© Universiti Tun Hussein Onn Malaysia Publisher's Office

JSCDM

http://penerbit.uthm.edu.my/ojs/index.php/jscdm

e-ISSN : 2716-621X

Journal of Soft Computing and Data Mining

Customer Churn Prediction of Telecom Company Using Machine Learning Algorithms

Angela Yi Wen Chong[1], Khai Wah Khaw[1*], Wai Chung Yeong[2], Wen Xu Chuah[1]

[1]School of Management,
Universiti Sains Malaysia, 11800 USM, Penang, MALAYSIA

[2]School of Mathematical Sciences,
Sunway University, Petaling Jaya, MALAYSIA

*Corresponding Author

DOI: https://doi.org/10.30880/jscdm.2023.04.02.001
Received 18 February 2023; Accepted 21 August 2023; Available online 04 October 2023

Abstract: We can't escape the fact that using telecommunications has become a significant part of our everyday lives. Since the Covid-19 pandemic, the telecommunication industry has become crucial. Hence, the industry now enjoys growth opportunities. In this study, KNN, Random Forest (RF), AdaBoost, Logistic Regression (LR), XGBoost, and Support Vector Machine (SVM) are 6 supervised machine learning algorithms that will be used in this study to predict the customer churn of a telecom company in California. The goal of this study is to identify the classifier that predicts customer churn the most effectively. As evidenced by its accuracy of 79.67%, precision of 64.67%, recall of 51.87%, and F1-score of 57.57%, XGBoost is the overall most effective classifier in this study. Next, the purpose of this study is to identify the characteristics of customers who are most likely to leave the telecom company. These characteristics were discovered based on customers' demographics and account information. Lastly, this study also provides the company with advice on how to retain customers. The study advises company to personalize the customer experience, implement a customer loyalty program, and apply AI in customer relationship management in retaining customers.

Keywords: Machine learning, supervised machine learning, customer churn prediction, XGBoost

1. Introduction

In this era of globalization, we can't deny that the use of telecommunications has become an important part of our daily life. Telecommunication is the term used to describe the conveyance of a communication or message across a distance through different techniques, including using the telephone, cable, and other methods. Telecommunications also had a distinct meaning before the growth of the Internet and other data networks, which is the public switched telephone network (PSTN) that offered telephone service, making it possible for people to speak to one another across long distances [1]. It was born out of people's need to communicate across longer distances than were possible with human voice alone. Telecommunications have undergone a significant transformation due to technical advancement during the past three decades. For instance, people communicated with drums and smoke signals in the early days whereas today people use digital wireless networks to communicate. Later in the paper, telecommunication will be shortened to telecom.

The World Health Organization (WHO) identified Covid-19 as a worldwide public health issue on January 30, 2020, and a pandemic on March 11, 2020 [2]. The Covid-19 pandemic is drastically affecting people's health, lifestyle, economy, society and so on. When the pandemic first began, the government implemented various preventative measures to stop its spread, including a nationwide lockdown to prevent people from crowding. Therefore, the telecom industry

*Corresponding author: khaiwah@usm.my

2023 UTHM Publisher. All rights reserved.

has been playing an extra significant role in connecting people and delivering information and message during the lockdown. Consequently, the increased usage of the internet, video conferencing, and telephone services have created chances for the telecom industry to grow. According to [3], the revenue growth rate of California telecom companies such as Zoom Video Communications has increased dramatically from -7.1% in 2020 to 3.5% in 2022, which has grown 149.30%. Overall, telecommunication includes a variety of service providers such as telephone, cable system operators, internet service, wireless providers, and satellite operators. The widespread acceptance and understanding of the societal significance of telecommunications are demonstrated by its almost universal penetration and usage. Communications across society (including families, businesses, and the government) rely on telecommunication as a technological backbone. The evidence is the increasing integration of telecommunications into our daily activities, from Web browsing to making phone calls and instant messaging.

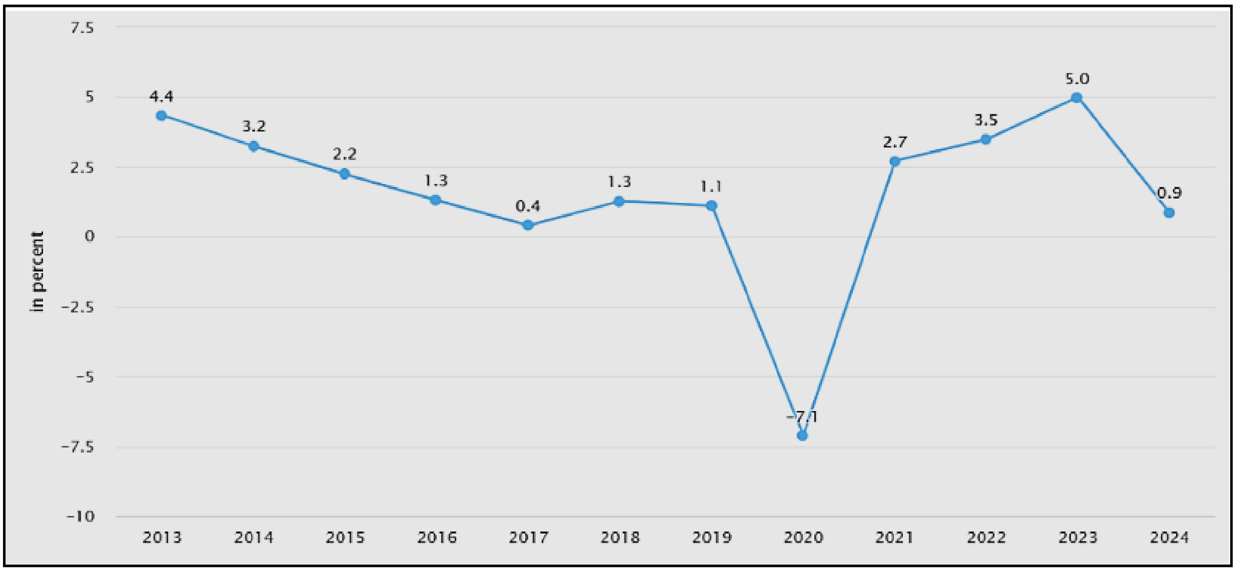


Fig. 1 - Revenue growth rate of telecom companies in California [2]

Lastly, it is also critical to understand the research topic: customer churn, after looking at the telecommunications sector. The process of customers switching service providers, whether they are post-paid or prepaid, is known as customer churn [4]. Even large telecom companies like Vodafone continue to see customer churn rates of 20% or greater in developed countries like Germany [5]. It is a serious issue for telecommunication companies as it can hinder the company's growth and revenue. Furthermore, the telecommunication industry has recently undergone numerous changes including market liberalization, new technologies, and market saturation, which has dramatically increased competition from other competitors. Consequently, service providers are increasingly tempted to expand their customer base, but it is much more challenging to retain customers.

One of the objectives of this paper is to predict the customer churn of a telecom company in California by using the selected 6 machine learning algorithms. Data mining approaches have evolved to address the difficult issues of customer churn in the telecommunications industry ([6]; [7]). In most cases, the company is unaware of why a customer is leaving and analyzing the data in this situation can assist in understanding and predicting customer churn to prevent it from happening. Furthermore, building an efficient and accurate customer churn model is crucial for businesses that looking to manage customer churn successfully [6]. Therefore, this study also aims to find out what are the characteristics of churn customers to understand them better. Finding the customers most likely to discontinue using a service or product is the goal of churn analysis [8]. Finally, the goal of this study is to advise the company in retaining customers.

Customer churns prediction and understanding the reason for customer churn is necessary for a telecom company. Firstly, customer churn prediction can be used as a business metric in assessing a company's performance. One of the key tactics for retaining customers is churn prediction, which has been a hot topic in the telecom business and research area [6]. A company's performance is usually evaluated by its own business metrics, which differ for every industry. Business metrics are crucial since they monitor a company's progress toward its objectives and encourage improvement to the way things are currently done. With prediction, companies can use this information to reduce the probability of customers churning and extend the retention period of existing customers. Secondly, a company can target its customer retention efforts only on high-risk categories if it could predict why customers are most likely to leave in advance. Using a customer churn prediction model will allow the company to use the retention strategy on a certain category instead of customers on an individual basis. The study by [9] shows that due to the increased market competition, companies have spent their efforts on keeping existing consumers rather than gaining new ones.

2. Literature Review

2.1 Supervised Machine Learning

This study will use supervised learning algorithms to develop customer churn prediction models. This is because the study's target to predict is already known; whether the customer has churned, which is "Yes" or "No". The function of SML is to match input to output using examples of input-output pairings from the training dataset. Therefore, this study is also a classification problem. Using a collection of known values (inputs), a classification model predicts unknown values (outputs). The problem is known as a classification problem when the output is in categorical form. A classification algorithm or classifier learns from the training dataset and assigns each new data point to a specific class. In this study, 6 supervised machine learning algorithms will be used, namely KNN, SVM, RF, LG, XGBoost and AdaBoost. The following section will briefly review the background of each supervised machine learning algorithm and justify the reason they are being selected for this study.

2.1.1 k-Nearest Neighbour (KNN)

In 1951 unpublished research from the US Air Force School of Aviation Medicine, Fix and Hodges created a non-parametric method for pattern classification, which has currently known as the k-Nearest Neighbour [10]. KNN is one of the most used supervised machine learning techniques that is widely used to solve classification problems. With N training vectors, the KNN approach determines the k-nearest neighbours of an unknown feature vector whose class has to be identified [11]. The formula for calculating Euclidean distance in KNN is as follows:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} \quad (xi - yi)^2} \tag{1}$$

The reason why KNN is selected to use in this study is due to it is one of the most used SML algorithms due to its simple yet effective implementation. This is because KNN is a lazy learning method since it memorises the training data and then utilises it to classify new input. According to [12], the KNN algorithm is easy to implement since it is based solely on calculating the distance (typically Euclidean distance) between the features. The authors of [13] show that the KNN can deal with data that is both large and incrementally multiclass data in which the study includes KNN as one of the algorithms to detect intrusion in a Cloud environment.

