Telco Customer Churn Prediction

Report

Author name: Sei Ryun Kim

Student Number: 500286795

Supervisor’s name: Dr. M. Bilgehan Erdem

Submission date: November 8, 2021

Ryerson University

Table of Contents

Abstract

Customer churn is one of the main problems that the telecommunications industry may face. Some customers switch to other telecommunications companies to receive better phone and internet services. Some customers switch to other companies for paying less service fee charges. Telecommunications companies gather information, analyze data, and find solutions to prevent customer churn.

This project profounds several research questions that may solve the problem regarding the telecommunications customer churn. This project focuses on what factors of the demographic may affect customer churn in telecommunications industry. This project also focuses on what types of customer account information may affect customer churn in telecommunications industry. This project also focuses on what types of the phone and internet services affect telecommunications customer churn. This project also focuses on how the service fee charges affect telecommunications customer churn. This project also focuses on what types of exploratory data analysis can help data analysts and researchers work on research for telecommunications customers churn prediction. Analyzing the data by using those research questions may help solve the problem regarding the telecommunications customer churn.

The themes chosen for this project are classification, regression, and predictive analytics. Classification may help assigning observations to certain categories. Classification may also help predict the probable outcomes by using independent variable and dependent variables. Regression may help analyze the relationship between independent variable and dependent variables. Predictive analytics is an important theme for data mining, predictive modelling, and machine learning to predict future by using information from data and calculating the probabilities.

The data used in this project is Telco Customer Churn Dataset. Telco Customer Churn Dataset is originally made by IBM Cognos Analytics. IBM Cognos made a sample data about a fictional telecommunications company's customer churn based on various factors. The dataset used in this project is Telco Customer Churn Dataset that is updated by BlastChar from Kaggle website. Telco Customer Churn Dataset updated by BlastChar from Kaggle website has 7,043 observations and 21 variables.

The techniques and the tools that are proposed to solve the stated problem are Python 3 in JupyterHub server and GitHub repository. Python can help data analysts and researchers provide clear code for projects and research. JupyterHub server provides cloud setting to help data analysts and researchers use Python 3 languages for their projects and research. GitHub can help data analytics and researchers provide repositories for collaboration to work on projects with other data analytics and researchers.

This project profounds a research proposal that will use several techniques for customer churn prediction. This project will perform initial data analysis, exploratory data analysis, predictive models, and evaluation. This project may use several techniques for building machine learning models including decision trees, logistic regression, and k-nearest neighbours. This project may use the technique to evaluate predictive models including a confusion matrix to calculate accuracy, sensitivity, and specificity.

Introduction

Telecommunications Customer Churn

Customer churn is one of the main problems that the telecommunications industry may face. Some customers switch to other telecommunications companies to receive better phone and internet services. Some customers switch to other companies for paying less service fee charges. Telecommunications companies gather information, analyze data, and find solutions to prevent customer churn.

Research Question

This project profounds several research questions that may solve the problem regarding the telecommunications customer churn:

- What factors of the demographic may affect customer churn in telecommunications industry?
- What types of customer account information may affect customer churn in telecommunications industry?
- What types of phone and internet services the telecommunications companies provide may affect customer churn in telecommunications industry?
- How do the service fee charges affect customer churn in telecommunications industry?
- What types of exploratory data analysis can help data analysts and researchers work on research for telecommunications customers churn prediction?

Analyzing the data by using those research questions may help solve the problem regarding the telecommunications customer churn. Exploratory data analysis may help data analysts and researchers to work on research for telecommunications customers churn prediction.

Research Proposal

This project profounds a research proposal that will use several techniques for customer churn prediction. Before building predictive models, data preparation and exploratory data analysis will be performed. The techniques that may use for building machine learning models are decision trees, logistic regression, k-nearest neighbours, support vector machine, and random forest. The technique that may use for evaluation of predictive models is confusion matrix to calculate accuracy, precision, F1 score, sensitivity, and specificity.

Literature Review

According to Wu et al. (2021), the telecommunications industry focuses two main elements for customer analytics: churn prediction and customer segmentation. Wu et al. (2021) asserted that customer churn prediction should be combined with customer segmentation to support telco marketers to make better decisions.

Wu et al. (2021) displayed that the purpose of the research is to obtain efficient company resource allocation and upgrade customer retention with an integrated customer analytics framework for churn management in telecommunications industry. Wu et al. (2021) proposed the five steps of the integrated telecommunications customer analytics framework for telecommunications customer churn. The first step of the integrated telecommunications customer analytics framework was data preprocessing (Wu et al., 2021). The techniques used for data preprocessing were data cleaning, data transformation, and data normalization (Wu et al., 2021). The second step of the integrated telecommunications customer analytics framework was exploratory data analysis (Wu et al., 2021). The analyses used for exploratory data analysis were univariate analysis and bivariate analysis (Wu et al., 2021). The third step of the integrated telecommunications customer analytics framework was churn prediction (Wu et al., 2021). The techniques used for churn prediction were logistic regression, decision tree, random forest, naïve Bayes, AdaBoost, and multi-layer perceptron (Wu et al., 2021). The fourth step of the integrated telecommunications customer analytics framework was factor analysis (Wu et al., 2021). The analyses used for factor analysis were Bayesian and logistic regression (Wu et al., 2021). The fifth step of the integrated telecommunications customer analytics framework was customer segmentation (Wu et al., 2021). The technique used for customer segmentation was K-means (Wu et al., 2021). The last step of the integrated telecommunications customer analytics framework was customer behaviour analytics (Wu et al., 2021).

The integrated telecommunications customer analytics framework for telecommunications customer churn that this article displayed can help data analysts and researchers to prepare the process of approach for the project. Moreover, the techniques used in this article might be useful for this project as well to predict telecommunications customer churn. Decision trees, naïve Bayes, and logistic regression can be great examples to use in this project.

Lin et al. (2014) displayed that the purpose of the research is to perform feature selection and data reduction and to examine the performance of feature selection and data reduction for telecommunications customer churn prediction. The purpose of feature selection was to delete the most unrelated and unnecessary items from the selected dataset (Lin et al., 2014). The techniques used for feature selection were principal component analysis (PCA) and association rules (Lin et al., 2014). The purpose of data reduction was to reduce into a smaller dataset although the principle of the original dataset remained (Lin et al., 2014). The technique used for data reduction was self-organizing map (SOM) (Lin et al., 2014). The techniques used for telecommunications customer churn prediction were artificial neural networks, decision trees, and logistic regression (LR) (Lin et al., 2014).

The techniques used in this article might be useful for this project as well to predict telecommunications customer churn. Principle component analysis (PCA), decision trees, and logistic regression can be great examples to use in this project.

Idris et al. (2012) found that there was a poor performance for customer churn prediction on classification algorithms due to skewed data distribution and high dimensionality. For the telecommunications customer churn prediction, Idris et al. (2012) used Chr-PmRF approach which was based on particle swamp optimization (PSO), mRMR, and random forest (RF) because they believed Chr-PmRF approach might provide the better results than sampling, feature reduction, and classification techniques.

At first, Idris et al. (2012) used data preprocessing to solve the problems of missing values and nominal values in the dataset. Then, Idris et al. (2012) used undersampling methods including RUS and PSO to evaluate the effectiveness of the performance of the customer churn prediction. Idris et al. (2012) used classification techniques including random forest and k-nearest neighbours (KNN) to evaluate sampling and feature selection methods by using area under the curve (AUC), sensitivity, and specificity.

This article helps data analysts and researchers understand evaluation of performance of customer churn prediction by using under the curve (AUC), sensitivity, and specificity. This project may use the confusion matrix to calculate accuracy, sensitivity, and specificity.

Zhu et al. (2017) stated that customer churn prediction associated with binary classification and multiclass classification algorithms from the research fields of machine learning methods including logical regression, decision trees, neural networks, support vector machines, partial least squares (PLS) and hazard models.

Zhu et al. (2017) performed two steps of data preparation before constructing the predictive model: solving problems with missing value and calculating the Fisher score for feature selection. Then, Zhu et al. (2017) performed sampling methods with two classifiers including decision tree and support vector machine (SVM). Zhu et al. (2017) performed two types of statistical tests for comparison: a Wilcoxon signed-rank test for paired comparison and a Friedman test with Iman-Davenport extension for multiple comparison.

This article helps data analysts and researchers focus on data preprocessing including data cleaning for missing values and finding correlations of variables. In this project, data preprocessing including data cleaning for missing values is important to evaluate the effectiveness of the performance better. Moreover, correlations of variables are also important to understand the significance of the relationship between variables.

Chen et al. (2012) built a framework with ensemble techniques and used a hierarchical multiple kernel support vector machine (H-MK-SVM) as a data mining technique to build a model with static and longitudinal behavioural data. Chen et al. (2012) explained how the framework with ensemble techniques worked. After data collection, data preprocessing was used for demographic data and transformation was used for transactional data (Chen et al., 2012). After data preprocessing and transformation, an ensemble learning method was used for ensemble classifier construction with classifier for static data and classifier for longitudinal data (Chen et al., 2012).

