Seminar Report

On

Prediction of customer churn using Machine learning.

By

Shubhendu Kulkarni

7182662F

Under the guidance
of

Prof. Shubhada Mone

DEPARTMENT OF COMPUTER ENGINEERING
Marathwada MitraMandal's College of Engineering
Karvenagar
Savitribai Phule Pune University
2019-2020

Marathwada Mitra Mandal's
College of Engineering
Karvenagar, Pune
Accredited with 'A' Grade by NAAC

CERTIFICATE

This is to certify that **Shubhendu Kulkarni** from **Third Year Computer Engineering** has success- fully completed his/her seminar work titled **"Prediction of customer churn using Machine learning"** at Marathwada Mitra Mandal's College of Engineering, Pune in the partial fulfillment of the Bachelors Degree in the Engineering.

Date:
Place:

Name of Guide

Guide | Head of the Department | Principal

Acknowledgment

I take this to express my deep sense of gratitude towards my esteemed guide Prof. Shubhada Mone for giving me this splendid opportunity to select and present this seminar and also providing facilities for successful completion.

I thank Dr. H. K. Khanuja, Head, Department of Computer Engineering, for opening the doors of the department towards the realization of the seminar, all the staff members, for their indispensable support, priceless suggestion and for most valuable time lent as and when required. With respect and gratitude, I would like to thank all the people, who have helped me directly or indirectly.

Shubhendu Kulkarni

Roll no.:136 Class:TE-1

Abstract

The purpose of this seminar to employ Machine Learning by creating a predictive model that predicts customer churn.

Machine learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks. Therefore, I decided to explore different Machine learning algorithms and techniques to incorporate in my customer churn predictive model.

Customer churn analysis and prediction in telecom sector is an issue now a days because it's very important for telecommunication industries to analyze behaviors of various customer to predict which customers are about to leave the subscription from telecom company. So data mining techniques and algorithm plays an important role for companies in todays commercial conditions because gaining a new customer's cost is more than retaining the existing ones.

The broad objective of this project is to create a customer churn prediction model using Telco Customer churn dataset. I used binary classification to model churned customers, pandas for data crunching and mat-plot-lib for visualizations. We will do all of that above in Python. The code can be used with another data-set with a few minor adjustments to train the baseline model.

This problem statement falls under the domain of Machine learning. Incorporating Machine learning techniques in this project was quite easy. I used logistic regression and binary classifier to train and test my model. The data-set I used was the Telco Customer Churn data-set that was available on Kaggle website.

I updated my guide Prof. Shubhada Mone regularly about the project progress and problems I was facing and she provided valuable advice and suggestions to help me. I was also helped by other faculty and staff members of the Computer department for the timely completion and implementation of this project.

In this way with the collaborative efforts of all the stakeholders, I was successful in incorporating Machine learning in my project.

Contents

1 Technical Keywords **1**
1.1 Domain Name 1
1.2 Technical Keywords 1

2 Introduction **2**
2.1 Domain Description 2
2.2 Problem Defifinition 2
2.3 Motivation 3

3 Literature Survey **4**
3.1 Existing Methods/Tools/Techniques 4
3.2 Literature Survey 6

4 Mathematical Model **7**

5 Proposed System Architecture **8**
5.1 System Architecture 8
5.2 Design with UML Diagrams 9
5.3 Algorithms 12
5.4 Implementation/Proof of Concept 13
5.5 Important Source Code 14
5.6 Testing Code 14
5.7 Result screenshot, tables and Analysis 14

6 Advantages / Disadvantages **23**

7 Applications **25**

8 Conclusion and future work **26**

9 Appendix **29**
9.1 Log Report 29
9.2 Internship letter(if any) 29
9.3 report of Internship Project(if any) 29
9.4 hard copy of screenshot of online internship feedback given after internship 29

List of Tables

3.2 Literature Survey 14

List of Figures

3.1 Logistic Regression model 13
4.1 Logistic Regression formula 15
4.2 Logistic Regression versus Linear Regression 16
5.1 System architecture diagram 17
5.2.1 Class diagram 18
5.2.2 Activity diagram 19
5.2.3 Use case diagram 20
5.2.4 Component diagram 21
5.3.1 Algorithm 22
5.7.1 Important modules 23
5.7.2 Reading the data-set 24
5.7.4 Churn rate by payment method 25
5.7.5 Churn rate by paperless billing, streaming movies, device protection, phone service 25
5.7.6 Monthly charge vs churn rate and Total charge vs churn rate 26
5.7.7 Tenure cluster 26
5.7.8 Tenure cluster vs churn rate 26
5.7.9 Monthly charge cluster 27
5.7.10 Monthly charge cluster vs churn rate 27
5.7.11 Total charge cluster 28
5.7.12 Total charge cluster vs churn rate 28
5.7.13 Linear model regression results 29
5.7.14 Linear model regression results 29
5.7.15 Linear model regression results 30
5.7.16 Linear model regression results 30
5.7.16 ROC curve 31
5.7.17 Accuracy of the model 32
5.7.18 Detailed analysis of accuracy of the model 32
5.7.19 Result customer churn probability of each customer in the data-set 33

Chapter-1 Technical Keywords

1.1 Domain Name

Machine learning.
Machine learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