2.1.2 Support Vector Machine (SVM)

SVM is one of the SML algorithms that are usually used for classification, regression, and pattern recognition. SVM started as binary classifiers that were not probabilistic. However, they are increasingly used in multi-class problems as well. SVM in this scenario constructs n-dimension hyperplanes that partition data into n groups/classes optimally [14]. SVM's primary objective is to use a surface that separates several classes in the training data by optimizing the margin between them [15]. The distance between the decision surface (hyperplane) and the nearest data point sets the classifier's margin.

The reason SVM is selected because this algorithm performs well when with higher dimension data. The most significant benefit of SVM is that it avoids overlearning and excessive dimension, which both contribute to computational complexity and local extremum [16]. Next, SVM always has high accuracy. Because of its excellent accuracy, SVM is another well-known and commonly used ML technique for tackling classification and regression problems. This classifier finds the most effective boundaries for distinguishing between positive and negative training examples [17].

2.1.3 Random Forest Classifier (RF)

Random Forest are a collection of tree predictors where each tree's value is decided by a random vector that is evenly and randomly chosen across the forest [18]. Random Forest is a machine learning algorithm that uses two methods at once. First, the bootstrap aggregating approach relies on building several decision trees and polling them all together while the random subspace approach, involves picking m-th randomly chosen attributes from a collection of all available attributes [19]. Then, each newly generated tree associates the item with one of the classes to determine the winner; the class with the most votes from the trees wins. This voting process is used to determine the final choice.

Random Forest is a powerful prediction tool due to the Law of Large Numbers prevents them from overfitting [18]. Next, for mining highly dimensional data, Random Forest offer a fast, efficient, and reliable solution [20]. They function effectively even when there are many features and few observations. Furthermore, there are methods in the random forest that can be used to choose features easily and measure the significance of variables [19]. In particular, the first step involves the usage of Random Forest to choose the most important features. Lastly, Random Forest is an effective classifier. The study of [21] used Random Forest to increase accuracy since it has a lower classification error than other traditional classification methods. Hence, these are the reasons why the Random Forest algorithm is selected to use in this study.

2.1.4 Logistic Regression (LG)

In comparison to “ordinary” linear regression, Logistic Regression goes further. It is applied in situations when the dependent variable, Y, is categorical [22]. Logistic Regression was used to analyze data to identify relationships between the dependent variable and multiple independent variables [23]. Multinomial and binary logistic regression models are the two types of logistic regression models. Binary logistic regression is a regression model in which the target variable has just two values, 0 or 1 [24]. Hence, binary logistic regression will be used in this study since the outcome of the prediction is in binary form (Yes or No).

First, the most popular method for a binary outcome is Logistic Regression. The study of [25] proposed that Logistic Regression is becoming more popular in the nursing field because a binary dependent variable and several independent variables can be modelled. Also, Logistic Regression is simple to use because it doesn’t need any hyperparameters to be optimized. Lastly, even with a small sample size, few events, and few simple predictors, logistic regression can still be effective. This is proved by the study of [26] where the result shows Logistic Regression outperforms ML models in terms of predicting the risk of major chronic illnesses. Logistic Regression also has the clear benefit of being able to quickly select suitable candidates from untested resources while also saving computation time and expense [27].

2.1.5 XGBoost

XGBoost was developed by Tianqi Chen as a research project for the Distributed (Deep) Machine Learning Community (DMLC) [28]. XGBoost, sometimes known as an open-source software library called Extreme Gradient Boosting offers a regularising gradient boosting framework for multiple platforms, including Python. This algorithm stands for a method of machine boosting, or more specifically, applying to boost to machines. It is created and tuned to be effective, adaptable, and portable [29].

The reason that XGBoost is being selected to use in this study is because of its execution speed and excellent model performance [28]. First, XGBoost is faster than other gradient boosting implementations. Next, as for model performance, XGBoost is now among the most accurate machine learning algorithms [29]. It performs well in classification and regression; the training algorithm performance is also good due to the underlying gradient-based approach that promotes trees, which are better able to adapt to imbalanced data sets [30]. Furthermore, scalability in all situations is the primary driver of XGBoost’s success. The system is more than ten times faster than common solutions on a single computer and scalable to billions of samples in distributed or memory-constrained scenarios [31]. Lastly, in comparison to the conventional approach, the XGBoost-based method can take into account more features [32].

2.1.6 AdaBoost

AdaBoost is a machine learning algorithm that enhances the performance of other machine learning algorithms. This kind of learner focuses more of their attention during training on samples that were incorrectly classified, adjusts the sample distribution, and repeats the process until the weak classifier has undergone a predetermined number of training, at which point the learning is complete [33].

The advantage of the AdaBoost method is that it can reclassify the classified error data in the subsequent training run by altering the weight of the error data [33]. Next, AdaBoost has a low own boosting overhead. The core algorithms’ training times almost entirely decide how long it takes to construct the final image [19]. The research of [34] has shown that AdaBoost is a popular and effective boosting algorithm while Decision Tree is a good weak learner in which the findings suggest that this approach performs well in fingerprint classification.

2.2 Related Works Regarding Customer Churn Prediction of Telecom Company

The section below shows some of the research that used machine learning algorithms to predict customer churn in the telecom company. The summarized literature table can be viewed in Table 1. To begin with, [35], researchers analyze the features that are related to churn and used Python to apply the four machine learning classifiers, Logistic Regression, KNN, SVM, and Random Forest, for predicting customer churn. The obtained result shows that total charges (0.9), monthly contract (0.7), and fiber optic internet service (0.5) are the top three features that contribute the most to customer churn. The study also proposed the accuracy and AUC value after applying recursive feature elimination (RFE), which removes the weakest feature and prevents over-fitting of the data while also improving accuracy. According to the findings of the research, each classifier has a favourable accuracy with an accuracy of over 85% after RFE, while the best result is obtained by Logistic Regression (100%).

In addition, [36], authors tried to increase the accuracy of customer churn prediction using KNN, Random Forest and XGBoost. The result shows that the performance of XGBoost outperforms both KNN and RF in terms of accuracy of 79.80% and an F-score of 58.2%. The study also uses XGBoost to discover the characteristic of customers that has a greater probability for churn and shows customers that who use fiber optic options that pay higher monthly charges have a stronger impact on churn.

Next, [37], numerous machine learning models were used by researchers to predict customer churn in the telecom sector, in which the obtained results confirmed that AdaBoost has the highest accuracy in comparison to other algorithms, scoring 81.71% for the churn prediction of the customer. All performance metrics, namely accuracy, precision, F-measure, recall, and AUC score were higher with them than with other methods.

Furthermore, [38], researchers applied two efficient machine learning algorithms which are Random Forest (RF) classifiers and Support Vector Machine (SVM) which select the important features, improving the performance of the currently existing system that trained blindly with many irrelevant features using the LDT and UDT algorithms in predicting churn customers. The findings show that RF achieved the highest efficiency of 96% whereas SVM achieved 85%. Both efficiencies are indicated in the form of a confusion matrix.

Also, [39], the churn prediction problem was addressed by authors using KNN and Logistic Regression (LR). KNN (88.5%) initially performed almost as well as LR (86.5%), but the performance is better than LR by 2% in terms of accuracy. The study concluded that KNN is more successful in predicting customer churn than LR out of 15 important parameters.

Then, [40], researchers evaluate various SVM kernels to predict customer churn on a telecom dataset with an uneven distribution of churn and non-churn customers. The results reveal that simple linear kernels outperform sigmoid, polynomial, and RBF kernels in terms of hit rate, F1-score and AUC. Choosing the right kernel parameters is essential for increasing a model’s accuracy. The experiment also shows that, instead of using complicated non-linear transformations of input data, it is possible to find a good linear hyperplane that effectively separates customer data that has been churn and non-churn when enough input features of customer data are available.

However, [41], authors encountered the problem of an imbalanced dataset where customers’ churn only accounts for 5% of the entries. This is a significant issue in churn prediction problems where a considerable negative impact on the final models might result from the target class being imbalanced. Undersampling or using tree techniques: Decision Tree (DT), Random Forest (RF), AdaBoost and XGBoost that are not impacted by this problem were used to solve the issue. The study adopted Area Under Curve (AUC) as the standard performance measure and XGBoost having the highest AUC value followed by AdaBoost, RF and DT.

Table 1 - Summary of related works

Authors	KNN	LR	RF	SVM	AdaBoost	XGBoost
[35]		Acc (RFE): 100% Acc (No RFE): 80.05%	Acc (RFE): 98.44% Acc (No RFE): 77.73%			
[36]	Acc: 75.40% F-score: 49.50%		Acc: 77.50% F-score: 50.60%			Acc: 79.80% F-score: 58.20%
[37]	Acc: 79.64% Recall: 79.1% Precision: 78.38% F-measure: 77%	Acc: 80.45% Recall: 80.23% Precision: 79.11% F-measure: 78.89%	Acc: 78.04% Recall: 78.68% Precision: 77.54% F-measure: 77.91%	Acc: 80.21% Recall: 80.64% Precision: 79.66% F-measure: 78.89%	Acc: 81.71% Recall: 81.21% Precision: 80.14% F-measure: 80.28%	Acc: 80.8% Recall: 80.7% Precision: 80.3% F-measure: 78.7%
[38]			Efficiency: 96%	Efficiency: 85%		
[39]	Acc: 88.47% Precision: 88.47%	Acc: 86.5% Precision: 87.5%				
[40]				Linear kernel Hit rate: 79.73% F-score: 80.27% AUC: 0.9278		
[41]			AUC: 78.47%		AUC: 90.89%	AUC: 93.30%

3. Methodology

Figure 2 shows the methodological framework of this study.

