UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI

APPLIED MACHINE LEARNING FOR CUSTOMER CHURN PREDICTION IN TELECOM INDUSTRY

Subject: Machine Learning and Data Mining II

Student group 3:

Pham Hai Nam (BI12 307)
Phung Duc Thai (BI12 396)
Hứa Hải Minh (BI12-272
Nguyễn Anh Quân (BI12-365)
Phạm Xuân Trung (BI12-458)

Hanoi, March 22th, 2023

ABSTRACT

Customer churn refers to the pace at which customers quit a company. Churn may be caused by a variety of circumstances, such as moving to a competitor, canceling their subscription owing to bad customer service, or ending all interaction with a business due to a lack of touchpoints. Long-term consumer connections are more effective than attempting to acquire new customers. A 5% improvement in customer satisfaction leads to a 95% increase in sales. Companies may forecast future revenue by analyzing historical behavior. The purpose of this research is to determine which characteristics in the Telecommunication dataset impact customer attrition in the California telecoms business. The aim of this study was to use Machine Learning algorithms to identify the elements that contribute to customer churn and to offer a churn prediction framework that is presently absent in the telecoms sector.

I. Introduction and theoretical backgrounds

In more economically developed countries, the telecommunications industry has grown to be one of the most significant (Lew Sook Ling, 2021). Because of technical developments and a growth in the number of operators, the level of competition has increased. Moreover, in this competitive market, thousands of companies have to be working diligently to keep those businesses operating by applying a number of both short and long-term strategies, including lengthen the customer retention period as the most lucrative strategy (Kabu Khadka, 2017). This is because maintaining an existing client is significantly less expensive than obtaining a new one, and it is also thought to be much easier than the upselling strategy (Lew Sook Ling, 2021). Companies must reduce the probability of client churn, often known as "the customer transfer from one supplier to another," in order to execute more effectively (Dalvi, Khandge, Deomore, Bankar, & Kanade, 2016). As a result, a plethora of innovative solutions in both conventional and practical applications are presented in order to address this serious problem. Machine learning algorithms are preferred to anticipate churn in competitive service industries, assisting the organization in gaining a significant income stream. In this study, churn prediction is studied and researched using a dataset from the telecommunications industry that may provide promising outcomes.

In the telecommunications industry, market competitiveness is measured by the churn rate. Telecommunication normally includes telephone, internet, and mobile services. One Internet Service Provider (ISP), for example, with 20 users, will cancel, resulting in a 5% decrease in income; on a daily basis, the telecom industry loses 20-40% of its consumers (Ming Zhao, 2021). Customer retention and satisfaction represent their relevant feedback in the positivity with a company’s products, services and capabilities. Customer retention and satisfaction show their positive feedback on a company's products, services, and capabilities. Essentially, the buyer experience has a significant impact on customer happiness, which leads to an increase in sales and recommendations (Daniel Hattenbach, 2004). To anticipate customer demands, staff support should be necessary to manage what customers desire, which includes not only personal wants, but also the overall consumer inclination to purchasing patterns. Moreover, exceptional services are defined as greater services with high-quality help offered by organizations that match consumers' pleasure and expectations. Ability, attitude, appearance, attentiveness, action, and accountability are some examples. When a company handles its customers well, it not only helps suppliers attract consumers but also saves a significant amount of money throughout business execution (Katherine N. Lemon, 2016).

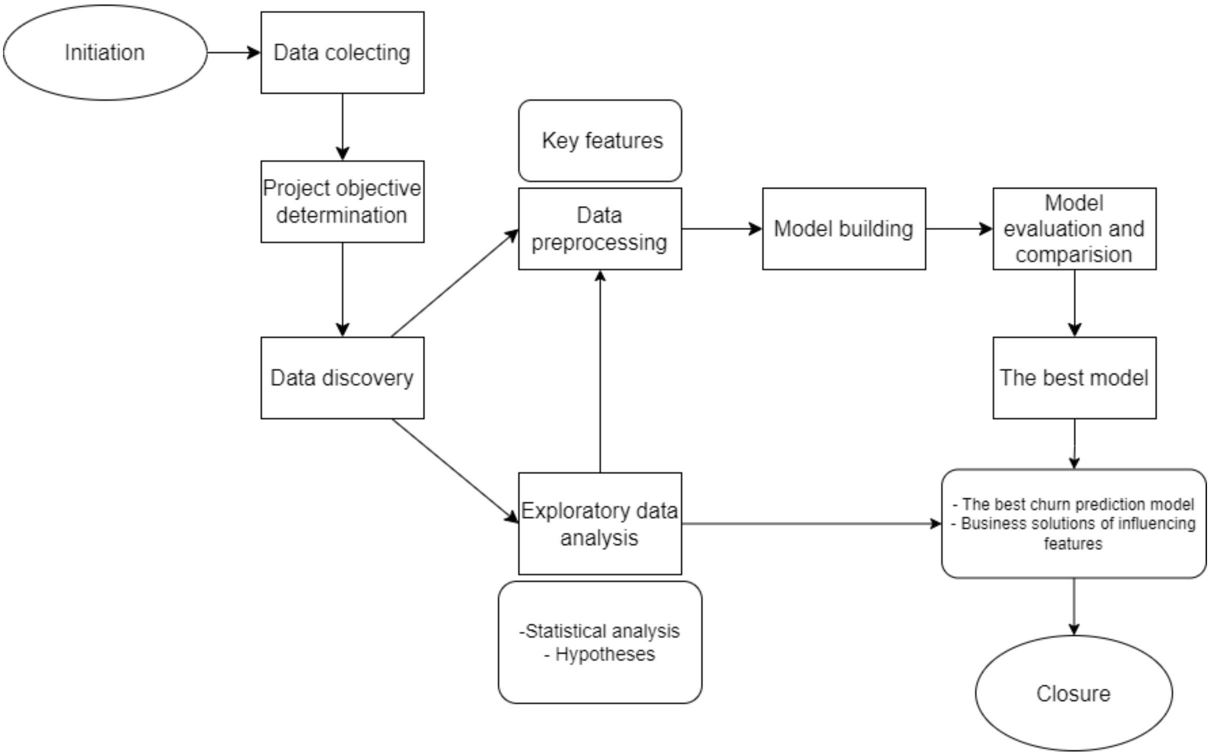

Initiation
Data colecting
Project objective determination
Data discovery
Key features
Data preprocessing
Model building
Model evaluation and comparision
The best model
Exploratory data analysis
-Statistical analysis
- Hypotheses
- The best churn prediction model
- Business solutions of influencing features
Closure

Image: Scientific outlines

II. Objectives

- Construct a churn prediction algorithm to help telecom providers forecast customers who are likely to be churners.
- Locate potential consumers.
- Provide optimal solutions to maximize loyal consumers.

According to objectives, there could be some particular questions needed to answer:

+ Which applied models have the best prediction model accuracy?
+ What components of customer data are responsible for becoming potential?
+ What should businesses do after successfully assessing improved offerings that significantly improve customer experiences and retention?

III. Experimental protocols (*Image*)

a. Data Preprocessing

The dataset was processed in order to convert it from its raw state to features for use in machine learning methods. Values of columns per month were aggregated (average, count, sum, max, min, etc) from each numerical column per customer and the count of distinct values for categorical columns. Before any further analysis, all the missing values of the data set would be detected and handled. It is possible to accomplish this by removing or replacing missing data with a substituted value. The data set consists of quantitative variables, so that any categorical data must be converted into numerical values before further analysis. To establish how the data attributes, connect to one another, a correlation matrix is generated. If a feature has a high correlation coefficient with another feature in the matrix, either of them will be eliminated. Eventually, only the most crucial features are preserved. The properties of the data set and the labels would be separated (Salvador García, 2016).

b. Exploratory Data Analysis (EDA)

In order to observe the distribution of every data in the dataset, bar charts used to picture categorical distribution, along with binary classification. On the other hand, kernel density estimates (KDE) and histogram are preferred to construct both the fluctuation and distribution of numerical data according to the periods of time (David M. Levine, 2010). Moreover, missing values are also major problems if existing in the dataset. Therefore, using "missing values" checks to examine whether the entire dataset has any missing values or not.

By observing using three methods, such as bar charts, KDE, and histogram, the dataset first must select certain properties that appear to be important to pre-analyze and then observe them using bar charts or pie charts depicting the distribution. Furthermore, the heat map is a further phase that describes the correlation between each sector in the dataset, leading to specialized feature selection.

Last but not least, based on the results achieved while applying algorithms, statistical analysis should be another step to demonstrate the final results to strengthen conclusions. Bar and pie forms are used for evaluating the proportions of every data after choosing the best ones through important features extracted by studied algorithms. Figure 1 shows the process of this study while performing Exploratory Data Analysis.

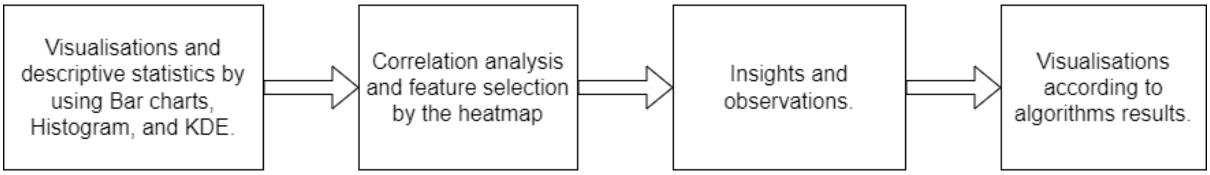

Visualisations and descriptive statistics by using Bar charts, Histogram, and KDE.
Correlation analysis and feature selection by the heatmap
Insights and observations.
Visualisations according to algorithms results.

Figure 1: The process of this study in EDA

c. Model Building - Experimental protocols

General steps explanation

Figure 2 depicts the process utilized in this study to evaluate the prediction model as well as identify critical factors directly connected to the outcome, whether customers are churners, in detail. To begin, P-values were used to determine the correlation between each characteristic and result by developing null and alternative hypotheses. Finally, to demonstrate whether or not null hypotheses are valid, certain specific aspects pertaining directly to the outcome were displayed. Following data pre-processing, three methods, including Decision Tree, Random Forest, and Multilayer Perceptron, were applied to the dataset. Moreover, a confusion matrix was created to properly test algorithm accuracy. Moreover, combined with K-fold cross validation and Area under the Curve (AUC) results of accuracy, the predictive models can be compared leading to the exact results of which model should be more distinctive and more accurate than others. Additionally, the important features were also looked for based on the results of three algorithms; the results of three most important features, along with less important or supporting features that seem to be considered to analyze in order to have better prediction were taken to assume what relates to real-life backgrounds these days.

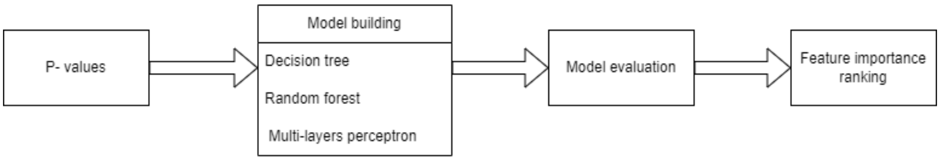


Figure 2: The process of model building and evaluation

Detailed functions of algorithms

Algorithm	Description
P-values	P- value is calculated to validate a hypothesis on a dataset. In this study, p-value is used to determine whether a data set feature has a significant impact on the output value. The lower p-value is, the more significant the features could be. (Gelman, 2013) + $p > 0.10$ → "not significant" + $p \leq 0.10$ → "marginally significant" + $p \leq 0.05$ → "significant" + $p \leq 0.01$ → "highly significant."
K-fold cross validation	The technique is applied when the data set is not massive enough for testing models. The idea is splitting the data into a k number of folds of equal size. In this method, k testing data sets are randomly picked from the original data set (Xinyu Zhang, 2022).
Area under the curve (AUC)	Area under the curve or area under the receiver operating characteristics is an evaluation metric used for measuring a classification model performance. The curve is plotted by the true positive rate as y-axis and the false positive rate as x axis. The higher AUC, the better the model (Elise F. Zipkin, 2012).
Decision tree	Decision tree is a classification and regression method. A decision tree has the structure of a tree, each internal node represents a condition, followed by the outcomes after each condition, denoted by the branches. At the end of the tree are the leaf nodes, representing the final decisions or data labels. (Chidanand Apté, 1997)
Random forest	Random forest is the use of multiple classifier models for classifying the same data set, all the outcomes are combined to generate a single result. By following the large number principle, a random forest classifier is considered to reduce the error and give better performance, compared to a single classifier (Andreas Ziegler, 2013).
Perceptron classifier	Perceptron is an algorithm used for binary classification. It is considered as the simplest type of neural network model. The input for a perceptron model is the sum of all the input values multiplied by the model coefficients, called weighted sum. The value of the weighted sum will be evaluated whether it is greater or less than a given threshold to determine which class the output should be. (Mohammad Ridwan Ismail, 2015)

Table 1: Churn prediction techniques

Experimental protocols

No	Models	Protocols
1	**Decision tree model**	The classification model is built by using the Decision tree function which is available on the Sklearn library, with the min sample leaf of the tree to 1, the minimum number of required samples to split an internal node to 2. Gini is used as a function to measure the quality of the split. **Time complexity** (Tjen-Sien Lim, 2000)**:** Complexity = O (n × m × log(n)) *Whereas:* n: number of samples m: number of features
2	**Random forest model**	The Random forest classifier is built by calling a function called RandomForestClassifier from the package ensemble, with 100 estimators by default, which means there are 100 different trees in the forest. **Time complexity** (Tjen-Sien Lim, 2000)**:** Complexity = O (n × m × k × n × log(n)) *Whereas:* k: number of trees m: number of features n: number of samples
3	**Multi-layer perceptron**	The MLP model is trained by using the MLPClassifier from scikit-learn with one hidden layer of 100 neurons, the relu activation function, the Adam solver, and a random seed of 42. All 4 activation functions, including “ relu”, “identity”, “logistic” and ”tank” and all 3 solvers, which are “sgd”, “ adam”, “lbfgs” are tried until the accuracy score is as stable as possible. **Time complexity** (Tjen-Sien Lim, 2000)**:** Complexity= O (n × m × h × k × o × i) *Whereas:* n: training samples m: number of features k: number of hidden layers h: number of neurons o: output neuron i: number of iterations If number of neurons in each layer is different calculate the average neurons. $h_{avg}=(h_1 \times h_2 \times \ldots \times h_k)^{1/k}$

Table 2: Churn prediction protocols and time complexity calculation

4. Ranking feature importance

To calculate feature importance, every node's importance of the decision tree has to be calculated as:

Node's importance = $\dfrac{\%\mathrm{SR} \times \mathrm{I} \times \%\mathrm{SRLS} \times \mathrm{ILS} - \%\mathrm{SRRS} \times \mathrm{IRS}}{100}$

Whereas:

SR: Sample reaching
I: Impurity
SRLS: Sample reaching left subtree
SRRS: Sample reaching right subtree
ILS: Impurity left subtree
IRS: Impurity right subtree

Now the importance of all features can be calculated as:

Feature importance = $\dfrac{\sum Node\ importance\ splitting\ on\ the\ feature}{N}$

Whereas:
N = All node's importance

IV. Research model - Customer churn determinants

H1: Demographic Characteristic

The connections between personal backgrounds and churn rate may affect collaborations and adaptability in customer service, which may involve arranging exclusive events for customers. Additionally, the dataset emphasizes the importance of identifying whether customers have partners or dependents, or are single, as well as other attributes such as gender and age range, with a particular focus on those under the age of 65.