This article helps data analysts and researchers learn more about a hierarchical multiple kernel support vector machine. In this project, data preprocessing used for demographic data is important to understand the significance of data preprocessing before building a predictive model.

Kim and Lee (2012) explained that there were several examples of the supervised and semi-supervised prediction models: neural networks, logistic regression, support vector machines, Gaussian processes, decision trees, k-nearest neighbour classifiers, and Bayesian networks. Kim and Lee (2012) used two prediction models for their research: neural networks and logistic regression. For neural networks, Kim and Lee (2012) used the multi-layer perceptron

(MLP) with input, hidden, and output layers. For logistic regression, Kim and Lee (2012) used two possible output values, zero and one, in order to predict the possibility of occurrence of an event with the simple variation.

This article helps data analysts and researchers learn more about supervised and semi-supervised prediction models. The techniques used in this article including decision trees, naïve Bayes, and logistic regression can be great examples in this project.

Data Description

Dataset

The data used in this project is Telco Customer Churn Dataset. Telco Customer Churn Dataset is originally made by IBM Cognos Analytics. IBM Cognos made a sample data about a fictional telecommunications company's customer churn based on various factors. The dataset used in this project is Telco Customer Churn Dataset that is updated by BlastChar from Kaggle website. Telco Customer Churn Dataset updated by BlastChar from Kaggle website has 7,043 observations and 21 variables. Table 1 shown below helps to describe variables displayed in the dataset (BlastChar, 2018; IBM Samples Team, 2019):

Table 1. Description of variables in the dataset.

Variable	**Type**	**Description**
customerID	Categorical	A unique ID to identify each customer.
gender	Categorical	The gender of each customer: Male or Female.
SeniorCitizen	Boolean	Indication whether the age of a customer is 65 or older: 0 as No or 1 as Yes.
Partner	Boolean	Indication whether a customer has a partner: Yes or No.
Dependents	Boolean	Indication whether a customer has a dependent: Yes or No.
tenure	Numeric	The total number of months that a customer has stayed with the company.
PhoneService	Boolean	Indication whether a customer has a phone service: Yes or No.
MultipleLines	Categorical	Indication whether a customer has multiple telephone lines: Yes, No, or No phone service.

InternetService	Categorical	Internet service provider that a customer has: DSL, Fiber optic, or No.
OnlineSecurity	Categorical	Indication whether a customer has online security: Yes, No, or No internet service.
OnlineBackup	Categorical	Indication whether a customer has online backup: Yes, No, or No internet service.
DeviceProtection	Categorical	Indication whether a customer has device protection: Yes, No, or No internet service.
TechSupport	Categorical	Indication whether a customer has tech support: Yes, No, or No internet service.
StreamingTV	Categorical	Indication whether a customer has streaming TV: Yes, No, or No internet service.
StreamingMovies	Categorical	Indication whether a customer has streaming movies: Yes, No, or No internet service.
Contract	Categorical	A customer’s contract term: Month-to-month, One year, or Two year.
PaperlessBilling	Boolean	Indication whether a customer has paperless billing: Yes or No.
PaymentMethod	Categorical	A customer’s payment method: Electronic check, Mailed check, Bank transfer (automatic), or Credit card (automatic).
MonthlyCharges	Numeric	A customer’s monthly charges.
TotalCharges	Numeric	A customer’s total charges.
Churn	Boolean	Indication whether a customer churns: Yes or No

Demographics

According to IBM Samples Team (2019), the variables to describe demographics in this dataset were customerID, gender, SeniorCitizen, Partner, and Dependents. This project will focus on whether those variables may affect the demographics to telecommunications customer churn.

Services

According to IBM Samples Team (2019), the variables to describe services are customerID, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, and TotalCharges. This project will focus on whether those variables may affect telecommunications customers to churn due to services that the telecommunications company provides.

Status

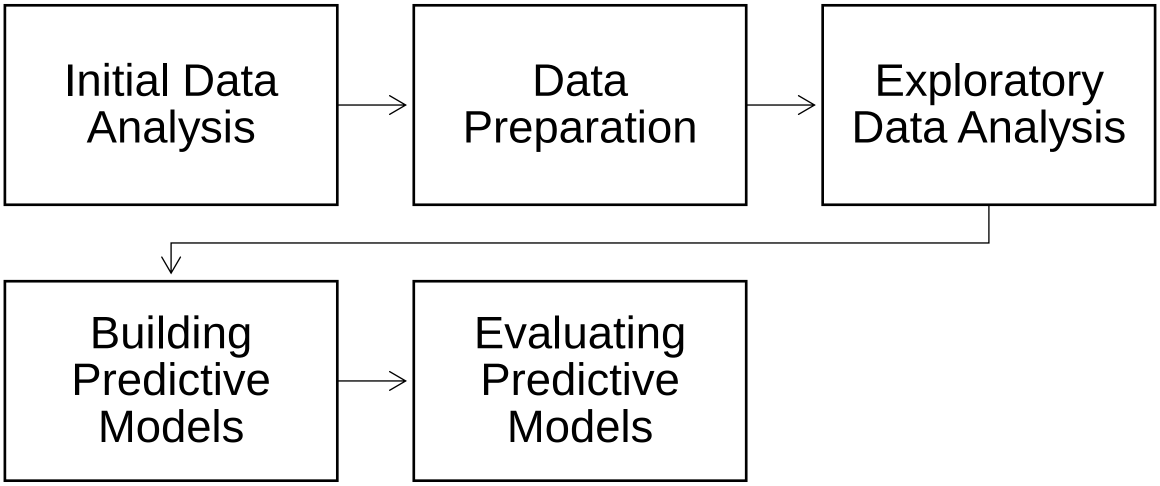
According to IBM Samples Team (2019), the variables to describe status are customerID and Churn. Churn is a dependent variable to predict telecommunications customer churn by using several techniques for classification, regression, and predictive methods.

Approach

In this project, the techniques and the tools that are proposed to solve the stated problem are Python 3 in JupyterHub server and GitHub repository. Python can help data analysts and researchers provide clear code for projects and research. JupyterHub server provides cloud setting to help data analysts and researchers use Python 3 languages for their projects and research. GitHub can help data analytics and researchers provide repositories for collaboration to work on projects with other data analytics and researchers. For this project, a link to a repository on GitHub is shown as following: https://github.com/seiryunkim/big-data-analytics-project.

This project will perform five steps of approach for telecommunications customer churn prediction. Figure 1 shown below is a graph of approach displaying the tentative overall methodology:

Figure 1. Approach for the tentative overall methodology.


Initial Data Analysis
Data Preparation
Exploratory Data Analysis
Building Predictive Models
Evaluating Predictive Models

Step 1: Initial Data Analysis

The first step of approach for a tentative overall methodology regarding telecommunications customer churn prediction is an initial data analysis. For initial data analysis, this project will display the information of the dataset to find the types of variables. Moreover, this project will perform finding the mean, median, and standard deviation in this dataset.

Step 2: Data Preparation

The second step of approach for a tentative overall methodology regarding telecommunications customer churn prediction is a data preparation. For data preparation, this project will focus on finding any missing values and outliers in this dataset. It is important to perform data preprocessing including data cleaning before exploratory data analysis because missing values and outliers may affect the results.

Step 3: Exploratory Data Analysis

The third step of approach for a tentative overall methodology regarding telecommunications customer churn prediction is an exploratory data analysis. For exploratory data analysis, this project will perform several visualizations including histograms, box plots, and scatter plots to understand the data and to summarize the characteristics of variables for preparing building predictive models and evaluating predictive models. Moreover, this project will also perform types of correlations of variables in this dataset and display a visualization of correlations between variables to compare the results among various types of correlations. This project may use Pearson and Spearman as types of correlation.

Step 4: Building Predictive Models

The fourth step of approach for a tentative overall methodology regarding telecommunications customer churn prediction is building predictive models. This project will perform data normalization first before building predictive models. The techniques that may use for building predictive models are decision trees, naïve Bayes, logistic regression, k-nearest neighbours, support vector machine, random forest, and AdaBoost.

Step 5: Evaluating Predictive Models

The last step of approach for a tentative overall methodology regarding telecommunications customer churn prediction is evaluating predictive models. The technique that may use for evaluation of predictive models is confusion matrix to calculate accuracy, precision, F1 score, sensitivity, and specificity. This project will compare various types of predictive models and see whether these predictive models have been shown the effectiveness to predict telecommunications customer churn.