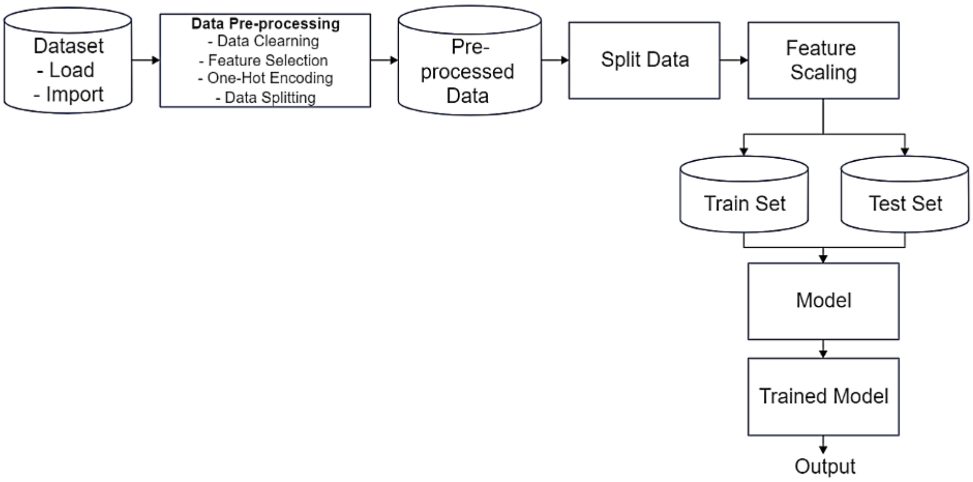


Fig. 2 - Methodological framework

3.1 Tools

In this study, the tool that is used in developing the machine learning classifier to predict customer churn is Python from Jupyter notebook. The reason to use Python is that it is supported by a large library and hence, we can just simply import functions or packages from Python's library.

3.2 Dataset

The telecom customer dataset was acquired from the Kaggle, which was uploaded by the user Blastchar 5 years ago. According to the uploader, the telecom customer data is based on the telecom company in California, USA. Table 2 displays a detailed dataset's data dictionary. This dataset has a total of 21 columns, which includes 7043 rows (each of which represents a different customer), 19 features and 1 target column (Churn) in the combination with categorical and numerical features. Table 2 displays a detailed dataset's data dictionary.

Table 2 - Data dictionary

Features/Target	Description
Customer ID	The id of customers.
Gender	Male or Female.
Senior Citizen	1= Senior citizen; 0=Not senior citizen.
Partner	Whether the customer has a partner (Yes or No).
Dependents	Whether the customer has dependents (Yes or No).
Tenure	Number of months the customer has been with the company.
Phone Service	Whether the customer has a phone service (Yes or No).
Multiple Lines	Whether the customer has multiple lines (Yes or No).
Internet Service	Internet service provider for the customer (DSL; Fiber optic; No).
Online Security	Whether the customer has online security (Yes; No; No internet service).
Online Backup	Whether the customer has online backup (Yes; No; No internet service).

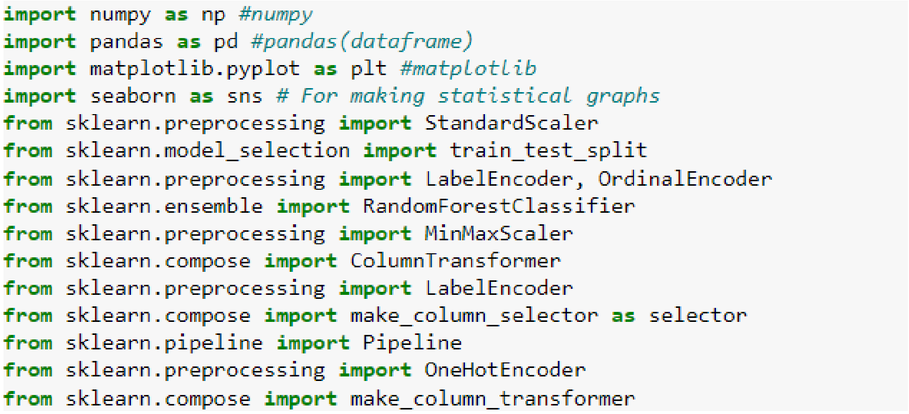
Device Protection	Whether the customer has device protection (Yes; No; No internet service).
Tech Support	Whether the customer has tech support (Yes; No; No internet service).
Streaming TV	Whether the customer has streaming TV (Yes; No; No internet service).
Streaming Movies	Whether the customer has streaming movies (Yes; No; No internet service).
Contract	Customer contract term (Month-to-month; One year; Two year).
Paperless Billing	Whether the customer has paperless billing (Yes or No).
Payment Method	Payment method used by the customer (Electronic check; Mailed check; Bank transfer (automatic); Credit card (automatic)).
Monthly Charges	The monthly amount charged to the customer.
Total Charges	The total charge made to the customer.
Churn	Whether the customer churned (Yes or No).

3.3 Loading Dataset

The telecom customer churn prediction dataset was loaded into Python by using forward slashes after retrieving the URL file path from the file saved on the laptop's desktop.

3.4 Importing The Required Libraries

The prewritten code in a Python library can be used by programmers to simplify tasks. The library of reused code is intended to solve certain problems. Hence, libraries will be used until data pre-processing steps were imported in advance to access the code to solve the problem.



```
import numpy as np #numpy
import pandas as pd #pandas(dataframe)
import matplotlib.pyplot as plt #matplotlib
import seaborn as sns # For making statistical graphs
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, OrdinalEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import MinMaxScaler
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import LabelEncoder
from sklearn.compose import make_column_selector as selector
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import make_column_transformer
```

Fig. 3 - Imported libraries

3.5 Importing Dataset

The dataset is given in CSV format and consists of tabular data. To generate a data frame from the provided dataset, the *read_csv()* function of the pandas' package was used.

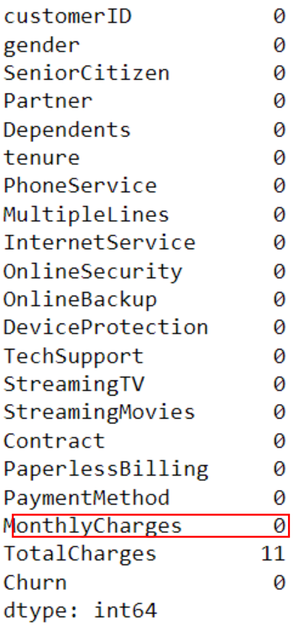
3.6 Data Pre-Processing

3.6.1 Data Cleaning

One of the important steps in data pre-processing is cleaning the data before building a machine learning model. Handling missing values and dropping unnecessary ones were the two procedures that were taken in this study to clean the data. The following section shows the details of the process of data cleaning.

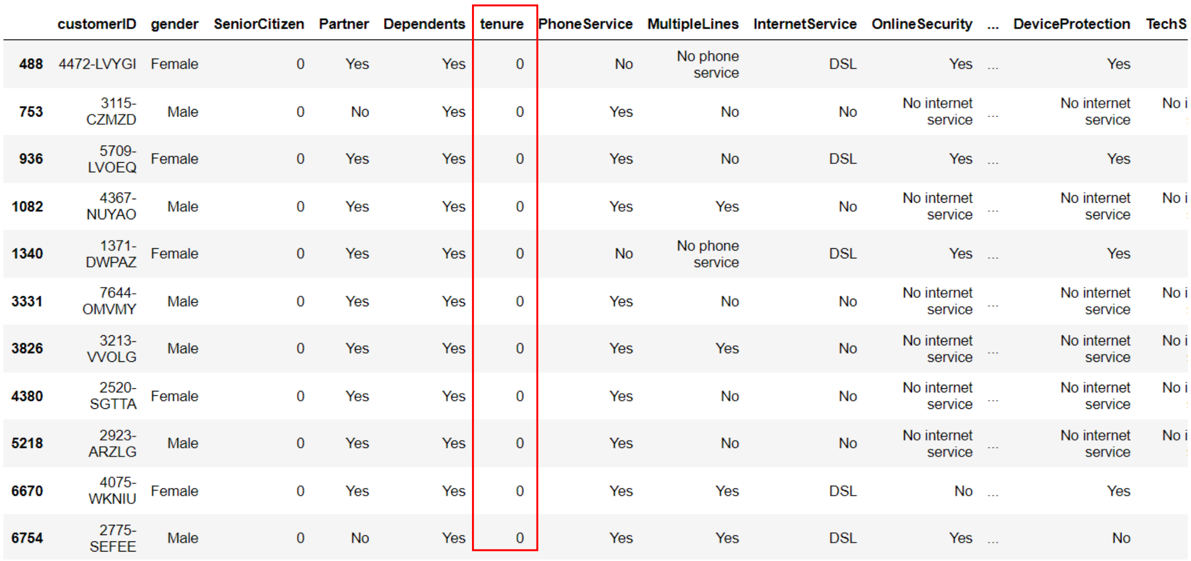
i. Handling Missing Values

Dealing with missing values was performed as part of the data cleaning. Firstly, the. dtypes function was used to check the datatype of the columns in the datasets, and it was revealed that the datatype of the numeric column "Total Charges" is in categorical or object type. Since the data in this column are in numerical values, *pd.to_numeric()* is used to convert the column into a numerical datatype. Then, the *.isnull().sum()* function is used to detect whether the dataset had any missing values. The result reveals that "Total Charges" has 11 missing values.



```
customerID           0
gender               0
SeniorCitizen        0
Partner              0
Dependents           0
tenure               0
PhoneService         0
MultipleLines        0
InternetService      0
OnlineSecurity       0
OnlineBackup         0
DeviceProtection     0
TechSupport          0
StreamingTV          0
StreamingMovies      0
Contract             0
PaperlessBilling     0
PaymentMethod        0
MonthlyCharges       0
TotalCharges        11
Churn                0
dtype: int64
```