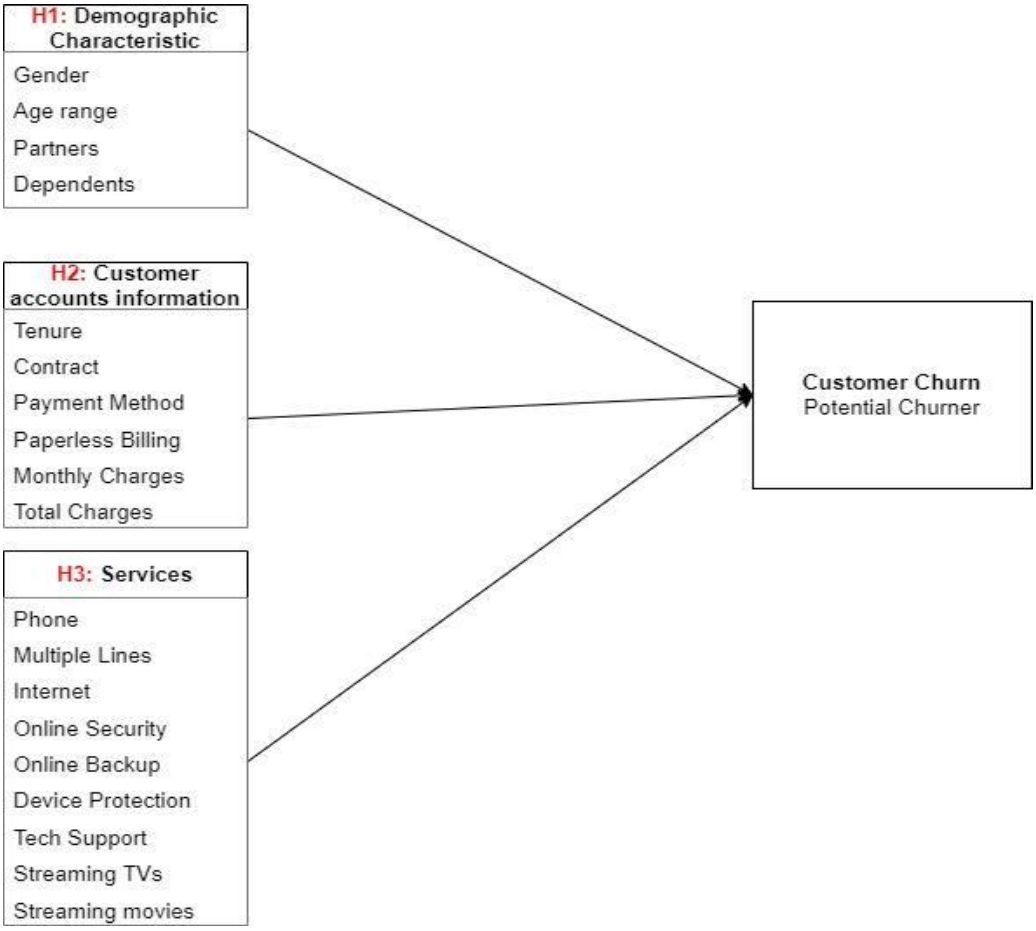


Figure 3: A conceptual model for the prediction of potential churners.

H2: Customer account information

The retention of customers is critical for businesses, and high retention rates are typically associated with satisfied customers who are willing to do business with a company again. Customer satisfaction is also linked to a willingness to pay more for products or services, with highly satisfied customers being more willing to pay than moderately satisfied or dissatisfied customers. The paragraph goes on to describe some of the factors that the H2 data analysis focuses on, including the length of time customers have used the services, payment methods, billing methods, and charges monthly and its total amount.

H3: Services

Customer service, including technical support and service staff responsiveness, is an important determinant factor in ISP selection. However, in this study, internet services also used to examine whether those affect customers' retention and behaviors or not. 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.

H1 a	Gender is positively associated with the customer churn probability
H1 b	SeniorCitizen is positively associated with the customer churn probability
H1 c	Partner is positively associated with the customer churn probability
H1 d	Dependents is positively associated with the customer churn probability

Table 3: H1 Demographic Characteristics

H2 a	Tenure is positively associated with the customer churn probability
H2 b	Contract is positively associated with the customer churn probability
H2 c	Payment method is positively associated with the customer churn probability
H2 d	Paperless billing is positively associated with the customer churn probability
H2 e	Monthly charges is positively associated with the customer churn probability
H2 f	Total charges is positively associated with the customer churn probability

Table 4: H2 Customer accounts information

H3 a	Phone is positively associated with the customer churn probability
H3 b	Multiple lines is positively associated with the customer churn probability
H3 c	Internet is positively associated with the customer churn probability
H3 d	Online security is positively associated with the customer churn probability
H3 e	Online backup is positively associated with the customer churn probability
H3 f	Device protection is positively associated with the customer churn probability
H3 g	Tech support is positively associated with the customer churn probability
H3 h	Streaming TVs is positively associated with the customer churn probability
H3 i	Streaming movies is positively associated with the customer churn probability

Table 5: H3 Services

	No	Column	Data type	Details
H1	1	gender	qualitative	The customer's gender (male or female)
	2	SeniorCitizen	qualitative	Whether the customer is a senior citizen (1) or not (0)
	3	Partner	qualitative	Whether the customer has a partner (Yes or No)
	4	Dependents	qualitative	Whether the customer has dependents (Yes or No)
H2	5	tenure	quantitative	The number of months the customer has been with the company
	6	Contract	qualitative	The type of contract the customer has (Month-to-month, One year, or Two year)
	7	PaymentMethod	qualitative	The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), or Credit card (automatic))
	8	PaperlessBilling	qualitative	Whether the customer has paperless billing (Yes or No)
	9	MonthlyCharges	quantitative	The amount charged to the customer monthly
	10	TotalCharges	quantitative	The total amount charged to the customer
H3	11	PhoneService	qualitative	Whether the customer has phone service (Yes or No)
	12	MultipleLines	qualitative	Whether the customer has multiple lines (Yes or No)
	13	InternetService	qualitative	The type of internet service the customer has (DSL, Fiber optic, or No)
	14	OnlineSecurity	qualitative	Whether the customer has online security (Yes or No)
	15	OnlineBackup	qualitative	Whether the customer has online backup (Yes or No)
	16	DeviceProtection	qualitative	Whether the customer has device protection (Yes or No)
	17	TechSupport	qualitative	Whether the customer has tech support (Yes or No)
	18	StreamingTV	qualitative	Whether the customer has streaming tv (Yes or No)
	19	StreamingMovies	qualitative	Whether the customer has streaming movies (Yes or No)

Table 6: Dataset details.

V. Results and achievements

a. Data collecting

There are 7043 rows and 21 columns in the original Telco customer churn data set in California. Each row represents a customer, and each column represents a customer attribute. 21 columns, including one for targeted value and 20 for consumer demographics, customer account information, and services. The customer churn determinants are likely to be built using a conceptual model that will assist providers in learning more about the individual customer data in the overall dataset. Figure 3 displays the conceptual model, which depicts mechanisms that may directly impact customer churn.

b. Exploratory data analysis

Figure 4 and 5 visualize some of the basic categorical and numerical features to give more insight on the difference between churn and non-churn classes in the given dataset.

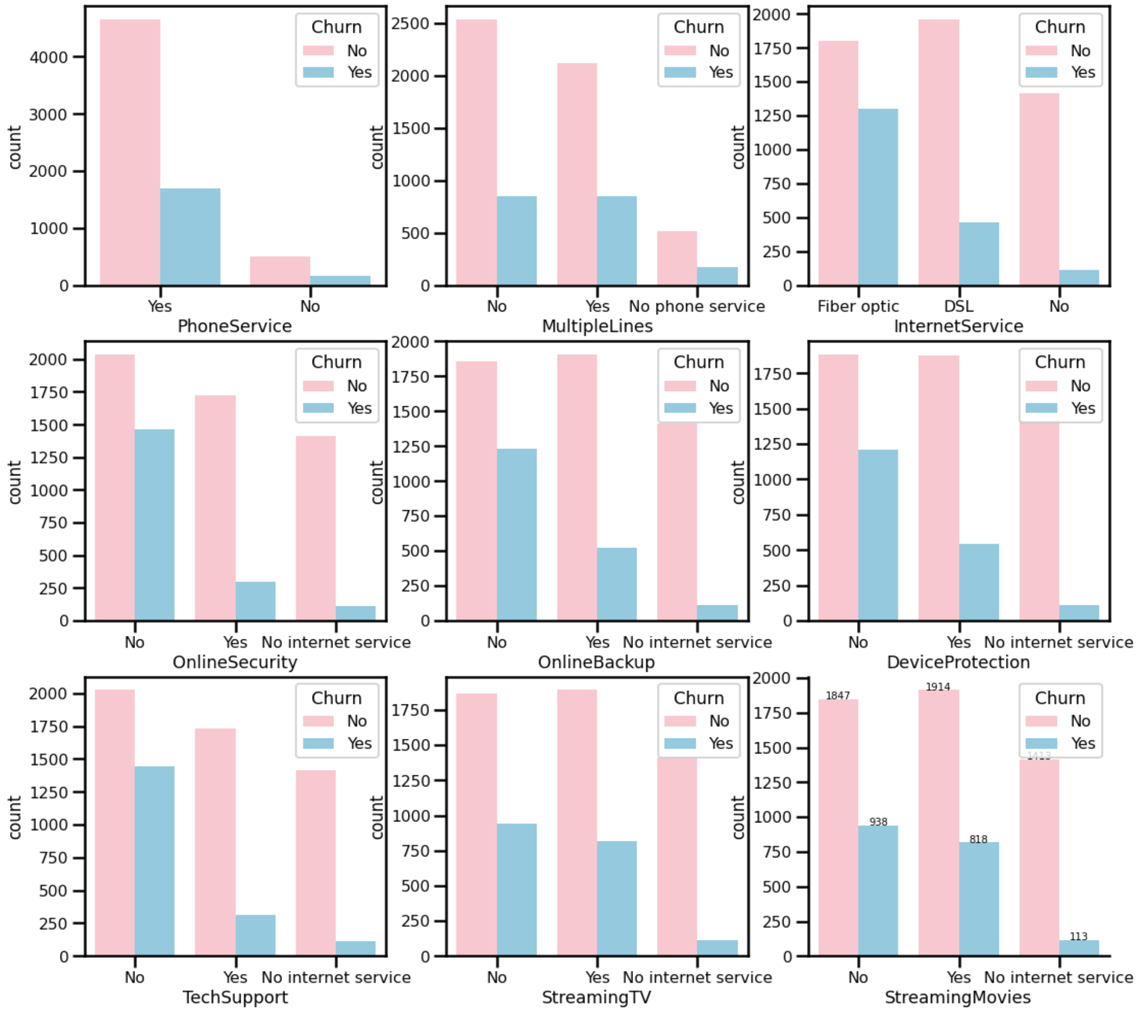


Figure 4: Distribution of some main categorical features

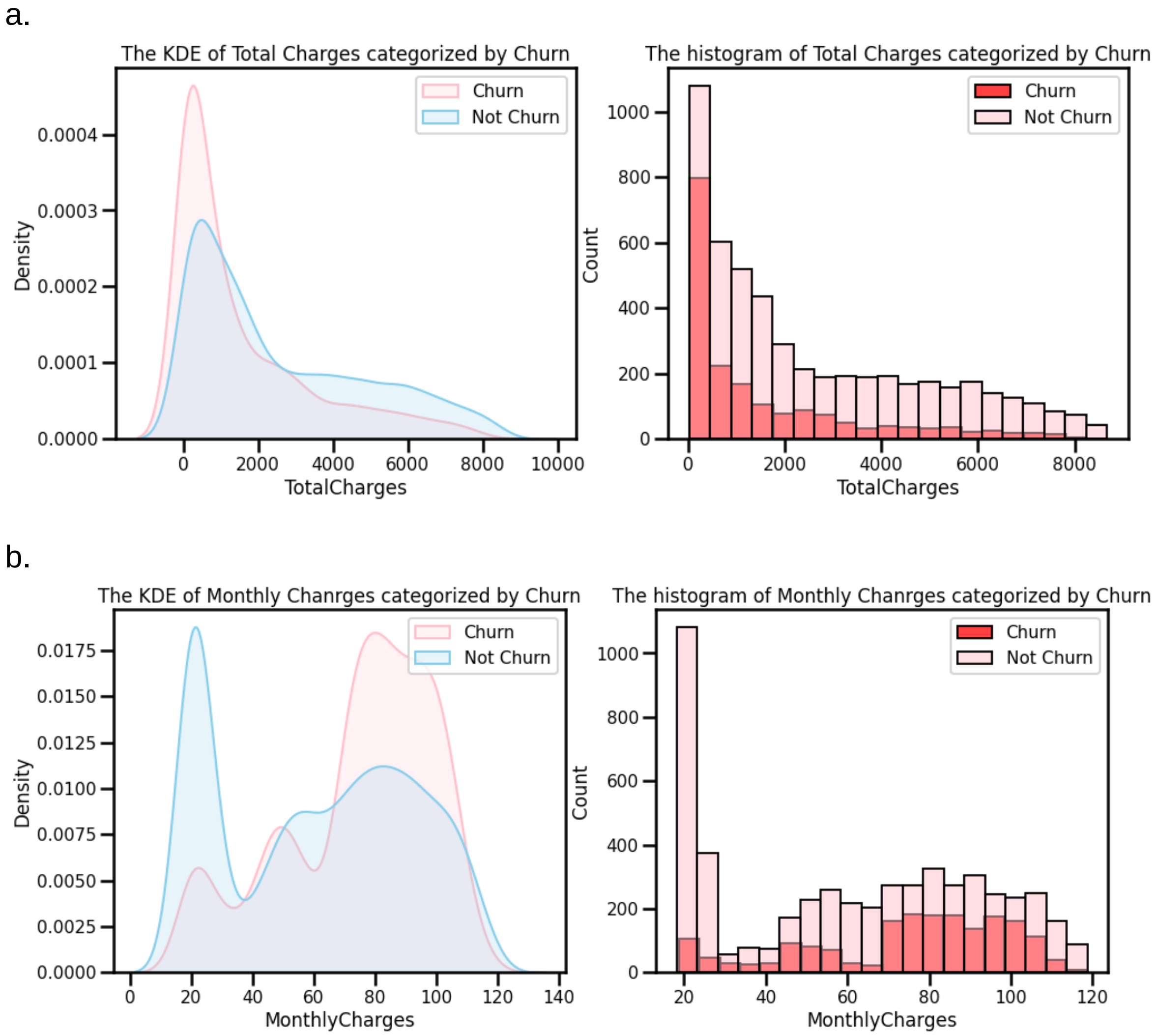


Figure 5: Feature distribution for some main numerical features. The left-handed side figures are kernel density estimates (KDE) representing churn as light pink color and light blue as non-churn. The right-handed side figures are histograms with the bold red as churn and another is non-churn. (a) Visualizes the distribution of total charges according to churn's classification. (b) also illustrates the distribution of charges monthly.