Results

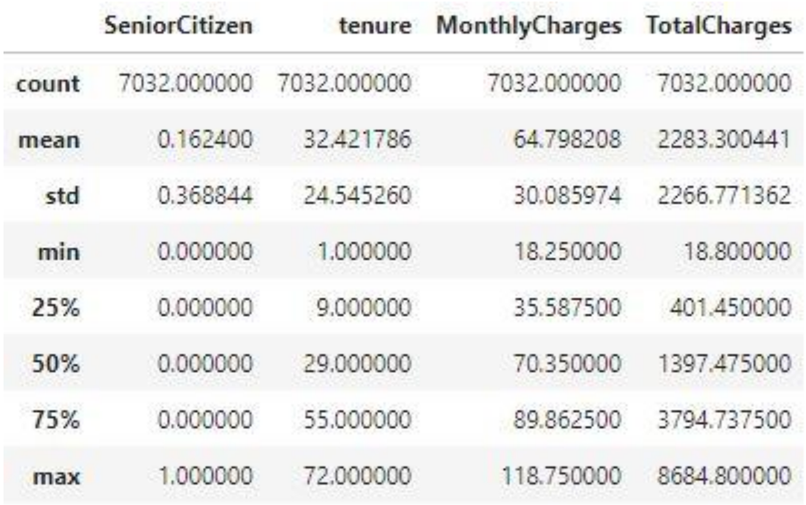
The results of the projects can answer the research questions to solve the problems regarding telecommunications customer churn.

Initial Data Analysis and Data Preparation

In the data preparation, there were 11 observations that needed to be removed due to missing values or null values in total charges. After data cleaning, there were 7,032 observations and 21 variables to analyze the data for the project.

In the initial data analysis, there was a data description for statistical measures to find the mean, minimum, maximum, and standard deviation.

Table 2. Data description for statistical measures.



	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
count	7032.000000	7032.000000	7032.000000	7032.000000
mean	0.162400	32.421786	64.798208	2283.300441
std	0.368844	24.545260	30.085974	2266.771362
min	0.000000	1.000000	18.250000	18.800000
25%	0.000000	9.000000	35.587500	401.450000
50%	0.000000	29.000000	70.350000	1397.475000
75%	0.000000	55.000000	89.862500	3794.737500
max	1.000000	72.000000	118.750000	8684.800000

For tenure, the mean was 32.42 months. The minimum was 1 month, and the maximum was 72 months. The first quartile or the 25th percentile was 9 months. The second quartile or the 50th percentile was 29 months. The third quartile or the 75th percentile was 55 months. The standard deviation for tenure was 24.54.

For monthly charges, the mean was $64.80. The minimum was $18.25, and the maximum was $118.75. The first quartile or the 25th percentile was $35.59. The second quartile or the 50th percentile was $70.35. The third quartile or the 75th percentile was $89.86. The standard deviation for monthly charges was 30.08.

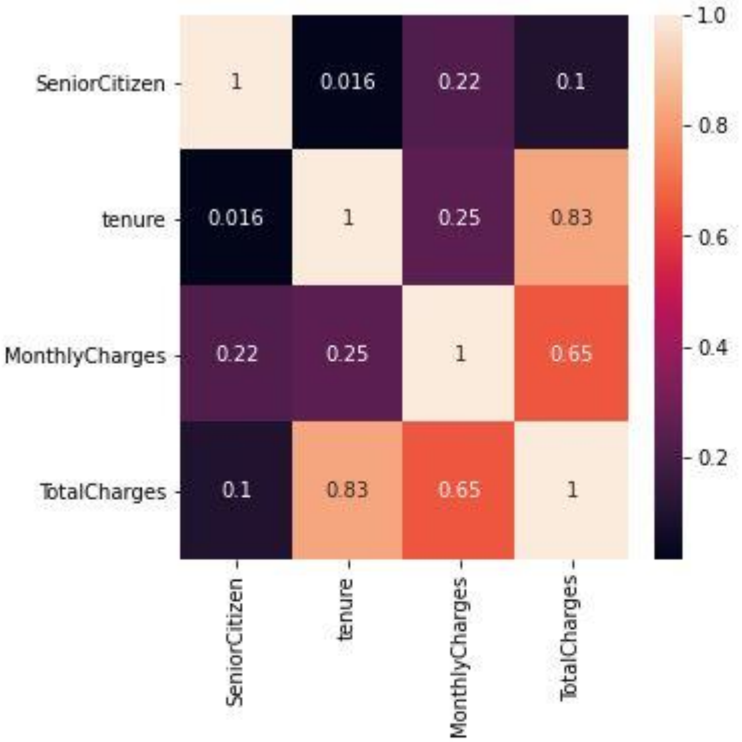
For total charges, the mean was $2,283.30. The minimum was $18.80, and the maximum was $8,684.80. The first quartile or the 25th percentile was $401.45. The second quartile or the 50th percentile was $1,397.48. The third quartile or the 75th percentile was $3,794.74. The standard deviation for total charges was 2266.77.

Exploratory Data Analysis

In the exploratory data analysis, there were several visualizations to display how demographic, customer account information, services, and fee charges affect customer churn.

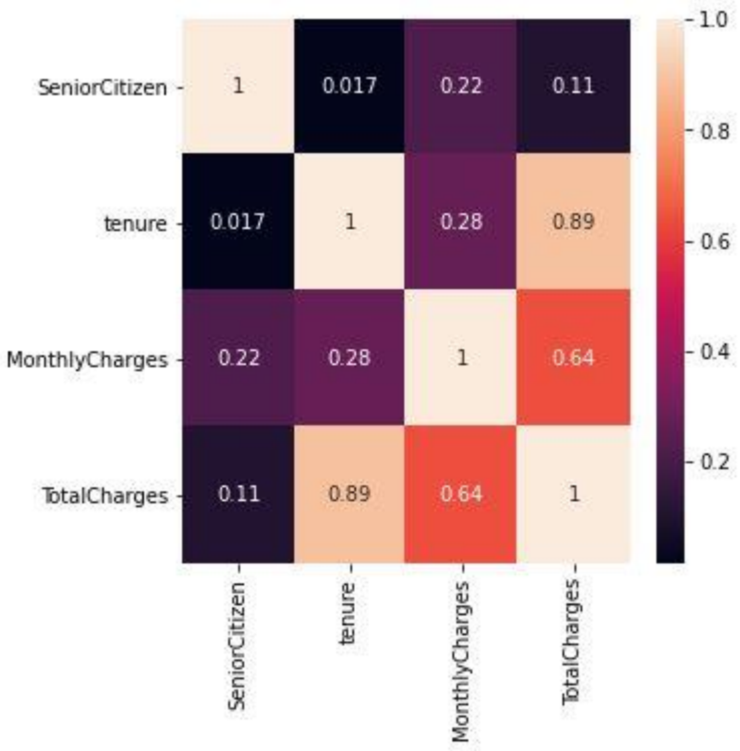
Types of correlation

Figure 2. Pearson correlation.



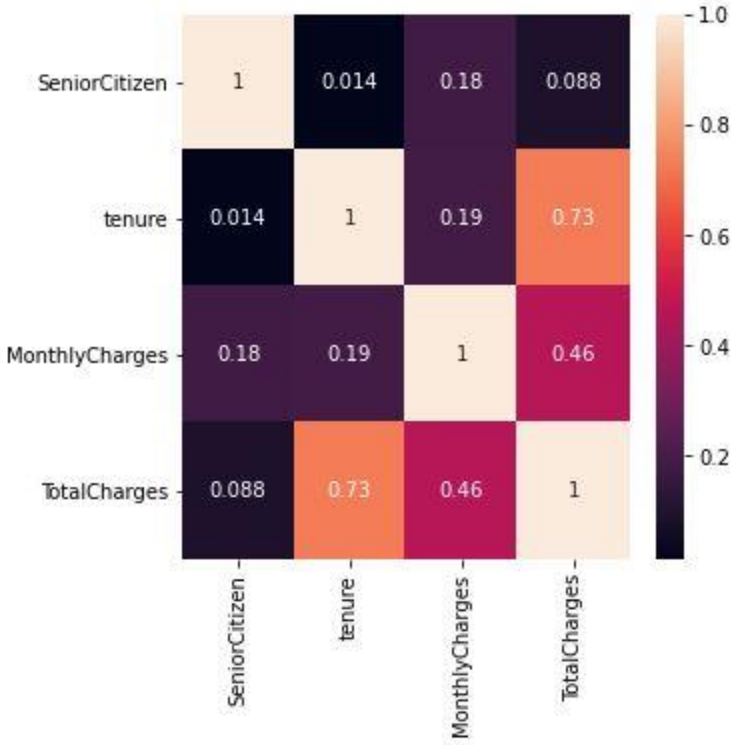
In the Pearson correlation, the relationship between tenure and total charges had a strong positive correlation because the correlation coefficient was 0.83. On the other hand, the relationship between tenure and monthly charges had a weak positive correlation because the correlation coefficient was 0.25.

Figure 3. Spearman correlation.



In the Spearman correlation, the relationship between tenure and total charges had a strong positive correlation because the correlation coefficient was 0.89. On the other hand, the relationship between tenure and monthly charges had a weak positive correlation because the correlation coefficient was 0.28.