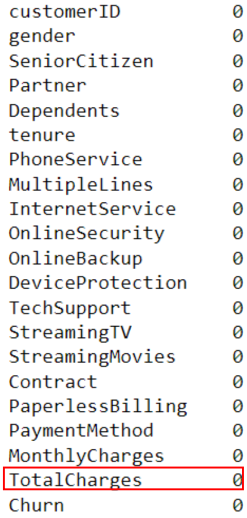
Fig. 4 - Missing values



	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechS
488	4472-LVYGI	Female	0	Yes	Yes	0	No	No phone service	DSL	Yes	...	Yes	
753	3115-CZMZD	Male	0	No	Yes	0	Yes	No	No	No internet service	...	No internet service	No i
936	5709-LVOEQ	Female	0	Yes	Yes	0	Yes	No	DSL	Yes	...	Yes	
1082	4367-NUYAO	Male	0	Yes	Yes	0	Yes	Yes	No	No internet service	...	No internet service	No i
1340	1371-DWPAZ	Female	0	Yes	Yes	0	No	No phone service	DSL	Yes	...	Yes	
3331	7644-OMVMY	Male	0	Yes	Yes	0	Yes	No	No	No internet service	...	No internet service	No i
3826	3213-VVOLG	Male	0	Yes	Yes	0	Yes	Yes	No	No internet service	...	No internet service	No i
4380	2520-SGTTA	Female	0	Yes	Yes	0	Yes	No	No	No internet service	...	No internet service	No i
5218	2923-ARZLG	Male	0	Yes	Yes	0	Yes	No	No	No internet service	...	No internet service	No i
6670	4075-WKNIU	Female	0	Yes	Yes	0	Yes	Yes	DSL	No	...	Yes	
6754	2775-SEFEE	Male	0	No	Yes	0	Yes	Yes	DSL	Yes	...	No	

Fig. 5 - Empty values in tenure columns

It can also find that the “Tenure” column contains an empty value along with the 11 missing values. Hence, a more detailed lookup needs to be done on the “Tenure” column to see if there are any other empty values. By using *df[df['tenure'] ==0]. index*, the result shows that there are not any other empty values in the “Tenure” column except the above 11 rows. It is impossible to have customers that have 0 tenure because this indicates they never stay with the company. Therefore, the next step is to remove the 11 missing values by using the *df.dropna (inplace=True)* function.

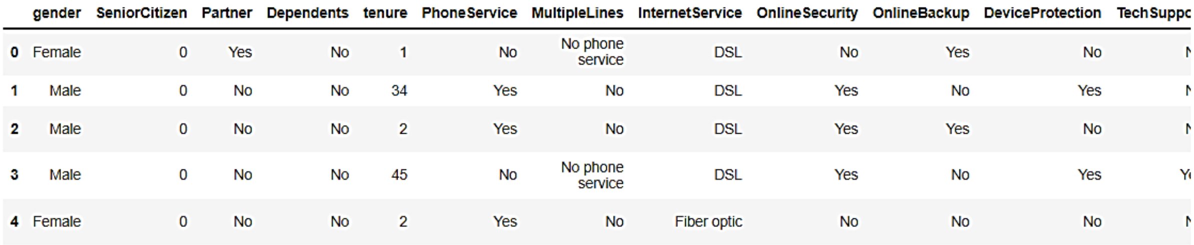


```
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges        0
Churn               0
```

Fig. 6 - After missing values were dropped

ii. Dropping Column

Since “Customer ID” can’t be use in predicting customer churn, it was dropped from the dataset using Pandas’*.drop()* method.



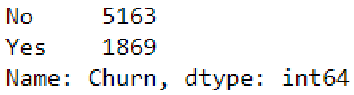
	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSuppc
0	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	N
1	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	N
2	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	N
3	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Y
4	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	N

Fig. 7 - Column dropped

3.6.2 Feature Selection

Feature selection is an essential part of the data pre-processing process since it increases the speed of machine learning algorithm training and simplifies the model’s creation and interpretation. In this stage, the important features are chosen, and the irrelevant ones are removed. The dataset was reduced to 20 columns after the column “Customer ID” was removed, indicating that it still has many features that will be used to train the model. Hence, it is unclear which features will have the most influence. Due to there being both categorical and numerical variables in the dataset, the method of using Random Forest to calculate feature importance will be used for feature selection.

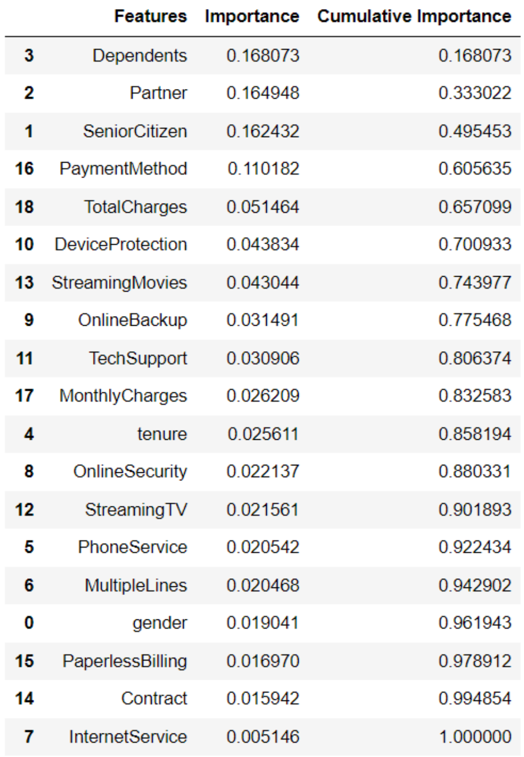
First, an instance of “features” was assigned by a list of features from the dataset. Next, the method of *ColumnTransformer* was used to convert our data into a format that machine learning can accept and use it. In the transformation, the *OrdinalEncoder* was used to transform categorical features while the *MinMaxScaler* was used to transform numerical features. Then, an instance of the Random Forest classifier, *“rf_fi”* was created. After that, a pipeline method was used to automate the process of transformation. However, by using *df[‘Churn’].value_counts()*, the dataset confronts the issue of unbalanced data. The output reveals that there are much more customers who do not churn (5163) than those who do (1869).



```
No      5163
Yes     1869
Name: Churn, dtype: int64
```

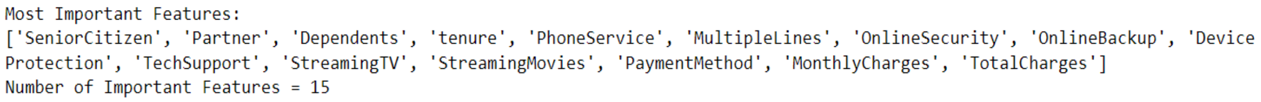
Fig. 8 - Customer churn counts

Hence, the settings of *class_weight = 'balanced'* to the classifier can help to deal with imbalanced data. Then, 30% of the dataset will be used as test sets, while 70% will be used for training. The machine learning model will be fitted using the training data. After that, the output of the classifier was checked using the *feature_importances* method. Moving on, the importance and cumulative importance of features were displayed as shown in figure 9. A loop is designed to cut off any feature that contributes less than 0.02 to the overall importance.



	Features	Importance	Cumulative Importance
3	Dependents	0.168073	0.168073
2	Partner	0.164948	0.333022
1	SeniorCitizen	0.162432	0.495453
16	PaymentMethod	0.110182	0.605635
18	TotalCharges	0.051464	0.657099
10	DeviceProtection	0.043834	0.700933
13	StreamingMovies	0.043044	0.743977
9	OnlineBackup	0.031491	0.775468
11	TechSupport	0.030906	0.806374
17	MonthlyCharges	0.026209	0.832583
4	tenure	0.025611	0.858194
8	OnlineSecurity	0.022137	0.880331
12	StreamingTV	0.021561	0.901893
5	PhoneService	0.020542	0.922434
6	MultipleLines	0.020468	0.942902
0	gender	0.019041	0.961943
15	PaperlessBilling	0.016970	0.978912
14	Contract	0.015942	0.994854
7	InternetService	0.005146	1.000000

Fig. 9 - Feature selection (i)



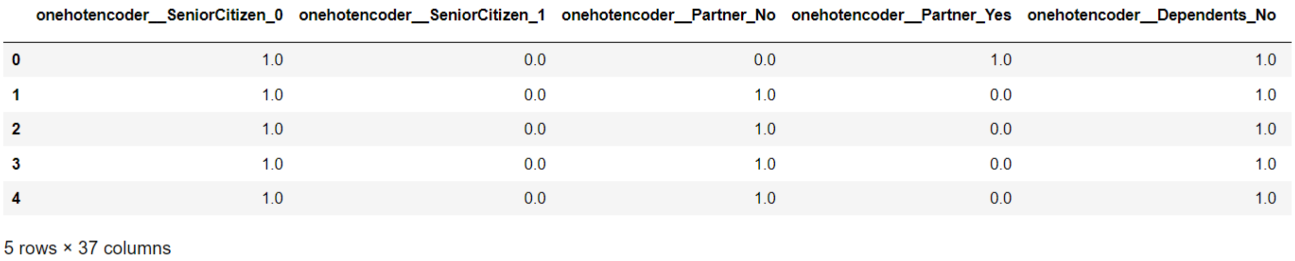
```
Most Important Features:
['SeniorCitizen', 'Partner', 'Dependents', 'tenure', 'PhoneService', 'MultipleLines', 'OnlineSecurity', 'OnlineBackup', 'Device
Protection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'PaymentMethod', 'MonthlyCharges', 'TotalCharges']
Number of Important Features = 15
```

Fig. 10 - Feature selection (ii)

Finally, based on the loop, the features that were selected from the loop function were printed by using the print function. This process has selected the 15 most important features out of 19 features that can be used to fit in our machine learning model later. As shown in figure 9 and 10, the 15 important features are "Senior Citizen", "Partner", "Dependents", "Tenure", "Phone Service", "Multiple Lines", "Online Security", "Online Backup", "Device Protection", "Tech Support", "Streaming TV", "Streaming Movies", "Payment Method", "Monthly Charges" and "Total Charges" as they contribute the importance that over 0.02. In addition, these 15 features together account for approximately 94% of the total importance—almost 100%.