Moreover, the missing values with other values derived from either the same features and other features. This method is preferable so that it enables us to use the information in most features for the training process. As a result, there may not be missing values in the entire dataset, following the illustration depicted in figure 6.

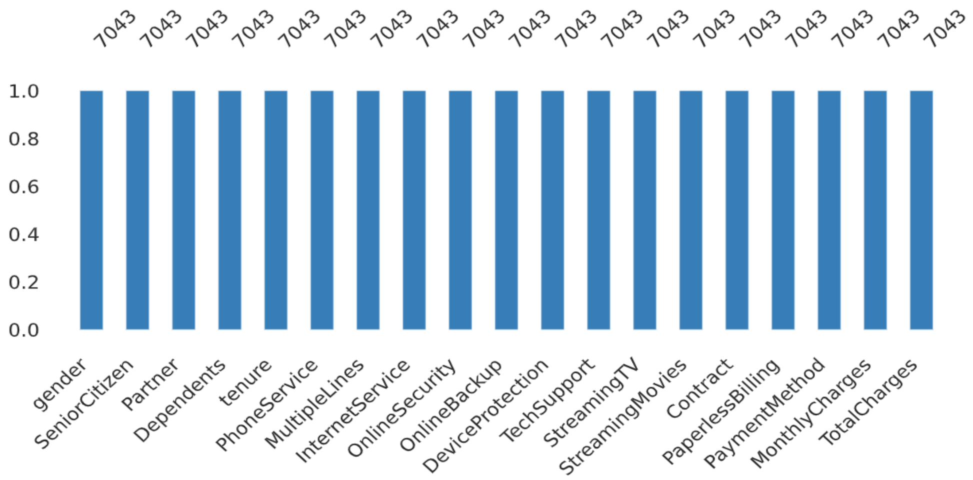


Figure 6: The entire features in the dataset for missing values checking.

The correlation between attributes also was performed in order to determine and measure the possibility of correlated features in the given dataset. In figure 7, this is because those given attribute values may be almost continuous leading to the less correlation, these represented through the number of lighter colors in this figure. On the other hand, there are some specific attributes that may be better correlated to each other, including monthly charges and total charges.

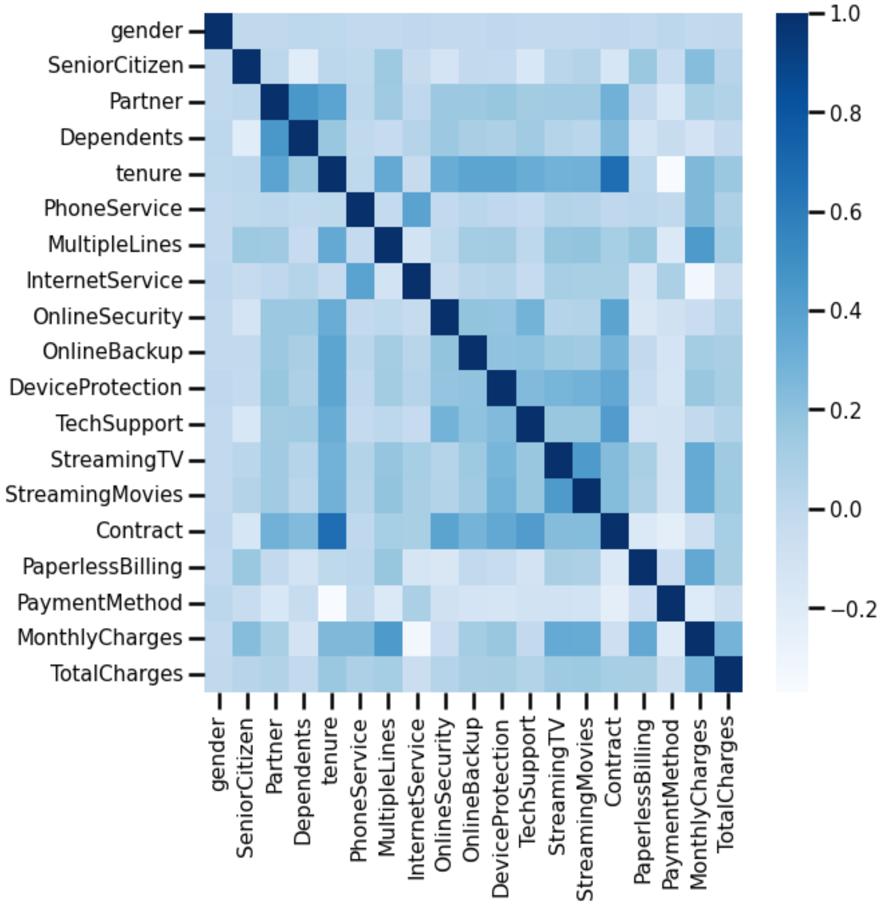


Figure 7: The heat map represents the correlation among attribute values. The bolder objects illustrate, the better correlated those show.

c. Statistical analysis

The excel tools are used in order to give the null hypotheses to the model. Based on the figure 3 giving the results to have initial hypotheses before exactly using methodology to analyze. The purpose of this analysis is to base on the technical uses in order to measure which part plays an important role in the prediction results, but also help to explain carefully about the results and future prediction of the dataset.

Senior citizens

The demographic characteristic is used to describe the personal information of customers. The service section is then reviewed to determine which service should company managers prioritize in order to improve customer satisfaction and experience. The senior citizens feature is utilized to determine which age group is more likely to benefit from the service. Figure 8 shows that adults above the age of 65 (approximately 1142 people) are less likely to register for and use this service than those under the age of 65 (about 5901 people). Briefly, those under the age of 65 are around 5 times larger than others.

In the left-handed side of the figure 9, customers who are smaller than 65 years old should be analyzed based on their percentage of staying and leaving. Figure 9 depending on their proportion of remaining and departing. It may be straightforward to figure out the fact that there are roughly 83% customers who maintain using the service, whilst 17% users abandon service after 1-month registration.

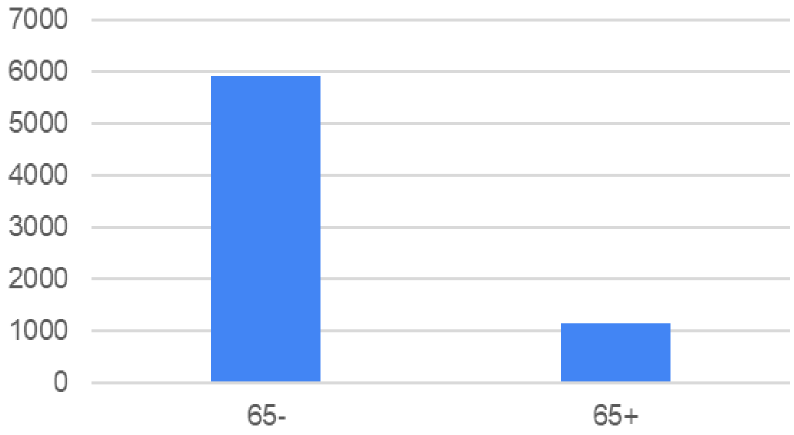
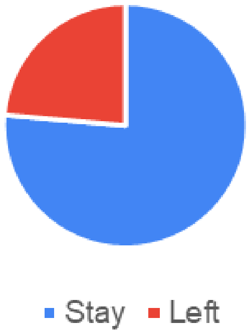
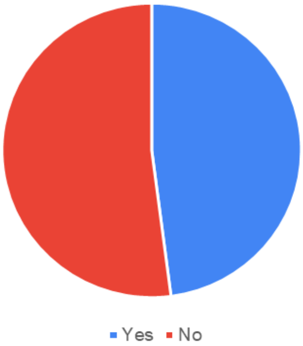


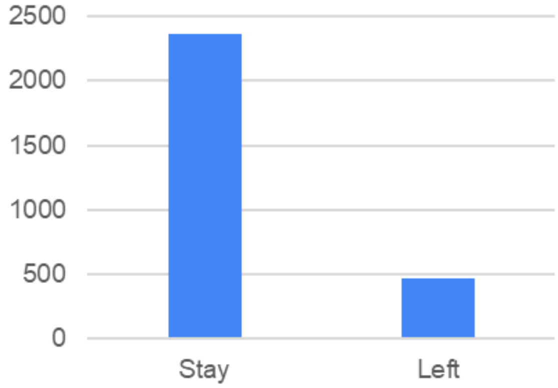
Figure 8: The distribution of both above and under the age of 65 in the service



The number of customer under 65 years old in the service.



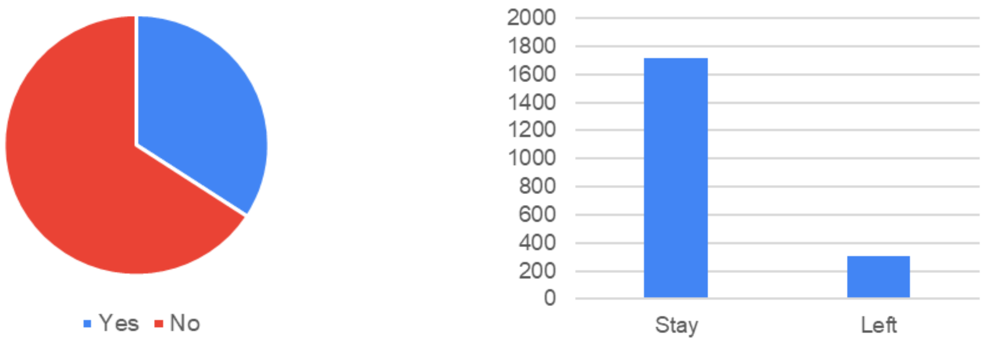
The number of customers under 65 years old with partners in the service.



The number of customer under 65 years old with partners in the service.

Figure 9: The right-handed side chart: the proportion of customers smaller than 65-year-old in staying and leaving in the entire dataset. The left-handed side chart: The percentage of people, who are smaller than 65 years old, have partners. The below chart: The number of people having partners, who stay and leave the service.

The charts compare the percentages of people who are smaller than 65 years old equivalent to their proportion of whether they get married or not, then their proportion of leaving and staying in the service. In the left-handed side chart, the percentage of people smaller than 65 years old use the service more than those greater than 65 years old about 3 times. In comparison to both the right-handed side chart and below bars chart, the number of people having partners is smaller than people who do not but the proportion of people staying is much higher. This is illustrated by showing the fact that the percentage of people having partners less than the rest is about 4 to 5% but in the bar chart, those both getting married and staying in the service are greater about 5 times compared with those both getting married and leaving the service.



The number of customers under 65 years old with dependents in the service.

The number of customers under 65 years old with dependents in the service.

Figure 10: The left-handed side pie chart: showing the proportion of people having dependents as a means of having members of the family. The right-handed side bar chart: Showing the number of people having dependents leaving and staying.

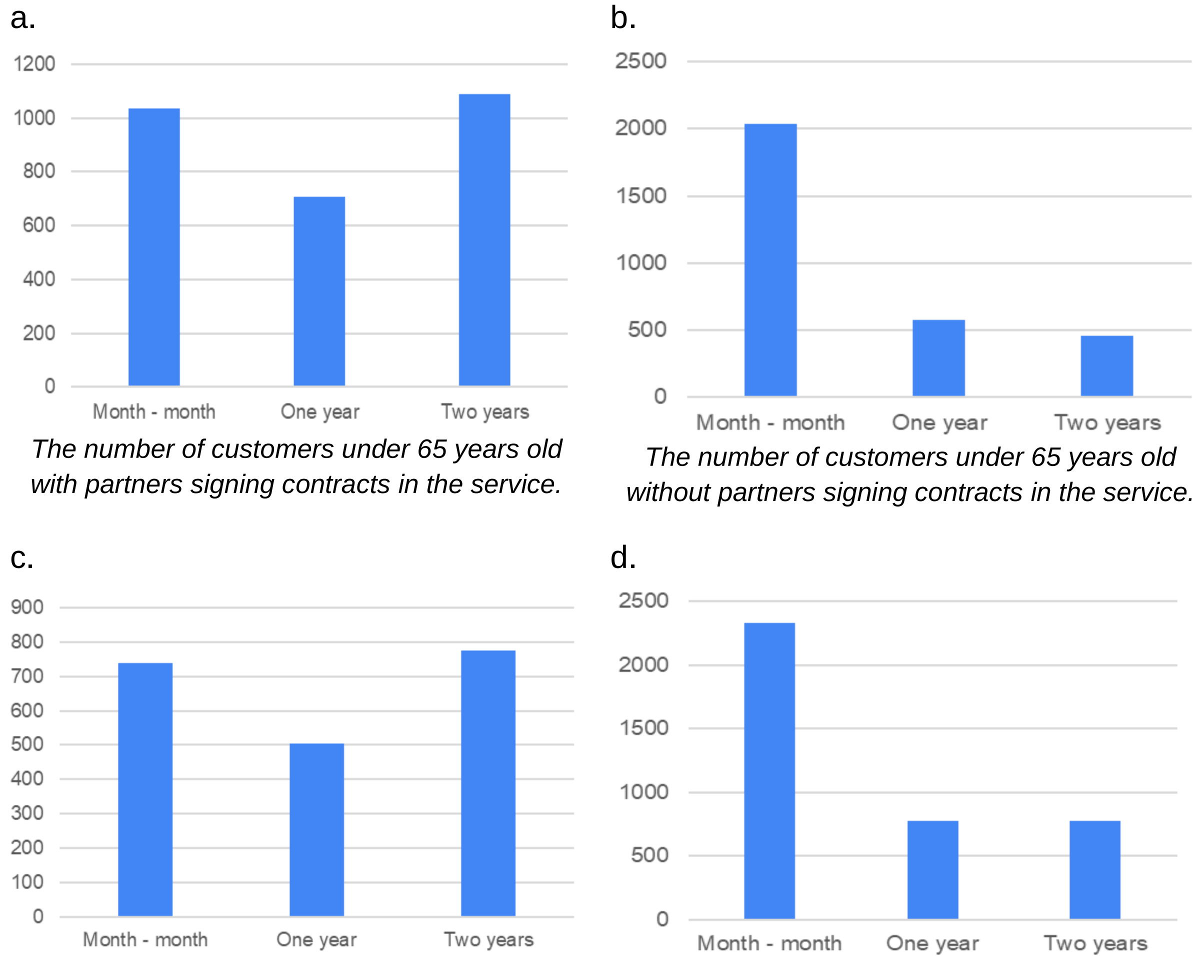
Equivalently, the percentage of persons who have dependents and are under 65 years old is lower when compared to those under 65 who have no dependents and are younger. The magnitude of the staying percentage on the right-hand side figure is approximately 8 times bigger than the departing one.