1.2 Technical Keywords

1. Machine learning
2. Customer churn
3. Supervised learning
4. Unsupervised learning
5. Artificial Neural Networks
6. Support Vector Machines
7. Decision Trees
8. Naive Bayes
9. Logistic Regression Analysis

Chapter-2 Introduction

2.1 Domain Description

Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so. For the early tasks that humans wanted computers to accomplish, it was possible to create algorithms that enabled the machine to execute all the steps needed to solve the problem in hand. So on the computer's part, no learning was needed. For certain advanced tasks, facial recognition for example, it is not easy to create the needed algorithms, partly as it's not easy for humans to precisely define how we recognize faces. Abundant face related data exists however. So far, compared to the difficulty in directly creating the required algorithms, it's turned out in practice to be easier to assist computers to learn themselves how to recognize faces from available data. The discipline of machine learning develops various approaches for computers to learn to accomplish tasks for which no algorithm exists.
Early classifications for machine learning approaches sometimes divided them into three broad categories, depending on the nature of the "signal" or "feedback" available to the learning system. These were:
Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.
Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself.
In this project, I incorporated Machine learning techniques algorithms to build a predictive model.

2.2 Problem Definition

The broad objective of this project is to develop a predictive model which can predict customer churn. To build a predictive model we first need to train it on data-set so that it can make accurate predictions in the future to train our model we use a Telco Customer Churn data-set that was available on Kaggle website. The basic layer for predicting future customer churn is data from the past. We look at data from customers that already have churned (response) and their characteristics / behaviour (predictors) before the churn happened. By fitting a statistical model that relates the predictors to the response, we will try to predict the

response for existing customers. This method belongs to the supervised learning category.

By using a tool called Pyspark we will deploy the application on the internet where the user has to input the customer ID and other important characteristics and the model will predict the chrun probablity of that individual customer. If you want to predict the customer churn probablity of multiple customers just give the entire data-set as input to the model and the model will predict the churn probality of each customer present in the data-set.

2.3 Motivation

Customer churn (or customer attrition) is a tendency of customers to abandon a brand and stop being a paying client of a particular business. The percentage of customers that discontinue using a company's products or services during a particular time period is called a *customer churn (attrition) rate*. One of the ways to calculate a churn rate is to divide the number of customers lost during a given time interval by the number of acquired customers, and then multiply that number by 100 percent. For example, if you got 150 customers and lost three last month, then your monthly churn rate is 2 percent. Churn rate is a health indicator for businesses whose customers are subscribers and paying for services on a recurring basis.

Customer retention is one of the primary growth pillars for products with a subscription-based business model. Competition is tough in the SaaS market where customers are free to choose from plenty of providers even within one product category. Several bad experiences – or even one – and a customer may quit. And if droves of unsatisfied customers churn at a clip, both material losses and damage to reputation would be enormous. This was the motivation behind a building a customer churn prediction model.

Chapter-3 Literature Survey

3.1 Existing Methods/Tools/Techniques

There are various ways to implement the customer churn prediction model some of them are as follows:-

1. Artificial Neural Network

Artificial Neural Networks (ANN) is a popular approach to address complex problems, such as the churn prediction problem. Neural networks can be hardware-based (neurons are represented by physical components) or software-based (computer models), and can use a variety of topologies and learning algorithms. One popular supervised model is the Multi-Layer Perceptron trained with variations of the Back-Propagation algorithm (BPN). BPN is a feed-forward model with supervised learning. In the case of the customer churn problem, Results have shown that neural networks achieve better performance compared to Decision Trees.

2. Support Vector Machines

Support Vector Machines (SVM), also known as Support Vector Networks, introduced by Boser, Guyon, and Vapnik, are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. SVM is a machine learning technique based on structural risk minimization. Kernel functions have been employed for improving performance. Research on selecting the best kernels or combinations of kernels is still under way. In the churn prediction problem, SVM outperform DT and sometimes ANN.

3. Decision Trees Learning

Decision Trees (DT) are tree-shaped structures representing sets of decisions capable to generate classification rules for a specific data set, or as Berry and Linoff noted “a structure that can be used to divide up a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules”. More descriptive names for such tree models are Classification Trees or Regression Trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. DT have no great performance on capturing complex and non-linear relationships between the attributes. Yet, in the customers churn problem, the

accuracy of a DT can be high, depending on the form of the data.

4. Naive Bayes

A Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be independent feature model. In simple terms, a Naive Bayes (NB) classifier assumes that the presence (or absence) of a particular feature of a class (i.e., customer churn) is unrelated to the presence (or absence) of any other feature. The NB classifier achieved good results on the churn prediction problem for the wireless telecommunications industry and it can also achieve improved prediction rates compared to other widely used algorithms.

5. Regression Analysis-Logistic Regression Analysis

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. In terms of customer churning, regression analysis is not widely used, and that is because linear regression models are useful for the prediction of continuous values. On the other hand, Logistic Regression or Logitic Regression analysis (LR) is a type of probabilistic statistical classification model. It is also used to produce a binary prediction of a categorical variable (e.g. customer churn) which depends on one or more predictor variables (e.g. customers' features).