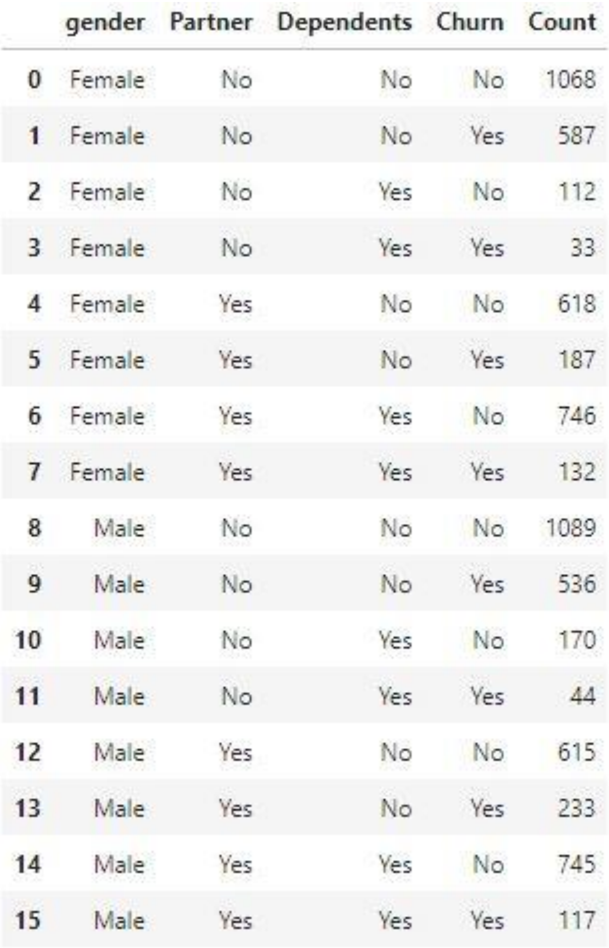
Figure 4. Kendall correlation.



In the Kendall correlation, the relationship between tenure and total charges had a strong positive correlation because the correlation coefficient was 0.73. On the other hand, the relationship between tenure and monthly charges had a weak positive correlation because the correlation coefficient was 0.19.

What factors of the demographic may affect customer churn in telecommunications industry?

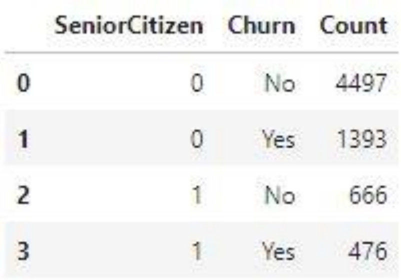
Table 3. Demographics affects to churn.



	gender	Partner	Dependents	Churn	Count
0	Female	No	No	No	1068
1	Female	No	No	Yes	587
2	Female	No	Yes	No	112
3	Female	No	Yes	Yes	33
4	Female	Yes	No	No	618
5	Female	Yes	No	Yes	187
6	Female	Yes	Yes	No	746
7	Female	Yes	Yes	Yes	132
8	Male	No	No	No	1089
9	Male	No	No	Yes	536
10	Male	No	Yes	No	170
11	Male	No	Yes	Yes	44
12	Male	Yes	No	No	615
13	Male	Yes	No	Yes	233
14	Male	Yes	Yes	No	745
15	Male	Yes	Yes	Yes	117

Table 3 displayed how the demographics might affect telecommunications customer churn. 1,068 women who had no partner and dependents, did not churn. 587 women who had no partner and dependents, churned. 112 women who had dependents but no partner, did not churn. 33 women who had dependents but no partner, churned. 618 women who had a partner but no dependents, did not churn. 187 women who had a partner but no dependents, churned. 746 women who had a partner and dependents, did not churn. 132 women who had a partner and dependents, churned. 1,089 men who had no partner and dependents, did not churn. 536 men who had no partner and dependents, churned. 170 men who had dependents but no partner, did not churn. 44 men who had dependents but no partner, churned. 615 men who had a partner but no dependents, did not churn. 233 men who had a partner but no dependents, churned. 745 men who had a partner and dependents, did not churn. 117 men who had a partner and dependents, churned.

Table 4. Senior citizen affects to churn.



	SeniorCitizen	Churn	Count
0	0	No	4497
1	0	Yes	1393
2	1	No	666
3	1	Yes	476

Table 4 displayed how the senior citizen might affect telecommunications customer churn. 4,497 people who were not senior citizens did not churn. 1,393 people who were not senior citizens churned. 666 people who were senior citizens did not churn. 476 people who were senior citizens churned.

What types of customer account information may affect customer churn in telecommunications industry?

Figure 5. Tenure and churn.

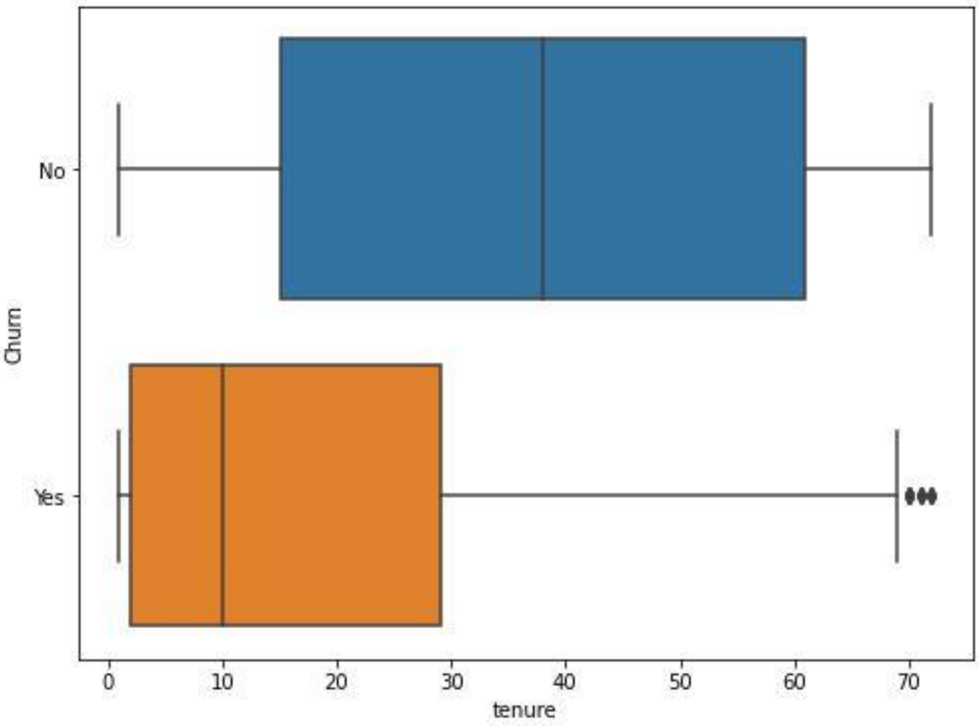


Figure 5 displayed a box plot to show how tenure affected telecommunications customer churn. The box plot showed that people who had tenure from 1 to 30 months tended to churn. On the other hand, the box plot showed that people who had tenure from 15 to 60 months tended not to churn.

What types of phone and internet services the telecommunications companies provide may affect customer churn in telecommunications industry?

Figure 6. Phone and internet services and churn.