3.6.3 One-Hot Encoding

Categorical data cannot be directly processed by machine learning algorithms. Hence, one-hot encoding is needed in encoding categorical data into binary form, which is 0 and 1. To begin with, the. drop *()* function from Pandas was used to drop the columns that are not selected in the previous step. The. replace *()* function from Pandas was used next to replace the data in the target column "Churn" with binary form (Yes =1, No=0). One-hot encoding is not used at the target's column because it will create problems while specifying X and Y since it expands the column. After that, the categorical columns (except column “Churn”) were encoded and transformed by using *OneHotEncoder* and *make_column_transformer*, while the other columns were skipped by using the *remainder = ‘passthrough’* setting.



	onehotencoder__SeniorCitizen_0	onehotencoder__SeniorCitizen_1	onehotencoder__Partner_No	onehotencoder__Partner_Yes	onehotencoder__Dependents_No
0	1.0	0.0	0.0	1.0	1.0
1	1.0	0.0	1.0	0.0	1.0
2	1.0	0.0	1.0	0.0	1.0
3	1.0	0.0	1.0	0.0	1.0
4	1.0	0.0	1.0	0.0	1.0

5 rows × 37 columns

Fig. 11 - Columns encoded by OHE

3.6.4 Data Splitting

First, the X and Y of the dataset were specified before splitting the dataset into training and test dataset. X represents the independent variable, whereas Y represents the dependent variables. By using the. *drop* function from Pandas, the column of “Churn” will be dropped from the features when specifying the Y while only the “Churn” column will be used for X. Next, the dataset is split into two subsets: the training set in training machine learning models, and the test set, which is used to measure how well the trained model performed with the ratio is chosen to be 70:30 by using the *train_test_split* function. Other than the *random_state* set as 0, the target proportion in the train and test datasets are kept the same as it was in the original dataset by setting the parameter *stratify=Y*.

3.6.5 Feature Scaling

The dataset was scaled as necessary for training the KNN machine learning model as the final step in the data pre-processing process by using the *StandardScaler()* function. The X_train set was fitted and transformed using the *sc.fit.transform()* method, while the x_test set was transformed using the *sc.transform()* method.

3.7 Model Training

In this study, there are 6 classifiers (KNN, RF, AdaBoost, LR, XGBoost and SVM) will be used to predict the telecom customer churn. The parameters applied to each of the classifiers in this study will be listed in more detail in the following section.

3.7.1 KNN

To begin with, the KNeighborsClassifier function was imported from the sklearn.neighbors so that the classifier can be fitted to the training set. Next, a loop between the range of 1-21 was created to determine the best *k* values for KNN. As shown in figure 12, the best *k* value for KNN to get the best accuracy will be *k*=20. Hence, the k value that is assigned to the parameter *n_neighbors* will be *k*=20.

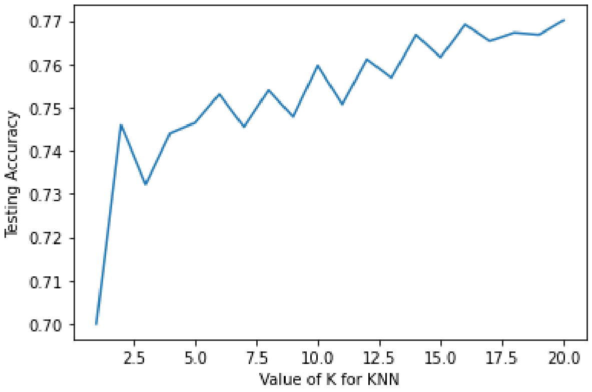


Fig.12 - Graph value of K for KNN

3.7.2 Support Vector Machine (SVM)

The Support Vector Machine was fitted to the training test using the *SVC* function, which was imported from the *sklearn.svm* library. The kernel parameter was a polynomial kernel, represented as *kernel = 'poly'*, with a default setting of *degree = 3*.

3.7.3 Random Forest Classifier (RF)

In fitting the Random Forest classifier to the training test, the *RandomForestClassifier* function was imported from the *sklearn.ensemble* library. 6 parameters were used to increase the speed and predictability of the classifier and can be viewed in the table below.

Table 3 - Parameters of RF

Parameters	Reason
n_estimator = 1,000	It represents how many trees the Random Forest needs. Here, the *n_estimators* are set as 1,000 because a higher number might result in better accuracy.
random_state = 50	The random state of the model was set to 50.
max_features = 'sqrt'	The total number of features for each run is calculated as its square root using the *'sqrt'* parameter.
oob_score = True	The model was validated using *oob_score* since it reduces leakage and produces a better model with lower variance.
n_jobs = -1	This parameter specifies the engine how many processors it can use, and the value of -1 is set to indicate there are no restrictions to it.
max_leaf_nodes = 30	30 was set on the splitting of nodes to prevent the model from underfitting or overfitting.

3.7.4 Logistic Regression (LR)

The *LogisticRegression* function was imported from *sklearn.linear* model library and used to fit the AdaBoost classifier to the training test. No parameter was used in this classifier other than *random_state=0*.

3.7.5 XGBoost

The *XGBClassifier* function was imported from the *xgboost* in order to fit the XGBoost classifier to the training test. Only two parameters were used in this classifier. The *learning_rate* was set to 0.3 by default to decrease the feature weights and make the boosting procedure more cautious. To restrict the depth of the tree, a value of 3 was applied to the *max_depth* parameter.

3.7.6 AdaBoost

In fitting the AdaBoost classifier to the training test, the *AdaBoostClassifier* function was imported from *sklearn.ensemble* library. For both the *n_estimators* and *learning_rate* parameters, the default values were set to 50 and 1, respectively.

3.8 Model Training

There are 4 performance metrics that based on the confusion matrix will be used in evaluating the performance of the classifiers in this study. The results of a machine learning model's prediction are summarized in a confusion matrix as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Observations that were accurately predicted (both true positives and true negatives) and observations that were wrongly predicted (both false positives and false negatives) can both be determined from the confusion matrix. Each of the metrics will be explained in the section that follows.

Table 4 - Confusion matrix

		Predicted Class	
		Customer that Churn (1)	**Customers That No Churn (0)**
Actual Condition	**Customer that Churn (1)**	True Positive (TP)	False Negative (FN)
	Customers That No Churn (0)	False Positive (FP)	True Negative (TN)

3.8.1 Classification Report

i. Accuracy

Accuracy is one of the performance metrics by displaying its results as a percentage and indicates whether the test data were correctly predicted. It is the measure of how closely a measurement matches the actual or true value. In this study, accuracy is the percentage of all predictions that were accurate.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \times 100\% \tag{2}$$

ii. Precision

Precision measures how well a model can predict a certain category. Precision gives the percentage of true positives. It determines the percentage of positively predicted samples that were accurately identified. It also describes how closely measurements of the same object match one another. In this study, precision measures how many of the predicted churns per the model the customer actually churns.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \times 100\% \tag{3}$$

iii. Recall

A recall is a metric used to determine the percentage of correct positive predictions among all potential positive predictions. In other words, it is the proportion of true positives to all true positives. The percentage of accurately detected positive samples among all positive samples was calculated in the recall. Contrary to precision, which only makes observations about the accurate positive predictions among all positive predictions, recall provides a measure of missed positive predictions. In this study, recall is the percentage of customers the model properly predicts as being likely going to churn of all those actually churn.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \times 100\% \tag{4}$$

iv. F1-Score

The performance of binary classification is frequently evaluated using the F1-score for both balanced and imbalanced datasets. Combining precision and recall, it summarizes the prediction performance of a model. A high F1-score indicates few false positives and few false negatives. F1-score can be calculated by using the equation below.

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})} \tag{5}$$

4. Result and Discussion

4.1 The Result of Exploratory Data Analysis Regarding the Characteristics of Churned Customers

The characteristics of the churned customer will be explored and visualized in a graph in Python with the use of Seaborn and matplotlib. To begin with, the dataset with only churn customers is being selected by implementing *df[df[‘Churn’]==‘Yes’]* for most of the parts in the exploratory data analysis section.

4.1.1 Customer Churn by Demographic Characteristics

Gender, Senior Citizens or not, presence of Dependents or Partner are included in the demographic characteristics of churn customers.

i. Churn Customers by Gender

The gender distribution of churn customers is nearly uniform, as seen by the pie chart. Male customers (50.2%) churn 0.4% more frequently than female customers (49.8%).

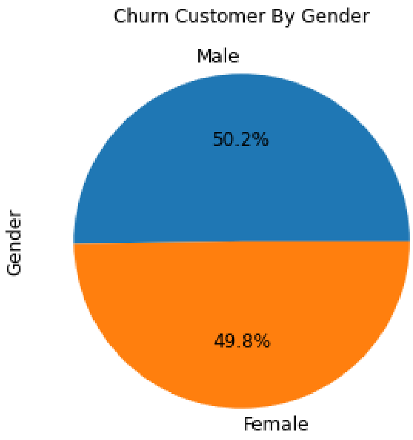


Fig. 13 - Churn customer by gender

ii. Churn Customers by Senior Citizen

There are only 25.5% of churn customers are senior citizens while 74.5% of the rest are younger churn customers, as seen from the pie chart. This shows that younger customers make up the majority of customers that churn.

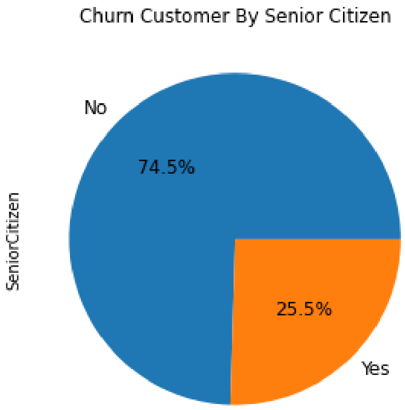


Fig. 14 - Churn customer by senior citizen

iii. Churn Customers by Dependents

According to figure 15, customers that churn usually do not have dependents.