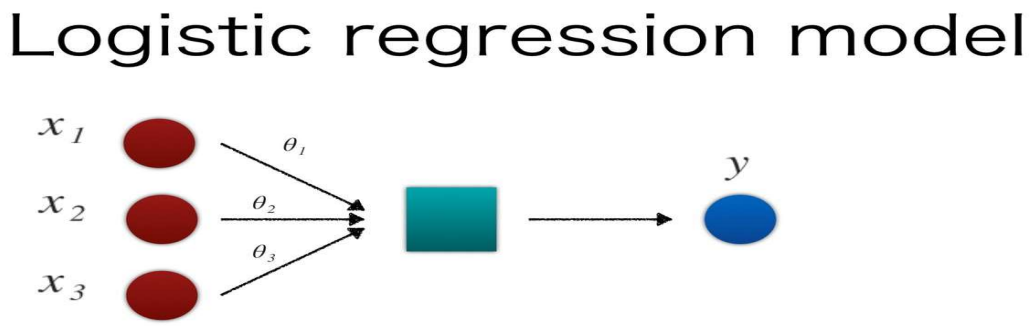

Figure 3.1 Logistic Regression model.

3.2 Literature Survey

Table 3.2 Literature Survey

Sr. No	Title of paper	Year	Authors	Advantages	Limitations	Scope of improvement
1	Customer churn analysis in telecom industry.	02/09/2015	Kiran Dahiya ; Surbhi Bhatia	Logistic regression is used which is very easy to implement.	It can only be used to produce binay prediction.	The model fitting phase of the project must be further optimized.
2	A comparative study of customer churn prediction in telecom industry using ensemble based classifiers.	28/06/2018	Abinash Mishra ; U. Srinivasulu Reddy	Ensemble based classifiers are used which can work on a fairly large data-set.	The accuracy is very low.	The accuracy of the model should be improved.
3	Predictive Analysis and Modeling of Customer Churn in Telecom using Machine Learning Technique.	10/10/2019	B.N.Krishna Sai ; T. Sasikala	Decision Trees are used here which have very low complexciy.	The decision tree classifiers are not very capable to handle large data-sets.	The capability of the model should be improved.
4	Customer Churn Prediction Modelling Based on Behavioural Patterns Analysis using Deep Learning.	19/11/2018	Sanket Agrawal ; Aditya Das ; Amit Gaikwad ; Sudhir Dhage	Deep learning and Naïve bayes classifiers are used here which improves accuracy of the model.	The complexcity of the model is very high.	The complexcity of the model must be decreased.

Chapter-4 Mathematical Model

System Description:

The system we are trying to build is a customer churn prediction model that will help us to detect the probability of churn of each individual customer in a dataset.

• Input:

The input for this model is the Telco Customer Churn data-set that is available on the Kaggle website and it has over 7000 data entries.

• Output:

The output of this model will be the churn probability of each customer in the data-set along with their customer ID.

• Functions and Modules used:

In this model we have used various modules and their associated function to accomplish tasks like data exploration and cleaning, model fitting, model training and testing. Some of the modules used in the model are as follows

1. pandas
2. numpy
3. matplotlib
4. seaborn
5. sklearn
6. xgboost

• Mathematical formulation:

The mathematical formula for logistic regression is as follows

$$P = \frac{1}{1 + e^{-(a+bX)}}$$

Figure 4.1 Logistic Regression formula

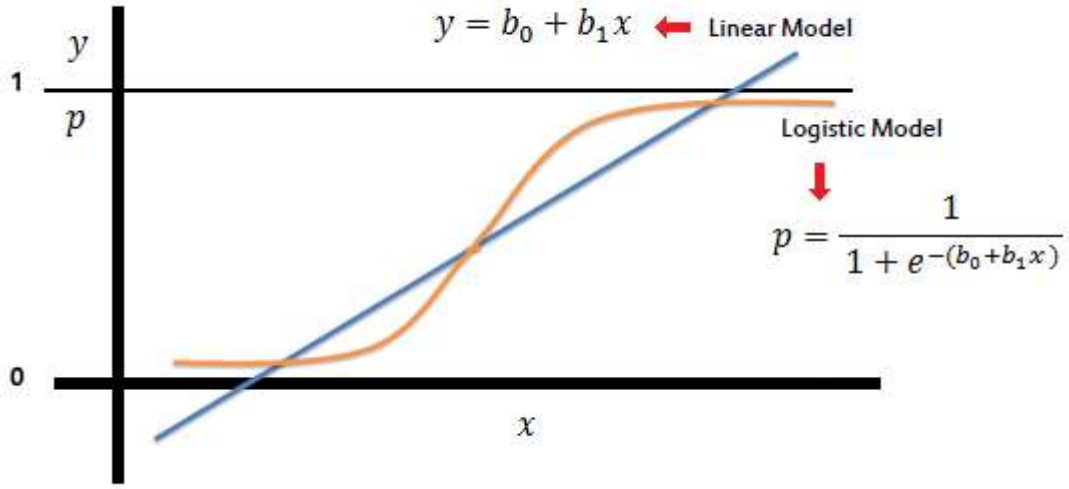

Figure 4.2 Logistic Regression versus Linear Regression

• Success Conditions:

The success of the model depends on the diligence with which the input data-set is cleaned and how well the selected model is fitted on the data-set. If the model is successful than the churn probabilities of all the users in the data-set is visible along with their customer ID.

