SCHOOL OF MATHEMATICAL & COMPUTER SCIENCES

DEPARTMENT OF Actuarial Mathematics and Statistics

DISSERTATION SUBMISSION

Please complete all of the following:

Programme	BSc (Hons) Statistical Data Science
Year of Study	Year 3, Malaysia, 2021-22, Semester 1
Surname	Tan
First Name	Yen Shen
ID No	H00336563
Course Code	F70DA
Course Title	Statistics Dissertation A
Title of Project	Customer Behaviour Analytics based on Machine Learning Classification Techniques
Supervisor(s)	Supervisor – Sarat Dass Co-supervisor – Benjamin Choong
Date of Submission	24th November 2021

I affirm that this dissertation is my own work, and I have not copied material without giving adequate references.	**Signed**__________________ **Dated**________24th November 2021________

Acknowledgements

Firstly, I would like to express my sincere gratitude to Dr Sarat Dass, my Main Supervisor, for the continuous guidance during my project, for his patience and immense knowledge.

In addition, I would like to thank Benjamin Choong, my Co-Supervisor, for his continuous support throughout my project, which has been incredibly helpful.

I look forward to working on my final piece in Semester 2 with the guidance of both parties.

Table of Contents

List of tables

List of Figures

Abstract

Customer churn is one of the most imperative issues that is connected with the existing cycle of businesses. The purpose behind churning observed from a business viewpoint is that customers leave and switch to another business causing a loss of revenue. In this thesis, we explore the fundamental issue - what are the types of customers who churn versus those who do not churn? Can we identify specific customer characteristics that determine a high likelihood of churning versus not churning?

The aim of this project is to explore the variables that affect churn based on the information provided from the Telcos dataset. Variables that carry information on customer demographics and behaviour, such as the kind of packages purchased by them and how long they have used the company's service, are explored to determine whether they affect churn. To comprehend this issue, we utilised two approaches: First, a model free variable selection method was developed to determine whether a variable affected churning, and second, a logistic linear regression model was developed in R to examine the types of variables that are significant enough to cause churn.

Keywords - Customer Churn, Variable Selection, Logistic Regression

1. Introduction

One of the biggest elements that businesses look into on determining the growth and competition of the company is customer churn. The goal of businesses is to first acquire new customers, and then engage with them at satisfactory levels of service thereby those existing customers. It has been revealed that gaining new customers has higher costs compared to maintaining and retaining existing customers. (J.Hadden 2006)

In this thesis, I will use Telco's Customer Churn data to identify customer characteristics that affect churn. Telco is a global leader in telecommunications with over 40 years of experience in the industry of design and development of network communication solutions. As defined by the Article 1.3 of the Radio Regulations (RR), telecommunication, or in short telecom, is defined as 'any transmission, emission or reception of signs, signals, writings, images and sounds or intelligence of any nature by wire, radio, optical or other electromagnetic systems.'

1.1 Basic Concepts

1.12 Customer Churn

Brought forward by Berson et.al (2000), 'customer churn' is a term used in the wireless telecom service industry to denote customer movement from one provider to another. 'Customer management' is a term that describes a company's process to retain profitable customers.

In a business model, the company segments customers based on profitability and focuses on retaining only those that are profitable. However the telecom service industry is yet to standardise a set of 'profitability' measurements.(see, for example, Hung et.al (2006) For example, current versus life-time, business unit versus corporate, etc.)

There are mainly two types of customer churn: voluntary and involuntary churners. Voluntary churners, also known as active churners, are customers who willingly choose to quit the service and move to the next provider. On the other hand, involuntary churners are customers who the company chooses to terminate the service provided to them, typically due to poor payment history (Blattberg et al 2008). It is important to note, that when referring to the customer churn problem, we are generally referring to the voluntary churners.

It was then further recognised by Hadden et al. (2006) that there are 2 types of voluntary churners, namely 'deliberate' and 'incidental'. 'Deliberate' voluntary churn occurs when the customer is unsatisfied or chooses to take on another better competitive deal in another company, and 'incidental' voluntary churn is when the customer chooses to cancel the service because they no longer require it or due to unforeseen circumstances rendering the services unusable.

As a result of customer churn, there are switching costs (Klemperer 1988). Switching costs include transaction cost (financial) and learning cost (non-financial). It's a type of cost that comes into play when customers choose to churn, in choosing one company's product over another. An example of a reason for the former switching cost is that it could be due to cheaper alternatives from identical products. An example of a reason for the latter is that having to re-learn a new product, time could come as a cost.This project aims to distinguish what are the types of variables that affect Churn in telecommunications, particularly Telco's.

1.13 How is Customer Churn calculated?

Customer churn is focused on the retention element, *r.* At the customer level, churn refers to the probability the customer leaves the company in a given time period. At the company level, churn is the percentage of the company's customer base that leaves in a given time period. Churn is thus one minus the retention rate:

$$\text{Churn} = c = 1 - \text{Retention Rate} = 1 - \text{r}$$

For businesses, the retention rate is defined as the percentage of customers that the business retains over a period of time. In other words, this percentage represents the loyal customers of the business.

The formula for retention rate is as follows

$$\text{Retention Rate} = \frac{Number\ of\ customers\ who\ continue\ business\ for\ this\ given\ time\ period}{Total\ number\ of\ customers\ at\ the\ beginning\ of\ that\ period} \times 100$$

1.14 Why is it important to calculate Customer Churn?

The elements of churn and retention rate are very important factors in determining the lifetime value (LTV) of a customer of that business. This in turn determines how viable the business is and it's growth. The Lifetime Value (LTV) is termed as 'how much revenue a customer is estimated to deliver across their entire time buying from the business'.

Robert et.al (2008) uses a simple retention model where the lifetime value of a customer is:

$$LTV = \sum_{t=1}^{\infty} \frac{m_t r^{t-1}}{(1+\delta)^{t-1}}$$

with:

$$m = Annual\ profit\ contribution\ per\ customer$$
$$c = Annual\ churn\ rate$$
$$\delta = Annual\ discount\ rate$$
$$r = retention\ rate$$
$$t = time$$

Companies that have a subscription service, like telecommunications companies, have a standardised way of measuring customers' value, which is coined as the Average Revenue Per User (ARPU). The calculation for ARPU would be calculated as

$$ARPU = Total\ Revenue\ /\ Average\ Customers$$

Despite this being a way for companies to measure the value of their customers, churn does not affect ARPU. Instead, reducing churn will drive up LTV. By decreasing the rate at which customers leave the service, we increase the chance of them staying in the service longer.

2. Literature Review

Understanding the factors that affect customer churn allows companies to make important business decisions for the company, because ultimately having customers, and especially loyal customers, is what keeps companies functioning based on higher profit.

There have been many studies conducted in the field of customer churn where different forms of analyses methods and different machine learning and data mining strategies were introduced. Machine Learning techniques like regression models on customer behaviour, decision tree classification and many more have been introduced into the studies of customer churn.

There is an abundant amount of data collected by telecommunications industries. Thus, to make sense of these data, a diverse use of machine learning techniques are introduced. In a particular study conducted, it proposed a binary classifier based on Naive Bayes and decision trees (Kirui et. al., 2013). On the other hand, there was a study that introduced a binomial logistic regression model (Ismail et al., 2015). There are many other studies that conducted various machine learning approaches such as logistic regression (Owczarczuk, 2010), support vector machine (Coussement & Poel, 2008), gradient boosting (Idris et.al 2012) , k-means (Hung et.al., 2006) and many more.

Huang et.al (2012) studies a customer churn prediction on telecommunication with seven techniques namely Logistic Regressions, Linear Classifications, Naive Bayes, Decision Trees, Multilayer Perceptron Neural Networks, Support Vector Machines and the Evolutionary Data Mining Algorithm, with special features of customer demographic profile, informations on bills and customer account. The study noted the effectiveness of each machine learning approach.

There have been other studies conducted based on churn prediction in the mobile telecom industry (Hadden et al., 2006, Hung et al., 2006, John et al., 2007, Luo et al., 2007, Wei and Chiu, 2002).

Based on all the studies mentioned above, a lot of research and analysis has been done on churn prediction and less on what are the types of variables that directly affect churn. We will bring to light in this current project, the types of variables that affect churn with the approach of proposing a model free variable selection approach and a Logistic Regression Method.

3. Data Source of Telco Customer Churn Dataset

Content of Telco Customer Churn data

An individual customer of Telco represents each row, and the customer's attributes are depicted by each column of the data.

The data set includes information about:

- Customers who left within the last month – this column is called Churn.
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
- Customer account information – how long they've been a customer- tenure, the type of contract, payment method, paperless billing, monthly charges, and the accumulation of the total charges.
- Demographic about customers – gender, whether they are seniors, and if they have partners and dependents.

<u>Definition of each column heading</u>

Variables are the column headings from the given Telco dataset.

Values are the output of the variables, it answers the question of - what kind of values are representing the given variable and what do they mean?

Type represents what kind the variables are - whether they are numerical, character or factor. This allows us to better understand the type of data variables in the given dataset.

Character - They are variables that are in the form of strings. An example is for variable customerID with values such as '5129-JLPIS' and '8865-TNMNX'.

Factor - They are character/numerical variables that are limited to a number of unique character strings or numbers. This is often represented as a categorical variable. An example could be for the gender variable which has levels 'Female' and 'Male'.

Numerical - They are variables that are numbered with or without decimals.

No	Variable	Values	Type
1	customerID	7043 unique values	Character
2	gender	**(Male, Female)**	Factor
3	SeniorCitizen	**(1,0)** Whether the customer is a Senior Citizen or not	Factor
4	Partner	**(Yes, No)** Whether customer has a Partner or not	Factor
5	Dependents	**(Yes, No)** Whether customer has Dependents or not	Factor
6	Tenure	Number of months the customer has stayed with the company	Numerical
7	PhoneService	**(Yes, No)**	Factor

		Whether customer has a PhoneService or not	
8	MultipleLines	**(Yes, No, No phone service)** Whether customer has a PhoneService or not	Factor
9	InternetService	**(DSL, Fiber Optic, No)** No means No Internet Service Customer's Internet Service Provider	Factor
10	OnlineSecurity	**(Yes, No, No internet service)** Whether the customer has online security or not	Factor
11	OnlineBackup	**(Yes, No, No internet service)** Whether the customer has online backup or not	Factor
12	DeviceProtection	**(Yes, No, No internet service)** Whether the customer has tech support or not	Factor
13	TechSupport	**(Yes, No, No internet service)** Whether the customer has tech support or not	Factor
14	StreamingTV	**(Yes, No, No internet service)** Whether the customer has streaming tv or not	Factor
15	StreamingMovies	**(Yes, No, No internet service)** Whether the customer has streaming movies or not	Factor
16	Contract	**(Month – to- month, One year, Two year)** The contract term of the customer	Factor
17	PaperlessBilling	**(Yes, No)** Whether the customer has paperless billing or not	Factor
18	PaymentMethod	**(Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))** The customer's payment method	Factor
19	MonthlyCharges	The amount charged to the customer monthly	Numerical
20	TotalCharges	The **total** amount charged to the customer	Numerical
21	Churn	**(Yes, No)** Whether the customer churned or not	Factor

Table 3.1 21 Variables value and their type

4. Exploratory Variables Analysis

For continuous variables:

Several statistical tests are given below with regard to testing whether a specific variable in the Telco Customer Churn dataset affects churn or not. The tests given below are (i) the Kolmogrov-Smirnov test, and (ii) the Mann-Whitney U test

4.11 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov goodness of fit test can be used to test if a given sample comes from a population with a specific distribution, commonly the normal distribution. It is used for samples that arises from continuous distributions, and does not require any underlying distribution in the data for it to be used (therefore, distribution-free)

Let $X_1, X_2, \dots, X_n$ be a random sample. The empirical distribution function $F_0(x)$ is a function of x, which equals the fraction of X_is that are less than or equal to x for each x, $-\infty < x < \infty$.The empirical distribution function $F_0(x)$ is used as estimator of $F(x)$, the unknown distribution function of the X_is. This statistic is suggested by Kolmogorov (1933) for the following hypotheses testing scenario:

$$H_0: F(x) = F_0(x) \qquad \text{vs} \qquad H_0: F(x) \neq F_0(x)$$

where $F_0(x)$ is a prespecified cumulative distribution function. The above test of hypotheses is carried out using the Kolmogorov-Smirnov test statistic.

$$D = sup_x \left| F_0(X) - S(X) \right|$$

where D is the greatest vertical distance between $S(X)$ and $F_0(X)$. Under H_0, the test statistic D follows the level α test which will reject H_0 if $D > D\{\alpha\}$ where $D\{\alpha\}$ is the upper tail of the significant level.