Then, this leads to the first conclusion that those under the age of 65 should be considered to be in the consumer segmentation. Furthermore, they should have partners or dependents to keep as many clients as feasible. All of the features listed below are in a services group that directly serves consumers in order to determine which service should be promoted by customers in order to improve customer experiences and attract more future potential customers.

Contract

The comparison between whether or not people sign the contract when they have partners or dependents.

The bar charts compare the different people having partners and dependents or not leading to their decision of the contracts signature, monthly, annually or two years. Overall, in comparison of charts when people have partners or dependents, the percentages in 3 types of contracts are fairly similar; however, it seems to be the fact that those having partners may register the contracts more than the others. This is because, for example, in the figure 11 a, the highest sector that shows about 1100 people in two-year contract; compared with it in figure 11 c, it is only nearly 800 people. However, both figures 11 b and d, it seems to figure out that people who don't have partners or dependents only use service monthly instead of annual or two-year contracts. That can be assumed that the number of people registering monthly contracts is significantly greater than both one or two-year contracts, about 2 or 3 times. On the other hand, the percentage of both contracts is relatively the same in figure 11 b and d.



The number of customers under 65 years old with partners signing contracts in the service.

The number of customers under 65 years old without partners signing contracts in the service.

The number of customers under 65 years old with dependents signing contracts in the service.

The number of customers under 65 years old without dependents signing contracts in the service.

Figure 11: The number of people who sign contracts, in terms of month, a year and 2 years, when having partners or dependents.

As can be seen from the chart, for younger citizens than 65 years old that have partners or dependents, the number of contracts signed in all 3 types is fairly balanced. However, for those without partners and dependents, most will choose to sign a month to month contract.

Overall comment whether or not people sign the contract

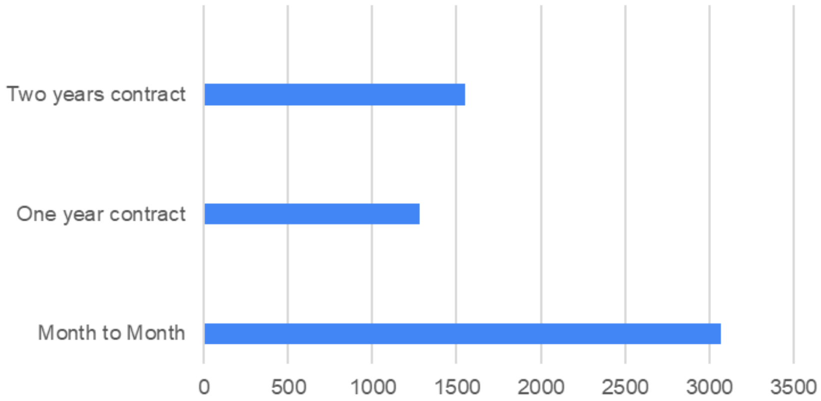
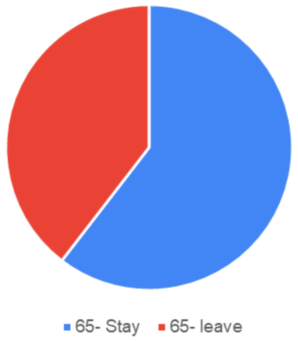


Figure 12: The number of people younger than 65 years old signing three types of contracts, including two-year, annual and monthly contracts.

a. Month to month contract

	People	Percentage
Total sign	3068	51.99%
Staying in the service	1854	60.43%
Leaving the service	1214	39.57%

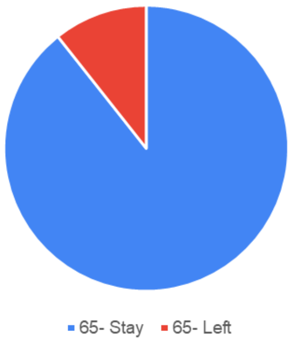


The number of customers under 65 years old signing month to month contracts in the service.

Table 7 and Figure 13: The number of people younger than 65 years old signing the month-to-month contract and their percentage.

b. One-year contract

	People	Percentage
Total sign	1283	21.74%
Staying in the service	1146	89.32%
Leaving the service	137	10.68%

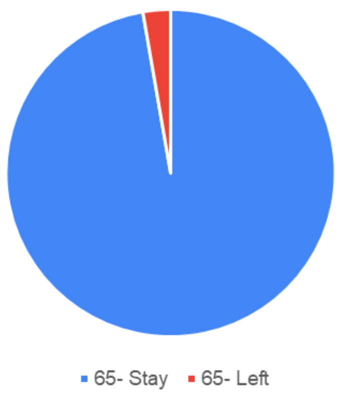


The number of customers under 65 years old signing one year contracts in the service.

Table 8 and Figure 14: The number of people younger than 65 years old signing the month-to-month contract and their percentage.

c. Two-year contract

	People	Percentage
Total sign	1550	26.27%
Staying in the service	1508	97.29%
Leaving the service	42	2.71%



The number of customers under 65 years old with signing two-year contracts in the service.

Table 9 and Figure 15: The number of people younger than 65 years old signing the month-to-month contract and their percentage.

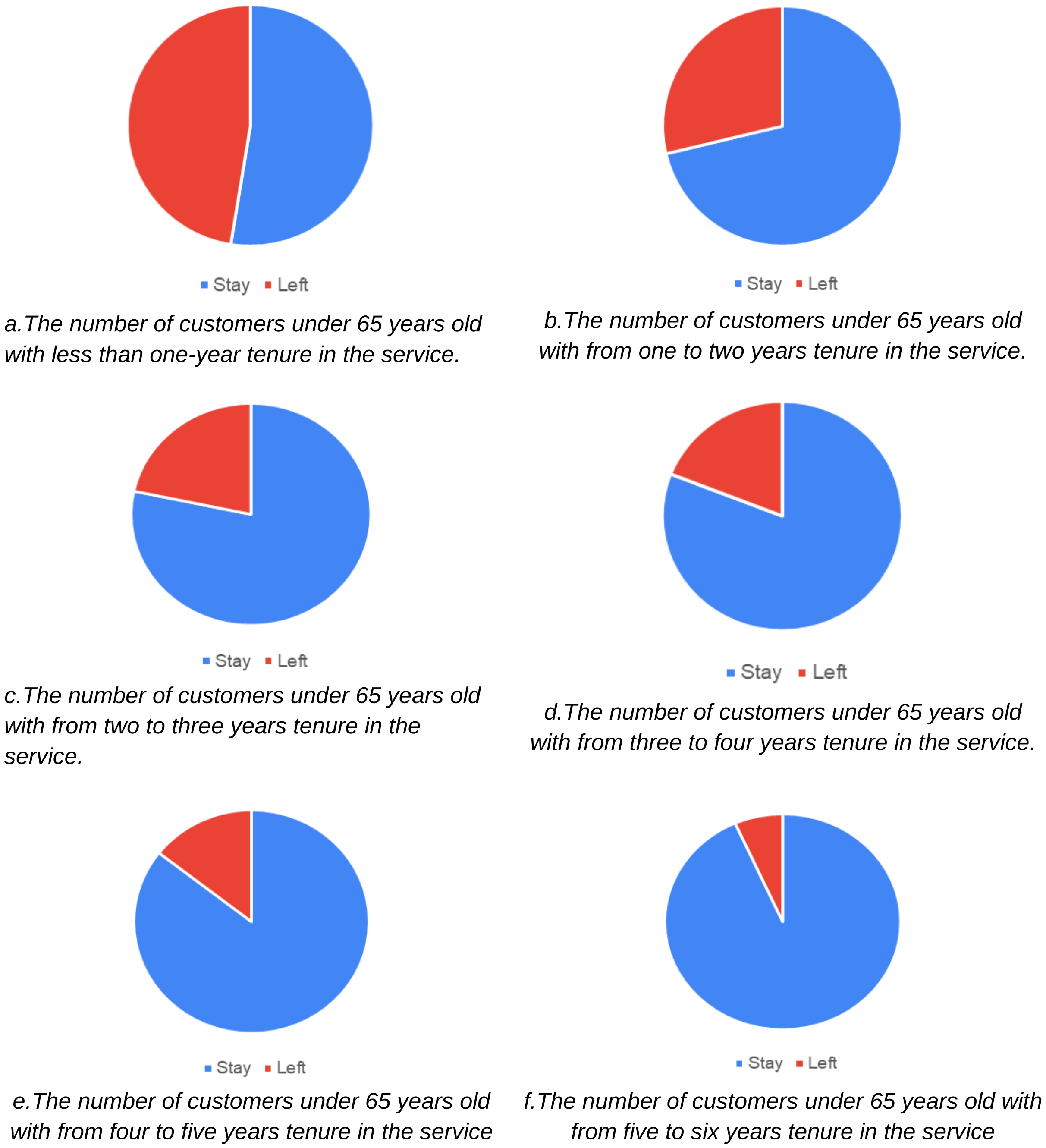
In general, the number of people registering the monthly contract seems to be greater than others, which is 3078 people compared with 1283 and 1550 in one-year and two-year contracts. However, those signing month-to-month contracts tend to leave the service quicker than other types of contract equivalent to nearly 40%. Compared to the leaving proportion with other contracts, those are approximately 11% and 3% annually and two years in statistics. Therefore, in order to segment and attract more customers, along with increasing the financial income, "contract" problems should be considered highly pay attention to, mostly in two-year contracts.

Tenure

Tenure is the number of months the customers have been with the company. More specifically, two years for a customer to generally have a short attention to the brand as one it can rely on. The table 10 shows information of how customers distribute with the time they stay in the telecommunication service. In general, as can be represented, the number of people in the service less than 1 year accounts for a higher number compared with other sectors; however, the leaving rate also experiences significantly greater than others resulting in more than a half people leaving the service. On the other hand, the middle categories providing information with every year staying in the service of customers shows that the number of people is all smaller than the category "< 1 year" and "5-6 years", whereas 3-4 years are when people in the service are the smallest. At the "left" category, the number of people leaving the service falls down annually. This means that based on the leaving rate, after about 3-year experience, customers tend to be loyal ones. The figure 16 also illustrates the proportions of customers staying or leaving according to pie forms.

	<1 year	**1 year - 2 year**	**2 year - 3 year**	**3 year - 4 year**	**4 year - 5 year**	**5 year - 6 year**
Stay	1149	730	652	617	712	1314
Left	1037	294	180	145	120	93
Total	2186	1024	832	762	832	1407

Table 10: The customer distribution in each tenure



a.The number of customers under 65 years old with less than one-year tenure in the service.

b.The number of customers under 65 years old with from one to two years tenure in the service.

c.The number of customers under 65 years old with from two to three years tenure in the service.

d.The number of customers under 65 years old with from three to four years tenure in the service.

e.The number of customers under 65 years old with from four to five years tenure in the service

f.The number of customers under 65 years old with from five to six years tenure in the service

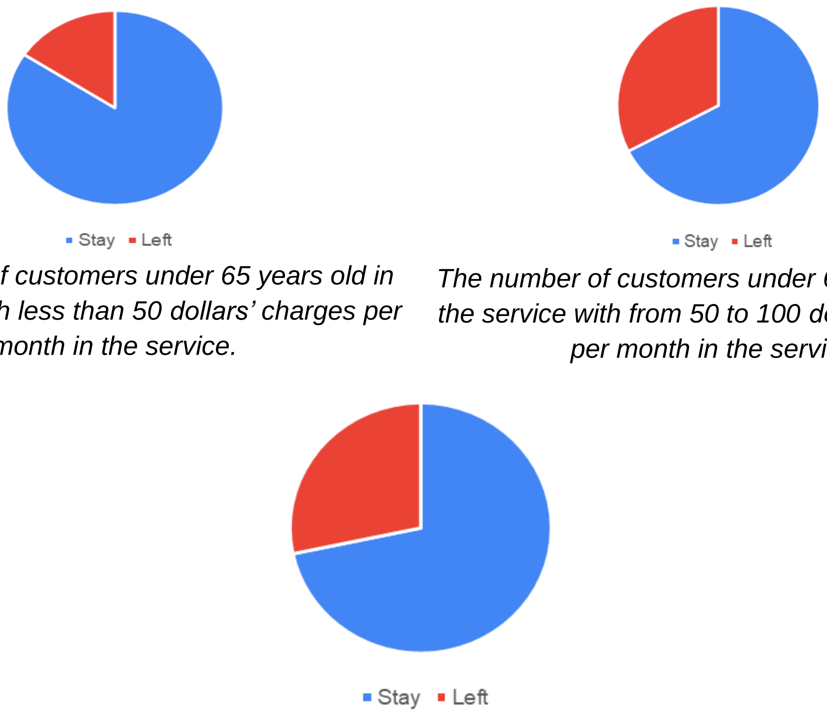
Figure 16: Proportions of customer in different tenures displaying with values in table 10

Monthly charges

Monthly charges may be referred to as the recurring billing as the type of charge a customer should spend for goods or services on a prearranged schedule. It requires the operators to gain the customers' information or permission. Any good or service that customers subscribe to with regularly scheduled payments may be a good candidate for recurring billing. The table 11 compares the number of monthly charges equivalent to three categories, including less than 50, 50 to 100 and more than 100. Overall, the number of people expected to stay in the service when they pay the least money. As can be seen in the table and illustrated by pie forms, the proportions of leaving in the "< 50" category, with 361 leaving people, seem to be less than two other sectors, 1251 people in "50-100" and 257 people in ">100". On the other hand, the figure relates to "50-100" that has the majority of people attending. This means that people are likely to pay the middle forms.

Monthly charges	**<50**	**50-100**	**>100**
Stay	1933	2590	651
Left	361	1251	257
Total	2294	3841	908

Table 11: Customer distribution corresponding to each monthly charge level



The number of customers under 65 years old in the service with less than 50 dollars' charges per month in the service.

The number of customers under 65 years old in the service with from 50 to 100 dollars' charges per month in the service.

The number of customers under 65 years old in the service with over 100 dollars charges per month in the service

Figure 17: The proportion of customers with different monthly charge levels

d. Customers churn determinants results

Table 12 shows the findings of the importance in the prediction of whether customers are churners or not, with statistical significance presented as a p-value less than 0.05. Attributes that have the p-value lower than 0.05 are considered as the important features and kept, the other attributes will be dropped due to the low impact on the customer churn rate. According to the results of this study, the predictors include SeniorCitizen, partner, dependents, tenure, multiple lines, internet service, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod and MonthlyCharges.Therefore, those play an important role in the model building, and evaluation and expected objectives.