• Failure Conditions:

The failure of the model depends on the parameters such as improper fitting on the model and improper and reckless cleaning of the data-set which may lead to deletion of data parameters which may be crucial in prediction. If the model fails the probabilities of churn are not visible for all users or it may be visible for only some users and not for all users.

Chapter-5 Proposed System Architecture

5.1 System Architecture

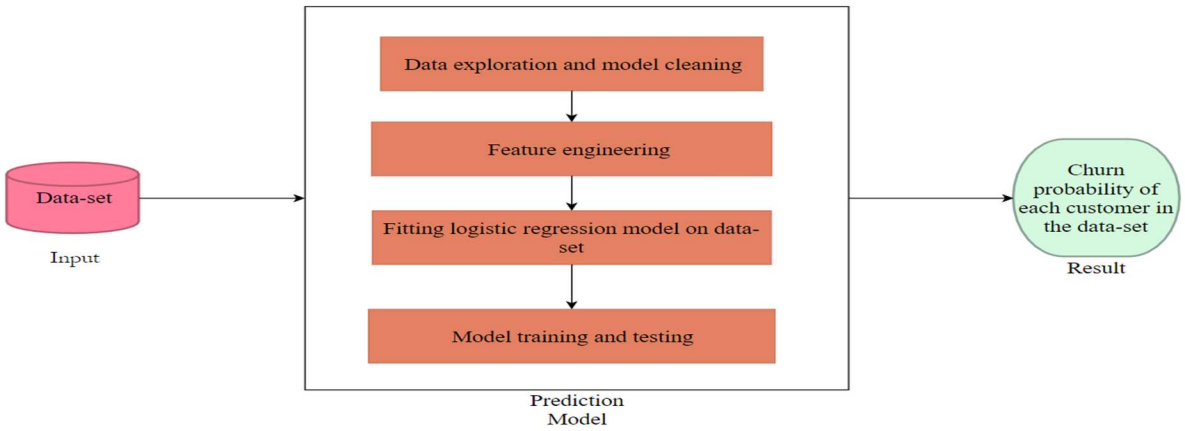

Figure 5.1 System architecture diagram

The system architecture consists of the following important stages they are as follows

1) Data collection & cleaning
With understanding the context it is possible to identify the right data sources, cleansing the data sets and preparing for feature selection or engineering. It sounds quite simple, but this is likely the hardest part. The predicting model is only as good as the data source. And especially Startups or small companies often have trouble finding enough data to train the model adequately.

2) Feature selection & engineering
With the third step we decide which features we want to include in our model and prepare the cleansed data to be used for the machine learning algorithm to predict customer churn.

3) Modelling
With the prepared data we are ready to feed our model. But to make good predictions, we firstly need to find the right model (selection) and secondly need to evaluate that the algorithm actually works. While this usually takes a few iterations, we will keep this quite simple and stop as soon

as the results fit our needs.

4) Insights and Actions
Last but not least we have to evaluate and interpret the outcomes. What does it mean and what actions can we derive from the results? Because predicting customer churn is only half of the part and many people forget that by just predicting, they can still leave. In our case we actually want to stop them from leaving.