4.12 Results

We shall perform manual calculation on the distribution of tenure given that they do not churn represented as x.

The hypothesis will follow as:

$$H_0: F(x) = F_0(x) \qquad \text{vs} \qquad H_0: F(x) \neq F_0(x)$$

where $F_0(x)$ is a normal distribution that follows mean and variance $N(0, 1)$

and it follows that $F(x) = P(X \leq x) = P(Z \leq x)$

$D_1 = F(x) - F_{n_1}(x)$ and $D_2 = F_{n_2}(x) - F(x)$

The values are calculated in R, and are tabulated below.

Table 4.12.1 Kolmogorov-Smirnov Table for tenure given that there is no churn

x	Freq	$F_{n_1}(x)$	$F(x)$	$F_{n_2}(x)$	D_1	D_2
0	11	0	0.5	0.01369863	0.5	-0.4863014
1	233	0.01369863	0.8413447	0.02739726	0.82764612	-0.8139475
2	115	0.02739726	0.9772499	0.04109589	0.94985261	-0.936154
3	106	0.04109589	0.9986501	0.05479452	**0.95755421**	-0.9438556
…	…	…	…	…	…	
…	…	…	…	…	…	…
70	108	0.95890411	1	0.97260274	0.04109589	-0.0273973
71	164	0.97260274	1	0.98630137	0.02739726	-0.0136986
72	356	0.98630137	1	1	0.01369863	0

Observing Table 4.12.1, the largest test statistic, D, is achieved at 0.9576. The table value of *D* at 5% significance level, which is calculated as

$$\frac{1.36}{\sqrt{n}} = \frac{1.36}{\sqrt{73}} = 0.159176$$

Since the calculated value of 0.9576, is greater than the critical value of 0.1592. Hence, we reject the null hypothesis and conclude that the distribution of variable tenure given no Churn does not follow a normal distribution.

For this given data set, we will test the normality for the other continuous variables for Tenure, Monthly Charges and Total Charges corresponding to whether they churn or don't churn. Using the Kolmogorov-Smirnov built-in function in R, we get the following test statistics, *D*, for each continuous variable.

Table 4.12.2 Output table from R of test statistic for Kolmogorov Smirnov test

Variables	No Churn	Yes Churn
Tenure	0.930 (0.9576)	0.841
Monthly Charges	1.000	1.000
Total Charges	0.998	1.00

The *p*-value of all variables was less than 2.2×10^{-16}.Under any significant level between 1% to 10%, the null hypothesis will be rejected. This suggests that the variables of Tenure, Monthly Charges and Total Charges, given that they did churn or did not, does not follow a normal distribution. It's important to note that the values of manual calculation and in R are different due to the presence of ties, and therefore are normally approximated in R. However, despite having different test statistics, *D*, the conclusion is the same.

Hence, equality of distribution given churn and not churn can not be performed using the standard two sample t-test. For this reason, we use the non-parametric Mann-Whitney U test to compare the equality between two arbitrary distributions.

4.21 Mann-Whitney U Test

The unpaired two samples Wilcoxon test also referred as Wilcoxon Rank Sum Test or Mann-Whitney test (Mann & Whitney 1947). This test is used to test the equality of means between two independent groups. Since this is a non-parametric test, it does not require the data to be normally distributed - which has been shown in Results in section 4.12, that the variables do not follow a normal distribution.

Given that we have two independent samples:
With reference from Shier 2004, the corresponding test statistic is defined as

$$U = min(U_1, U_2)$$

U is then given by :

$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$ where n_1 is the sample size of sample 1 and R_1 is the sum of he ranks in sample 1

$U_2 = R_2 - \frac{n_2(n_2+1)}{2}$ where n_2 is the sample size of sample 2 and R_2 is the sum of he ranks in sample 2

Note that the ranks of the given sample will require reassigning from the original data. With tied values, one has to assign a rank equal to the midpoint of the unadjusted rankings. For a large sample dataset, it is best to use programming languages in R to compute the order at which ranks are placed. The given hypothesis for the non-parametric test are stated as

H_0: *Populations are equal* *vs* H_1: *Populations are not equal*

Since the number of observations in our given dataset (7043 rows and 21 columns) is so large ($n_1 n_2 > 20$), normal approximation has been used with $\mu_x = \frac{n_1 n_2}{2}$, $\sigma_x = \sqrt{\frac{n_1 n_2 (N+1)}{12}}$, where $N = n_1 + n_2$

Dealing with ties: Normal approximation must be adjusted to the standard deviation,

$$\sigma_{ties} = \sqrt{\frac{n_1 n_2}{N(N-1)} \times \left[\frac{N^3 - N}{12} - \sum_{j=1}^{g} \frac{t_j^3 - t_j}{12}\right]}$$

where $N = n_1 + n_2$

g = the number of group of ties

t_j = the number of tied ranks in group j

For large samples in the presence of ties, U is approximately normally distributed. In this case, the standardised value is such

$$z = \frac{U - \mu_x}{\sigma_{ties}}$$

Referring to the normal distribution table for the standardised distribution, if the p-value is less than the significant value of 0.05, we reject the null hypothesis.

4.22 Results

Hand calculation can be very tedious as that process of ranking all the variables and reassigning the rank based on duplicates and the order. Therefore, we shall perform the calculation on the distribution of tenure corresponding to the churn variable using the built-in function in R.

The hypothesis will follow as:

H_0: The population of customers' tenure given that they churned or did not churn are equal

