and Operation Research.*

Garson, D. (2009), Logistic Regression with SPSS. Retrieved from https://faculty.chass.ncsu.edu/garson/PA765/logistic.htm

Huang, B., et al. (2012), Customer Churn Prediction in Telecommunications. *Expert Systems with Applications, vol. 39, no. 1, 2012,* 1414–1425. doi:10.1016/j.eswa.2011.08.024.

Nie, G., et al. Credit Card Churn Forecasting by Logistic Regression and Decision Tree. *Expert Systems with Applications, vol. 38, no. 12, 2011*, 15273–15285. doi:10.1016/j.eswa.2011.06.028.

Portela, S. & Menezes, R. (2011), Detecting customer defections: an application of continuous duration models, *Journal of Global Strategic Management (09)*, 22–30.

ResearchAndMarkets.com (2018), Global Telco Bundling Strategies 2018 - The Evolution of the Telco Bundle. Retrieved from https://www.businesswire.com/news/home/20180209005703/en/

Rodríguez, G. (2019), *GR's Website*. Retrieved from https://data.princeton.edu/wws509/r/mlogit

Wong, K. K.-K. (2011), Using Cox regression to model customer time to churn in the wireless telecommunications industry, *Journal of Targeting, Measurement and Analysis for Marketing 19(1)*, 37–43.

Appendix

Appendix 1: Summary of models for churn prediction

	Dependent variable: Churn		
	With all variables	stepAIC	Final model
tenure	-2.469***	-2.320***	
	(0.324)	(0.283)	
MonthlyCharges	-1.146	-0.579***	
	(1.225)	(0.217)	
TotalCharges	0.217		
	(0.212)		
AvgMonthlyCharges	0.382		
	(0.390)		
gender	0.005		
	(0.078)		
SeniorCitizen	0.313***	0.316***	0.311***
	(0.101)	(0.101)	(0.099)
Partner	0.006		
	(0.094)		
Dependents	-0.188*	-0.188*	-0.204**
	(0.108)	(0.099)	(0.097)
PhoneService	0.107		
	(0.790)		
MultipleLines	0.538**	0.536***	0.352***
	(0.216)	(0.104)	(0.095)
InternetService.xFiber.optic	1.700*	1.647***	1.302***
	(0.969)	(0.221)	(0.177)
InternetService.xNo	-1.623	-1.499***	-1.060***
	(1.009)	(0.225)	(0.160)
OnlineSecurity	-0.167	-0.181*	-0.322***
	(0.215)	(0.107)	(0.101)
OnlineBackup	-0.048		
	(0.212)		
DeviceProtection	0.085		
	(0.212)		
TechSupport	-0.095		
	(0.219)		
StreamingTV	0.609	0.571***	0.392***
	(0.394)	(0.119)	(0.107)
StreamingMovies	0.570	0.541***	0.367***
	(0.396)	(0.117)	(0.106)
Contract.xOne.year	-0.636***	-0.644***	-0.738***
	(0.130)	(0.129)	(0.127)
Contract.xTwo.year	-1.335***	-1.371***	-1.654***
	(0.218)	(0.215)	(0.214)
PaperlessBilling	0.370***	0.365***	0.333***
	(0.089)	(0.089)	(0.088)

PaymentMethod.xCredit.card..automatic.	0.039		
	(0.136)		
PaymentMethod.xElectronic.check	0.247**	0.235***	0.266***
	(0.114)	(0.084)	(0.083)
PaymentMethod.xMailed.check	0.003		
	(0.140)		
TenureBin.x1.2.years	0.391**	0.373**	-0.879***
	(0.189)	(0.188)	(0.114)
TenureBin.x2.3.years	1.107***	1.095***	-1.245***
	(0.313)	(0.313)	(0.135)
TenureBin.x3.4.years	2.204***	2.220***	-1.253***
	(0.447)	(0.445)	(0.148)
TenureBin.x4.5.years	2.836***	2.873***	-1.672***
	(0.578)	(0.574)	(0.165)
TenureBin.x5.6.years	3.727***	3.820***	-1.868***
	(0.722)	(0.713)	(0.196)
MonthlyChargesBin.x30.60	0.061		
	(0.308)		
MonthlyChargesBin.x60.90	-0.238	-0.343	-0.646***
	(0.420)	(0.211)	(0.181)
MonthlyChargesBin.x90.120	-0.427	-0.521*	-0.972***
	(0.514)	(0.316)	(0.267)
Constant	-3.975**	-3.762***	-0.603***
	(1.619)	(0.387)	(0.112)
Observations	4,922	4,922	4,922
Log Likelihood	-2,014.195	-2,016.474	-2,055.272
Akaike Inf. Crit.	4,094.390	4,076.948	4,150.544