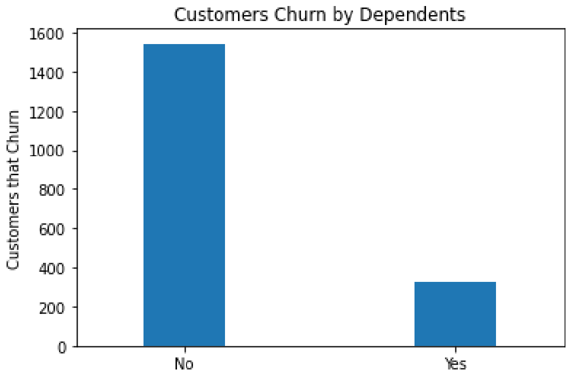


Fig. 15 - Churn customer by dependents

iv. Churn Customers by Partner

The EDA also discovers that most of the customers that churn does not have a partner.

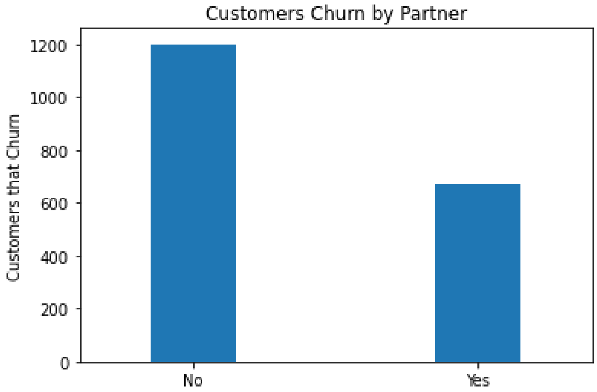


Fig. 16 - Churn customer by partner

4.1.2 Customer Churn by Customer Account Details Characteristics

Information on the customer's account includes the Contract, services offered by the telecom company, Tenure, Monthly Charges, and Total Charges.

i. Churn Customers by Contract

Customer churn is more likely to occur with month-to-month contracts than with one and two-year contracts, as shown in figure 17.

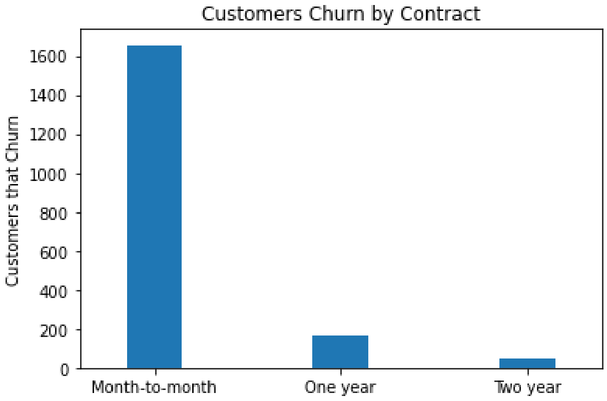


Fig. 17 - Churn customer by contract

ii. Churn Customers by Telecom Services

Most of the churn customers, as shown in the bar charts, use Phone Services, Multiple Lines services (with just one additional customer than the customer who does not use multiple lines services), and fiber optic Internet Service options. Furthermore, most churn customers do not use services like Online Security, Tech Support, Online Backup, Streaming Tv, Device Protection and Streaming Movies services provided by the telecom company.

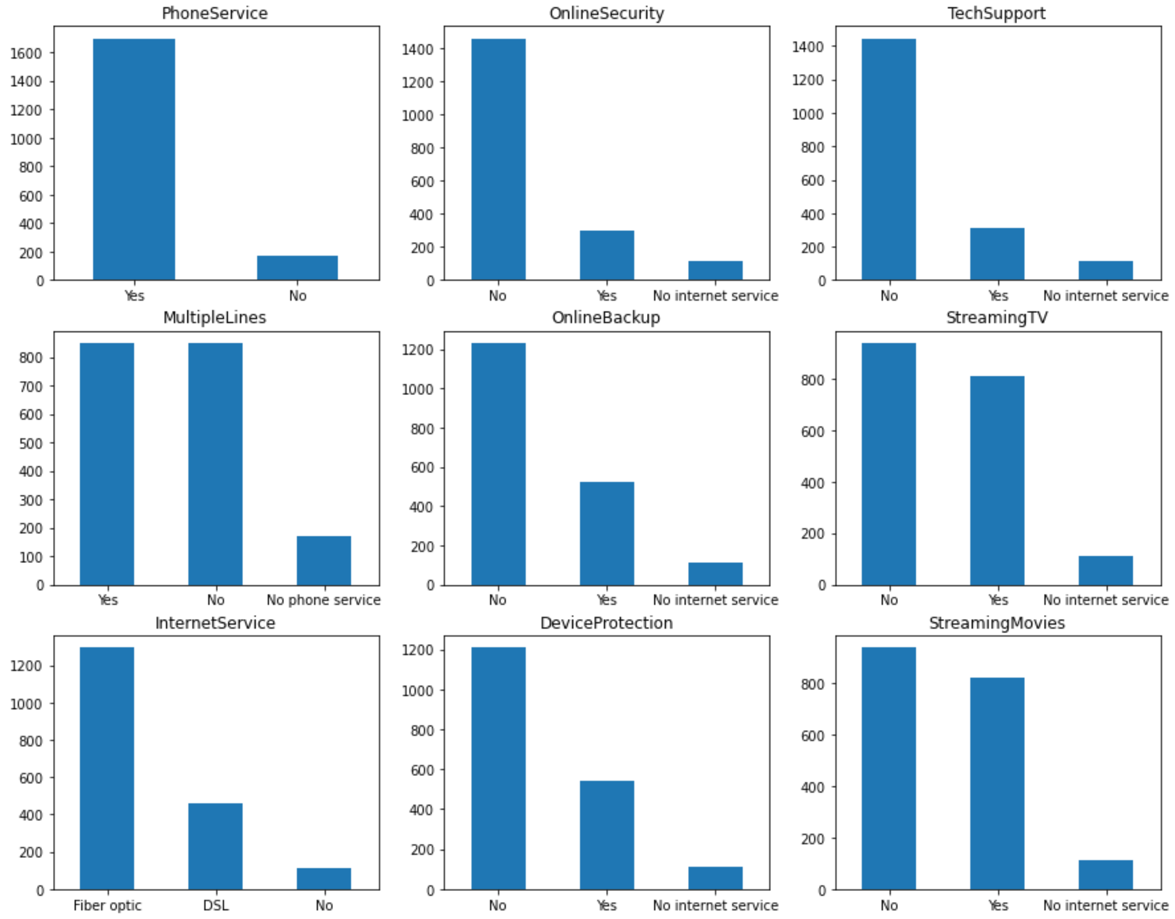


Fig. 18 - Churn customer by telecom services

iii. Churn Customers by Their Tenure

Figure 19 shows that many of the telecom company's customers who churn had just been customers for a month. More particularly, customers that churn are more likely to do so before 20 months of tenure. This might be because various clients have different contracts. As a result, depending on the contract, it may be simpler or harder for the customer to churn.

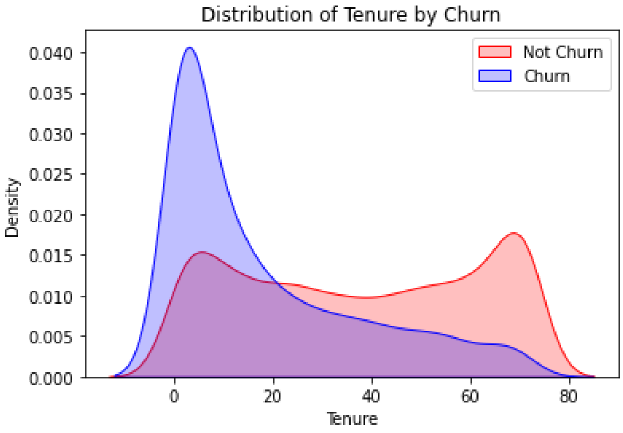


Fig. 19 - Churn customer by their tenure

iv. Churn Customers by Monthly Charges

The kernel density estimate (KDE) plot above shows the distribution of the monthly charges of churn customers. When customers' monthly charges increase, they are more likely to churn. More specifically, it is most common for customers to churn when their monthly charges exceed $60.

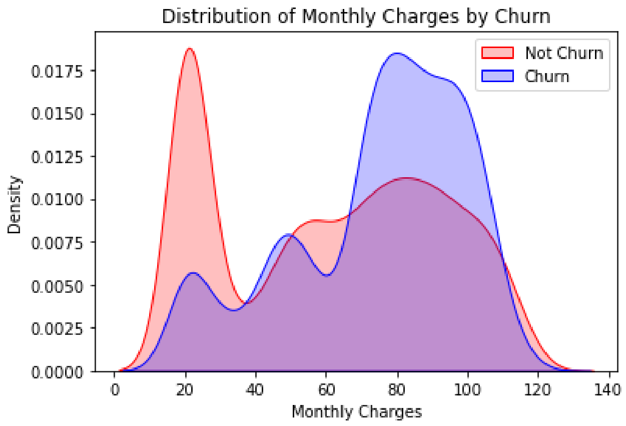


Fig. 20 - Churn customer by monthly charges

v. Churn Customers by Total Charges

As observed from figure 21, when the total charges are lower, the happen of customer churn is more likely.

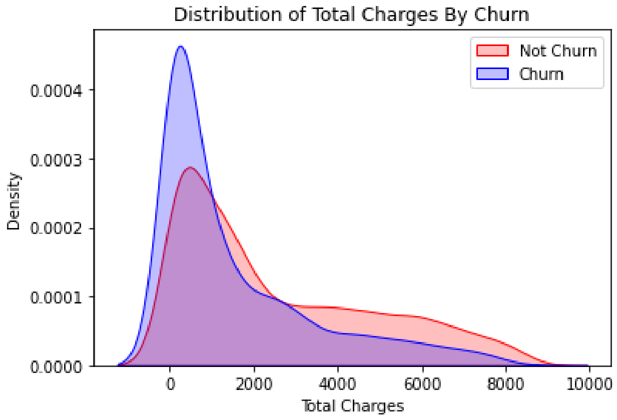


Fig. 21 - Churn customer by total charges

4.2 The Result of Comparison Performance Metrics

4.2.1 Accuracy

Figure 22 shows a bar chart that compares and displays the accuracy of each classifier. The accuracy ranged from 76% to 79% for all classifiers. As shown in figure 22, with XGBoost achieved the best accuracy (79.7%), followed by AdaBoost (79.5%), LR (79.2%), RF (79.1%), and SVM (78.1%). Meanwhile, the lowest accuracy of the 6 classifiers was achieved by KNN which achieved a 77% accuracy rate.