H_1: The population of customers' tenure given that they churned or did not churn are not equal

```
        Wilcoxon rank sum test with continuity
        correction

data:  data$tenure[data$Churn == "Yes"] and data$tenure[data$Churn == "No"]
W = 2515538, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Figure 4.22.1 Output figure from R

W in the Output Figure 4.22.1 mentioned above is the test statistic where, $W = min(U_1, U_2)$. The p-value is less than 2.2×10^{-16} which is less than the significant value, and thus, the null hypothesis is rejected. This implies that the population of customers' tenure given that they churned or did not churn are not equal.

This hypothesis statement will apply for the other numerical variables such as Monthly Charges and Total Charges.

Using the results from R, the table below shows the test statistics, U, for each continuous variable.

Table 4.22.1 Output table for Mann-Whitney U Test

Variables	**Test statistic, U**	***P*-value**	**Significant at 5%?**
Tenure	2 515 538	2.419636e-208	Yes
Monthly Charges	6 003 126	3.311628e-54	Yes
Total Charges	3 381 224	5.685034e-83	Yes

The p-value of all variables was less than 2.2×10^{-16}. The exact values are calculated above in Table 4.22.1. Under any significant level between 1% to 10%, the null hypothesis will be rejected. This suggests that the variables of the continuous data (Tenure, Monthly and Total Charges) had different populations for churned and not churned, thus implying that they could influence the outcome of churn in a prospective way.

For categorical variables:
Chi- Squared of Independence statistical test is given below with regard to testing whether a specific variable in the Telco Customer Churn Dataset affects churn or not.

4.31 Chi- Squared of Independence

Introduced by Pearson (1900), the chi-square (χ^2) test of independence is considered to be one of the most important fundamental test in modern statistics (William, 1952). This test is used to test the relationship between two categorical variables. This is a non-parametric test, thus is a distribution free test - data does not need to fulfill normality for this test to be used. It also assumes that the data is independent, and uses this fact to compute the expected values in the cells in a two-way contingency table.

When calculating the chi-squared of independence for 2 categorical variables, here is the hypothesis that we are testing

$$H_0: The\ 2\ categorical\ variables\ are\ independent$$
$$H_1: The\ 2\ categorical\ variables\ are\ dependent$$

$$\text{Chi-Square Test Statistic: } \chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

$$\text{Expected Cell Value: } Expected = \frac{Row\ Total \times Column\ Total}{n}$$

$$\text{Degree of Freedom:} \quad df = (no.\ of\ rows - 1)(no.\ of\ columns - 1)$$

The Observed value is the frequency of that value. If the test statistic is large according to the chi-squared distribution with the given degree of freedom, then reject the null hypothesis of independence.

4.32 Results

We shall perform the manual calculation on the distribution of Gender corresponding to the churn variable

The hypotheses is as follows:

H_0: There is no relationship between Gender and whether someone has churned or not (they are independent)

H_1: There is a relationship between Gender and whether someone has churned or not (they are dependent)

Observed and expected counts are presented together in Table 4.32.1.

First, we calculate the expected values for each cell. The calculation is as follows:

$$E_{Female,No} = \frac{3488 \times 5174}{7043} = 2562.39,\ E_{Female,Yes} = \frac{3488 \times 1869}{7043} = 925.6101$$

$$E_{Male,No} = \frac{3555 \times 5174}{7043} = 2611.61,\ E_{Male,Yes} = \frac{3555 \times 1869}{7043} = 943.3899$$

Note that all expected values are at least 5, thus assumption of the χ^2 test independence has been met. The observed and expected counts are represented in Table 4.32.1. The expected values are presented in parentheses.

Table 4.32.1 Contingency chi-squared table of independence on Gender and Churn

	Has this customer churned?		
	No	**Yes**	**Sum of Row**
Female	2 549 (2 562.39)	939 (925.6101)	3 488
Male	2 625 (2 611.61)	930 (943.3899)	3 555
Sum of Column	5 174	1 869	

Next, we will use Yate's correction of continuity on the Chi-Square Test Statistic, which is introduced to prevent overestimation of statistical significance as it is aims to correct the error made by assuming that the discrete probabilities of frequencies in the table in the table can be approximated by continuous distribution.

$$\chi^2_{Yates} = \sum \frac{(|Observed - Expected| - 0.5)^2}{Expected} = \frac{(|2\,549 - 2\,562.39| - 0.5)^2}{2\,562.39} + \frac{(930 - 943.3899)^2}{943.3899}$$

$$+ \frac{(|939 - 925.6101| - 0.5\,)^2}{925.6101} + \frac{(|2\,625 - 2\,611.61| - 0.5)^2}{2\,611.61} = 0.4840829$$

The degree of freedom is, $df = (2 - 1)(2 - 1) = 1$

Calculating the p-value, we have that $P(\chi^2_{Yates,\, df = 1} > 0.4841) = 0.4866$. Since this value is more than the significant level at 0.05. Therefore there is no evidence against the null hypothesis and implies that Gender and churn are independent of each other. This means that we retain the null hypothesis and reject the alternative hypothesis. In other words, Gender does not affect churn.

For this given data set, we will test the normality for the other categorical variables like Senior Citizen, Partner, etc… corresponding to whether they churn or don't churn. Using the Chi-Squared of Independence built-in function in R, we get the following test statistic, χ^2, degree of freedom, *df* and p-value for each categorical variable.

Table 4.32.2 Table output on the chi-squared of independence

	Variable	**Test Statistic, χ^2**	**Degree of Freedom, *df***	***P*-value**	**Significant at 5% ?**
1	gender	0.484	1	0.487	**No**
2	SeniorCitizen	159.426	1	1.51E-36	Yes
3	Partner	158.733	1	2.14E-36	Yes
4	Dependents	189.129	1	4.92E-43	Yes
5	PhoneService	0.915	1	0.339	**No**
6	MultipleLines	11.330	2	0.003	Yes
7	InternetService	732.310	2	9.57E-160	Yes
8	OnlineSecurity	849.999	2	2.66E-185	Yes
9	OnlineBackup	601.813	2	2.08E-131	Yes
10	DeviceProtection	558.419	2	5.51E-122	Yes
11	TechSupport	828.197	2	1.44E-180	Yes
12	StreamingTV	374.204	2	5.53E-82	Yes
13	StreamingMovies	375.661	2	2.67E-82	Yes
14	Contract	1184.597	2	5.86E-258	Yes
15	PaperlessBilling	258.278	1	4.07E-58	Yes
16	PaymentMethod	648.142	3	3.68E-140	Yes

For the variables that were not significant in the 5% level, only variable gender and PhoneService. This suggests rejecting the null hypothesis and that variables gender and PhoneService are independent of Churn, and thus does not have an effect on Churn.

Whereas for the other variables, as stated in Table 4.32.1 that they are significant in the 5% significant level, and thus are dependent on Churn. More analysis will be done below after the proportion tables and figures and have been analysed.

5. Figures and Tables

This section gives the result of the variables in regards to Churn (excluding CustomerID because each value is unique). The variables that we are trying to explore (19 variables) are given in the subsection below and will be accompanied from Section 4's analysis on their *p*-value and hypothesis test for that given variable. The tables and figures are derived from R and the derivation can be found in the appendix.

5.1 Gender

For gender, the table 5.1.1 gives the numerical values responding to the probabilities. Figure 5.1.1 gives the corresponding figure to the conditional distribution of Gender given Churn is either No or Yes.

Table 5.1.1 Gender Proportions

Gender	Churn	
	No	Yes
Female	0.4927	0.5024
Male	0.5073	0.4976

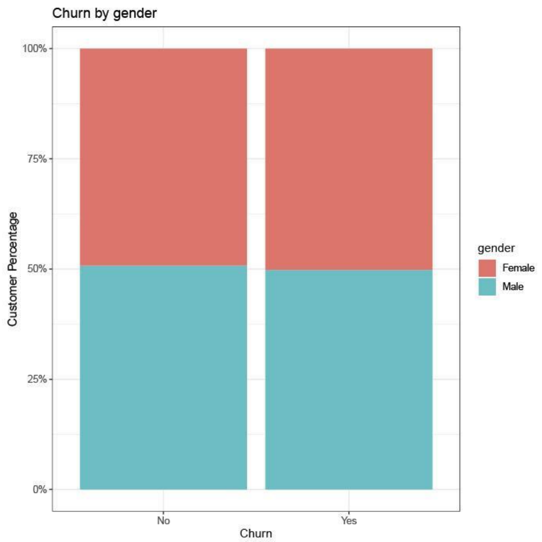

Figure 5.1.1 Figure showing the conditional distribution of gender given churn outcome

From Section 4.32, the hypotheses is as follows:

H_0: There is no relationship between Gender and Churn (they are independent)

H_1: There is a relationship between Gender and Churn (they are dependent)

From the table and the figure above, and with a chi-squared *p*-value of 0.4866 from Table 4.32.2. Therefore there is no evidence against the null hypothesis and implies that Gender and churn are independent of each other. This means that we retain the null hypothesis and reject the alternative hypothesis. In other words, Gender does not affect churn.

This is a demographic factor, based on the results, it is clear that Telcos products do not have an influence over more or less than a particular gender. This helps us better understand the target groups that use telecommunications services- that it is a service used by all.

5.2 Senior Citizen

For Senior Citizen, the table gives the numerical values responding to the probabilities. Figure 5.2.1 gives the corresponding figure to the conditional distribution of Senior Citizen given Churn is either No or Yes.

Table 5.2.1 Senior Citizen Proportions

Senior Citizen (0 = No, 1= Yes)	Churn	
	No	Yes
0	0.8713	0.7453
1	0.1287	0.2547

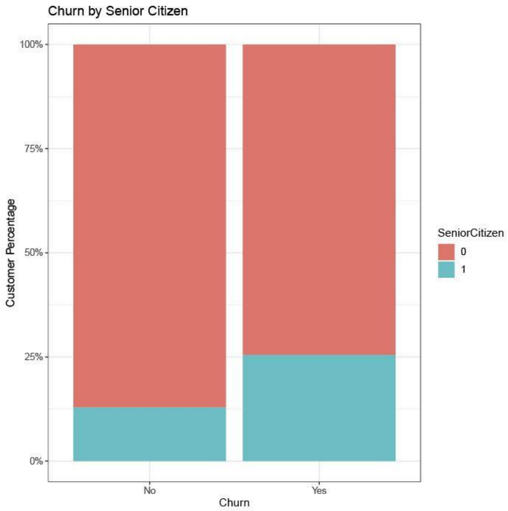

Figure 5.2.1 Figure showing the conditional distribution of Senior Citizen given churn outcome

From Section 4.32, the hypotheses is as follows:

H_0: There is no relationship between Senior Citizen and Churn (they are independent)

H_1: There is a relationship between Senior Citizen and Churn (they are dependent)

With a chi-squared p-value of 1.51×10^{-36} from Table 4.32.2, there is very strong evidence against H_0. This implies that Senior Citizen does affect churn, thus suggesting that the variables are dependent.

Looking at Figure 5.2.1, those that churn most are Non-Senior Citizens. They make up 74.5% of those that churn. The reasoning could be that because the younger generation citizens are more aware of what's available in the market, and thus would choose for other better alternatives.

Note that, on the other data set of not churning, Non-Senior Citizens make up 87.1%. This could primarily be because Non-Senior Citizens are the majority of the customer base - older citizens would make a small proportion as they are less exposed to technology over the given lifetime of technological advancement. And thus, making Non-Senior Citizens the larger proportions on both sections.

5.3 Phone Service

For Customers with phone service, the table gives the numerical values responding to the probabilities. Figure 5.3.1 gives the corresponding figure to the conditional distribution of phone service given Churn is either No or Yes.

Table 5.3.1 Phone Service Proportions

Phone Service	Churn	
	No	Yes
No	0.0990	0.0910
Yes	0.9010	0.9090

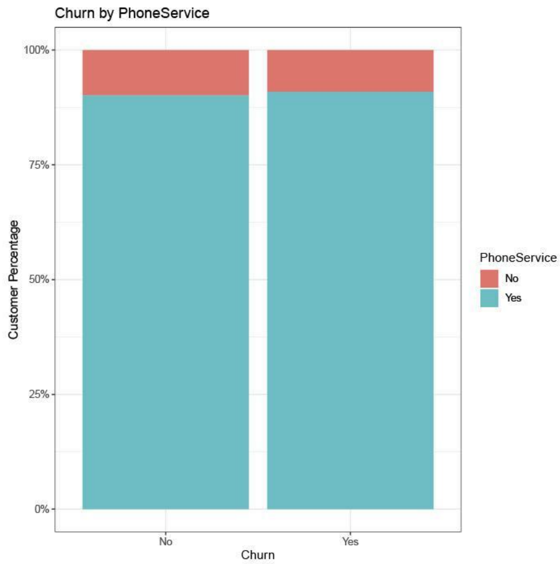

Figure 5.3.1 Figure showing the conditional distribution of Phone Service given churn outcome

From Section 4.32, the hypotheses is as follows:

H_0: There is no relationship between Phone Service and Churn (they are independent)

H_1: There is a relationship between Phone Service and Churn (they are dependent)

With a chi-squared p-value of 0.3388 from Table 4.32.2, there is no evidence against H_0. This implies that Phone Service does not affect churn, thus suggesting that the variables are independent from one another.

Whether they churn or not, those that have a phone service makes up 90% of the distribution. This could be because those that have phone services are exposed to the telecommunication sector of technology. These individuals usually make up the overall customer base and thus, can be backed up that they have the most activity when it comes to churning or not churning. This has the same argument for non-senior citizens in Section 5.2 in regards to a particular group of customers having more activity in each category.

5.4 Partner