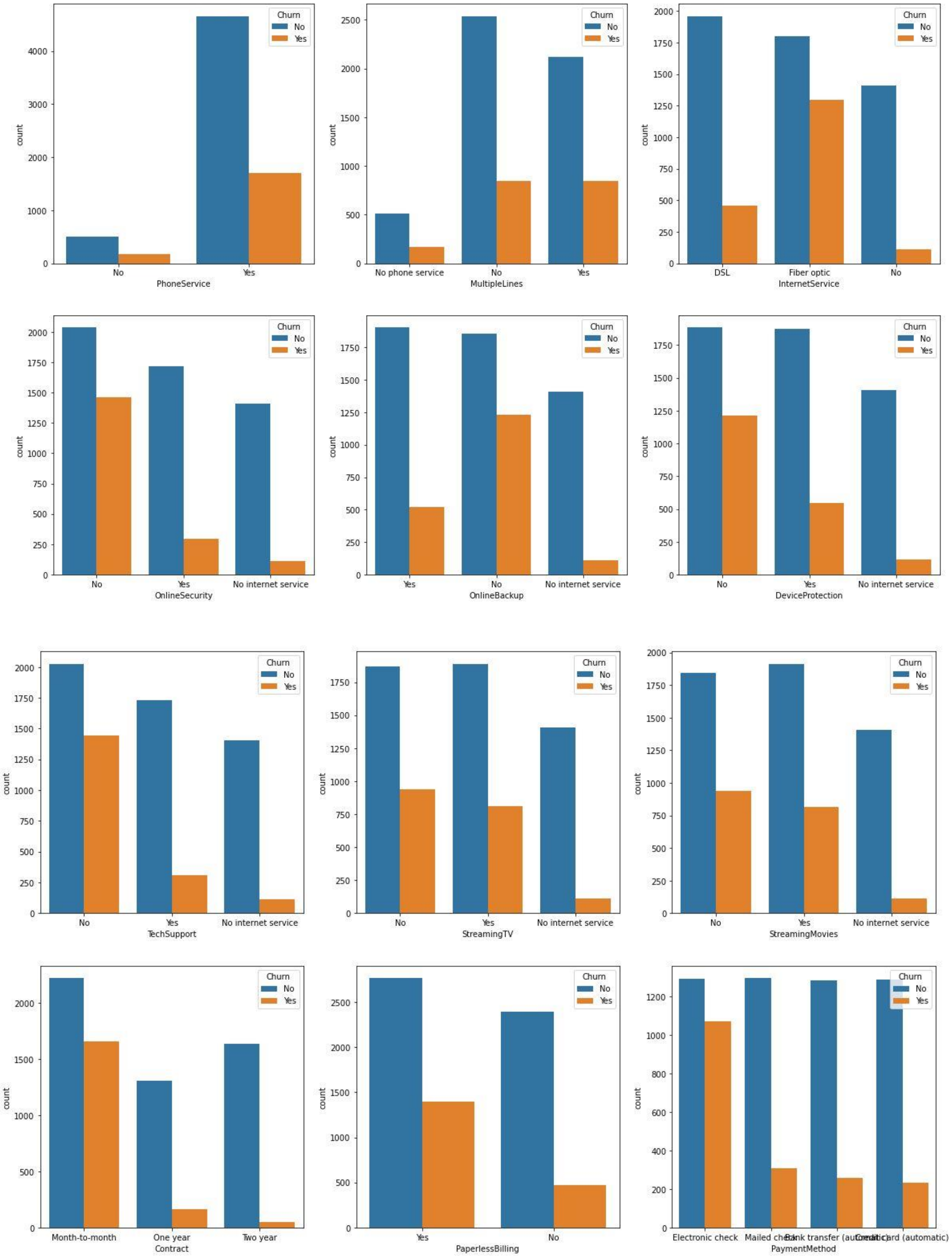

Churn
No
Yes
count
4000
3000
2000
1000
0
No
Yes
PhoneService
2500
2000
1500
1000
500
0
No phone service
No
Yes
MultipleLines
2000
1750
1500
1250
1000
750
500
250
0
DSL
Fiber optic
No
InternetService
No
Yes
No internet service
OnlineSecurity
Yes
No
No internet service
OnlineBackup
No
Yes
No internet service
DeviceProtection
No
Yes
No internet service
TechSupport
No
Yes
No internet service
StreamingTV
No
Yes
No internet service
StreamingMovies
Month-to-month
One year
Two year
Contract
Yes
No
PaperlessBilling
1200
1000
800
600
400
200
0
Electronic check
Mailed check
Bank transfer (automatic)
Credit card (automatic)
PaymentMethod

Figure 6 displayed bar graphs to show how phone and internet services, contract, paperless billing, and payment method affected telecommunications customer churn. The bar graph of phone service showed that people who had phone service tended to churn more than people who did not have phone service. The bar graph of multiple lines showed that people who had no multiple lines tended to churn more than people who had multiple lines and people who did not have phone service. The bar graph of internet service showed that people who had fiber optic tended to churn more than people who had DSL and people who did not have internet service. The bar graph of online security showed that people who did not have online security tended to churn more than people who had online security and people who did not have internet service. The bar graph of online backup showed that people who did not have online backup tended to churn more than people had online backup and people who did not have internet service. The bar graph of device protection showed that people who did not have device protection tended to churn more than people who had device protection and people who did not have internet service. The bar graph of tech support showed that people who did not have tech support tended to churn more than people who had tech support and people who did not have internet service. The bar graph of streaming television showed that people who did not have streaming television tended to churn more than people who had streaming television and people who did not have internet service. The bar graph of streaming movies showed that people who did not have streaming movies tended to churn more than people who had streaming movies and people who did not have internet service. The bar graph of contract showed that people who had month-to-month contract tended to churn more than people who had one-year contract and people who had two-year contract. The bar graph of paperless billing showed that people who had paperless billing tended to churn more than people who did not have paperless billing. The bar graph of payment method showed that people who paid in electronic check tended to churn more than people who paid in mailed check, people who paid bank transfer (automatic), and people who paid in credit card (automatic).

How do the service fee charges affect customer churn in telecommunications industry?

Figure 7. Monthly charges and churn.

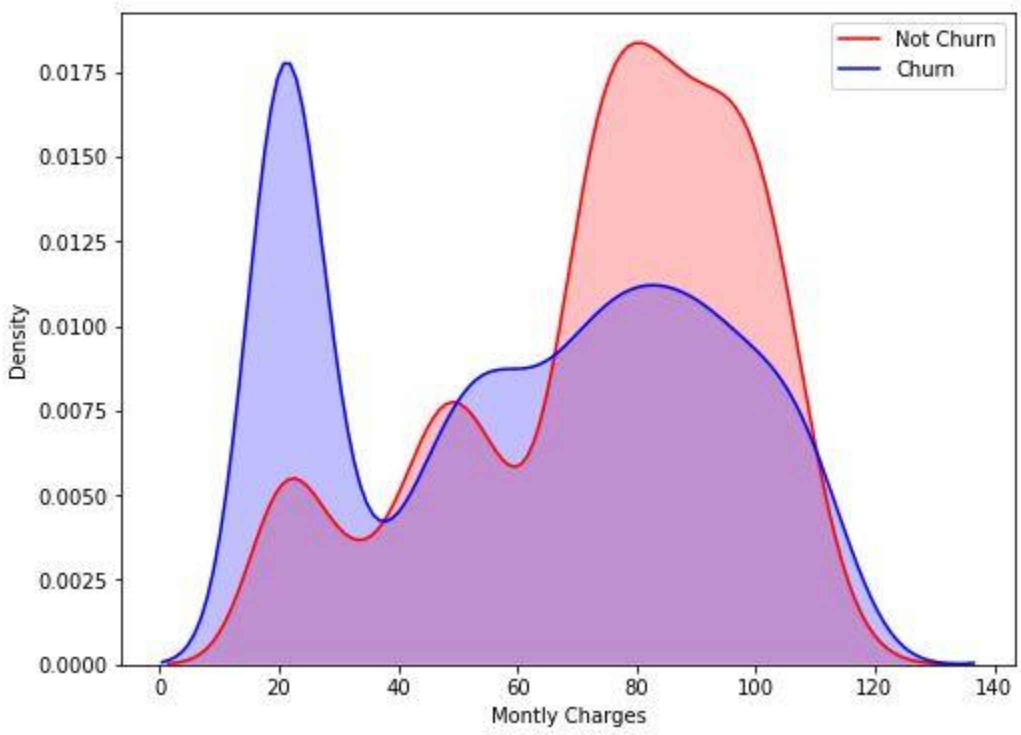


Figure 7 displayed kernel density estimation (KDE) plot to show how monthly charges affected telecommunications customer churn. The KDE plot showed that $20.00 of monthly charges had the highest density to churn by 0.0175. On the other hand, the KDE plot showed that $80.00 of monthly charges had the highest density not to churn by higher than 0.0175.

Figure 8. Total charges and churn.

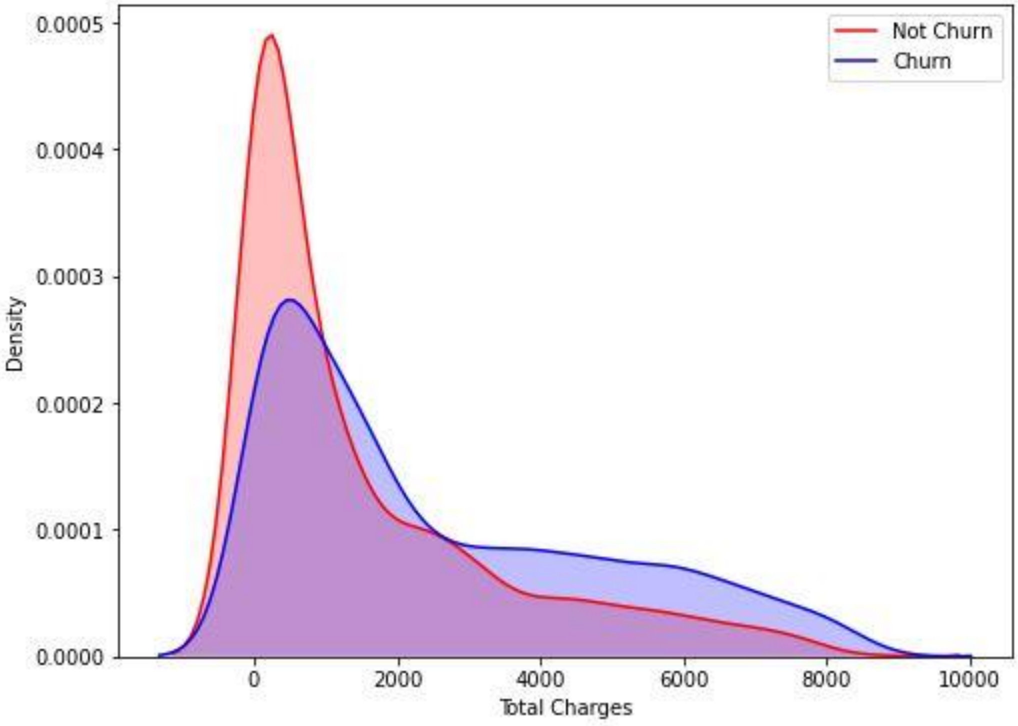


Figure 8 displayed kernel density estimation (KDE) plot to show how total charges affected telecommunications customer churn. The KDE plot showed that total charges more than $0 and less than $2000 had higher density not to churn by around 0.0005 rather than to churn by around 0.0003.