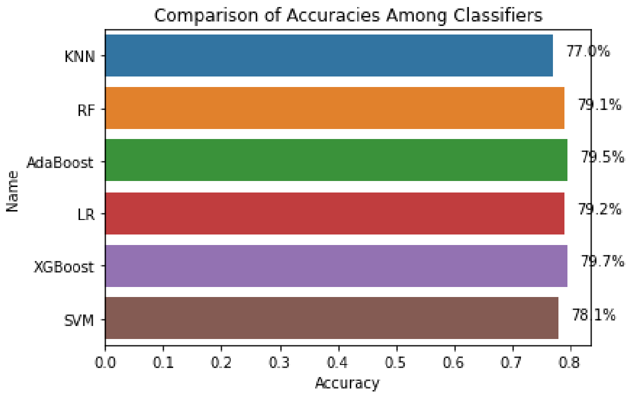


Fig. 22 - Comparison of accuracies

4.2.2 Classification Report

Table 5 shows the summary table of classification reports for each classifier. Other than accuracy, precision, recall, and F1-score of predicted labels 1 (customer that churns) are the performance indicators in the classification report that are used to assess the performance of classifiers. The classification report's support value of each classifier, where label 0 (No) is 1,549 and label 1 (Yes) is 561, is the same. Thus, this also highlights the problem of an imbalanced dataset.

Table 5 - Summary Results from Classification Report

	KNN (%)	RF (%)	AdaBoost (%)	LR (%)	XGBoost (%)	SVM (%)
Precision	59.95	**67.15**	64.97	63.80	64.67	64.08
Recall	40.82	41.53	49.91	50.27	**51.87**	39.75
F1-score	48.57	51.32	56.45	56.23	**57.57**	49.07
Accuracy	77.01	79.05	79.53	79.19	**79.67**	78.06
Average	56.59	59.76	62.72	62.37	**63.45**	57.74

First, the RF had the best precision (67.15%), followed by AdaBoost (64.97%), XGBoost (64.67%), SVM (64.08%), LR (63.80%) and KNN (59.95%), which had the lowest precision for predicted label 1 (presence of customer churn in the telecom company). In terms of how well the classifiers performed in precision, RF outperformed the other 5 classifiers with a precision of 67.15%.

Next, the recall score of 51.87% for XGBoost was the highest and followed by LR (50.27%), AdaBoost (49.91%), RF (41.53%), KNN (40.82%) and SVM (39.75%). Lastly, the F1-score for XGBoost was 57.57%, followed by F1-scores for AdaBoost at 56.45%, LR at 56.23%, RF at 51.32% and SVM at 49.07%. Meanwhile, the F1-score for KNN of 48.57% was the lowest.

The overall best classifier for predicting customer churn in the telecom company was determined by averaging the score of precision, recall, F1-score, and accuracy for each classifier. The table shows that XGBoost had the highest average score of 63.45% and is the overall most effective classifier in this study, whereas KNN received the lowest average score of 56.59%.

4.3 Discussion

The characteristics of churn customers from the telecom company in California have been discovered after exploratory data analysis. First, the demographic characteristics of customers who churned were examined. Since the gender distribution of churn customers was almost equal, there was no trend indicating which gender will be more likely to leave the telecom company. In terms of age, younger customers tend to churn more than elderly customers. Additionally, most customers that churn do not have dependents or partner.

Moving on to the next section regarding the account details of churn customers. Customer churn happens more often for customers with month-to-month contracts. Furthermore, the majority of the churn customers use phone services, multiple lines services and fiber optic internet service options. This result is in line with earlier research by [35], which discovered that monthly contracts and fiber optic internet service are two of the top three factors that lead to customer churn. To add up, most of the churn customers also do not use services like online security, tech support, online backup, streaming tv, device protection and streaming movies provided by the telecom company. Next, the tenure for customers that churn is usually for a month, and customer churn usually happens before 20 months. This might be because of numerous customers having various contracts. As a result, depending on the contract, it may be simpler or harder for customers to churn. Furthermore, customers are more likely to leave when their monthly charges rise. Specifically, customers are most likely to churn when their monthly charges exceed $60. This finding is consistent with prior research done by [36], which found that churn was more strongly influenced by fiber optic customers who pay higher monthly fees. Lastly, customer churn is more likely to happen when the total charges are lower.

Next, all 6 classifiers built in this study can be used to predict the customer churn of telecom companies after being evaluated by the 4 performance metrics. First, the best accuracy of the prediction model was achieved by using XGBoost (79.67%), then followed by AdaBoost (79.53%), LR (79.19%), RF (79.05%), SVM (78.06%) and KNN (77.01%). This result is consistent with prior research by [36] where a similar accuracy score was achieved in using KNN, RF, and XGBoost to predict customer churn in a telecom company in achieving an accuracy of 75.40%, 77.50% and 79.80% respectively. Regarding how well the classifiers performed in precision, RF outperformed the other 5 classifiers with a precision of 67.15%. Contrary to what [37] discovered, they were able to achieve the highest precision of 80.14% by using AdaBoost rather than RF. In addition, a better accuracy (77.54%) was also achieved by the authors using RF. The obtained result also contrasted with the study from [39], which found that KNN performed best in terms of accuracy and precision, but KNN performed poorly in terms of both metrics in this study.

Then, the highest recall score was obtained by XGBoost (51.87%) in this study. However, this contrasts with the study of [37], which found that AdaBoost had the highest recall score (81.21%). The recall score obtained by XGBoost differs significantly from the recall scores obtained by this study (51.87%) and from the authors (80.7%). Besides that, the study found that the F1-scores obtained in this study for KNN (48.57%), RF (51.32%), and XGBoost (57.57%) were similarly close to the F1-scores obtained by [36], in which KNN scored 49.50%, RF scored 50.60%, and XGBoost scored 58.20%. After averaging the accuracy, precision, recall, and F1-score values revealed that XGBoost was the overall best classifier in predicting customer churn of the telecom company in this study. It received a 63.45% average score, which is the highest of the 6 classifiers.

Finally, this study runs into an issue with an imbalanced dataset, where there are more customers who remain with the telecom company than there are customers who leave, with the amount of 5163 for customers who do not churn and 1869 for customers that churn. This is the same as the research done by [40] and [41]. Both authors met the issue of an unbalanced dataset where the distribution of churn and non-churn customers are uneven. This is a major issue with churn prediction because an imbalance in the target class may have a substantial negative effect on the final models.

5. Recommendations and Conclusion

5.1 Potential Recommendations to the Company

This study also seeks to offer some feasible suggestions to the company in addition to developing machine learning models to predict customer churn in a telecom company in California. Thus, this section will provide some recommendations that can be made to the telecom company to retain customers.

Firstly, the telecom company can create personalized experiences for its customers. To accomplish the goal of customer retention, the company can issue a telecom strategy towards the characteristics of the customers who are more likely to churn as they are in the high-risk category of churning. For instance, the company shall develop a solution to situations where customers tend to churn when their monthly charges exceed $60 or before 20 months of tenure. Besides that, the telecom company can segment its customer base into several meaningful groups in order to provide various segments with different personalized services. The segmentation of customers into customer value, customer behaviour, customer life cycle and customer migration as segmentation techniques were considered in the study of [42]. This provides precise targeting as well as a grasp of each customer's retention and value drivers. The study of [42] showed a well-segmented customer base is effectively used to increase customer growth and retention. In this study, it was found that 15 features are contributing more than 0.02 importance in predicting customer churn. Hence, management can take note of these features of customers and segment them effectively.

Secondly, a customer loyalty program can help the telecom company retain its customers. Customers are encouraged to remain with the company by participating in customer loyalty programs, which are organized incentive schemes. For instance, the company can offer monthly incentives like gift cards to retain the customers with the month-to-month contract since they are more likely to churn. The study by [43] demonstrates increasing customer retention levels in the mobile telecom sector results from customers being satisfied with the loyalty program. In other words, customer relationship satisfaction is influenced significantly by loyalty program satisfaction.

Last but not least, the telecom company can adopt artificial intelligence (AI) in managing customer relationships to retain customers. The ability of the company to choose which customers to invest in is one of the implications of using AI-CRM [44]. Personalization provided by AI-CRM allows companies to target more customers which will provide the company with a strategic advantage, saving the company money from investing in the incorrect customer base. Hence, the telecom company can use AI-CRM to retain customers.

In short, to retain customers of the telecom company is advised that the management of the telecom company personalize the customer experience, implement a customer loyalty program, and apply AI in customer relationship management to retain customers.

5.2 Limitations

Each study has potential limitations that need to be noted and acknowledged. Firstly, the imbalanced dataset that was used to train the machine learning model in predicting customer churn, with a ratio of 100:36 between the majority class of customers (those who stay) and the minority class of customers (those who churn), was a limiting factor in this study. This makes it difficult for the models to learn from the minority class, which makes it harder to predict. Additionally, the outcome could potentially be skewed in favour of the majority class.

Additionally, another factor that restricted this study was the used dataset only includes some features of churn customers in the telecom company, which is very limited. There are also plenty of additional features customers of the telecom company may have left out and which might contribute to customer churn. Thus, the results cannot be used to generalize other datasets in various situations.

Lastly, the findings of this study were limited as only 6 machine learning algorithms are being selected for developing models. However, there are many more machine learning algorithms available that could be used to create models and may perform better.