For Customers with a partner, the table gives the numerical values responding to the probabilities. Figure 5.4.1 gives the corresponding figure to the conditional distribution of Gender given Churn is either No or Yes.

Table 5.4.1 Partner Proportions

Partner	Churn	
	No	Yes
No	0.4718	0.6421
Yes	0.5282	0.3579

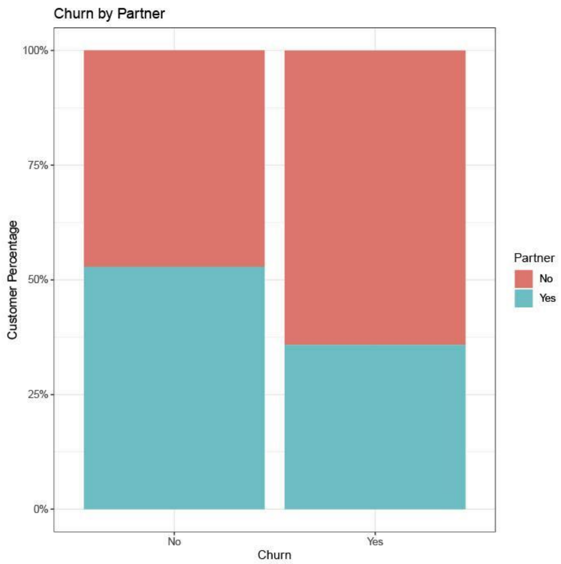

Figure 5.4.1 Figure showing the conditional distribution of Partner given churn outcome

From Section 4.32, the hypotheses is as follows:

H_0: There is no relationship between Partner and Churn (they are independent)

H_1: There is a relationship between Partner and Churn (they are dependent)

With a chi-squared *p*-value of 2.14×10^{-36} from Table 4.32.2, there is very strong evidence against H_0. This implies that variable Partner does affect churn, thus suggesting that the variables are dependent.

64% of customers who churn are those with no partners. This could be a result of not being dependent to keep in contact with their significant other/partners since we live in an era of technology where connection can be made through telecommunication. Also, we have that those with partners make up 52% of those that do not churn, this supports the argument that those with partners would want to stay connected, thus churning less.

For the variable Dependent, with a chi-squared *p*-value of 4.92×10^{-43}, it has the same results as variable Partner - that Dependent and Churn are dependent. For variable Dependent, it comes in the form of family members like children/parents that are under their care – those that don't have dependents churn more as a result of not needing this service to stay connected.

5.5 **Tenure**

For the customer's tenure, the table gives the numerical values responding to the probabilities. Figure 5.5.1 gives the corresponding figure to the conditional distribution of their tenure on the organisation given Churn is either No or Yes.

Table 5.5.1 Summary Table for Tenure

	Churn	
	No	Yes
Minimum	0.0	1.000
1st Quartile	15.00	2.00
Median	38.00	10.00
Mean	37.57	17.98
3rd Quartile	61.00	29.00
Maximum	72.00	72.00
Standard dev	24.11	19.53

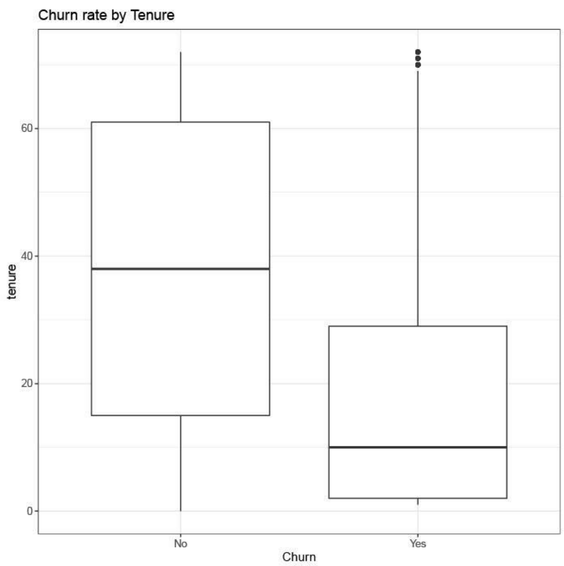

Figure 5.5.1 Figure showing the boxplot of tenure given churn outcome

For normality, by observing Figure 5.5.1 and the p-value with less than 2.2×10^{-16} from Section 4.22 Results. The distributions do not follow a normal distribution.

To test for the independence on the variables, with Mann-Whitney U test of p-value of 2.42 2.42×10^{-208} from Table 4.22.1, there is a very strong evidence against H_0. This implies that the population of customer's tenure given that they churned or did not churn are different with the analyses and hypotheses stated in Section 4.22 Results.

This variable answers the question of, how long has a customer used the company's product at that given time point? It is noticeable that people who use the product longer are least likely to churn with a median of 38 months compared to those who churn with a median of 17.98 months. This could be due to the loyalty or means of convenience that the customer holds for the company that results in them staying.

5.6 **Multiple Lines**

For Customers with multiple lines, the table gives the numerical values responding to the probabilities. Figure 5.6.1 gives the corresponding figure to the conditional distribution of multiple lines with churn is either No or Yes.

Table 5.6.1 Multiple Lines Proportions

	Churn	
Multiple Lines	No	Yes
No	0.4911	0.4543
No phone service	0.0990	0.0910
Yes	0.4099	0.4548

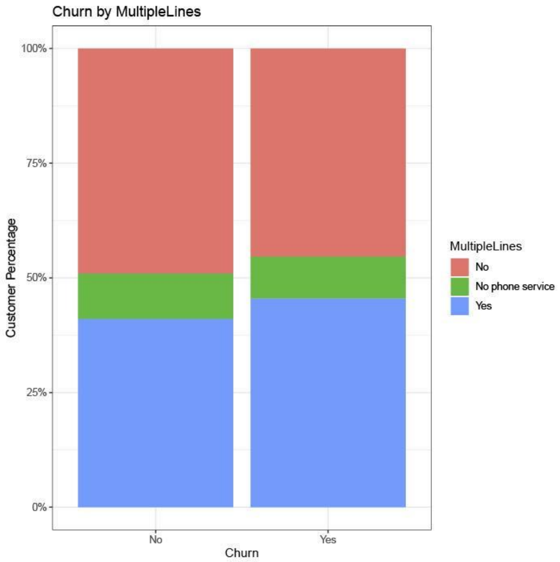

Figure 5.6.1 Figure showing the conditional distribution of Multiple Lines given churn outcome

From Section 4.32, the hypotheses is as follows:

H_0: There is no relationship between Multiple Lines and Churn (they are independent)

H_1: There is a relationship between Multiple Lines and Churn (they are dependent)

With a chi-squared p-value of 0.003464 from Table 4.32.1, there is some evidence against H_0 under the 5% significant level. Which means there is a possibility for Multiple Lines and Churn to be independent.

However, when looking at Figure 5.6.1, one would argue that Multiple Lines do not have an effect on churn based on how close the proportions are. If the significance level is at 0.3%, then there would be no evidence against the null hypothesis and it can be suggested that Multiple Lines and Churn are independent.

However, based on the test and the graph, having multiple phone lines does affect churn. Those that churn the most are those that do not have Multiple Lines, in this case customers who only have one phone line. A reasoning could be a switch to other alternatives are more flexible with customers who have only one phone line, 5.7 **Internet Service**

For Customers with internet service, the table gives the numerical values responding to the probabilities. Figure 5.7.1 gives the corresponding figure to the conditional distribution of internet service given Churn is either No or Yes.

Table 5.7.1 Internet Service Proportions

Internet Service	Churn	
	No	Yes
DSL	0.3792	0.2456
Fiber optic	0.3477	0.6940
No	0.2731	0.0605

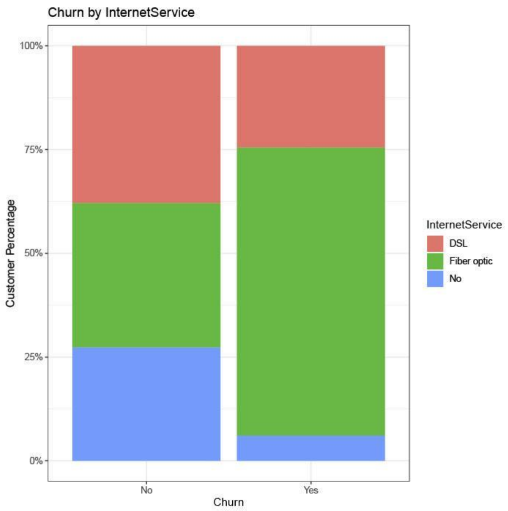

Figure 5.7.1 Figure showing the conditional distribution of Internet Service given churn outcome

From Section 4.32, the hypotheses is as follows:

H_0: There is no relationship between Internet Service and Churn (they are independent)

H_1: There is a relationship between Internet Service and Churn (they are dependent)

With a chi-squared p-value of 9.57×10^{-160}, from Table 4.32.2, there is very strong evidence against H_0. This implies that Internet Service does affect churn, thus suggesting that the variables are dependent.

Those who use Fiber Optic (69.4%) churn almost three times as those who use DSL (24.6%). This is because Fiber Optic is more expensive and could be more readily available and affordable in richer neighbourhoods. To add on, these neighbourhoods usually have a bigger pool of other telecommunication products (i.e Fiber Optic alternatives like TIME, Unifi and Maxis compared to other areas with only one DSL line with one provider like Streamyx). Thus, implying that richer customers churn in favour of better alternatives.

However, for those that do not churn, DSL has the larger portion of 37.9% and Fiber Optic at 34.8%. This is backed up by the statement earlier that DSL churn less depending on the targeted group and area due to the availability of alternatives and the income of neighbourhoods in that area.

5.8 **Online Security**

For Customers with online security, the table gives the numerical values responding to the probabilities. Figure 5.8.1 gives the corresponding figure to the conditional distribution of online security given Churn is either No or Yes.

Table 5.8.1 Online Security Proportions

Online Security	Churn No	Churn Yes
No	0.3937	0.7817
No internet service	0.2731	0.0605
Yes	0.3332	0.1578

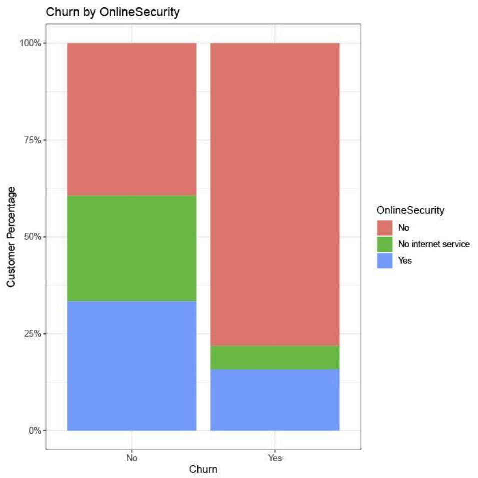

Figure 5.8.1 Figure showing the conditional distribution of Online Security given churn outcome

From Section 4.32, the hypotheses is as follows:

H_0: There is no relationship between Online Security and Churn (they are independent)

H_1: There is a relationship between Online Security and Churn (they are dependent)

With a chi-squared p-value of 2.66×10^{-185}, from Table 4.32.2, there is very strong evidence against H_0. This implies that Online Security does affect churn, thus suggesting that the variables are dependent.

People who churn are people who do not have online security which makes up 78%. Customers with online security are least likely to churn could be due to the stickiness to the product. Stickiness means users of the product will keep coming back to use the product due to all the facilities/packages like variable Online Security attached along with the product.

Other variables like Online Backup, Device Protection and Tech Support have the same conditional distribution and results as Online Security. These are all variables that affect customer churn due to dependency on the product as a result of product stickiness.

5.9 Streaming TV

For Customers with Streaming TV, the table gives the numerical values responding to the probabilities. Figure 5.9.1 gives the corresponding figure to the conditional distribution of Streaming TV given Churn is either No or Yes.

Table 5.9.1 Streaming TV Proportions

Streaming TV	Churn	
	No	Yes
No	0.3610	0.5040
No internet service	0.2731	0.0605
Yes	0.3659	0.4355

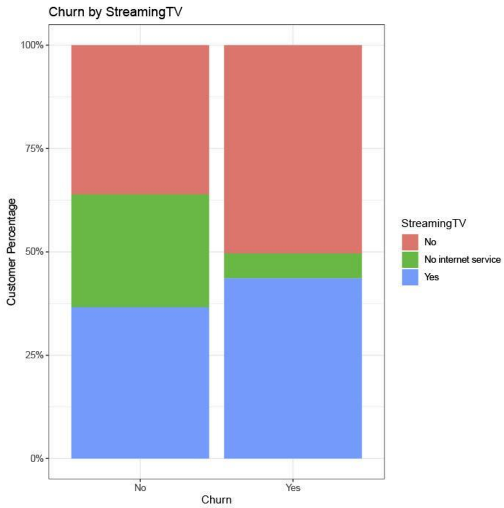

Figure 5.9.1 Figure showing the conditional distribution of Streaming TV given churn outcome

From Section 4.32, the hypotheses is as follows:

H_0: There is no relationship between Streaming TV and Churn (they are independent)

H_1: There is a relationship between Streaming TV and Churn (they are dependent)

With a chi-squared p-value of 5.53×10^{-82}, from Table 4.32.2, there is very strong evidence against H_0. This implies that Streaming TV does affect churn, thus suggesting that the variables are dependent.

People who churn the most on those who don't stream TV. Same reasoning on product stickiness with the other variables mentioned in the previous page. Also, the variable Streaming Movies, have similar distribution and outcome of the p-value of 2.67×10^{-82} and analysis and thus can be concluded that results are similar to Streaming TV - variable Streaming Movie and Churn are dependent.

5.10 Contract

For Customers with Contract, the table gives the numerical values responding to the probabilities. Figure 5.10.1 gives the corresponding figure to the conditional distribution of Contract given Churn is either No or Yes.

Table 5.10.1 Contract Proportions

Contract	Churn	
	No	Yes
Month-to-month	0.4291	0.8855
One year	0.2526	0.0888
Two year	0.3183	0.0257

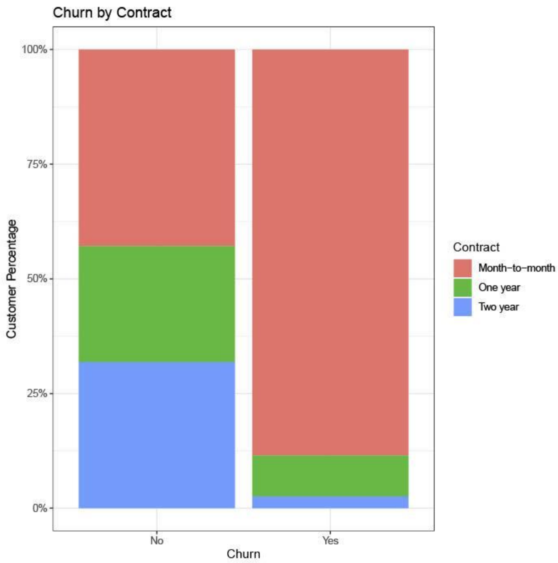

Figure 5.10.1 Figure showing the conditional distribution of Contract given churn outcome

From Section 4.32, the hypotheses is as follows:

H_0: There is no relationship between Contract and Churn (they are independent)

H_1: There is a relationship between Contract and Churn (they are dependent)

With a chi-squared p-value of 5.86×10^{-258}, from Table 4.32.2,therefore there is very strong evidence against H_0. This implies that Contract does affect churn, thus suggesting that the variables are dependent.

Those that are on a Month-to-month basis rather than yearly contracts churn most. Around 80% of them, on shorter contracts, churned. For long term contracts, yearly contracts, customers invest their time and money into the product which results in the customer choosing to continue rather than find other better alternatives. In doing the latter, it would result in higher cost in terms of low satisfaction, more time and higher cost, thus resulting in customers that are committed to stay in the plan.

5.11 Paperless Billing

For Customers with Paperless Billing, the table gives the numerical values responding to the probabilities. Figure 5.11.1 gives the corresponding figure to the conditional distribution of Paperless Billing given Churn is either No or Yes.

Table 5.11.1 Paperless Billing Proportions

Paperless Billing	Churn	
	No	Yes
No	0.4644	0.2509
Yes	0.5356	0.7491

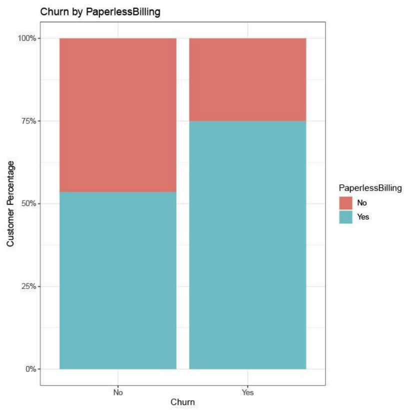

Figure 5.11.1 Figure showing the conditional distribution of Paperless Billing given churn

From Section 4.32, the hypotheses is as follows:

H_0: There is no relationship between Paperless Billing and Churn (they are independent)

H_1: There is a relationship between Paperless Billing and Churn (they are dependent)