Figure 9. Monthly charges and total charges affect churn.

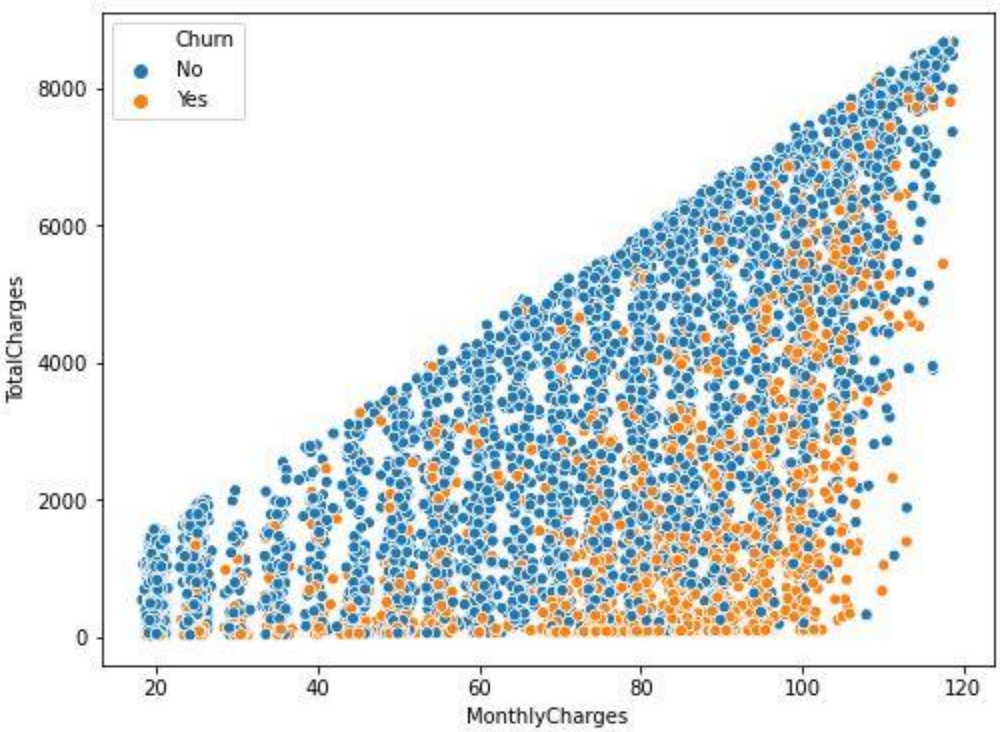


Figure 9 displayed a scatter plot to show how monthly charges and total charges affected telecommunications customer churn. The scatter plot showed that monthly charges from $80 to $100 and total charges from $1 and $2000 had more tendency to churn.

How do the services and monthly charges affect customer churn in telecommunications industry?

Figure 10. Phone and internet services and monthly charges affect churn.

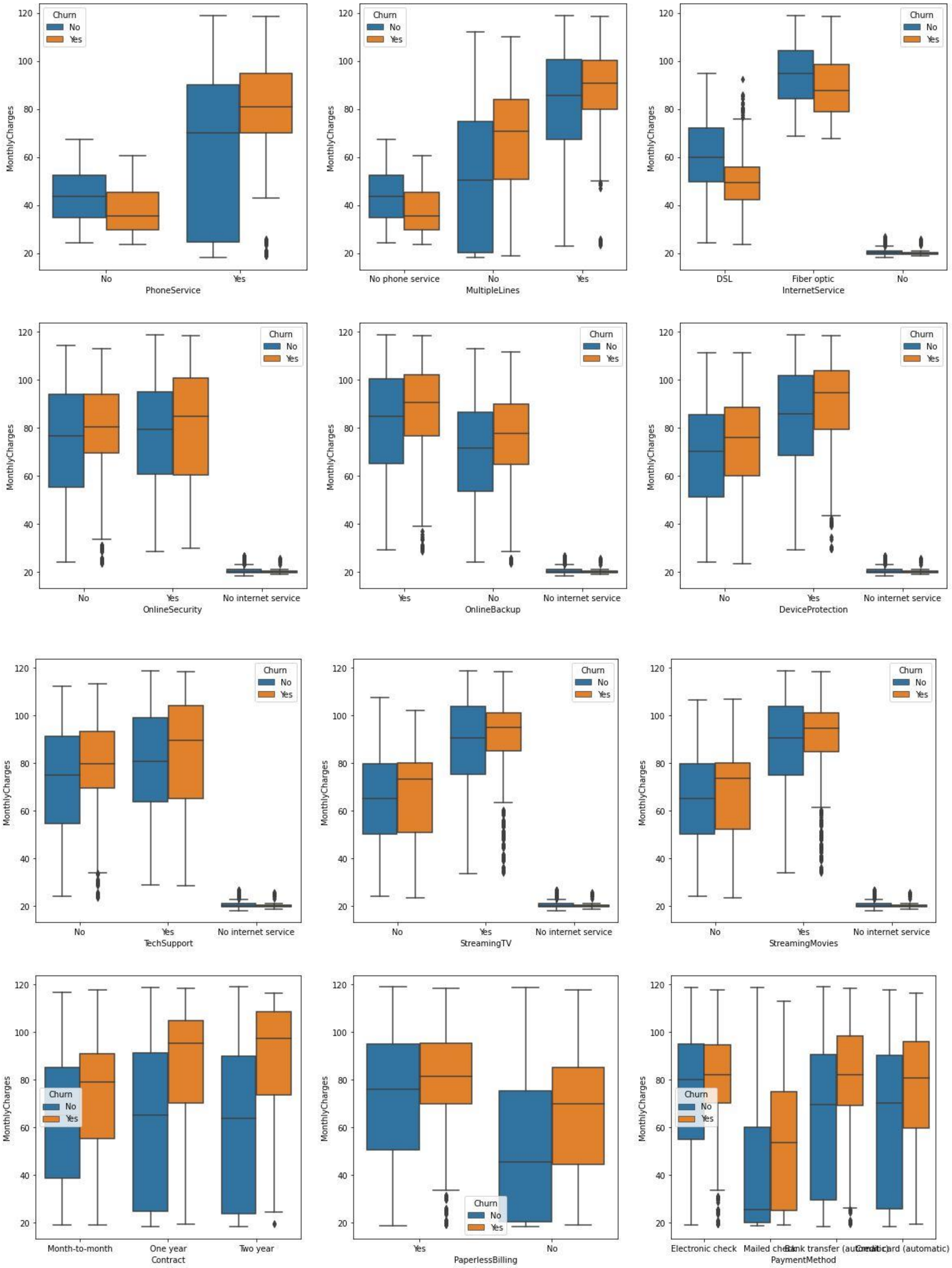

Churn
No
Yes
MonthlyCharges
120
100
80
60
40
20
PhoneService
No
Yes
MultipleLines
No phone service
No
Yes
InternetService
DSL
Fiber optic
No
OnlineSecurity
No
Yes
No internet service
OnlineBackup
Yes
No
No internet service
DeviceProtection
No
Yes
No internet service
TechSupport
No
Yes
No internet service
StreamingTV
No
Yes
No internet service
StreamingMovies
No
Yes
No internet service
Contract
Month-to-month
One year
Two year
PaperlessBilling
Yes
No
PaymentMethod
Electronic check
Mailed check
Bank transfer (automatic)
Credit card (automatic)