Feature name	p-value	Predictor
gender	0.46990	No
SeniorCitizen	3.839866e-37	Yes
Partner	6.22073e-37	Yes
Dependents	9.14043e-44	Yes
tenure	7.99905e-205	Yes
PhoneService	0.31631	No
MultipleLines	0.00140	Yes
InternetService	7.17724e-05	Yes
OnlineSecurity	7.41749e-136	Yes
OnlineBackup	1.22340e-61	Yes
DeviceProtection	2.65126e-51	Yes
TechSupport	2.35111e-129	Yes
StreamingTV	0.00213	Yes
StreamingMovies	0.00123	Yes
Contract	3.66667e-264	Yes
PaperlessBilling	2.35655e-59	Yes
PaymentMethod	2.07510e-19	Yes
MonthlyCharges	2.70664e-60	Yes
TotalCharges	0.22438	No

Table 12: P-values of attributes

H1 b	SeniorCitizen is positively associated with the customer churn probability
H1 c	Partner is positively associated with the customer churn probability
H1 d	Dependents is positively associated with the customer churn probability
H2 a	Tenure is positively associated with the customer churn probability
H2 b	Contract is positively associated with the customer churn probability
H2 d	PaperlessBilling is positively associated with the customer churn probability
H2 c	PaymentMethod is positively associated with the customer churn probability
H2 d	MonthlyCharges is positively associated with the customer churn probability
H3 b	MultipleLines is positively associated with the customer churn probability
H3 c	InternetService is positively associated with the customer churn probability
H3 d	OnlineSecurity is positively associated with the customer churn probability
H3 e	OnlineBackup is positively associated with the customer churn probability
H3 f	DeviceProtection is positively associated with the customer churn probability
H3 g	TechSupport is positively associated with the customer churn probability
H3 h	StreamingTV is positively associated with the customer churn probability
H3 i	StreamingMovies is positively associated with the customer churn probability

Table 13: Attributes relate to the prediction of potential customer's churn

e. Data set splitting

In order to avoid the overfitting problem, the data set needs to be split into a training set and test set before training data. In the training process, the model learns from the training set. Meanwhile, the training groups consist of 95% of the data set with an aim to train the algorithms by comparing the outcome of the model running on the testing set with the true outcome of the test set. On the other hand, the other 5% is used to test the algorithms. The results of dividing groups are illustrated through figure 19.

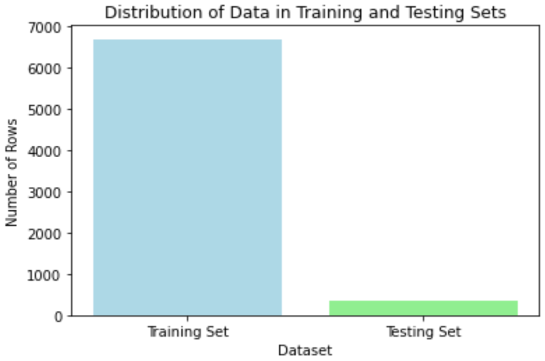


Figure 18: The distribution of data in training set and testing set

f. Model building and evaluation

The classification models are built in order to predict customer churn rate and determine what features that affect this. In this study, there are some particular models built using Decision tree, Random forest and Multi-layer perceptron evaluated by K-fold cross validation and Area under the Curves (AUC) to optimize the hyperparameters of those algorithms. The value of k was 100.

Decision Tree model

Decision tree algorithm is trained and optimized for the depth and the maximum number of node hyperparameters with the optimized number of nodes that was 41 nodes in the tree and the depth value that was 5.

Feature	**contract**	**tenure**	**Online security**	**Monthly charges**	**Internet service**	**Streaming movies**	**Senior citizen**	**Streaming TV**
Importance rate	0.312506	0.189313	0.170356	0.056679	0.047464	0.029776	0.028615	0.023879
Feature	**partner**	**dependents**	**Multiple lines**	**Online backup**	**Device protection**	**Tech support**	**Paperless billing**	**Payment method**
Importance rate	0.021987	0.020777	0.019627	0.018519	0.016972	0.016216	0.01419	0.013125

Table 14: Feature importance rate

The table 14 shows information about the feature importance rate with a purpose of predicting which feature should have a higher impact than the others. Overall, the monthly charges sector experiences the highest position, equivalent to 0.313 as the point to obtain a higher impact compared with others. Next, tenure and contract respectively are in the second and third rank of the dataset with the points, 0.189 and 0.170. Therefore, the usage of a decision tree algorithm to figure out which attributes have influences on whether customers are churners or not contains three first sectors, including contract, tenure and monthly charges. Those are also illustrated through figure 20. Based on the decision tree classification model, three hypotheses could be selected.

Rank	**H1, H2, or H3**	**Content**
1	H2 e	MonthlyCharges is positively associated with the customer churn probability
2	H2 a	Tenure is positively associated with the customer churn probability
3	H2 b	Contract is positively associated with the customer churn probability

Table 15: Variable selection

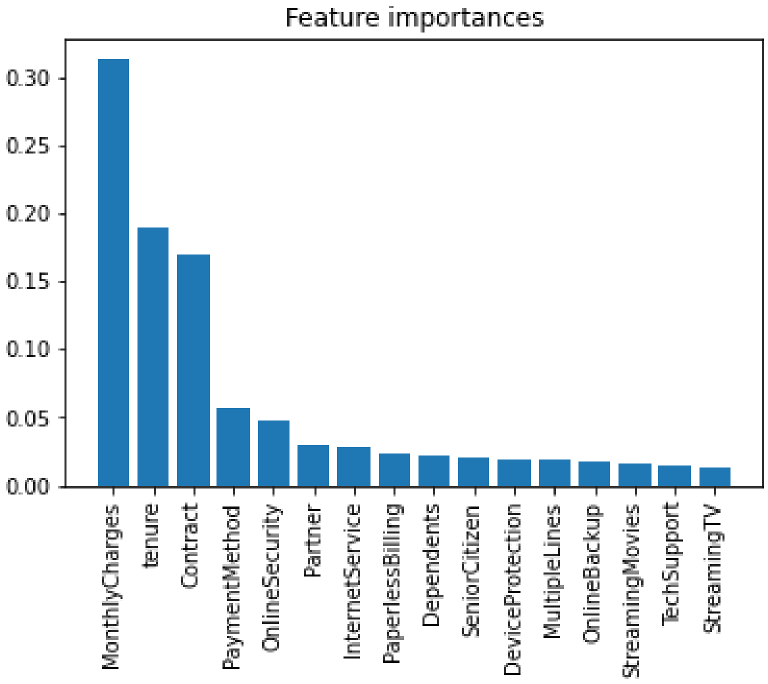


Figure 19: Importance rate generated by Decision Tree algorithm

Model evaluation of the Decision Tree algorithm

The result evaluation based on the predictive results and actual labels through confusion matrices of table 16. The confusion matrix helps evaluate how the model performed with values equivalent to true positive, true negative, false positive and false negative, respectively 215, 46, 43 and 49, with 73.9% of accuracy. Moreover, the confusion matrix is also used to depict AUC results. In detail, the parts related to true positive and false positive are taken to illustrate the figure 21. As can be seen in this figure, the perfect classifier is the 1.0 point, with the curve nearly approaching this perfect classification leading to the result of 67% of accuracy. This means that the model built using the decision tree algorithm obtains a good predictor. On the other hand, another procedure to evaluate this model is K-fold cross validation giving the result that is 72.6% leading to the higher believable model.

		Actual	
		Churn (+)	**Non-churn (-)**
Predict	**Churn (+)**	215	43
	Non-churn (-)	49	46

Table 16: Confusion matrix evaluating the Decision Tree model

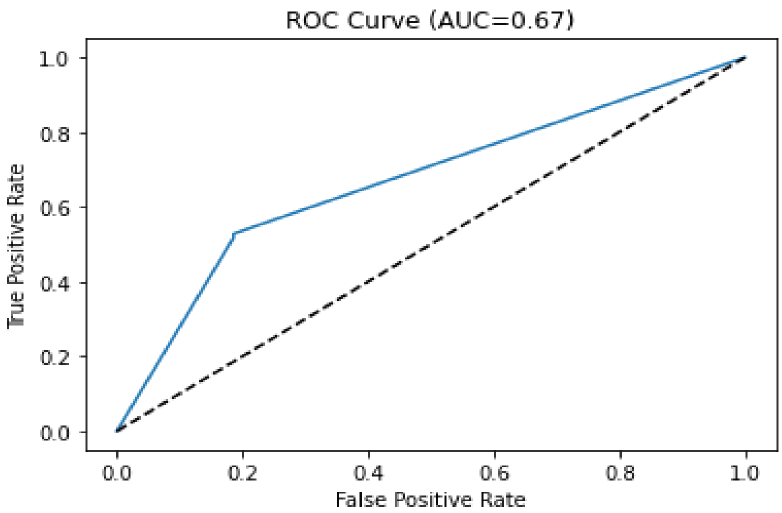


Figure 20: AUC figure to evaluate the Decision Tree model

Procedure	Score
Accuracy	73.9%
K-fold cross validation	72.6%
Area under the curve (AUC)	67%

Table 17: The procedures used to evaluate the model and their scores.

Time complexity = *O (n × m × log(n)) = O (16 × 7043 × 95% × log (7043))*
= O (411916)
Whereas:
n = number of samples = 7043
m = number of features = 16

Random Forest model

Random Forest algorithm was also trained, it optimized the number of trees hyper parameter. The model was built by changing the values of this parameter every time in 100 and 200. The proper results show that the best number of trees was 100 trees.

Feature	**Monthly Charges**	**tenure**	**Contract**	**Payment Method**	**Online Security**	**TechSupport**	**Paperless Billing**	**Multiple Lines**
Importance rate	0.2586	0.2388	0.0865	0.0616	0.0612	0.038	0.0302	0.0288
Feature	**Partner**	**Internet Service**	**Online Backup**	**Device Protection**	**Dependents**	**Senior Citizen**	**Streaming Movies**	**Streaming TV**
Importance rate	0.0287	0.0281	0.0274	0.0262	0.0235	0.0232	0.02	0.0196

Table 18: Importance rate generated by Random Forest algorithm

Rank	H1, H2, or H3	Content
1	H2 e	Monthly charges is positively associated with the customer churn probability
2	H2 a	Tenure is positively associated with the customer churn probability
3	H2 b	Contract is positively associated with the customer churn probability

Table 19: Variables selection

In the table 18, it displays details about the feature importance rate in order to predict which feature may have a greater impact than the others. All in all, the monthly charges category has the highest position, equivalent to 0.258 as the point to achieve a significant impact when compared to others. Following that, tenure and contract are ranked second and third in the dataset, with points of 0.238 and 0.086 respectively. As a result, the use of a random forest algorithm to determine which attributes influence whether customers churn or not contains four first segments: monthly charges, tenure, and contract depicted in figure 22.

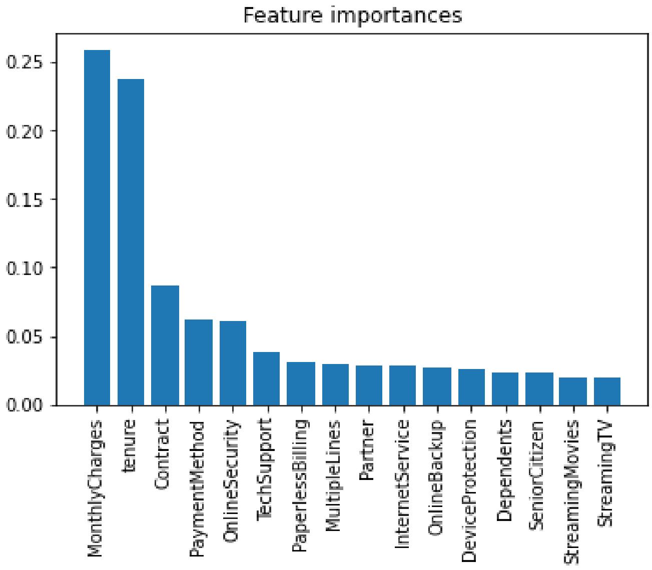


Figure 21: Importance rate generated by Random Forest algorithm

Model evaluation of Random Forest algorithm

The outcome evaluation is according to the prediction results and actual labels using confusion matrix representing through table 20. This evaluates how the random forest algorithm executes with values corresponding to true positive, true negative, false positive and false negative, which are respectively 242, 42, 47 and 22 leading to the accuracy that is 81.5%. Then further, AUC findings are captured based on the confusion matrix. To demonstrate the chart, the elements relevant to true positive and false positive in the figure 23. yielding an accuracy of 84.1%. This suggests that the random forest algorithm produced a decent prediction, along with the outcome of the K-fold cross validation result with 80%.

		Actual	
		Churn (+)	**Non-churn (-)**
Predict	**Churn (+)**	242	47
	Non-churn (-)	22	42

Table 20: Confusion matrix evaluating Random Forest algorithm

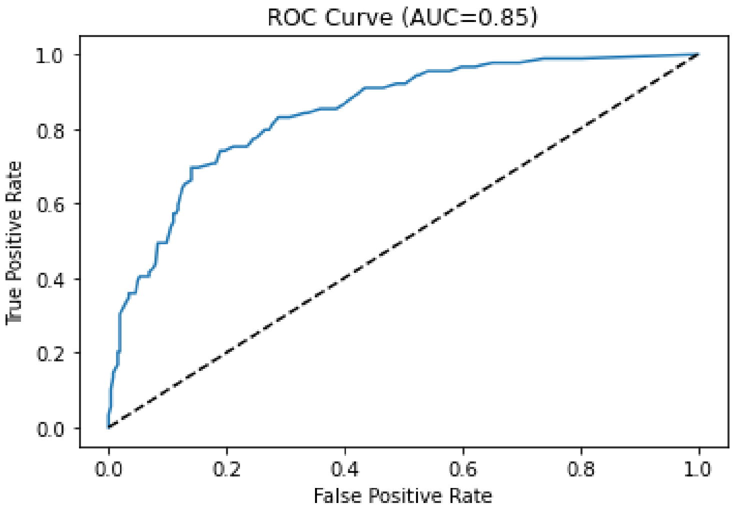


Figure 22: AUC figure to evaluate the Random Forest model

Procedure	**Score**
Accuracy	81.5%
K-fold cross validation	80%
Area under the curve (AUC)	84.1%

Table 21: The procedures used to evaluate the model and their scores.

Time complexity $= O\ (n \times m \times k \times n \times \log(n)) =$
$O\ (100 \times 16 \times 7043 \times 95\% \times \log\ (7043)) = O\ (41191631)$
Whereas:
k *= number of trees = 100*
n *= number of samples = 7043*
m *= number of features = 16*

Multi-Layer Perceptron model

Multi-Layer perceptron (MLP) algorithm is applied with the selected attributes after extracting the data by P-values methods in the input layer and (16, 8) as the hidden layer size. This means that there are two layers of classifiers, including 16 and 8 neurons, for first and second layers, respectively.