With a chi-squared p-value of 4.07×10^{-58}, from Table 4.32.2, therefore there is very strong evidence against H_0. This implies that Paperless Billing does affect churn, thus suggesting that the variables are dependent. Those that churn most are those that use Paperless Billing(74.9%). This method of billing, are forms of online payments/transactions that are commonly used during this era to pay bills.

Now we ask the question, could this variable be dependent on another variable? We introduce the Senior Citizen variable.

Table 5.11.2 Paperless Billing and Senior Citizen

Paperless Billing (PB)	Senior Citizen		Non-Senior	
	No Churn	Yes Churn	No Churn	Yes Churn
No PB	188	78	**2215**	**391**
Yes PB	478	398	**2293**	**1002**

Looking at the figure above, Paperless Billing is dependent on the Senior Citizen Variable. This is because Senior Citizens are least likely to churn and those that are younger are twice as likely to churn - same reasoning for Senior Citizens on younger generations being exposed to better alternatives.

Since Paperless Billing is dependent on the variable Senior Citizen, Paperless Billing is a response variable to churn rather than a predictor - which implies that this variable is a correlation, rather than a causation.

5.12 Payment Method
For Customers with Payment Method, the table gives the numerical values responding to the probabilities. Figure 5.12.1 gives the corresponding figure to the conditional distribution of Payment Method given Churn is either No or Yes.

Table 5.12.1 Payment Method Proportions

	Churn	
Payment Method	No	Yes
Bank transfer (automatic)	0.2486	0.1380
Credit card (automatic)	0.2493	0.1241
Electronic check	0.2501	0.5730
Mailed check	0.2520	0.1648

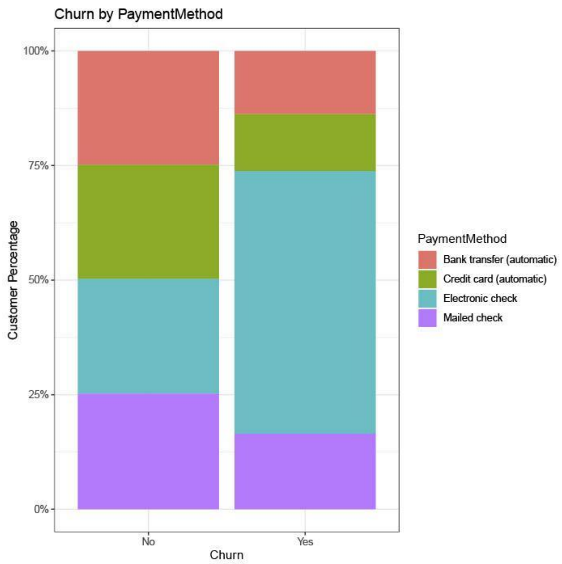

Figure 5.12.1 Figure showing the conditional distribution of Payment Method given churn outcome

From Section 4.32, the hypotheses is as follows:

H_0: There is no relationship between Payment Method and Churn (they are independent)

H_1: There is a relationship between Payment Method and Churn (they are dependent)

With a chi-squared p-value of 3.68×10^{-140} from Table 4.32.1,therefore there is very strong evidence against H_0. This implies that the Payment Method does affect churn, thus suggesting that the variables are dependent.

The subset of customers who churn most, making up 57% of those who churn, is electronic check. By observing the other payment methods, we have Bank Transfer (automatic), Credit card(automatic) and mailed check - electronic check is one of the newest and most convenient choices for customers. It can be implied that those that are more tech savvy or more exposed to other technological alternatives would rather seek for better products thus resulting in them churning more.

5.13 Monthly Charges

For Customers' Monthly Charges, the table gives the numerical values responding to variables.. Figure 5.13.1 gives the corresponding figure to the statistical summary values of Monthly Charges given Churn is either No or Yes.

Table 5.13.1 Summary Table for Monthly Charges

	Churn	
	No	Yes
Minimum	18.25	18.85
1st Quartile	25.10	56.15
Median	64.42	79.65
Mean	61.27	74.44
3rd Quartile	88.40	94.20
Maximum	118.8	118.35
Standard dev	31.09	24.67

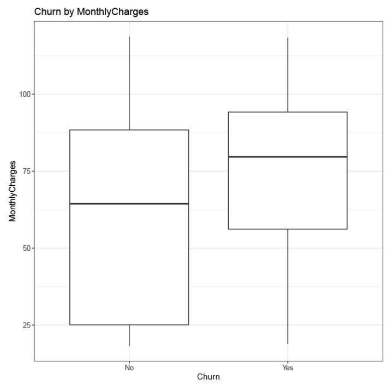

Figure 5.13.1 Figure showing the boxplot of Monthly Charges given the churn outcome

For normality (Kolmogorov Smirnov-Test), by observing figure 5.13.1 and the *p*-value with less than 2.2×10^{-16} from the Results Section 4.22. The distributions do not follow a normal distribution.

The Mann-Whitney U hypothesis will follow as:

H_0: The population of customers' monthly charges given that they churned or did not churn are equal

H_1: The population of customers' monthly charges given that they churned or did not churn are not equal

To test for the independence on the variables, the *p*-value of 3.311×10^{-54} from Table 4.22.1, suggests that there is a very strong evidence against H_0. This implies that the population of customer's Monthly Charges given that they churned or did not churn are different.

Those that are most likely to churn are those that have high monthly charges. This could be because of a high-cost package that they purchased and were not satisfied in it, which results in them churning. The mean of the Monthly Charges that churn is 79.65, compared to those that didn't which is 64.42.

5.14 Total Charges

For Customers' Total Charges, the table gives the numerical values responding to variables.. Figure 5.14.1 gives the corresponding figure to the statistical summary values of Total Charges given Churn is either No or Yes.

Table 5.14.1 Summary Table for Total Charges

	Churn	
	No	Yes
Minimum	0.0	18.85
1st Quartile	572.9	134.5
Median	1680	703.6
Mean	2550	1532
3rd Quartile	4263	2331
Maximum	8673	8685
Standard dev	2330	1890

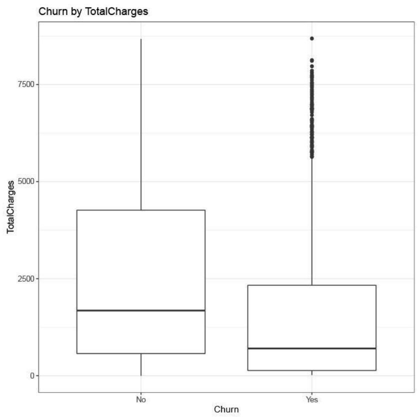

Figure 5.14.1 Figure showing the boxplot of Total Charges given the churn outcome

For normality (Kolmogorov-Smirnov Test), by observing figure 5.14.1 and the *p*-value with less than 2.2×10^{-16} from Results Section 4.22. The distributions do not follow a normal distribution.

The hypothesis will follow as:

H_0: The population of customers' total charges given that they churned or did not churn are equal

H_1: The population of customers' total charges given that they churned or did not churn are not equal

To test for the independence on the variables, with Mann-Whitney U test of *p*-value of 5.685×10^{-83} from Table 4.22.1, there is a very strong evidence against H_0. This implies that the population of customer's Total Charges given that they churned or did not churn are different.

Those that are most likely to churn are those that have lower total charges. This variable is dependent on the Tenure and Monthly Charges variable. With the calculation of how long they used the package multiplied by monthly charges. Because Total Charges is dependent on the other explanatory variables. This variable is a response variable to churn. This variable is a correlation, rather than a causation.

6. Results

6.1 Exploratory Variables Analysis

After exploring the 19 variables' independence, similarity and normality, corresponding to the variable of Churn. Backed up with statistical tests, figures and tables, it is summarised in Table 6.1.1 below on whether the given variable in Telcos Dataset affects churn or not based on the column on whether it is significant under the 5% level.

The analysis has been done for each and every variable in Section 5. We found that there are 2 variables, gender and Phone Service, that do not affect churn and the other 17 variables affect churn. It's important to highlight that out of the 17 variables that are significant, Paperless Billing and Total Charges are not predictor variables of Churn but rather response variables as they depend on other variables in the dataset like Senior Citizen and Monthly Charges to make an impact on churn.

Table 6.1.1 Table Summary table on the variables that affect churn

	Variable	**P-value**	**Significant at 5%?**
1	gender	0.48657874	**No**
2	SeniorCitizen	1.51E-36	Yes
3	Partner	2.14E-36	Yes
4	Dependents	4.92E-43	Yes
5	tenure	2.42E-208	Yes
6	PhoneService	0.33878254	**No**
7	MultipleLines	0.00346438	Yes
8	InternetService	9.57E-160	Yes
9	OnlineSecurity	2.66E-185	Yes
10	OnlineBackup	2.08E-131	Yes
11	DeviceProtection	5.51E-122	Yes
12	TechSupport	1.44E-180	Yes
13	StreamingTV	5.53E-82	Yes
14	StreamingMovies	2.67E-82	Yes
15	Contract	5.86E-258	Yes
16	PaperlessBilling	4.07E-58	Yes
17	PaymentMethod	3.68E-140	Yes
18	MonthlyCharges	3.312E-54	Yes
19	TotalCharges	5.69E-83	Yes

6.2 Fitting a Multiple Logistic Regression (MLR) Function

In this section, we use a model based approach to predict customer churn based on the independent variables in the Telco Customer Churn dataset. The model based approach used is logistic regression.

A logistic regression is used to predict the class (or category) of a set based on one of the multiple predictor variables (x). It is used to model a binary outcome. that is a variable, which can only have two possible outcomes: 0 or 1, yes or no, diseased or non-diseased.

A logistic regression function is used for predicting the outcome of an observation given a predictor variable (x). In this case, given the variables in the dataset, will the outcome be churn or no churn. We have a logistic regression as follows:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \rightarrow \frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X}$$

For a multiple logistic regression, we have the following equation

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 \cdots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 \cdots + \beta_n X_n}}$$

By taking the logarithm of both sides, the following logit equation is obtained:

$$log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \cdots + \beta_n X_n,$$

where $X = (X_1, X_2, \ldots, X_p)$ are p predictors

We then introduce the maximum likelihood method to estimate $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$

$$l(\beta_0, \beta_1, \ldots, \beta_n) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

For the first fit on the model, called Model 1, we will fit all the variables in the dataset. The following output derived from R from Model 1 of coefficients are as follows:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.1240574387	9.124513e-01	1.23190957	2.179829e-01
genderMale	0.0123956025	7.269436e-02	0.17051671	8.646038e-01
SeniorCitizen	0.2848350196	9.539294e-02	2.98591286	2.827332e-03
PartnerYes	-0.0084204583	8.763875e-02	-0.09608145	9.234559e-01
DependentsYes	-0.1053000303	1.004687e-01	-1.04808781	2.945982e-01
tenure	-0.0642584033	7.010619e-03	-9.16586716	4.915020e-20
PhoneServiceYes	0.0585678377	7.241401e-01	0.08087915	9.355381e-01
MultipleLinesYes	0.3619994847	1.974573e-01	1.83330499	6.675723e-02
InternetServiceFiber optic	1.6059563449	8.907324e-01	1.80296169	7.139423e-02
InternetServiceNo	-1.6763703516	9.009059e-01	-1.86076073	6.277797e-02
OnlineSecurityYes	-0.2899254100	2.000352e-01	-1.44937203	1.472337e-01
OnlineBackupYes	0.0091260681	1.962994e-01	0.04649055	9.629193e-01
DeviceProtectionYes	0.0843490212	1.973466e-01	0.42741562	6.690766e-01
TechSupportYes	-0.2814454498	2.014357e-01	-1.39719729	1.623543e-01
StreamingTVYes	0.6423647570	3.652251e-01	1.75881885	7.860828e-02
StreamingMoviesYes	0.5299217681	3.635885e-01	1.45747673	1.449848e-01
ContractOne year	-0.6100943131	1.206666e-01	-5.05603141	4.280709e-07
ContractTwo year	-1.2511390101	1.929871e-01	-6.48302047	8.990430e-11
PaperlessBillingYes	0.3422177301	8.388820e-02	4.07945022	4.514233e-05
PaymentMethodCredit card (automatic)	-0.1055739473	1.285189e-01	-0.82146653	4.113806e-01
PaymentMethodElectronic check	0.2590344467	1.071309e-01	2.41792486	1.560930e-02
PaymentMethodMailed check	-0.1242628280	1.303269e-01	-0.95347026	3.403519e-01
MonthlyCharges	-0.0356175856	3.546998e-02	-1.00416133	3.153009e-01
TotalCharges	0.0003591742	7.949514e-05	4.51819064	6.237032e-06

Figure 6.2.1 Output of the Coefficients for Model 1

The estimates for Figure Output 6.2.1 gives the coefficient for the variables which can be equated in the line of equations below. For the second column, the measure of accuracy of the coefficient estimates are calculated by the standard errors. Followed by the *z*-statistic is associated with the normal distribution.

For Model 1, we have the following line of equations with the given variables. This line of equation for Model 1 had a fitting for all variables significant or insignificant.