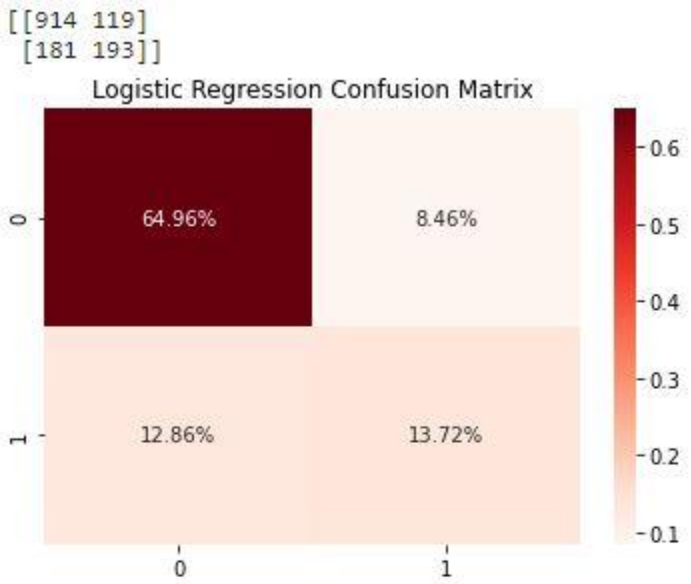
Figure 10 displayed box plots to show how services and monthly charges affected telecommunications customer churn. The box plot of phone service and monthly charges showed that people who had phone service and paid monthly charges from $70 to $90 had more tendency to churn. The box plot of multiple lines and monthly charges showed that people who did not have multiple lines and paid monthly charges from $50 to $80 had more tendency to churn. The box plot of internet service showed that people who had fiber optic and paid monthly charges from $80 to $100 had more tendency to churn. The box plot of online security and monthly charges showed that people who had online security and paid monthly charges from $70 to $110 had more tendency to churn. The box plot of online backup and monthly charges showed that both people who had online backup and people who did not have online backup had similar tendency to churn. The box plot of device protection and monthly charges showed that people who did not have device protection and paid monthly charges from $60 to $90 had more tendency to churn. The box plot of tech support and monthly charges showed that people who had tech support and paid monthly charges from $70 to $110 had more tendency to churn. The box plot of streaming television and monthly charges showed that people who did not have streaming television and paid monthly charges from $50 to $80 had more tendency to churn. The box plot of streaming movies and monthly charges showed that people who did not have streaming movies and paid monthly charges from $50 to $80 had more tendency to churn. The box plot of contract and monthly charges showed that people who had month-to-month contract and paid monthly charges from $50 to $90 had more tendency to churn. The box plot of paperless billing and monthly charges showed that people who did not have paperless billing and paid monthly charges from $50 to $90 had more tendency to churn. The box plot of payment method and monthly charges showed that people who paid monthly charges from $30 to $70 in mailed check had more tendency to churn.

Evaluating Predictive Models

The results of the projects can evaluate predictive models to solve the problems regarding telecommunications customer churn. There are several visualizations to display evaluating predictive models whether they predict telecommunications customer churn effectively by using classification report and confusion matrix.

Logistic Regression

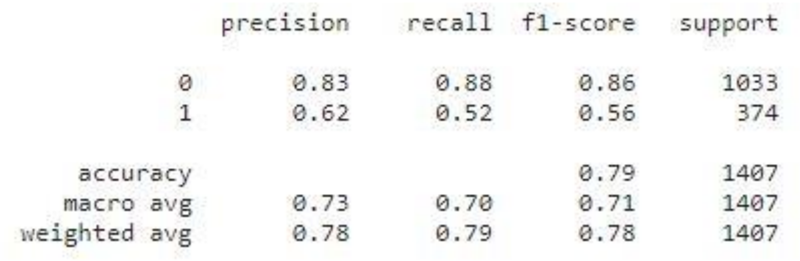
Figure 11. Confusion matrix of Logistic Regression.



```
[[914 119]
 [181 193]]
```

A confusion matrix of Logistic Regression displayed True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positives (TP) were 193 which were 13.72%. True Negatives (TN) were 914 which were 64.96%. False Positives (FP) were 119 which were 8.46%. False Negatives (FN) were 181 which were 12.86%.

Figure 12. Classification report of Logistic Regression.

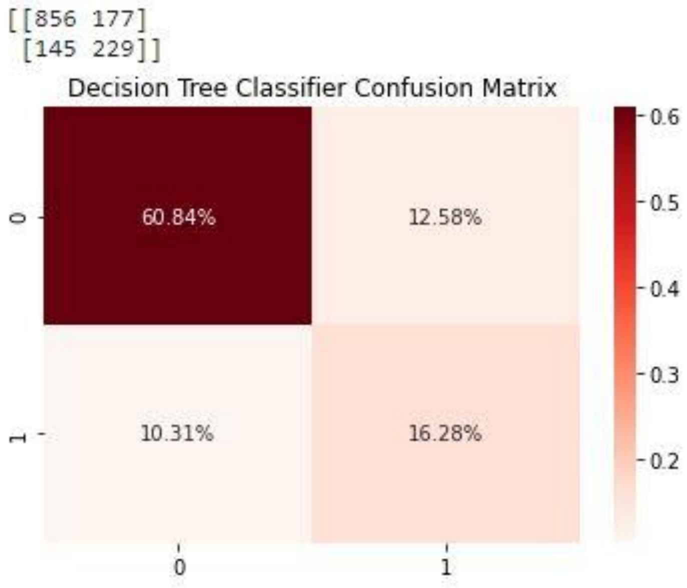


```
              precision    recall  f1-score   support

           0       0.83      0.88      0.86      1033
           1       0.62      0.52      0.56       374

    accuracy                           0.79      1407
   macro avg       0.73      0.70      0.71      1407
weighted avg       0.78      0.79      0.78      1407
```

A classification report of Logistic Regression displayed accuracy, precision (positive class), precision (negative class), sensitivity (recall of positive class), specificity (recall of negative class), F1 score (positive class), and F1 score (negative class). Accuracy was 0.79. Precision (positive class) was 0.62 and precision (negative class) was 0.83. Sensitivity (recall of positive class) was 0.52. Specificity (recall of negative class) was 0.88. F1 score (positive class) was 0.56 and F1 score (negative class) was 0.86.

Decision Tree

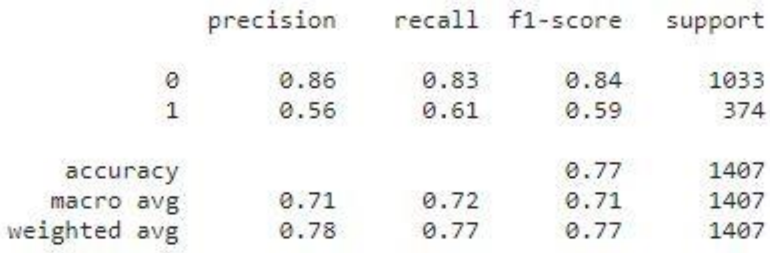
Figure 13. Confusion matrix of Decision Tree.



```
[[856 177]
 [145 229]]
```

A confusion matrix of Decision Tree displayed True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positives (TP) were 229 which were 16.28%. True Negatives (TN) were 856 which were 60.84%. False Positives (FP) were 177 which were 12.58%. False Negatives (FN) were 145 which were 10.31%.

Figure 14. Classification report of Decision Tree.

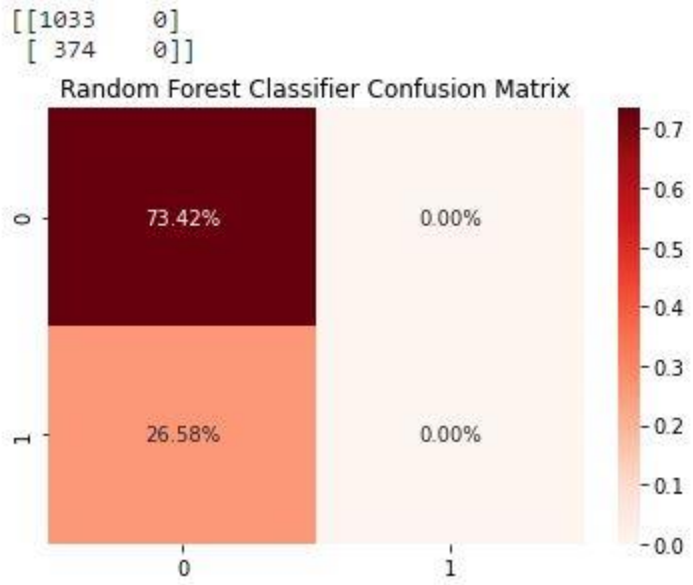


```
              precision    recall  f1-score   support

           0       0.86      0.83      0.84      1033
           1       0.56      0.61      0.59       374

    accuracy                           0.77      1407
   macro avg       0.71      0.72      0.71      1407
weighted avg       0.78      0.77      0.77      1407
```

A classification report of Decision Tree displayed accuracy, precision (positive class), precision (negative class), sensitivity (recall of positive class), specificity (recall of negative class), F1 score (positive class), and F1 score (negative class). Accuracy was 0.77. Precision (positive class) was 0.56 and precision (negative class) was 0.86. Sensitivity (recall of positive class) was 0.61. Specificity (recall of negative class) was 0.83. F1 score (positive class) was 0.59 and F1 score (negative class) was 0.84.

Random Forest

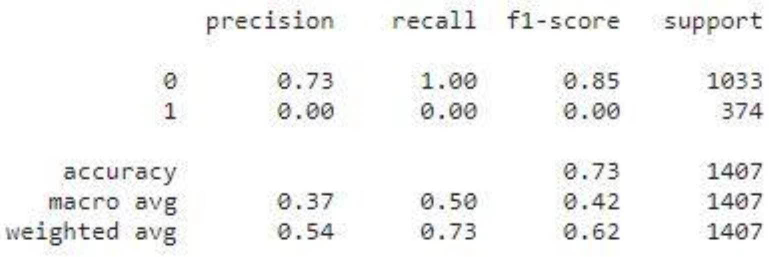
Figure 14. Confusion matrix of Random Forest.



```
[[1033    0]
 [ 374    0]]
```

A confusion matrix of Random Forest displayed True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positives (TP) were 0 which were 0%. True Negatives (TN) were 1033 which were 73.42%. False Positives (FP) were 0 which were 0%. False Negatives (FN) were 374 which were 26.58%.