6. Conclusion

This study aims to predict customer churn of a telecom company in California by developing a prediction model using the 6 selected supervised machine learning algorithms, which are KNN, Random Forest Classifier (RF), AdaBoost, Logistic Regression (LR), XGBoost and Support Vector Machine (SVM). Furthermore, this study also seeks to identify the most accurate or effective model for predicting customer churn, identify the characteristic of customers who are most likely to leave the telecom company, and offer potential recommendations to the company for customer retention. This study found that the demographic characteristics of churn customers are younger customers and customers that are without dependents or partner. As for the customer account details, the study discovered that customer churn happens more during month-to-month contracts. Besides that, most of the churn customers use phone services, multiple lines services and fiber optic internet service options. It was also discovered that most of the customers that churn doesn't use services like online security, tech support, online backup, streaming tv, device protection and streaming movies services provided by the telecom company. Then, the tenure customers that churn are usually customers for a month and customer churn mostly happens before 20 months of tenure. In addition, customers are more likely to churn when their monthly charges rise, specifically when exceeding $60. Customer churn is also more likely to happen when the total charges are lower. Finally, the results indicate that XGBoost is the overall most effective classifier in this study, followed by AdaBoost, LR, RF, SVM, and KNN. The overall performance of the 6 classifiers is evaluated using the average score of all performance measures. Some potential recommendations that can be taken by the management of the telecom company include personalizing the customer experience, implementing a customer loyalty program, and applying AI in customer relationship management for customer retention. Study limitation of this study was also addressed and needed to improve for the continuance of research.

Acknowledgement

Special thanks were extended to the School of Management, Universiti Sains Malaysia.

References

[1] National Research Council, Sciences, D. E. P., Board, C. S. T., Development, C. T. R., Council, N. R., Eisenberg, J., & Lucky, R. W. (2006). *Renewing U.S. Telecommunications Research*. Amsterdam University Press.

[2] World Health Organization (WHO). (2020, March 11). *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020*. World Health Organization. Retrieved November 15, 2022, from https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

[3] Statista. (2022a, June). *Telecommunications - California | Statista Market Forecast*. Retrieved November 15, 2022, from https://www.statista.com/outlook/io/information-communication/telecommunications/united-states/california

[4] Berson, A., Smith, S., & Thearling, K. (2000). *Building Data Mining Applications for CRM*. McGraw-Hill Education.

[5] Statista. (2022b, June 13). *Churn rate of Vodafone in Germany Q1 2018/19 - Q4 2021/2022*. Retrieved November 1, 2022, from https://www.statista.com/statistics/483007/vodafone-churn-rate-germany/

[6] Bin, L., Peiji, S., & Juan, L. (2007). Customer Churn Prediction Based on the Decision Tree in Personal Handyphone System Service. *2007 International Conference on Service Systems and Service Management*. https://doi.org/10.1109/icsssm.2007.4280145

[7] Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems With Applications*, *39*(1), 1414–1425. https://doi.org/10.1016/j.eswa.2011.08.024

[8] Özdemir, O., Batar, M., & Işık, A. H. (2020). Churn Analysis with Machine Learning Classification Algorithms in Python. *Artificial Intelligence and Applied Mathematics in Engineering Problems*, 844–852. https://doi.org/10.1007/978-3-030-36178-5_73

[9] Van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research*, *157*(1), 196–217. https://doi.org/10.1016/s0377-2217(03)00069-9

[10] Taunk, K., De, S., Verma, S., & Swetapadma, A. (2019). A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. https://doi.org/10.1109/iccs45141.2019.9065747

[11] Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, *1142*, 012012. https://doi.org/10.1088/1742-6596/1142/1/012012

[12] De Leonardis, G., Rosati, S., Balestra, G., Agostini, V., Panero, E., Gastaldi, L., & Knaflitz, M. (2018). Human Activity Recognition by Wearable Sensors : Comparison of different classifiers for real-time applications. *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. https://doi.org/10.1109/memea.2018.8438750

[13] Xu, B., Chen, S., Zhang, H., & Wu, T. (2017). Incremental k-NN SVM method in intrusion detection. *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. https://doi.org/10.1109/icsess.2017.8343013

[14] Carvalho, T. P., Soares, F. A. A. M. N., Vita, R., Francisco, R. D. P., Basto, J. P., & Alcalá, S. G. S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers &Amp; Industrial Engineering*, *137*, 106024. https://doi.org/10.1016/j.cie.2019.106024

[15] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215. https://doi.org/10.1016/j.neucom.2019.10.118

[16] Tao, Z., Huiling, L., Wenwen, W., & Xia, Y. (2019). GA-SVM based feature selection and parameter optimization in hospitalization expense modeling. *Applied Soft Computing*, *75*, 323–332. https://doi.org/10.1016/j.asoc.2018.11.001

[17] Ahmad, M., Aftab, S., Muhammad, & S. S. (2017). Machine Learning Techniques for Sentiment Analysis: A Review. *International Journal of Multidisciplinary Sciences and Engineering*, *8*(3), 27.

[18] Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

[19] Nazarenko, E., Varkentin, V., & Polyakova, T. (2019). Features of Application of Machine Learning Methods for Classification of Network Traffic (Features, Advantages, Disadvantages). *2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon)*. https://doi.org/10.1109/fareastcon.2019.8934236

[20] Ziegler, A., & König, I. R. (2013). Mining data with random forests: current options for real-world applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*(1), 55–63. https://doi.org/10.1002/widm.1114

[21] Farnaaz, N., & Jabbar, M. (2016). Random Forest Modeling for Network Intrusion Detection System. *Procedia Computer Science*, *89*, 213–217. https://doi.org/10.1016/j.procs.2016.06.047

[22] Fritz, M., & Berger, P. D. (2015). Will anybody buy? Logistic regression. *Improving the User Experience Through Practical Data Analytics*, 271–304. https://doi.org/10.1016/b978-0-12-800635-1.00011-2

[23] 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)*. https://doi.org/10.1109/cdan.2016.7570883

[24] Artetxe, A., Beristain, A., & Graña, M. (2018). Predictive models for hospital readmission risk: A systematic review of methods. *Computer Methods and Programs in Biomedicine*, *164*, 49–64. https://doi.org/10.1016/j.cmpb.2018.06.006

[25] Park, H. A. (2013). An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. *Journal of Korean Academy of Nursing*, *43*(2), 154. https://doi.org/10.4040/jkan.2013.43.2.154

[26] Nusinovici, S., Tham, Y. C., Chak Yan, M. Y., Wei Ting, D. S., Li, J., Sabanayagam, C., Wong, T. Y., & Cheng, C. Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*, *122*, 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002

[27] Liu, Y., Esan, O. C., Pan, Z., & An, L. (2021). Machine learning for advanced energy materials. *Energy and AI*, *3*, 100049. https://doi.org/10.1016/j.egyai.2021.100049

[28] Wade, C., & Glynn, K. (2020). *Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python*. Packt Publishing.

[29] Zhao, W. P., Li, J., Zhao, J., Zhao, D., Lu, J., & Wang, X. (2020). XGB Model : Research on Evaporation Duct Height Prediction Based on XGBoost Algorithm. *Radioengineering*, *29*(1), 81–93. https://doi.org/10.13164/re.2020.0081

[30] Peng, Z., Huang, Q., & Han, Y. (2019). Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm. *2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT)*. https://doi.org/10.1109/icait.2019.8935894

[31] Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785

[32] Chen, M., Liu, Q., Chen, S., Liu, Y., Zhang, C. H., & Liu, R. (2019). XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System. *IEEE Access*, *7*, 13149–13158. https://doi.org/10.1109/access.2019.2893448

[33] Zhang, Y., Ni, M., Zhang, C., Liang, S., Fang, S., Li, R., & Tan, Z. (2019). Research and Application of AdaBoost Algorithm Based on SVM. *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*. https://doi.org/10.1109/itaic.2019.8785556

[34] Liu, M. (2010). Fingerprint classification based on Adaboost learning from singularity features. *Pattern Recognition*, *43*(3), 1062–1070. https://doi.org/10.1016/j.patcog.2009.08.011

[35] Mohammad, N. I., Ismail, S. A., Kama, M. N., Yusop, O. M., & Azmi, A. (2019). Customer Churn Prediction In Telecommunication Industry Using Machine Learning Classifiers. *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing*. https://doi.org/10.1145/3387168.3387219

[36] Pamina, J., Beschi Raja, J., Sathya Bama, S., Soundarya, S., Sruthi, M. S., Kiruthika, S., Aiswaryadevi, V. J., & Priyanka, G. (2019). An Effective Classifier for Predicting Churn in Telecommunication. *Journal of Advanced Research in Dynamic and Control Systems*, *11*(01-Special Issue), 221–229.

[37] Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2021). Customer churn prediction system: a machine learning approach. *Computing*, *104*(2), 271–294. https://doi.org/10.1007/s00607-021-00908-y

[38] Geetha, V., Punitha, A., Nandhini, A., Nandhini, T., Shakila, S., & Sushmitha, R. (2020). Customer Churn Prediction In Telecommunication Industry Using Random Forest Classifier. *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*. https://doi.org/10.1109/icscan49426.2020.9262288

[39] Bhatnagar, A., & Srivastava, S. (2019). A Robust Model for Churn Prediction using Supervised Machine Learning. *2019 IEEE 9th International Conference on Advanced Computing (IACC)*. https://doi.org/10.1109/iacc48062.2019.8971494

[40] Hossain, M. M., & Miah, M. S. (2015). Evaluation of different SVM kernels for predicting customer churn. *2015 18th International Conference on Computer and Information Technology (ICCIT)*. https://doi.org/10.1109/iccitechn.2015.7488032

[41] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0191-6

[42] Bayer, J. (2010). Customer segmentation in the telecommunications industry. *Journal of Database Marketing &Amp; Customer Strategy Management*, *17*(3–4), 247–256. https://doi.org/10.1057/dbm.2010.21

[43] Ammari, N. B., & Bilgihan, A. (2019). Customer retention to mobile telecommunication service providers: the roles of perceived justice and customer loyalty program. *International Journal of Mobile Communications*, *17*(1), 82. https://doi.org/10.1504/ijmc.2019.096518

[44] Libai, B., Bart, Y., Gensler, S., Hofacker, C. F., Kaplan, A., Kötterheinrich, K., & Kroll, E. B. (2020). Brave New World? On AI and the Management of Customer Relationships. *Journal of Interactive Marketing*, *51*, 44–56. https://doi.org/10.1016/j.intmar.2020.04.002