$$\begin{aligned}\widehat{Churn}_{Model\ 1} = {} & 1.167 - 0.0219\, I(Gender = Male) + 0.216\, I\,(SeniorCitizen) \\ & - 0.00195\, I(Partner = Yes) - 0.152\, I\,(Dependent = Yes) \\ & - 0.0603\, I(tenure) + 0.178\, I(PhoneService = Yes) \\ & + 0.447\, I(MultipleLine = Yes) \\ & + 1.752\, I(InternetService = Fiber\ Optic) \\ & - 1.794\, I\,(InternetService = No) - 0.205\, I(Online\ Security = Yes) \\ & + 0.0255\, I(Online\ Backup = Yes) + 0.147\, I(Device\ Protection = Yes) \\ & - 0.180\, I(Tech\ Support = Yes) + 0.591\, I(StreamingTV = Yes) \\ & + 0.603\, I(StreamingMovie = Yes) - 0.665\, I(Contract = One\ year) \\ & - 1.380\, I(Contract = Two\ year) - 0.342\, I(Paperless\ Billing = Yes)) \\ & + 0.305\, I(PaymentMethod = Electronic\ Check) \\ & - 0.0579\, I(PaymentMethod = Mailed\ Check) \\ & - 0.0405\, I(Monthly\ Charges) - 0.000328\, I(Total\ Charges)\end{aligned}$$

Note that some variables have an indicator function. So for example for the equation below, for the 3rd variable,if the customer is dependent then, *I*(Dependent =Yes) = 1, if else then it will equate to 0.

We shall interpret the line of the equation. For the variable given that the gender is male, it has a coefficient , β_1, of -0.0219; this indicates that the probability of not churning and the gender given male has a negative correlation. In other words, for that given customer who is male for gender, the probability of not churning will decrease. Similarly, for a numerical variable, Monthly Charges, it has a coefficient of -0.0405; this implies a negative relationship, that by increasing monthly charges, it will result in a lower rate for not churning.

For Model 2, we fit the model based on variable selection based on our results in Section 4 on Explanatory Variable Analysis. We will only fit the variables that were significant in our exploratory data analysis in Table 4.32.1 - which leaves out variables gender and Phone Service.

The results of the logistic regression line is as below

$$\begin{aligned}\widehat{Churn}_{model\ 2} = {} & 1.343 - 0.216\, SeniorCitizen - 0.158\, I\,(Dependent = Yes) \\ & - 0.0603\, I(tenure) + 0.448\,(MultipleLine = Yes) \\ & + 1.758\, I(InternetService = Fiber\ Optic) \\ & - 1.799\, I\,(InternetService = No) + 0.593\, I(StreamingTV = Yes) \\ & + 0.605\, I(StreamingMovie = Yes) - 0.665\, I(Contract = One\ year) \\ & - 1.380\, I(Contract = Two\ year) + 0.342\, I(Paperless\ Billing = Yes)) \\ & + 0.305\, I(PaymentMethod = Electronic\ Check) \\ & - 0.000328\, I(Total\ Charges)\end{aligned}$$

Finally, for our 3rd Model, we will fit the significant variables except for Paperless Billing and Total Charges which were claimed to be response rather than predictor variables- variables that correlated to Churn rather than predictors.

The third model fit gives:

$$\begin{aligned}\widehat{Churn}_{model\ 3} = \ & 1.112 + 0.231\, I(SeniorCitizen) - 0.176 I(Dependent) \\ & - 0.034\, I(tenure) + 0.494\, I(MultipleLine) \\ & + 1.848\, I(InternetService = Fiber\ Optic) \\ & - 1.771\, I\,(InternetService = No) + 0.651\, I(StreamingTV = Yes) \\ & - 0.692\, I(Contract = One\ year) - 1.398\, I(Contract = Two\ year) \\ & + 0.328\, I(PaymentMethod = Electronic\ Check)\end{aligned}$$

6.21 Comparing the models

This section hopes to explore the question of - Which model is a better fit? Thus, we fit a predict function in R for Models 1, 2 and 3 to see which one is a better model. It's important to note that our baseline on what we are trying to fit is the probability of not churning, so when we are predicting the variables, we are getting the probabilities of the likelihood of them not churning. We then compared it with the true values of the Churn data in Telcos dataset, and found that the misclassification rates are as below.

	Model 1	Model 2	Model 3
Misc. Rate	0.522	0.523	0.516
Deviance	5828.284	5828.398	5871.63

Table 6.2.1 Misclassification rate and deviance for each models

Model 3 has the lowest misclassification rate of 0.516 compared to the other models when the response variables were removed. It is then followed by Model 1 with 0.522 and Model 2 with 0.523. Model 3 is the better fit compared to the other models due it to having a lower misclassification rate.

Deviance indicates the degree to which the likelihood of the saturated model exceeds the likelihood of the proposed model. The deviance can be used for a goodness of fit check. If the proposed model has a good fit, the deviance will be small. If the proposed model has a bad fit, the deviance will be large. In this case, Model 3 had the highest deviance which indicates that it is not a good fit.

To further compare 2 models, we shall fit the Likelihood Ratio Test in the built-in function of R to compare the models to see one which is a better fit based on it's significance. If it is statistically significant, less than the proposed significant level of 5%, then the proposed second fit has improved the fit. The codes can be found in the appendix.

First we compare Model 1 and Model 2. For Model 2, we only took significant values that were obtained from Section 4 of this project. The *p*-value output for this using the significance was achieved at 0.7356. This implies that Model 2 did not significantly improve the model.

Then we compare Model 2 and Model 3. For Model 3, we took out Paperless Billing and Total Charges due to them being response variables. The *p*-value obtained at 4.095×10^{-10}, this means that removing these 2 variables improved the model significantly.

Overall, the rates for these models are relatively high, which suggests that the linear logistic regression is not a good model for customer churn due to its overall high misclassification rate.

7.Conclusion and Discussion

Customer churn is one of the most crucial goals in the telecommunications industry, or any other industry in the market with a target group of consumers. The main objective of businesses is to reduce churn, to decrease the rate at which customers stop using the companies' service or product. We have identified in this project the many factors that affected churn such as Multiple Lines, whether they had dependents or not and the other variables to be found to be significant in Table 4.22.1 and Table 4.32.2 in Section 4 of Exploratory Data Analysis. On the other hand, the factors that had no significance to churn were gender and whether they had Phone Service or not.

The advice to businesses after this analysis would be to focus on these factors. For example, for a customer's type of Internet Service, we found that people who use Fiber Optic are more likely to churn rather than DSL. For customers that use Fiber Optic, we could find the root cause of why they would churn and suggest a solution. An example of a solution would be to have more marketing campaigns, improve the quality of service to satisfy these subsets of customers or even convince customers to switch to DSL since those subsets of customers churn less.

The goal is always to strive to reduce churn. However, there may be other factors that can't be influenced or changed, for example, certain demographic factors like whether someone has a partner or not, or whether they are senior citizens. These kinds of factors allow us to understand the causes of churn and allow us to focus on the targeted group, but may not be useful information when it comes to influencing them to switch up their behaviour in favour of the other variable (We can't influence them to get a partner or become a senior citizen to reduce churn like how we mentioned for Internet Service earlier in the paragraph)

In this project, we focused on figuring out the variables that affected churn and fit it to a machine learning classification technique which is a logistic regression. However, there are certain limitations to this technique. It assumes linearity, which may not be the case for this dataset. Observing the graph below, it shows that the probability to churn is higher compared to the probability of not churning. The Figure 7.1 backs up the argument that logistic regression is not a good machine learning technique model to use due to it's high misclassification rate as the threshold of probability is different at every point.

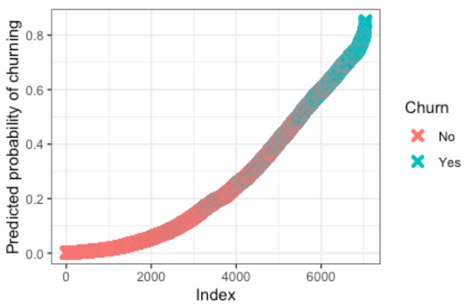

Figure 7.1 Graph on the predicted probability of churning

In Semester 2, we will explore the many different kinds of Machine Learning Techniques and test it's adequacy to the given dataset.

References

Berson, A., Smith, S., & Thearling, K. (2000). Building data mining applications for CRM. New York, NY: McGraw-Hill.

Bingquan H., Mohand T. K. , Brian B., (2012). *Customer churn prediction in telecommunications*, Expert Systems with Applications,Volume 39, Issue 1, Pages 1414-1425,

Blattberg, R.C., Kim, P., & Neslin, S.A. (2008). Database Marketing: Analyzing and Managing Customers.

Coussement K. and Poel. D. V. D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques, *Expert Syst. Appl.*, vol. 34, no. 1, pp. 313-327

C. Wei, I. Chiu (2002) Turning telecommunications call details to churn prediction: A data mining approach, Expert Systems with Applications, 23 , pp. 103-112

Hadden J. , Tiwari A. , Roy R. , Ruta D. (2007) Computer assisted customer churn management: state-of-the-art and future trends. Comput. Oper. Res., 34 (10), pp. 2902-2917

Ismail M. R., M. K. Awang, M. N. A. Rahman and M. Makhtar. (2015) "A multi-layer perceptron approach for customer churn prediction", *Int. J. Multimedia Ubiquitous Eng.*, vol. 10, no. 7, pp. 213-222.

Idris A. , Khan A. and Lee Y. S., (2012) Genetic programming and adaboosting based churn prediction for telecom, *Proc. IEEE Int. Conf. Syst. Man Cybern.*, pp. 1328-1332

James G. , Daniela W., Trevor H., and Robert T.. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

John, H., Ashutosh, T., Rajkumar, R., Dymitr, R. (2007). Computer assisted customer churn management: State-of-the-art and future trends.

Kirui C. , L. Hong, W. Cheruiyot and H. Kirui. (2013) "Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining", *Int. J. Comput. Sc. Iss.*, vol. 10, no. 2, pp. 165-172.

Klemperer, P. (1988). Welfare Effects of Entry Into Markets with Switching Costs. *The Journal of Industrial Economics*, *37*(2), 159–165.

Luo, B., Shao, P., Liu, J. 2007. Customer churn prediction based on the decision tree in personal handyphone system service. In *International conference on service systems and service management* (pp. 1–5).

Mann, Henry B.; Whitney, Donald R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics*. 18 (1): 50–60.

M. R. Ismail, M. K. Awang, M. N. A. Rahman and M. Makhtar. (2015) "A multi-layer perceptron approach for customer churn prediction", *Int. J. Multimedia Ubiquitous Eng.*, vol. 10, no. 7, pp. 213-222

M. Owczarczuk. (2010) "Churn models for prepaid customers in the cellular telecommunication industry using large data marts", *Expert Syst. Appl.*, vol. 37, no. 6, pp. 4710-4712.

Pearson K. (1900). "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling" (PDF). *Philosophical Magazine*. Series 5. 50 (302): 157–175.

Robert C. Blattberg, Kim B.D and Scott A. (2008) Database Marketing: Analyzing and Managing Customers. Springer

Rosie Shier. (2004) “Statistics: 2.3 The Mann-Whitney U Test”, *Mathematics Learning Support Centre*,

Stephens M.A. (1992) Introduction to Kolmogorov (1933) On the Empirical Determination of a Distribution. In: Kotz S., Johnson N.L. (eds) Breakthroughs in Statistics. Springer Series in Statistics (Perspectives in Statistics). Springer, New York, NY.

S. Y. Hung, D. C. Yen , H. Y. Wang, (2006) Applying data mining to telecom churn management, *Expert Syst. Appl.*, vol. 31, no. 3, pp. 515-524,

Shin-Yuan H, David C. Y, Hsiu-Yu W (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*. Volume 31, Issue 3

William C. G. (1952). "The Chi-square Test of Goodness of Fit". *The Annals of Mathematical Statistics*. 23 (3): 315–345.

Appendix