Note: *p<0.1; **p<0.05; ***p<0.01

Appendix 2: Summary of models for contract choice

	Dependent variable:					
	Model 1 (Full Model)		*Model 2*		*Model 3*	
	One year	Two year	One year	Two year	One year	Two year
genderMale	0.088	0.110	0.038	-0.007		
	(0.070)	(0.080)	(0.061)	(0.058)		
SeniorCitizen1	-0.293***	-0.571***			-0.304***	-0.561***
	(0.106)	(0.137)			(0.096)	(0.112)
PartnerYes	0.573***	1.132***			0.803***	1.431***
	(0.080)	(0.094)			(0.074)	(0.079)
DependentsYes	0.233***	0.304***			0.207**	0.253***
	(0.087)	(0.096)			(0.081)	(0.081)
tenure	-0.891***	-1.801***				
	(0.045)	(0.065)				
PhoneServiceYes	-0.642***	-1.620***				
	(0.070)	(0.092)				
MultipleLinesNo phone service	-0.250***	-0.181**				
	(0.074)	(0.089)				
MultipleLinesYes	-0.144*	0.177*				
	(0.082)	(0.095)				
InternetServiceFiber optic	-0.737***	-1.052***			-0.478***	-0.714***
	(0.102)	(0.126)			(0.077)	(0.084)
InternetServiceNo	0.0004	0.152***			0.378***	0.858***
	(0.014)	(0.016)			(0.093)	(0.090)
OnlineSecurityNo internet service	0.0004	0.152***				
	(0.014)	(0.016)				
OnlineSecurityYes	0.498***	0.799***				
	(0.079)	(0.098)				
OnlineBackupNo internet service	0.0004	0.152***				
	(0.014)	(0.016)				
OnlineBackupYes	0.289***	0.345***				
	(0.077)	(0.098)				
DeviceProtectionNo internet service	0.0004	0.152***				
	(0.014)	(0.016)				
DeviceProtectionYes	0.448***	0.709***				
	(0.080)	(0.104)				
TechSupportNo internet service	0.0004	0.152***				
	(0.014)	(0.016)				
TechSupportYes	0.494***	1.148***				
	(0.080)	(0.100)				

StreamingTVNo internet service	0.0004	0.152***				
	(0.014)	(0.016)				
StreamingTVYes	0.168*	0.165				
	(0.089)	(0.113)				
StreamingMoviesNo internet service	0.0004	0.152***				
	(0.014)	(0.016)				
StreamingMoviesYes	0.214**	0.168				
	(0.089)	(0.113)				
PaperlessBillingYes	-0.350***	-0.508***			-0.252***	-0.361***
	(0.077)	(0.088)			(0.070)	(0.073)
PaymentMethodCredit card (automatic)	0.131	0.125			0.116	0.146
	(0.105)	(0.112)			(0.096)	(0.092)
PaymentMethodElectronic check	-0.892***	-1.760***			-1.048***	-1.973***
	(0.100)	(0.127)			(0.091)	(0.106)
PaymentMethodMailed check	-0.573***	-0.887***			-0.789***	-1.203***
	(0.107)	(0.115)			(0.099)	(0.098)
MonthlyCharges	-0.891***	-1.801***				
	(0.045)	(0.065)				
agg_services	0.436***	0.851***				
	(0.018)	(0.026)				
Constant	-0.891***	-1.801***	-0.986***	-0.823***	-0.512***	-0.551***
	(0.045)	(0.065)	(0.044)	(0.041)	(0.096)	(0.098)
Akaike Inf. Crit.	9,892.495	9,892.495	14,076.320	14,076.320	12,006.240	12,006.240

Note: *p<0.1; **p<0.05; ***p<0.01

Appendix 3: Confusion matrices for Multinomial Logit models

Model 1

	Month-to-month	One year	Two year
Specificity	0.8907	0.1088	0.7224
Sensitivity	0.6417	0.96407	0.8372

Model 2

	Month-to-month	One year	Two year
Specificity	1	0	0
Sensitivity	0	1	1

Model 3

	Month-to-month	One year	Two year
Specificity	0.8632	0	0.6043
Sensitivity	0.4963	1	0.7985