Feature	**Senior Citizen**	**Contract**	**Internet Service**	**Multiple Lines**	**Streaming Movies**	**Streaming TV**	**Dependents**	**Online Security**
Importance	0.0877	0.0859	0.0785	0.0693	0.0654	0.0638	0.0632	0.0628
Feature	**Payment Method**	**Partner**	**Paperless Billing**	**TechSupport**	**Online Backup**	**Device Protection**	**tenure**	**Monthly Charges**
Importance	0.0608	0.0589	0.0588	0.0564	0.0519	0.0485	0.0476	0.0404

Table 22: Importance rate generated by Multi-Layer Perceptron algorithm

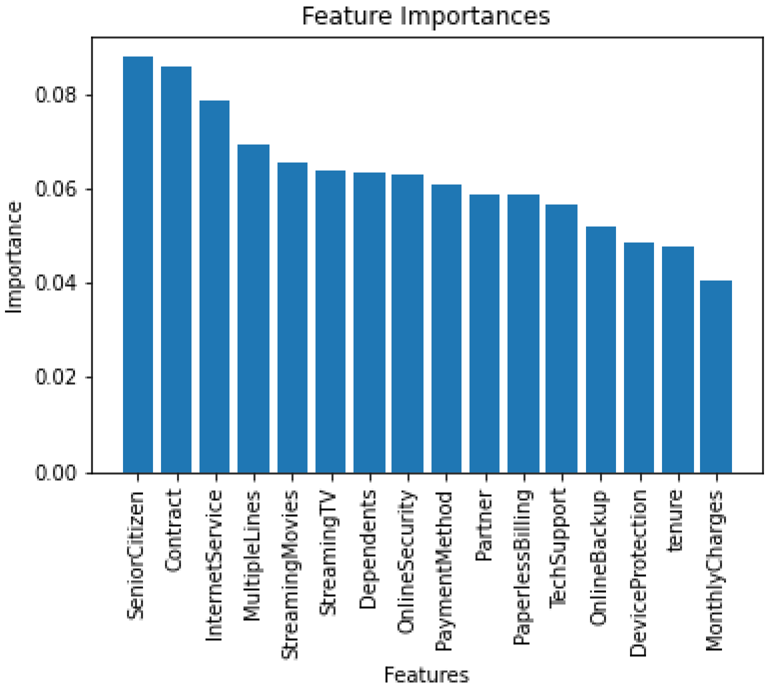


Figure 23: Importance rate generated by multi-layer perceptron algorithm

Rank	H1, H2, or H3	Content
1	H2 b	SeniorCitizen is positively associated with the customer churn probability
2	H2 e	Contract is positively associated with the customer churn probability
3	H1 d	InternetService is positively associated with the customer churn probability

Table 23: Variables selection

Model evaluation of Multi-Layer Perceptron algorithm

The evaluation of the outcomes according to the predictive and actual results shown through the confusion tables. The values comparable to true positive, true negative, false positive, and false negative respectively include 238, 49, 40 and 25, resulting in an accuracy of 81.6%. Furthermore, the confusion matrix is used to illustrate AUC values, with the accuracy of 86.4% and K-fold cross validation accuracy of 79.7%. Therefore, MLP (multi-layer perceptron) should be evaluated as the proper model.

		Actual	
		Churn (+)	**Non-churn (-)**
Predict	**Churn (+)**	239	40
	Non-churn (-)	25	49

Table 24: The confusion matrix to evaluate MLP model

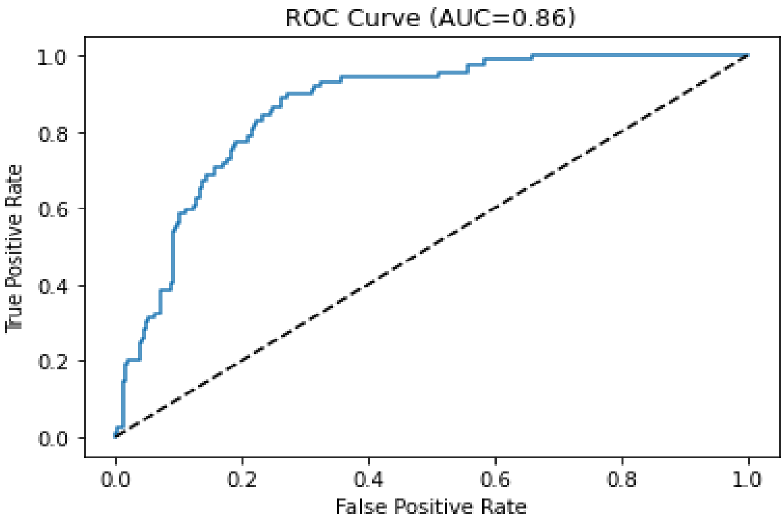


Figure 24: AUC values to evaluate the accuracy of MLP model

Procedure	**Score**
Accuracy	81.6%
K-fold cross validation	79.7%
Area under the curve (AUC)	86.4%

Table 25: The procedures used to evaluate the model and their scores.

Time complexity = $O (n \times m \times h^{k} \times o \times i) = O (95\% \times 7043 \times 16 \times 1^{100} \times 2 \times 125)$ = O (26763400)
Whereas:
n: training samples = 7043
m: number of features = 16
k: number of hidden layers = 100
h: number of neurons = 1
o: output neuron = 2
i: number of iterations = 125

Model comparison between Decision tree, Random Forest and Multi-Layer Perceptron.

	Decision tree	Random forest	Multi-layer perceptron
Accuracy score	73.9%	81.5%	81.6%
K-fold cv score	72.6%	80.0%	79.7%
AUC score	67%	84.1%	86.4%
Time complexity	O (411916)	O (41191631)	O(26763400)

Table 26: Machine learning algorithms comparison results

In general, all algorithms used to measure the accuracy of the model produce logical and impressive outcomes. The table 26 analyzed the scores given by evaluation procedures to equivalent algorithms, including Decision tree, Random Forest and MLP. Overall, as can be illustrated in this table, the accuracy values of Random Forest seem to be more reliable resulting in percentages.

The accuracy scores of three types of procedures in the Decision tree seem not to be corresponding. The fact that the accuracy calculated from the confusion matrix is 73.9%, whereas the K-fold validation category represents 72.6% and 67% of accuracy is the outcome of AUC findings. Next, in Multi-layer perceptron (MLP) measured scores also portray accuracy and K-fold accuracy with 81.6% and 79.7%, respectively. However, only the AUC score of MLP reaches the highest proportion, with 86.4%, compared to other categories. Random Forest was initially evaluated as the better algorithm than others. With the values of accuracy in three considering types, all are more than 80%. The greater value than others also belongs to AUC values, reaching about 84.1% leading to the pre-conclusion that Random Forest seems to be assumed as the best method to lean towards.

In time complexity given in table 26, the Decision Tree model is expected to be the fastest among the three models. However, it may not provide the best accuracy since decision trees tend to over-fit to the training data. The MLP model is expected to be slower than the decision tree model since it involves computing the gradients of the loss function for all the weights at each iteration. However, it can provide high accuracy for complex datasets with nonlinear relationships between the features and the target variable. The Random Forest model, on the other hand, is expected to the slowest among the three models since it trains multiple decision trees on subsets of the data. However, it can provide better accuracy by reducing overfitting through the ensemble method.

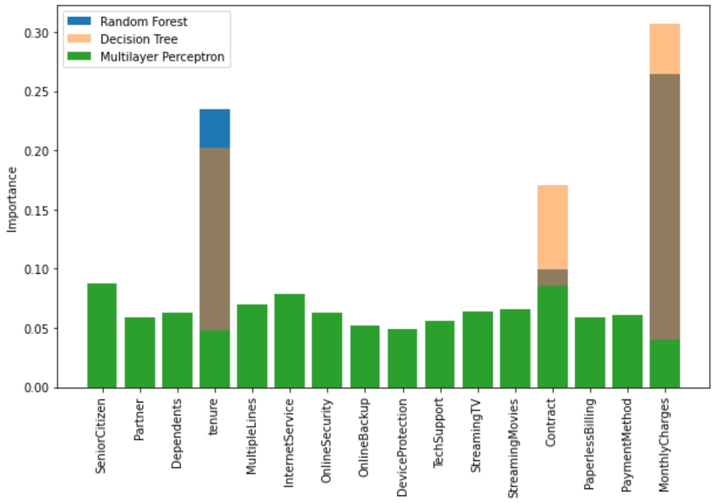


Figure 25: Model comparison between Decision Tree, Random Forest and Multi-layer perceptron (MLP)

The figure 26 shows the importance rate of each attribute in all 3 methods, including Decision tree, Random Forest and Multi-Layer Perceptron (MLP). In general, all rates equivalent to all verifying features were represented through every color. Random Forest is illustrated blue, while Decision Tree and Multi-layer perceptron are yellow and green, respectively. On the other hand, MLP seems to depict features similarly due to the fact of its importance rate between each attribute that are relatively same as others. However, Random Forest and Decision Tree may have better desired results thanks to the more detailed bar categories.

VI. Discussion and Suggestions to Telecommunication problems

This study aims to construct a churn prediction in order to provide telecommunication operators to predict whether or not customers become churners by developing models with Decision Tree, Random Forest and MLP. Overall, there are three important features that play a crucial role in becoming churners, including fixed-term contract, long tenure, and monthly charges as money paid for services. Operators, in detail, can base on those specific features to contradict trial and safe prediction of potential customers, as well as the possibility of those who can become churners. To enhance the customer loyalty and satisfaction, the important features should be predominantly concerned and concentrated on, together with producing comprehensive solutions related.

Understanding the operators' potential churner also assists how the company operates, whether it launches a high-quality product with outstanding customer service or needs to improve significantly to complete. Customers' contributions to the service or values of service should be acknowledged to improve the business future service. By approaching the customer behaviors and the period they stay in the service, company operators need to have deeper insight into those to catch the customers' satisfaction. Customers are satisfied completely with service, for example, businesses can offer more considerable benefits in order to let customers stay longer (Kabu Khadka, 2017). By contrast, the services provided seem to be a major problem increasing the customer's leaving rate, businesses should pay higher attention to make customers delighted or pleased again. Briefly, from company perspectives, it would be delightful if they meet all customers' needs after offering the highest quality services (Katherine N. Lemon, 2016). In fact, in this study using some specific algorithms extracts four key features. For each key feature, the hypothesis of customer churn and the solution to reduce it is proposed.

a. Contract

The contract type is a highly important factor in determining customer churn. Customers with month-to-month contracts are much more likely to churn than those with longer-term contracts. Based on the analysis of figure 14, given data shows information about when customers sign short-term contracts, month-to-month in detail, resulting in the after-a-month period, the rate of leaving customers is 39.57%. Instead, the ones, who sign long-term contracts in a year, increase the staying rate 89.32%, whereas two-year contracts registration has 97.29% of staying rate. This means that operators should enhance customers' experience while signing long-term contracts or incentives by offering more major benefits to reduce customers' churn.

Methods used to enhance the experience of long-term contracts registration:

- Land and expand: To increase the percentage of long-term contracts signed by customers, the business model should consider "Land and expand". Operators can leverage the relation with current customers in order to cross and upsell, together with rising the contract values and minimizing revenue stagnation to let customers recognize that they receive more major values than those who do not sign this contract.
- Built-in price uplifts: Instead of uplifting the price annually, the company can start with a higher price at the beginning and lower gradually after a couple of years. In real-life situations, customers can know exactly what price they annually pay. It leads to the fact that customers benefit from annual pricing reductions when signing long term contracts.

b. Tenure

Tenure is the time customers stay in the service. According to the table 10, customers in this telecommunication industry experience more than about 3 years leading to the decision to be loyal customers. This means that there should be a problem with the first one and two years resulting in the leaving rate of customers more than other periods of time. Therefore, operators should introduce some particular methods to attract more customers, along with letting them sign the contracts at the 1 or 2-year beginning of periods with the telecommunication service.

Some supporting solutions include:

+ The majority of the customers are people who have less than 1-year tenure, followed by over 5-year tenure customers. To reduce customer churn, the company needs to ensure they can keep new customers. This can be done by improving the quality of short-term services.

+ The company needs to pay better attention to those with a long period of tenure by providing monthly or quarter extra benefits for each customer.

c. Monthly charges

The money paid monthly seems to affect customers' behaviors. The number of people tends to spend less money than expected during the service according to the result shown in table 11. This means that operators should prolong the period of customers staying in the services by applying a number of methods.

+ Offering lower-priced plans or discounts for customers with high monthly bills may be effective in reducing churn. It gives information that providing a discounted renewal rate could push customers to stick around the service.
+ Offering customers to the service of long-term contracts with lower prices.
+ Pay more terms in the service with a free term with lower price to stimulate customers' satisfaction when receiving benefits.
+ Considering customer needs or demands with doing weekly, monthly, and annual surveys to understand more about customers' experiences, especially the ones who seem to get ready to cancel due to a lack of demands. Then, providers can let them pay for the service after they have met their needs.

d. Senior Citizens

To ensure the contracts that customers sign and proceed in the next one or two years, the determination of customers' ages should be one of considerable problems operators concern. According to figure 8 and 9 in this dataset evaluation, 5901 per 7043 persons are younger than 65 years old, whereas about 83% of people who keep using the service, whereas 17% leave service after 1-month registration. This means that people, who are smaller than 65 years old, have more potential to use the telecommunication service than another. Therefore, while having potential customers smaller than 65, business operators should take more care of those by offering numerous benefits related to what were analyzed above, including contracts, monthly charges, and payment methods, to attract as much as possible. On the other hand, providers also do a short survey related to total attributes, like internet service, to understand more about customer behaviors leading to the most correct and effective persuasions.

H1, H2, or H3	Content
H2 e	Monthly charges is positively associated with the customer churn probability
H2 a	Tenure is positively associated with the customer churn probability
H2 b	Contract is positively associated with the customer churn probability
H1 b	Senior citizens is positively associated with the customer churn probability

Table 27: Important features directly related to customer churn.

VII. Conclusion and further discussion

Data mining technology provides a novel and optimal solution for the prediction of whether or not customers in the telecommunication industry can be churners. Owing to the use of distinctive algorithms, along with evaluating procedures, the prediction model can be more accurate to meet the scientific demands of this study. Particularly, the individual attributes given by the dataset are considered to be logical, compared to real-life situations. The majority of customers, for instance, who subscribe in short-term periods are likely to terminate contracts due to a lack of personal demands. However, there are some features of the data set that are not as same as expected, including internet service. This means that it shows the proportion of people using the internet, along with its services less than reality. Markedly, the use of the internet and some online services is not as high as expected, opposed to its high demand in a modern society as in California.

In this study, three machine learning algorithms were used to predict the customer churn, which are Decision tree, Random forest and Multi-layer perceptron. Although the performance of the Random forest model seems to be slightly better, the model's accuracy score fluctuates after every running time. Moreover, in the implementation of Random Forest, due to the complexity, it seems to be highly arduous to implement the new model by adjusting parameters even though it took numerous efforts to execute. On the other hand, the classification model using Multi-Layer Perceptron is considered as the best model, due to the stable performance every time implementing the model regardless of how many times it is executed to examine the accuracy.