```
## 1
#Website link:
https://towardsdatascience.com/predict-customer-churn-the-right-way-using-pycaret-8ba6541
608ac
# Youtube link to ggplot tutorial = https://www.youtube.com/watch?v=49fADBfcDD4

file <-
"https://raw.githubusercontent.com/srees1988/predict-churn-py/main/customer_churn_data.cs
v"

data <- read.csv(file)
View(data) # view the data in a different tab

#library("dplyr")


column_names <- data

name_of_col <- colnames(data)

#structure of the data
str(data)
#        7043 obs. of  21 variables:

class(data)


### loop over the column headings and change characters to factors
  for (i in 1:length(column_names)) {

      if (lapply(data, class)[[i]] == "character") {
        data[,i] <- as.factor(data[,i])

      # data$column<- as.factor(data$column)
      }
  }



data$customerID<- as.character(data$customerID) # CustomerID must be Character because
it is Unique


## check that the data has changed to factor
#structure of the data
```

```
str(data)


# is.na.data.frame(data)

# to check null values in each column
cbind( lapply( lapply(data, is.na),sum) )

# replace NA columns with 0 values
data <- replace(data, is.na(data), 0)

## use ggplot or dply for visualisation
library(ggplot2)
library(scales) # for percentage scales

## 2
pdf('/Users/yenshen/Desktop/FYP/Data log/Output/plot_graphs.pdf') #create a pdf so that any
graphs that load can be saved inside
# please change the pathway to find a place to be downloaded in your own computer

myplots <- list()

 for (i in 1:length(column_names)) {

 if (class(data[,i]) == "factor") {

   plotz <- ggplot(data, aes(x= Churn, fill = data[,i])) +
          theme_bw() +
          geom_bar(position = 'fill') +
          labs(y = 'Customer Percentage',
               title = paste('Churn by', column_names[i])) +
     labs(fill= paste(column_names[i]) ) +  # rename the Legend title
     scale_y_continuous(labels = scales::percent)

   print(plotz)
   myplots[[i]] <- plotz

 }
   if (class(data[,i]) == "numeric") {

   plotw <- ggplot(data, aes(x= Churn, y= .data[[column_names[i]]])) +
     theme_bw() +
     labs( y= paste(column_names[i]) ,
          x = 'Churn',
          title = paste('Churn by', column_names[i])) +
     geom_boxplot()
     print(plotw)
```

```
      myplots[[i]] <- plotw
    } else {
      myplots[[i]] <- 0
    }
}

## For Senior Citizen (Exception)
vv <-table(data$SeniorCitizen, data$Churn)
frame <- as.data.frame(vv)
names(frame)[2] <- 'Churn'
names(frame)[1] <- 'SeniorCitizen'

ggplot(frame, aes(x=Churn, y = Freq, fill= SeniorCitizen)) +
        theme_bw() +
        geom_bar(stat="identity", position = 'fill') +
        scale_y_continuous(labels = scales::percent)+
  labs(title = 'Churn by Senior Citizen', y = 'Customer Percentage')

## For tenure (Exception)
ggplot(data, aes(x= Churn, y= tenure)) +
  theme_bw() +
  geom_boxplot() +
  labs( y= 'tenure',
        x = 'Churn',
        title = 'Churn rate by Tenure')

dev.off() # end the doc

## 3
####### Testing for Normality  #########
############ numerical  ####################
######## We shall use Kolmogorov Smirnov test ########

## tenure
ks_ten_no <- ks.test(data$tenure[data$Churn == 'No'], "pnorm")
ks_ten_yes <- ks.test(data$tenure[data$Churn == 'Yes'], "pnorm")

## Monthly Charges
ks_mc_no <- ks.test(data$MonthlyCharges[data$Churn == 'No'], "pnorm")
ks_mc_yes <- ks.test(data$MonthlyCharges[data$Churn == 'Yes'], "pnorm")

## Total Charges
ks_tc_no <- ks.test(data$TotalCharges[data$Churn == 'No'], "pnorm")
ks_tc_yes <- ks.test(data$TotalCharges[data$Churn == 'Yes'], "pnorm")



#### All of them do not follow a normal distribution ####
```

```
###### p value less than 2.2 x 10 ^ -16 ######


######  EXAMPLE KS TEST ######
###### TENURE WITH NO CHURN ######

ten_no <- data$tenure[data$Churn == 'No']

freq <- as.data.frame(table(ten_no))

x <- seq(0, 72 , 1)

seqq1 <- seq(0, 1 , 1/73)
seqq1 <- seqq1[-74]

seqq2 <- seq(0, 1 , 1/73)
seqq2 <- seqq2[-1]

f_x <- pnorm(x)

D_1 <- f_x - seqq1
D_2 <- seqq2 - f_x

##### mean 0 and variance 1
cbind(freq, seqq1, f_x, seqq2, D_1, D_2)
max(D_1,D_2)
#  0.9575542

# D at 5% level
1.36 / sqrt(length(x))
# 0.159176

ks_ten_no
# 0.93009




#------------------------------------------------------------------------------------------
################ Man Whitney Test ########################
################## numerical #############################

# can't be computed exact because it has ties

# For tenure
tenure_tab <- wilcox.test(x= data$tenure[data$Churn =='Yes'],
                    y=data$tenure[data$Churn =='No'],
                    exact = FALSE)
```

```
tenure_tab$p.value

# For MonthlyCharges
MC_tab <- wilcox.test(x= data$MonthlyCharges[data$Churn =='Yes'],
                      y=data$MonthlyCharges[data$Churn =='No'],
                      exact = FALSE)

MC_tab$p.value


#For TotalCharges
TC_tab <- wilcox.test(x= data$TotalCharges[data$Churn =='Yes'],
                      y=data$TotalCharges[data$Churn =='No'],
                      exact = FALSE)
TC_tab$p.value


#------------------------------------------------------------------------------------------
####### Chisquared of Independence #########
############ categorical variables ##################
df <- data

df <- subset (df, select = -Churn) # delete churn
df <- subset (df, select = -customerID) # delete off CustomerID

df_chiqsq <- df
## delete off the numerical variables ###
df_chiqsq <- subset (df_chiqsq, select = -tenure) # delete off tenure
df_chiqsq <- subset (df_chiqsq, select = -MonthlyCharges) # delete off MonthlyCharges
df_chiqsq <- subset (df_chiqsq, select = -TotalCharges) # delete off TotalCharges

row.names(df_chiqsq) <- NULL # reset the index

colname_chi <- colnames(df_chiqsq)