5.2 Design with UML Diagrams

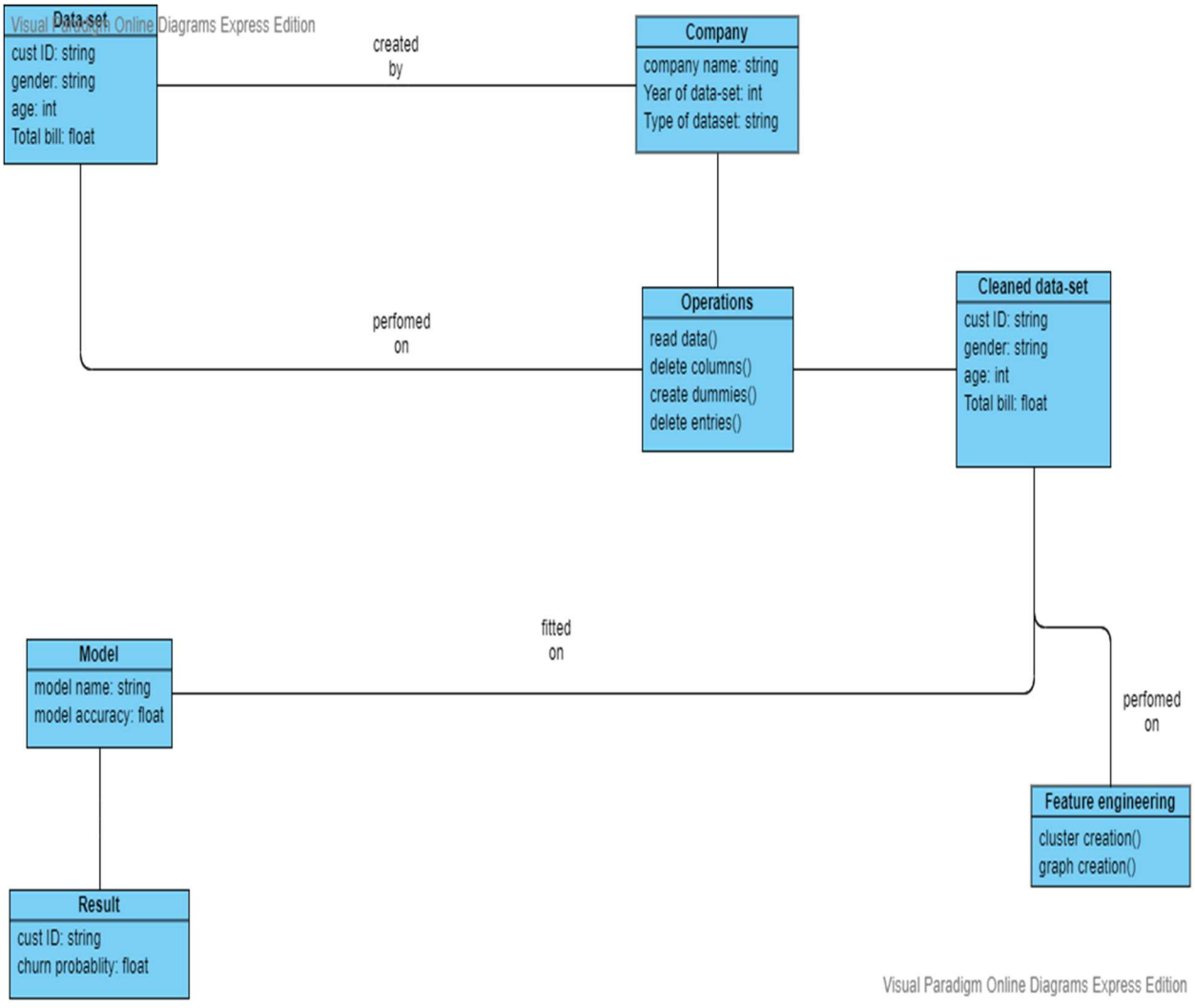

Figure 5.2.1 Class diagram

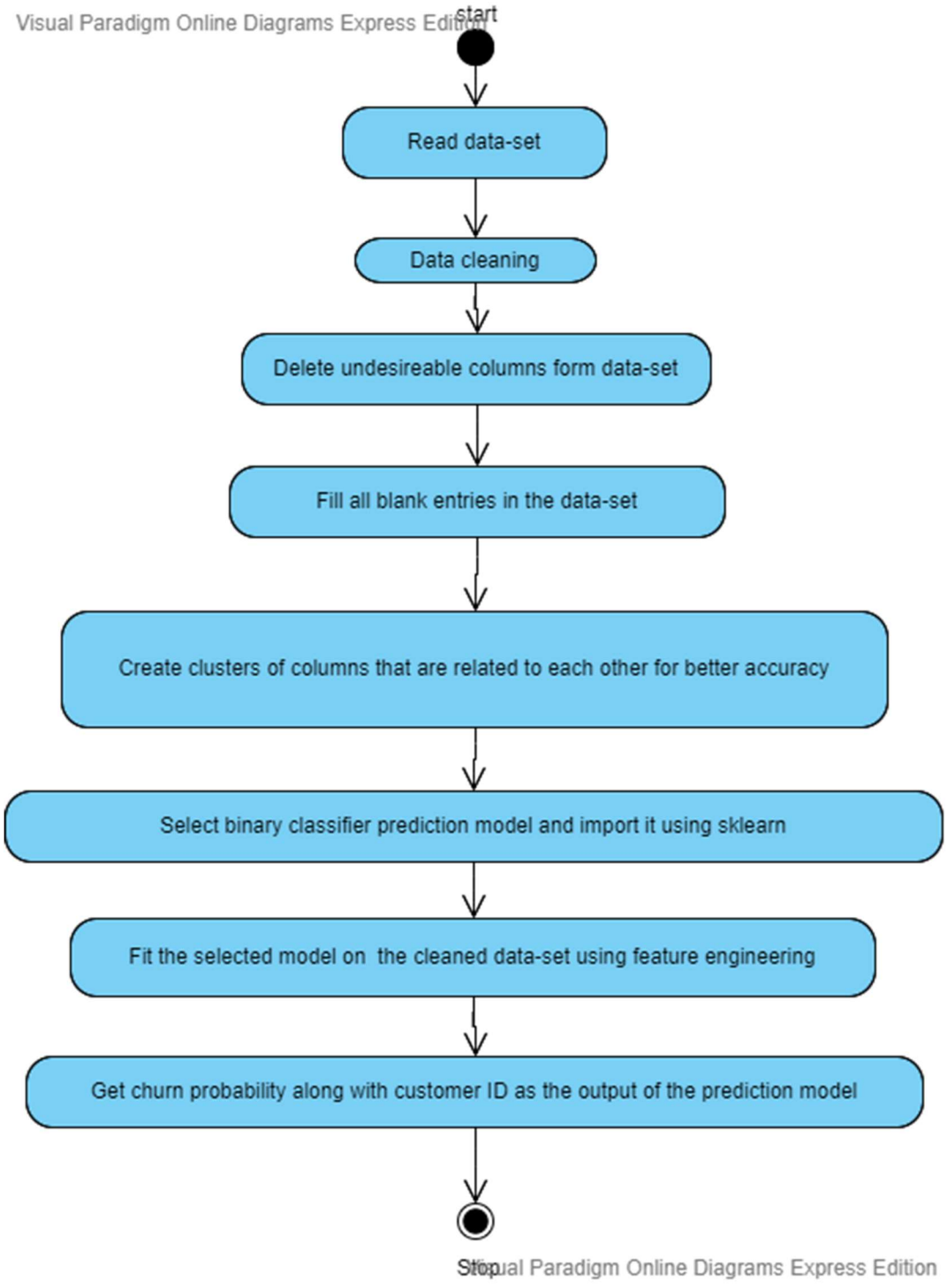

Figure 5.2.2 Activity diagram

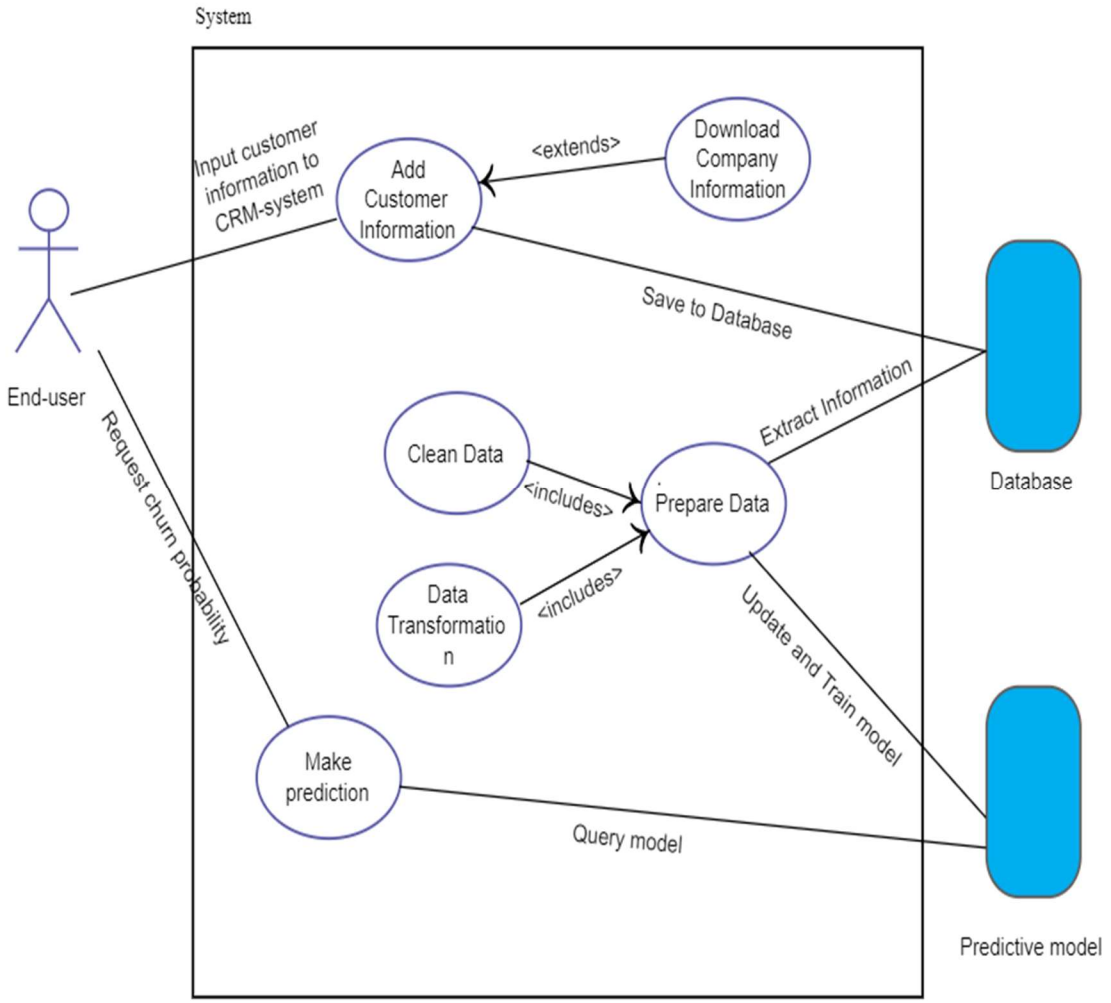

Figure 5.2.3 Use case diagram

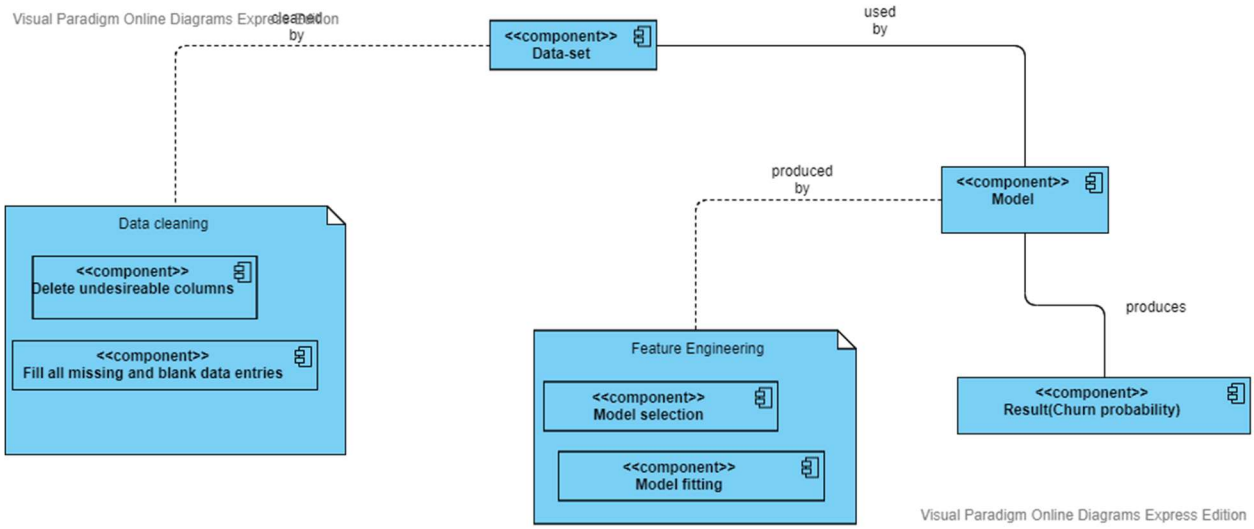

Figure 5.2.4 Component diagram

5.3 Algorithm

To create the customer churn prediction system we have to perform the following steps:-

1. Import all necessary modules like pandas, seaborn, sklearn, xgboost etc that are useful in various important operations

2. Open the data-set using the read function in pandas module and explore the data-set.

3. Once the data-set is explored start with data-cleaning stage.

4. Start to find the columns that are undesirable using various graphs and charts and then delete those columns.

5. Fill out all the empty entries in the data-set.

6. Delete all duplicate entries in the data-set.

7. Start forming clusters of data columns that are related to each other this will increase the accuracy of the model.

8. Now import logistic regression model from the sklearn module.

9. Fit the data-set on the logistic regression(binary classifier) model properly to ensure maximum accuracy.