Figure 15. Classification report of Random Forest.

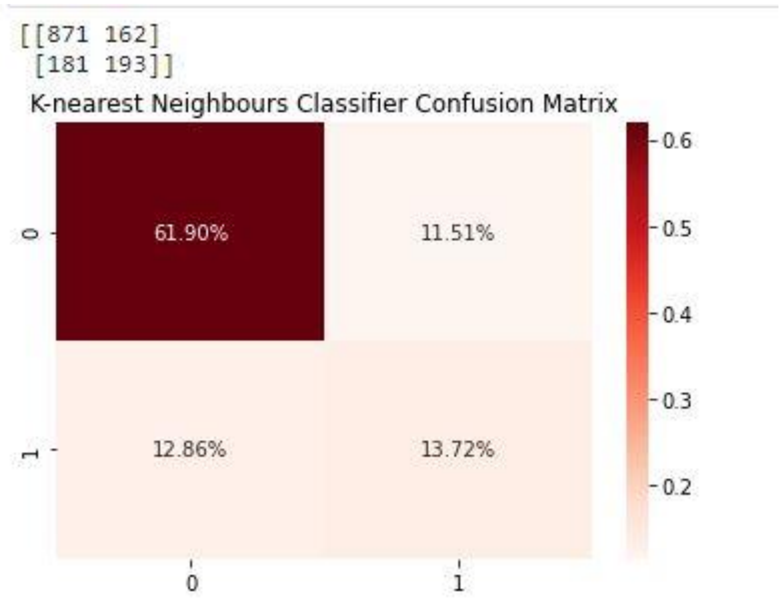


```
              precision    recall  f1-score   support

           0       0.73      1.00      0.85      1033
           1       0.00      0.00      0.00       374

    accuracy                           0.73      1407
   macro avg       0.37      0.50      0.42      1407
weighted avg       0.54      0.73      0.62      1407
```

A classification report of Random Forest displayed accuracy, precision (positive class), precision (negative class), sensitivity (recall of positive class), specificity (recall of negative class), F1 score (positive class), and F1 score (negative class). Accuracy was 0.73. Precision (positive class) was 0 and precision (negative class) was 0.73. Sensitivity (recall of positive class) was 0. Specificity (recall of negative class) was 1. F1 score (positive class) was 0 and F1 score (negative class) was 0.85.

K-nearest Neighbours

Figure 16. Confusion matrix of K-nearest Neighbours.

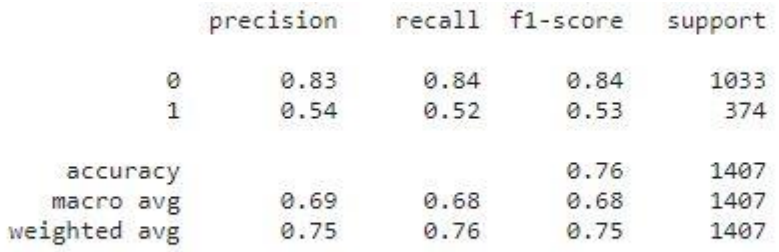


```
[[871 162]
 [181 193]]
```

K-nearest Neighbours Classifier Confusion Matrix

A confusion matrix of K-nearest Neighbours displayed True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positives (TP) were 193 which were 13.72%. True Negatives (TN) were 871 which were 61.90%. False Positives (FP) were 162 which were 11.51%. False Negatives (FN) were 181 which were 12.86%.

Figure 17. Classification report of K-nearest Neighbours.

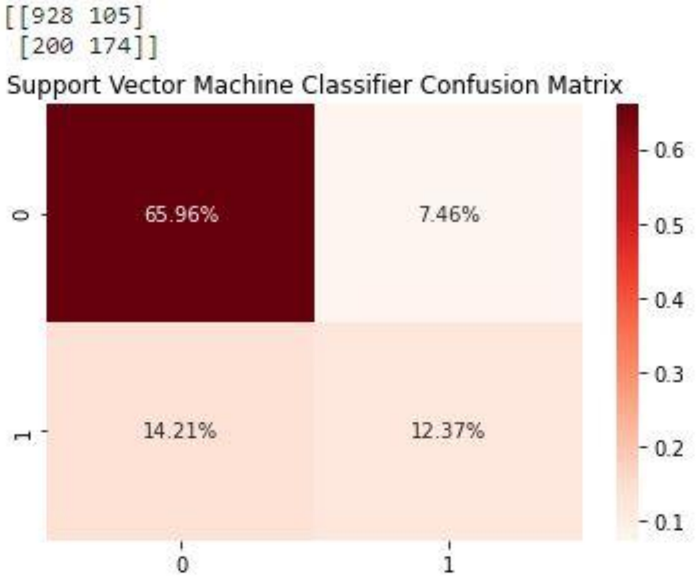


```
              precision    recall  f1-score   support

           0       0.83      0.84      0.84      1033
           1       0.54      0.52      0.53       374

    accuracy                           0.76      1407
   macro avg       0.69      0.68      0.68      1407
weighted avg       0.75      0.76      0.75      1407
```

A classification report of K-nearest Neighbours displayed accuracy, precision (positive class), precision (negative class), sensitivity (recall of positive class), specificity (recall of negative class), F1 score (positive class), and F1 score (negative class). Accuracy was 0.76. Precision (positive class) was 0.54 and precision (negative class) was 0.83. Sensitivity (recall of positive class) was 0.52. Specificity (recall of negative class) was 0.84. F1 score (positive class) was 0.53 and F1 score (negative class) was 0.84.

Support Vector Machine

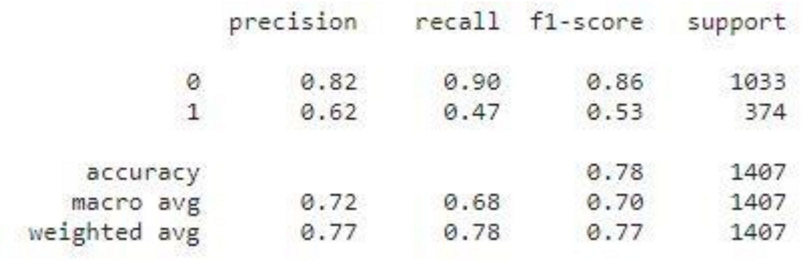
Figure 18. Confusion matrix of Support Vector Machine.



```
[[928 105]
 [200 174]]
```

A confusion matrix of Support Vector Machine displayed True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positives (TP) were 174 which were 12.37%. True Negatives (TN) were 928 which were 65.96%. False Positives (FP) were 105 which were 7.46%. False Negatives (FN) were 200 which were 14.21%.

Figure 19. Classification report of Support Vector Machine.



```
              precision    recall  f1-score   support

           0       0.82      0.90      0.86      1033
           1       0.62      0.47      0.53       374

    accuracy                           0.78      1407
   macro avg       0.72      0.68      0.70      1407
weighted avg       0.77      0.78      0.77      1407
```

A classification report of Support Vector Machine displayed accuracy, precision (positive class), precision (negative class), sensitivity (recall of positive class), specificity (recall of negative class), F1 score (positive class), and F1 score (negative class). Accuracy was 0.78. Precision (positive class) was 0.62 and precision (negative class) was 0.82. Sensitivity (recall of positive class) was 0.47. Specificity (recall of negative class) was 0.90. F1 score (positive class) was 0.53 and F1 score (negative class) was 0.86.

References

BlastChar. (2018, February 23). *Telco customer churn*. Kaggle. Retrieved September 26, 2021, from https://www.kaggle.com/blastchar/telco-customer-churn.

Chen, Z., Fan, Z., & Sun, M. (2012). A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. *Eur. J. Oper. Res., 223*, 461-472.

IBM. (2020, May 1). *Telco customer churn*. IBM Community. Retrieved September 26, 2021, from https://community.ibm.com/accelerators/catalog/content/Telco-customer-churn.

IBM Samples Team. (2019, July 11). *Telco customer churn (11.1.3+)*. IBM Community. Retrieved September 26, 2021, from https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113.

Idris, A., Rizwan, M., & Khan, A. (2012). Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies. *Comput. Electr. Eng., 38*, 1808-1819.

Kim, K., & Lee, J. (2012). Sequential manifold learning for efficient churn prediction. Expert Systems with Applications, 39(18), 13328–13337. doi:10.1016/j.eswa.2012.05.069

Kim, S. (2021, October 17). *big-data-analytics-project*. GitHub. https://github.com/seiryunkim/big-data-analytics-project.

Lin, W., Tsai, C., & Ke, S. (2014). Dimensionality and data reduction in telecom churn prediction. *Kybernetes, 43*, 737-749.

Wu, S., Yau, W., Ong, T., & Chong, S. (2021, January 1). *Integrated churn prediction and customer segmentation framework for telco business.* Institute of Electrical and Electronics Engineers. doi:10.1109/ACCESS.2021.3073776.

Zhu, B., Baesens, B., & Broucke, S.V. (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. *Inf. Sci., 408*, 84-99.