# For 2 categorial data, we shall perform chi square of independence test

store_here <- list() ## store the p-value here

for (i in 1:length(df_chiqsq)) {

  if (class(df_chiqsq[,i]) == "factor") {

    tab1 <- xtabs(~df_chiqsq[[colname_chi[i]]] + data$Churn )

    chisq_ <- chisq.test(tab1)

    store_here[i] <- paste(chisq_$p.value)
```

```
  }

  else{
    store_here[i] <- '-'
  }
}
# For SeniorCitizen
SC_tab <- xtabs(~data$SeniorCitizen + data$Churn)
SC_chisq_ <- chisq.test(SC_tab)
store_here[[2]]<- SC_chisq_$p.value

## store the test statistic here

test_stat_here <- list() ## store the test statistic here

for (i in 1:length(df_chiqsq)) {

  if (class(df_chiqsq[,i]) == "factor") {

    tab1 <- xtabs(~df_chiqsq[[colname_chi[i]]] + data$Churn )

    chisq_ <- chisq.test(tab1)

    test_stat_here[i] <- paste(chisq_$statistic)
  }
  else{
    test_stat_here[i] <- '-'
  }
}
# Senior Citizen Test Stat
test_stat_here[[2]]<- SC_chisq_$statistic

######## degree of freedom ############
df_here <- list() ## store the df here

for (i in 1:length(df_chiqsq)) {

  if (class(df_chiqsq[,i]) == "factor") {

    tab1 <- xtabs(~df_chiqsq[[colname_chi[i]]] + data$Churn )

    chisq_ <- chisq.test(tab1)

    df_here[i] <- paste(chisq_$parameter)
  }
  else{
    df_here[i] <- '-'
  }
```

```
}
# Senior Citizen DF
df_here[[2]]<- SC_chisq_$parameter

#### combine all the values into a table
chi_table <- data.frame(cbind(unlist(colname_chi), unlist(test_stat_here), unlist(df_here),
unlist(store_here) ))
colnames(chi_table) <- c('Variable', 'Test Statistic', 'DF', 'P_value')


P_value <- chi_table$P_value
P_value<- as.numeric(P_value)

isit = list()

for (i in 1: nrow(chi_table)) {
  if (P_value[[i]] < 0.05) { # If our significant value is 5%

    isit[i] <- 'Yes'
  }
  else{
    isit[i] <- 'No'
  }
}

p <- data.frame(matrix(unlist(isit), nrow=length(isit), byrow=TRUE))
chi_table<- cbind(chi_table, p)
names(chi_table)[5] <- 'Significant at 5% ?' # Is it significant to be rejected at this level?

row.names(chi_table) <- NULL # reset the index
chi_table

write.csv(chi_table,'/Users/yenshen/Desktop/FYP/Data log/Output/chi_table.csv')



############ EXAMPLE GENDER ############
############ chi_squared --- gender ##################

total_row <- apply(table_chi, 1, sum) # sum by row of female/male
total_col <- apply(table_chi, 2, sum) # sum by column of yes/no
total_n <- sum(table_chi) ## total of all men and women


O_male_no = table_chi[2,1]
O_male_yes = table_chi[2,2]
O_fem_yes = table_chi[1,2]
O_fem_no = table_chi[1,1]
```

```
# calculate the expected values
 # E = (row total x column total) / n

# expected of men, yes to churn
E_male_no = (3555* 5174)/total_n
E_male_yes = (3555* 1869)/total_n
E_fem_yes = (3488 * 1869)/ total_n
E_fem_no = (3488 * 5174)/ total_n


# chi squared = SUM [ ( (obs - exp)^2 / exp )]

gen_chi_sq_test =  ( (O_male_no - E_male_no)^2 / E_male_no ) +
                 ( (O_male_yes - E_male_yes)^2 / E_male_yes ) +
                 ( (O_fem_yes - E_fem_yes)^2 / E_fem_yes ) +
                   ( (O_fem_no - E_fem_no)^2 / E_fem_no )

gen_chi_sq_test

# Yate's corrected version
yate_gen_chi_sq_test =  ( ( abs(O_male_no - E_male_no) - 0.5)^2 / E_male_no ) +
  ( (abs(O_male_yes - E_male_yes) - 0.5 )^2 / E_male_yes ) +
  ( (abs(O_fem_yes - E_fem_yes) - 0.5 )^2 / E_fem_yes ) +
  ( (abs(O_fem_no - E_fem_no) - 0.5 )^2 / E_fem_no )

yate_gen_chi_sq_test


p_value_gen <- pchisq(0.4840829, 1, lower.tail = FALSE)
p_value_gen


table_chi <- xtabs(~data$gender + data$Churn )
chisq_gen <- chisq.test(table_chi)
chisq_gen


#-----------------------------------------------------------------------------------------
####### For numerical values #########
####  Find the mean and variance ####
## Tenure ## Monthly Charges ## Total Charges


######tenure
summary(data$tenure[data$Churn == 'No'])
summary(data$tenure[data$Churn == 'Yes'])
#std dev
```

```
sd(data$tenure[data$Churn == 'No'])
sd(data$tenure[data$Churn == 'Yes'])


#total charges
summary(data$TotalCharges[data$Churn == 'No'])
summary(data$TotalCharges[data$Churn == 'Yes'])
#std dev
sd(data$TotalCharges[data$Churn == 'No'])
sd(data$TotalCharges[data$Churn == 'Yes'])


##### monthly charges
summary(data$MonthlyCharges[data$Churn == 'No'])
summary(data$MonthlyCharges[data$Churn == 'Yes'])
#std dev
sd(data$MonthlyCharges[data$Churn == 'No'])
sd(data$MonthlyCharges[data$Churn == 'Yes'])

#------------------------------------------------------------------------------------------
################  Proportion tables #####################
##########  to be accompanied with the figures ##########

tab_here <- list()

for (i in 1:length(column_names)) {
  if (class(data[,i]) == "factor") {

    tab <- prop.table(xtabs(~ data[[column_names[i]]] + data$Churn ), 2)
    tab_here[[i]] <- tab
    write.csv(tab_here[[i]], file= paste(i, column_names[i], 'csv', sep = '.'))
    # for every proportion table, output it into a separate csv file
  }
  else{
    tab_here[[i]] <- 0
  }
}


######################################################################

### additional tables ######

## Paperless Billing & Senior Citizen
table_pb_sc <- xtabs(~data$PaperlessBilling+ data$Churn +data$SeniorCitizen)


table_dep_multiple <- xtabs(~data$Dependents+ data$Churn +data$MultipleLines)
```

```
##4

library(tidyverse)
library(caret)
theme_set(theme_bw())

################# PREPARING THE DATA ######################
## remove the CustomerID column
data_new <- subset(data, select = -customerID)

# one hot encoding is not needed as we have changed our categorical variables to factors


########### COMPUTE LOGISTIC REGRESSION ################

### FIT 1 ###
# fit the model with all the variables
model1 <- glm(formula = Churn~. , data = data_new, family = binomial)
# ~. means every possible combination of var

summary(model1)
summary(model1)$coef


### FIT 2 ### (with variables selection)
# fit the model with all the significant variables
data_new2 <- subset(data, select = -customerID)
data_new2 <- subset(data_new2, select = -gender) # not significant
data_new2 <- subset(data_new2, select = - PhoneService) # not significant

model2 <- glm(formula = Churn~., data = data_new2, family = binomial)

summary(model2)
summary(model2)$coef

### FIT 3 ###
# fit the model with all variables from Fit2 (variable selection) except for Paperless Billing
and Monthly Charges #
data_new3 <- data_new2
data_new3<- subset(data_new3, select = -PaperlessBilling)
data_new3 <- subset(data_new3, select = -TotalCharges)

model3 <- glm(formula = Churn~., data = data_new3, family = binomial)

summary(model3)
summary(model3)$coef
```

```
##### plot the predicted probability chart ############
predicted_data <- data.frame(prob_of_churn = model$fitted.values,
                             Churn = data_new$Churn)

predicted_data <- predicted_data[
  order(predicted_data$prob_of_churn, decreasing = FALSE), ] # sort from low to high prob

predicted_data$rank <- 1:nrow(predicted_data) # add a new column that ranks it from low to
high prob

ggplot(data = predicted_data, aes(x = rank, y = prob_of_churn)) +
  geom_point(aes(color = Churn), alpha = 1, shape = 4, stroke = 2) +
  xlab("Index") +
  ylab("Predicted probability of churning")

############## How do I determine which model is better? #####################
########################## PREDICTION PROBABILITY
##############################
## with a threshold of 0.5
# if it is more than 0.5, output 1 - No Churn
# if it is less than 0.5, ouput 0 - Yes Churn
# sum up the entries
# compare each model with each classification

Churn <-  data_new$Churn
Churn <- as.factor(Churn)
Churn <- as.numeric(Churn) # 1 as No, 2 as Yes
Churn <- as.factor(Churn)


############### for model 1 prediction ################################
value_1 <- predict(model1, type = 'response')
# baseline is No, so the probabilities are prob of not churning

list_1 <- list() # store into this list of model 1 outputs whether churn or not

for ( x in 1: length(value_1) ) {

  # baseline is No, so the probabilities are prob of not churning

  if (value_1[x] >= 0.05) {  #if more than 0.05
    list_1[x] = 1 # No to churn

  }
  else {
    list_1[x] = 2 # Yes to churn
}
```

```
}

### compare to the true value of churn
### Misclassification = 1
### No misclassification = 0

mis_rate_1 <- ifelse(list_1 == Churn, 0, 1)
# if same, output 0. If not, output 1
# 0 - correct
# 1 - misclassification

mean(mis_rate_1)
# 0.5222206

################################################################################
################
############### for model 2 ################################

value_2 <- predict(model2, type = 'response')
# baseline is No, so the probabilities are prob of not churning


############## for model  prediction ################################

list_2 <- list() # store into this list of model 2 outputs whether churn or not

for ( x in 1: length(value_2) ) {

  if (value_2[x] >= 0.05) { #if more than 0.05
    list_2[x] = 1 # No to churn

  }
  else {
    list_2[x] = 2 #Yes to churn
  }
}

mis_rate_2 <- ifelse(list_2 == Churn, 0, 1) # if same, output 0. If not, output 1

mean(mis_rate_2)
# 0.5232145

################################################################################
################
############### for model 3 ################################

value_3 <- predict(model3, type = 'response')
# baseline is No, so the probabilities are prob of not churning
```

```
############## for model 3 prediction ###############################

list_3 <- list() # store into this list of model 3 outputs whether churn or not

for ( x in 1: length(value_3) ) {

  if (value_3[x] >= 0.05) { #if more than 0.05
    list_3[x] = 1 # No to churn

  }
  else {
    list_3[x] = 2 #Yes to churn
  }
}

mis_rate_3 <- ifelse(list_3 == Churn, 0, 1)
# if same, output 0. If not, output 1

mean(mis_rate_3)
# 0.5156893

# deviance models
# This is a generic function which
# can be used to extract deviances for fitted models
deviance(model1)
deviance(model2)
deviance(model3)


# Conduct the likelihood ratio test on each model
anova(model1, model2, test = 'LRT')
anova(model2, model3, test = 'LRT') # significant
```