Therefore, the conclusions to answer the questions produced in the objectives:

- The prediction model was constructed by using three types of algorithms with high accuracy equivalent to each.
- Citizens who are under 65 years old should be more significantly cared about than others.
- To maximize the loyal customers in the initial steps of sales, providers should take higher care of their personal information, then introduce them to some more beneficial programs with long-term contracts to increase staying tenure and other lower-priced benefits by promotions or coupons.

Finally, the methods proposed in this study seem to have significant limitations while functioning algorithms, together with statistical analysis. The results of this study may have key roles for further research if possible, in terms of application scenarios. In further studies, they should focus on applications in more complex and wider production environments.

VII. Key references

1) Andreas Ziegler, I. R. (2013). Mining data with random forests: current options for real-world applications. *WIREs Data Mining and Knowledge Discovery*, 55-63.
2) Chidanand Apté, S. W. (1997). Data mining with decision trees and decision rules. *ScienceDirect*, 197-210.
3) Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. A. (2016). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, 1-4.
4) Daniel Hattenbach, J. W. (2004). *INNOVATIVE WAYS TO FINANCE TELECOMMUNICATION IN DEVELOPING COUNTRIES.* Ronneby.
5) David M. Levine, D. F. (2010). Presenting Data in Charts and Tables: Categorical and Numerical Variables. New Jersey.
6) Elise F. Zipkin, E. H. (2012). Evaluating the predictive abilities of community occupancy models using AUC while accounting for imperfect detection. *esajournals*, 1962-1972.
7) Gelman, A. (2013). Commentary: P Values and Statistical Practice. *Epidemiology*, 69–72.
8) Kabu Khadka, S. M. (2017). CUSTOMER SATISFACTION AND CUSTOMER LOYALTY. *Centria University of Applied Sciences Pietarsaari*.
9) Katherine N. Lemon, P. C. (2016). Understanding Customer Experience Throughout the Customer Journey.
10) Lew Sook Ling, S. F. (2021). Customer churn prediction for telecommunication industry: A Malaysian Case Study. *NCBI*.
11) Ming Zhao, Q. Z. (2021). A Prediction Model of Customer Churn considering Customer Value: An Empirical Research of Telecom Industry in China. *Discrete Dynamics in Nature and Society*.
12) Mohammad Ridwan Ismail, M. K. (2015). A Multi-Layer Perceptron Approach for Customer Churn Prediction. *International Journal of Multimedia and Ubiquitous Engineering*, 213-222.
13) Salvador García, J. L. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *ScienceDirect*.
14) Tjen-Sien Lim, W.-Y. L.-S. (2000). A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. *SpringerLink*, 203–228.
15) Xinyu Zhang, C.-A. L. (2022). Model averaging prediction by K-fold cross-validation. *ScienceDirect*.

VIII. Appendix

1. Statistical analysis

2. Feature name	t-value	p-value	Predictor
gender	-0.72267	0.46990	No
SeniorCitizen	12.80786	3.839866e-37	Yes
Partner	-12.76949	6.22073e-37	Yes
Dependents	-13.96958	9.14043e-44	Yes
tenure	-31.57955	7.99905e-205	Yes
PhoneService	1.00213	0.31631	No
MultipleLines	3.19401	0.00140	Yes
InternetService	-3.97269	7.17724e-05	Yes
OnlineSecurity	-25.36063	7.41749e-136	Yes
OnlineBackup	-16.72953	1.22340e-61	Yes
DeviceProtection	-15.19028	2.65126e-51	Yes
TechSupport	-24.71052	2.35111e-129	Yes
StreamingTV	-3.07159	0.00213	Yes
StreamingMovies	-3.23225	0.00123	Yes
Contract	-36.26415	3.66667e-264	Yes
PaperlessBilling	16.40076	2.35655e-59	Yes
PaymentMethod	9.03557	2.07510e-19	Yes
MonthlyCharges	16.53673	2.70664e-60	Yes
TotalCharges	1.21506	0.22438	No

Table 28: P values of attributes

As can be displayed in the figure 26, some features, such as “tenure”, “contract” and “monthly charges”, are significantly distinctive or noticeable compared with the other sectors, which seem to be considered as the three most important attributes that have the most impact on customer churn rate. Therefore, these features will be focused on to analyses the hypothesis of what leads to customer churn. Then further, three more vital attributes are shown in the table 32.

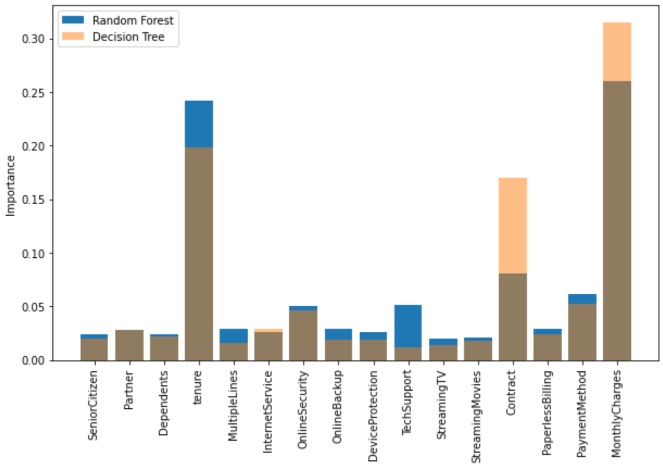


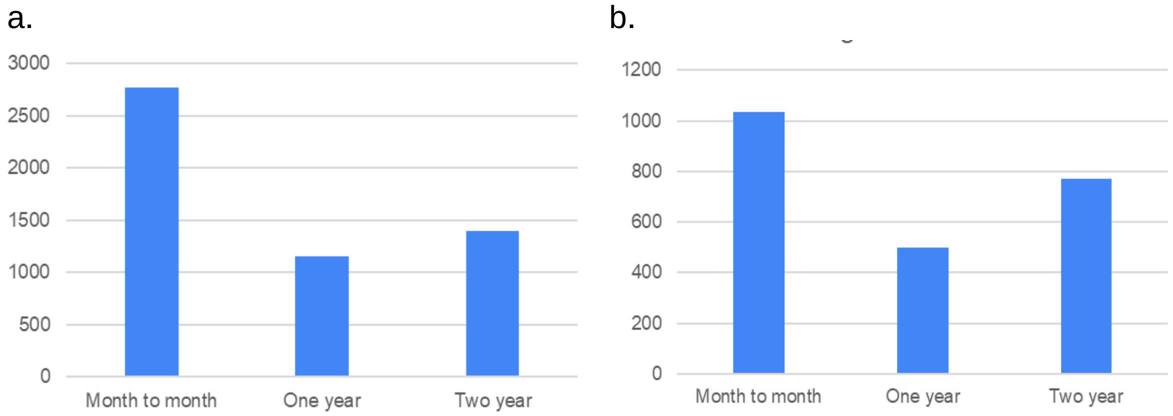
Figure 26: Model comparison between Random Forest and Decision Tree

Rank	Decision tree	Random forest	Multi-Layer Perceptron
1	MonthlyCharges	MonthlyCharges	SeniorCitizen
2	tenure	tenure	Contract
3	Contract	Contract	InternetService
4	PaymentMethod	PaymentMethod	MultipleLines
5	OnlineSecurity	OnlineSecurity	StreamingMovies
6	Partner	TechSupport	StreamingTV
7	InternetService	PaperlessBilling	Dependents
8	PaperlessBilling	MultipleLines	OnlineSecurity
9	Dependents	Partner	PaymentMethod
10	SeniorCitizen	InternetService	Partner
11	DeviceProtection	OnlineBackup	PaperlessBilling
12	MultipleLines	DeviceProtection	TechSupport
13	OnlineBackup	Dependents	OnlineBackup
14	StreamingMovies	SeniorCitizen	DeviceProtection
15	TechSupport	StreamingMovies	tenure
16	StreamingTV	StreamingTV	MonthlyCharges

Table 29: Comparison of importance ranks among three methods

Contract

The comparison between whether or not people sign the contract when they use phone service.



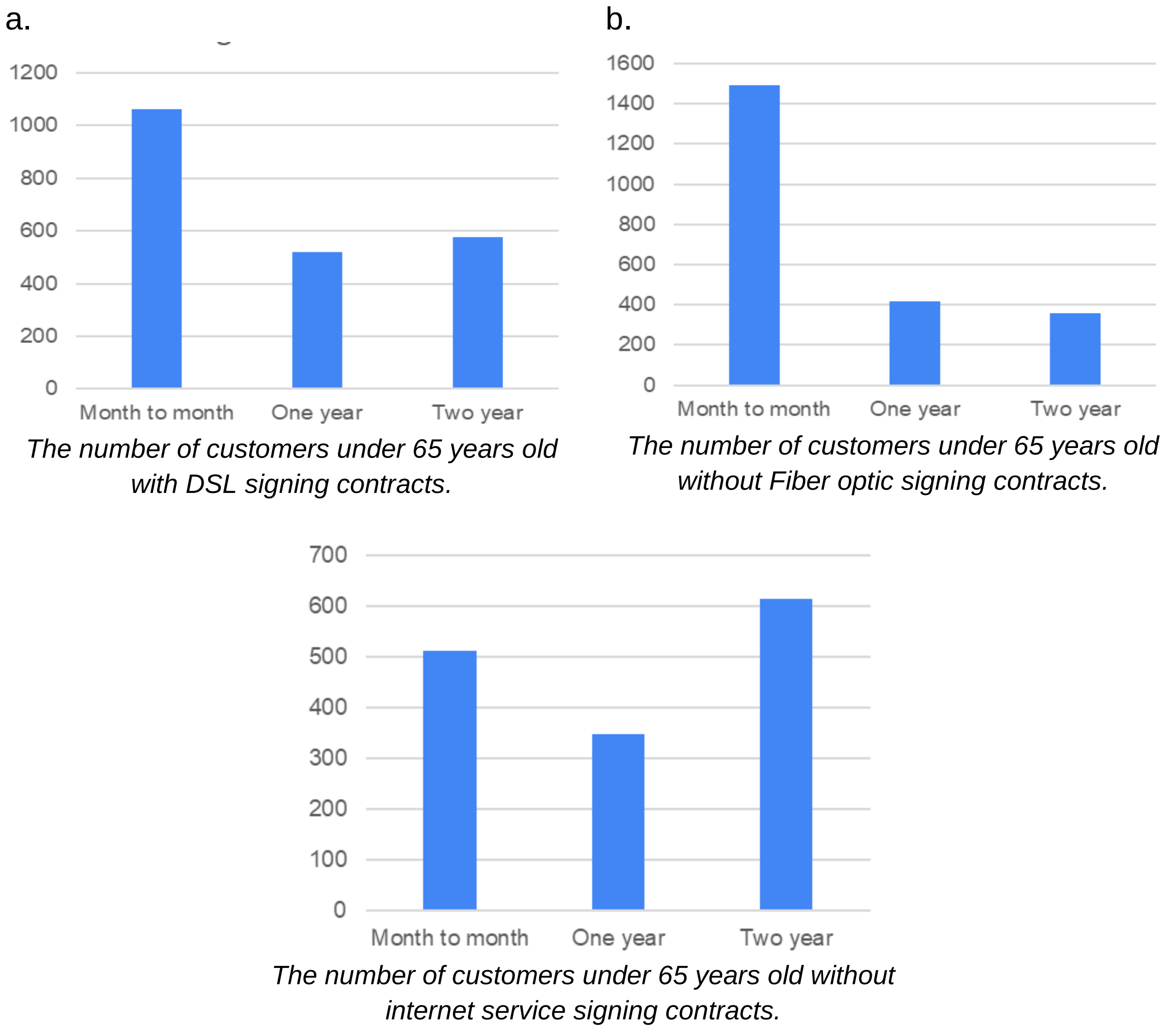
The number of customers under 65 years old with phone service signing contracts.

The number of customers under 65 years old without phone service signing contracts.

Figure 27: The number of people who sign contracts, in terms of month, a year and 2 years, when using phone service and multiple lines.

The figure 27 illustrates the number of people younger than 65 years old using phone service and multiple lines leading to the registration of signing a contract. Overall, the number of one-year contract registration seems to be smaller than others in figure 12 a and b, respectively 1200 and 500. On the other hand, two-year contracts experience the second position arrangement in both figures, whereas month-to-month contracts reach the highest position that more than others from 1.5 to 2 times. Therefore, most people who use phone service and multiple lines tend to sign month to month contracts.

The comparison between whether or not people sign the contract when they use DSL (Digital Subscriber Line), Fiber optic and no internet usage.



The number of customers under 65 years old with DSL signing contracts.

The number of customers under 65 years old without Fiber optic signing contracts.

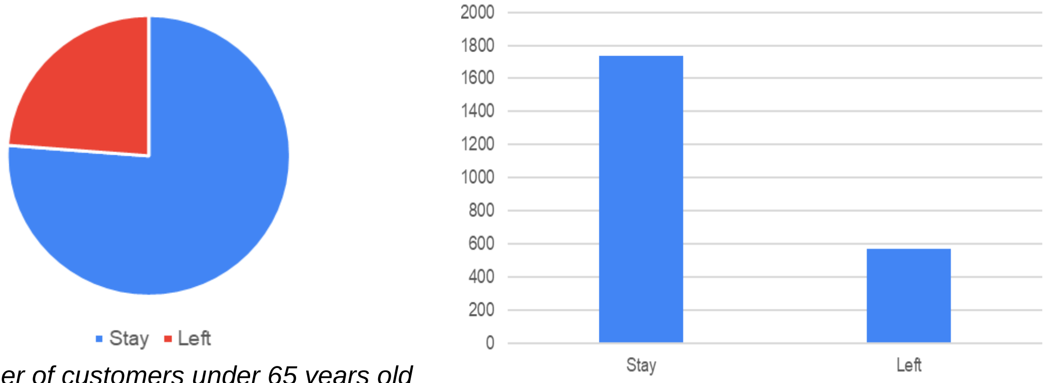
The number of customers under 65 years old without internet service signing contracts.

Figure 28: The number of people who sign contracts, in terms of month, a year and 2 years, when using DSL (Digital Subscriber Line), Fiber optic and no internet usage.

The number of people using DSL (Digital Subscriber Line), Fiber optic and no internet usage. For those who subscribe to internet services (DSL and Fiber optic), most will sign a month to month contract. In contrast, for those who do not use internet services, the amount is more evenly distributed across the three types of contracts.

Phone

A lot of people use phones nowadays thanks to the development of technology. According to recent data, the average person spends 3 hours and 15 minutes on their phone each day.



The number of customers under 65 years old with phone service in the service.

The number of customers under 65 years old with multi lines in the service.

Figure 29: The left-handed side pie chart: The number of people, who both are smaller than 65 years old and have phone services, leaving and staying.
The right-handed side bar chart: The number of people, who both are smaller than 65 years old and use multiple lines, leaving and staying.