Appendix 4: Type III ANOVA Test for Model 1

Response: Contract

	LR Chisq	Df	Pr(>Chisq)
gender	2.35	2	0.308666
SeniorCitizen	19.04	2	7.335e-05 ***
Partner	152.57	2	< 2.2e-16 ***
Dependents	11.46	2	0.003249 **
tenure	0.00	2	1.000000
PhoneService	0.00	2	1.000000
MultipleLines	0.00	4	1.000000
InternetService	82.79	4	< 2.2e-16 ***
OnlineSecurity	0.00	4	1.000000
OnlineBackup	0.00	4	1.000000
DeviceProtection	0.00	4	1.000000
TechSupport	0.00	4	1.000000
StreamingTV	0.00	4	1.000000
StreamingMovies	0.00	4	1.000000
PaperlessBilling	37.15	2	8.560e-09 ***
PaymentMethod	338.50	6	< 2.2e-16 ***
MonthlyCharges	0.00	2	1.000000
agg_services	0.00	2	1.000000

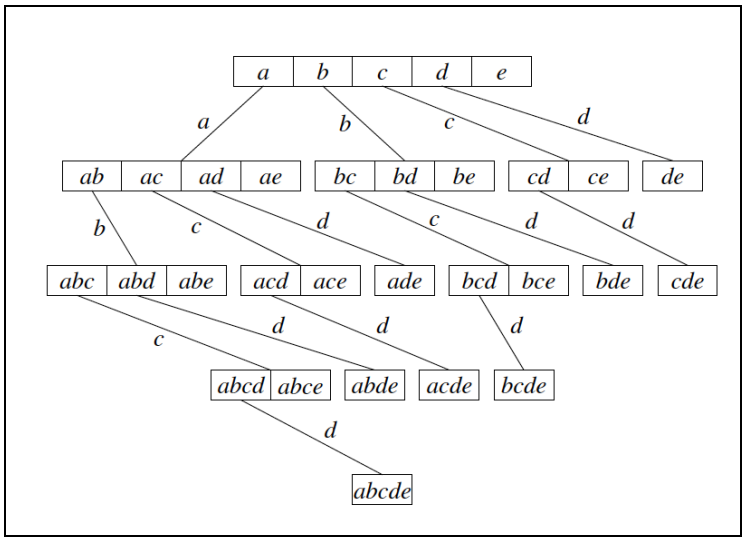
Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Appendix 5: Formula for number of possible preceding itemsets

Approximation of possible number preceding itemset combinations, n denotes number of total items that can be bought.

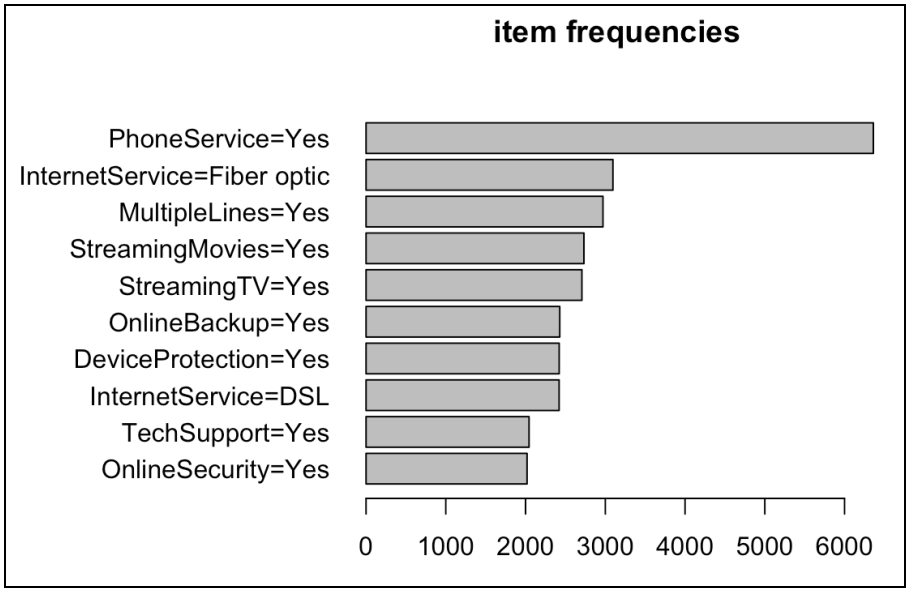
$$Approximation\ of\ antecedent\ itemset\ combinations\ possible\ =\ \sum_{i=2}^{n} nC(i-1)$$

Appendix 6: Prefix Tree in use for association rule mining



Itemset prefix tree, from Christian Borgelt (2003)

Appendix 7: Item Frequency in Transactions



Number of transactions that involves each item