10. Split the data-set into 80% and 20% where 80% of data-set is used for training the data-set and 20% of the data-set is used for testing the model.

11. Now that our model is trained and tested give another data-set as input to the model and Check the accuracy of the model.

12. The model gives customer ID along with churn probability for each customer in the data-set.

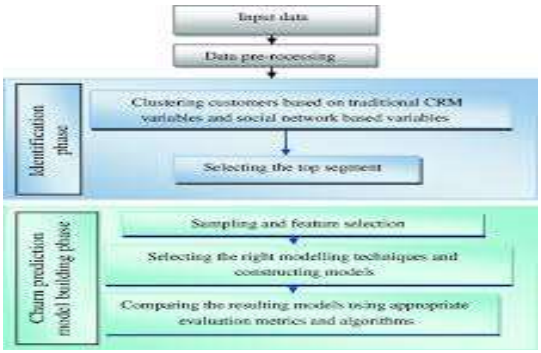

Figure 5.3.1 Algorithm

5.4 Implementation/Proof of Concept

The important tools and technologies used in this project are as follows:-

1. Python language

2. Jupyter Notebook IDE

3. Pyspark

4. sklearn and xgboost modules

5.5 Important Source Code

The important source material for this project was accessed using google and youtube videos the main source codes for this project were obtained from the websites “Towards data-science” and “Daitaku.com”.

5.6 Testing Code

The testing of the project was done using the Telco customer churn data-set and the output we obtained was the customer ID along with the churn probability of each customer in the data-set.

5.7 Result screenshots

```
In [1]: from datetime import datetime, timedelta,date
        import pandas as pd
        %matplotlib inline
        from sklearn.metrics import classification_report,confusion_matrix
        import matplotlib.pyplot as plt
        import numpy as np
        import seaborn as sns
        from __future__ import division
        from sklearn.cluster import KMeans

In [2]: import plotly.plotly as py
        import plotly.offline as pyoff
        import plotly.graph_objs as go

In [3]: import xgboost as xgb
        from sklearn.model_selection import KFold, cross_val_score, train_test_split
```

Figure 5.7.1 Important modules

```
In [7]: df_data = pd.read_csv('D:\TEproject\ChurnDataset.csv')
        df_data.head(10)
```

Out[7]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSup
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	
5	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	...	Yes	
6	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	...	No	
7	6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes	...	No	
8	7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	...	Yes	
9	6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes	...	No	