The number of people smaller than 65 years old having phone service stay in the tele service greater about 3 times compared to the rest. Equivalently, those, who use multiple lines, stay in the service also greater than those leaving about 3 times. In the phone attribute, the multiple lines should be considered more than others resulting from its offerings of more benefits compared to other attributes at the same title. Normally, people use multiple lines, which is not much but every person who registers this keeps registering more and more, shown by the figure that illustrates the number of people staying more than leaving about 3 or 4 times.

Internet

In the internet matter, there are 3 different types of selection, including no internet usage, using DSL (Digital Subscriber Line) and Fiber optic.

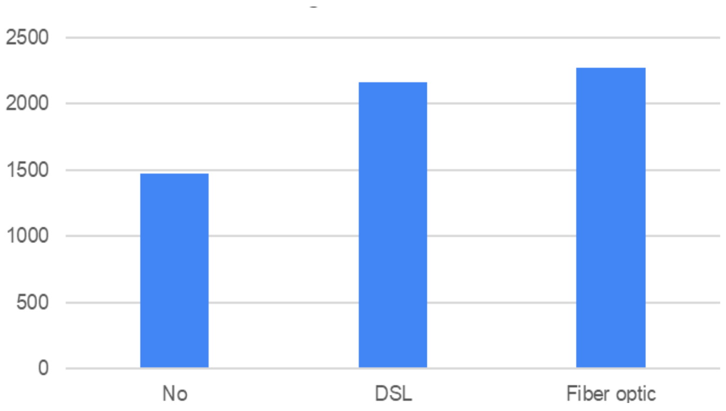
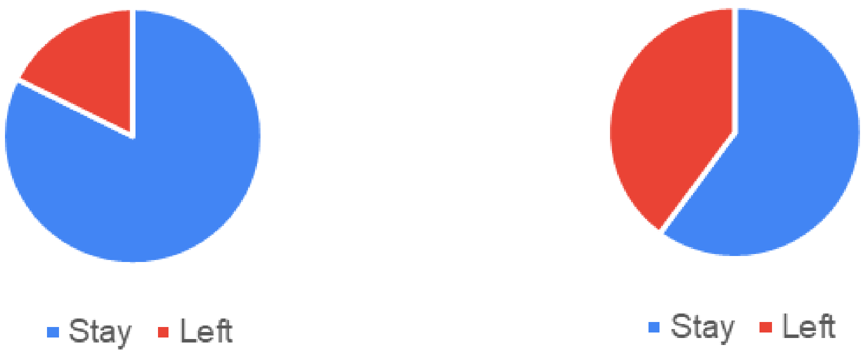


Figure 30: The distribution of people under 65 who do not use the internet, use DSL and Fiber optic.

Statistically, in percentage, the number of smaller-than-65-years-old people, who do not use the internet is about 1474 equivalent to 24.98%. Moreover, those using Fiber optic account for 2265, about 38.38%. Last but not least, 36.64% of people use DSL with the real number of people about 2162. In general, there is a short comment that the proportion of people who don't use the internet is more than expected even though the internet may be common with the community these days.
However, this study only focuses on those using the internet equivalent to both DSL and Fiber optic to segment the potential customers. Briefly, the proportion of users using both are relatively the same, only 2% in difference, according to figure 30.



The number of customers under 65 years old with DSL.

The number of customers under 65 years old with Fiber.

Figure 31: The left-handed side figure: The number of users, who are under 65 years old and use DSL, staying and leaving the service. The right-handed side figure: The number of users, who are smaller than 65 years old and use Fiber optic, staying and leaving the service.

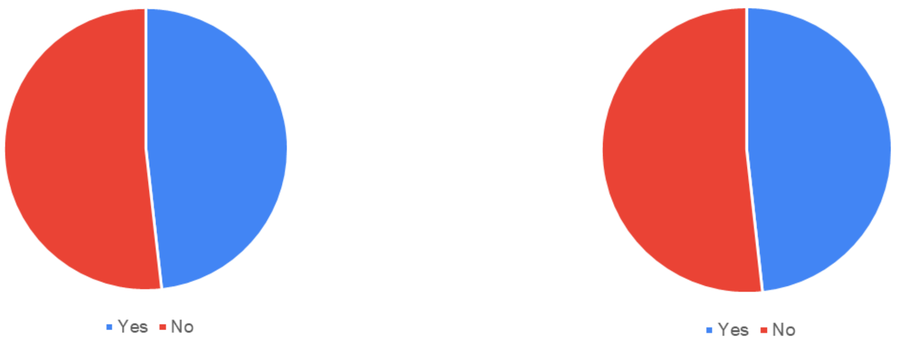
In general, people who use DSL, have the percentage of staying more than leaving. In fact, staying one accounts for 82.38%, whereas another about 17.62%. In contrast, with the same proportion in initial accounting of people using the Internet but those using Fiber optic leave this service more than those using DSL. Statistically, the number of smaller-than-65 users using Fiber leaving and staying the service, approximately 60.09% and 39.91%.

The table 29 shows the information of the number of people smaller than 65 years old, together with their 4 different types of internet usage in terms of whether or not they use the tele service. As can be seen from the table, the non-users of telecommunication services are almost greater than users. Significantly, it is about 1.5 times the difference in comparison to all 4 types of people's use. To conclude, those 4 types are not used to segment customers due to the greater number of non-users in telecommunication. In real-life situations, people also may not use those particular types of internet usage that much.

	Use	**Not use**
online Security	1737	2690
online Backup	1953	2474
device protection	1956	2471
Tech support	1784	2643

Table 30: The number of people with 4 types of internet usages dividing into users or non-users.

On the other hand, streaming TV and Movies should be considered to be analyzed due to its reasonableness. In general, the number of people seem to not use streaming TV or movies that much despite the technological advancement in the fourth industrial era. Furthermore, those sectors may be easier to treat compared with other more difficult factors, including online security, online backup, device protection or tech support.



The number of customers under 65 years old with Streaming TV

The number of customers under 65 years old with Streaming Movies

Figure 32: The left-handed side figure: The number of users, who are smaller than 65 years old and use streaming TV, staying and leaving the service.
The right-handed side figure: The number of users, who are smaller than 65 years old and use streaming movies, staying and leaving the service.

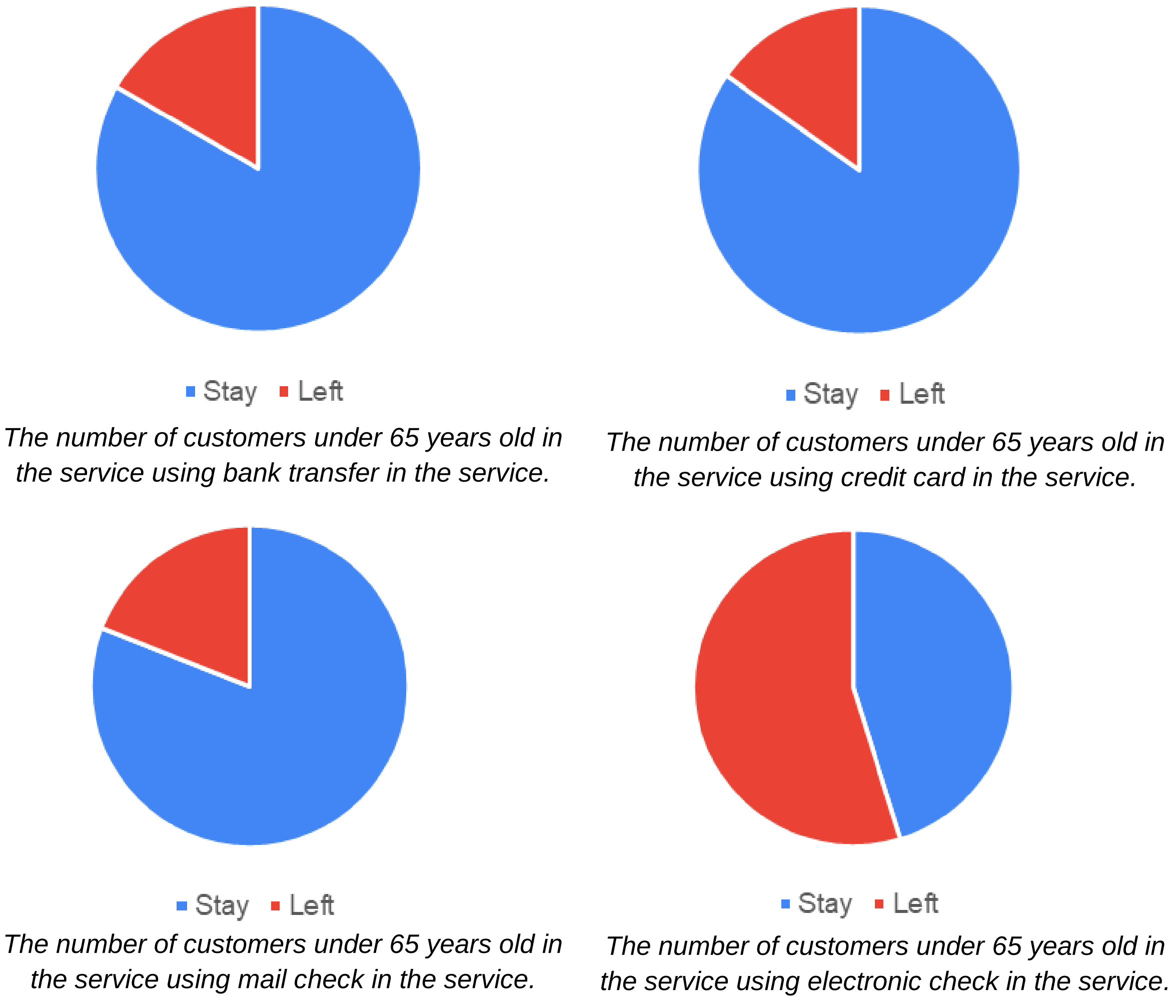
Overall, the difference between people who use both streaming TVs and Movies is not significant. This is explained by the number of people smaller than 65 years old using streaming TVs about 2135, others not using around 2292. Moreover, the number of people younger than 65 years old using streaming movies 2137, whereas others are 2290. However, the dataset has its own problems related to the proportions of streaming TVs, movies and internet less than expected.

Payment method

Payment method is the way of transaction that the company provides. From the table 30, the payment method that the customer prefers the most is Electronic Check, with the proportion of total customers is as twice as any of the other methods, with 2365 people using in total, whereas about 1500 people use other payment methods, including Bank Transfer, Credit card, and Mailed check. However, among these customers using electronic checks as payment methods, more than half of them leave the company shown by figure 32. On the other hand, for paying by Bank Transfer, Credit Card and Mailed Check, only one sixth of the customers are churn.

	Bank Transfer	Credit Card	Mailed Check	Electronic Check
Stay	1286	1290	1304	1071
Left	258	232	308	1294
Total	1544	1522	1612	2365

Table 31: Customer distribution with corresponding payment method



The number of customers under 65 years old in the service using bank transfer in the service.

The number of customers under 65 years old in the service using credit card in the service.

The number of customers under 65 years old in the service using mail check in the service.

The number of customers under 65 years old in the service using electronic check in the service.

Figure 33: Proportion of customer with different payment methods

3. Discussion and Suggestions to Telecommunication problems

In statistical analysis, attributes seem to be analytical and coherent to analyze and produce the predictive results. Furthermore, there may be some in-depth discussions to support the further experiments. Payment methods should be improved because humans are in technological advancement leading to the fact that numerous high-tech payment methods, including credit cards, mobile wallets or electronic check, are likely to be concentrated on to update the functions in order to enhance customer retention. Instead, some traditional payment ones, like paying in cash or mailed check, should be limited due to their inconvenience and annoyance. The internet is also one of attributes people may take more care of, by increasing the sales of the internet with clarifying its beneficial impacts. It assists people to have more opportunities to approach and use the internet these days. On the other hand, tenure or staying periods may be a subsequent factor that operators notably care about. Thanks to the time staying with customers, providers not only recognize what categories customers are concerned to improve, but they also obtain what features allow people to become loyal clients. Monthly charges may be the last but not the least one to be concerned, by using a variety of payment methods noticed above, customers may have another way to pay with instalment in order to have early experiences without paying the entire required fees for the telecommunication services.

a. Contract

To boost the rate of signing long-term contracts, there may be some supporting solutions:

- Electronic documents or electronic means: In the modern information and communication technology, electronic templates to put signatures created in order to attract more customers seem to popularly be used. By using electronic documents, customers can sign contracts quickly without meeting in person. This allows customers to save time and effort to purchase the company services. There are particularly three types of electronic signature that seem to be common in Vietnam, including digital signature, scanned signature and image signature.
- High sense of urgency: A sense of urgency in customer service is the feeling of empathy for what is driving operators' customer demeanor. Creating a sense of urgency in the mind of customers may allow providers to increase the financial income and high forecast accuracy. There are some particular methods relating to creating a sense of urgency or letting customers put this service in their priority in this study.
    - Set an inspiration date: to create a sense of urgency, the companies can set a time limit for registering. Customers will pay more attention to the service in that period to decide whether they should subscribe to it.
    - Offer a deal: Operators should offer some types of deal or added values to take customer retention. Those may include free samples, extra benefits or slight discounts.
- Follow customers up: Operators can increase the rate of signing contracts by taking more care of their customers in the first time when those customers are considering or thinking whether or not they should sign these long-term contracts. If customers have not signed yet, salespeople can follow up and ask within days to ensure they sign contracts on time.

b. Payment methods

Payment methods are the paths helping people transact money in order to pay for the service. From the table …, the leaving rate of electronic check seems to be higher than other categories even though in modern information and communication technology resulting in customers who pay by electronic check are more likely to churn. Therefore, some supporting solutions include to address those problems:

- Encouraging customers to pay using other methods, such as credit card or automatic bank transfer, may help reduce churn. Credit cards or automatic bank transfer payment sometimes helps customers receive some profits, including paying 1,000,000 VND and receiving 100,000 VND in return.
- Operators should advance the technology used for electronic check in order to improve customers' experience and retention. These days, it seems to be easy

to update and enhance some companies' applications to assist customers to solve some service's circumstances.

- Instalment should be another method to enhance the signing contracts of customers by using payment methods noticed in the attributes.

c. Internet service

Internet service is also an important impact in predicting customer churn. Among the customers who are under 65 years old, 24.8% of them do not use the internet, 38.4% of them use Fiber optics internet and the other 36.6% subscribe to DSL services. However, the proportion of customers who signed for the Internet service is quite low, especially when the internet is one of the vital resources in the modern world. Besides, customers who used the Fiber optics service are likely to leave. In order to reach more customers using internet service and keep them stay, it is necessary for the company to improve the Fiber optics service and launch more advertise campaigns with accompanied benefit and incentive packages.

H1, H2, or H3	**Content**
H2 c	Payment Method is positively associated with the customer churn probability
H3 c	Internet Service is positively associated with the customer churn probability
H1 b	Senior citizens is positively associated with the customer churn probability

Table 32: Supporting features also having association with the customer churn