10 rows × 21 columns

Figure 5.7.2 Reading the data-set

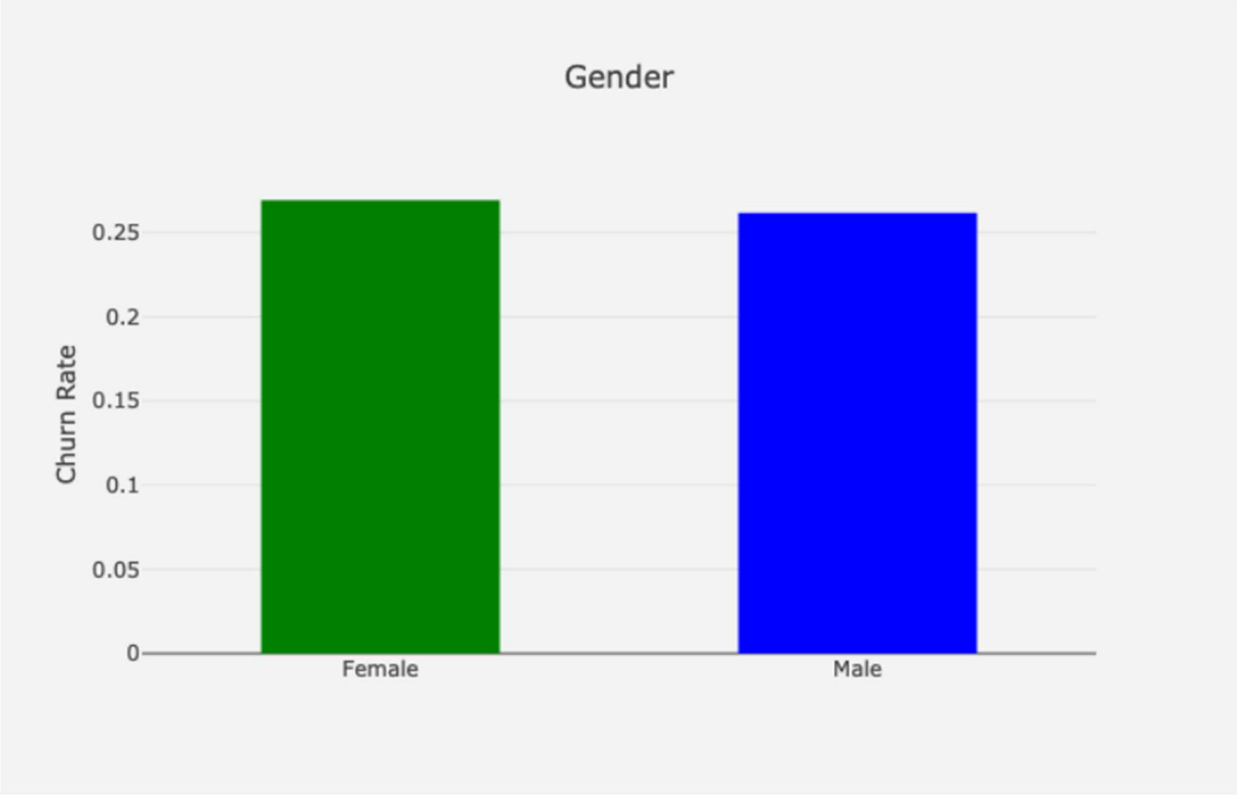

Figure 5.7.3 Churn rate by gender

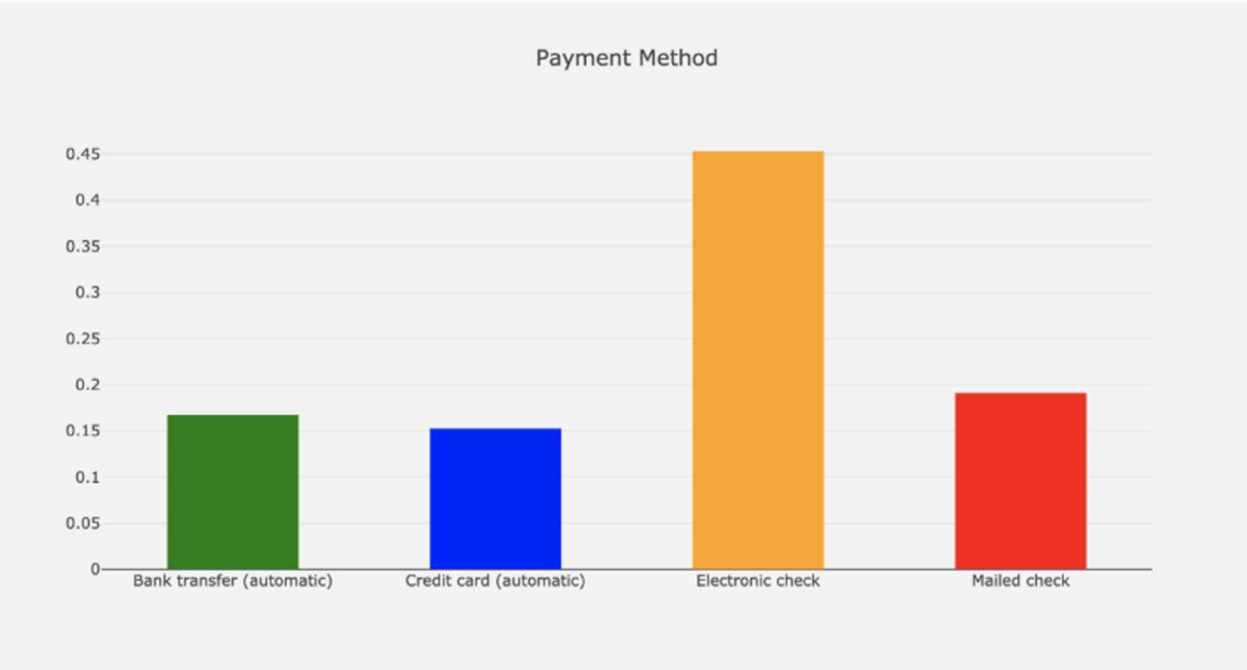

Figure 5.7.4 Churn rate by payment method

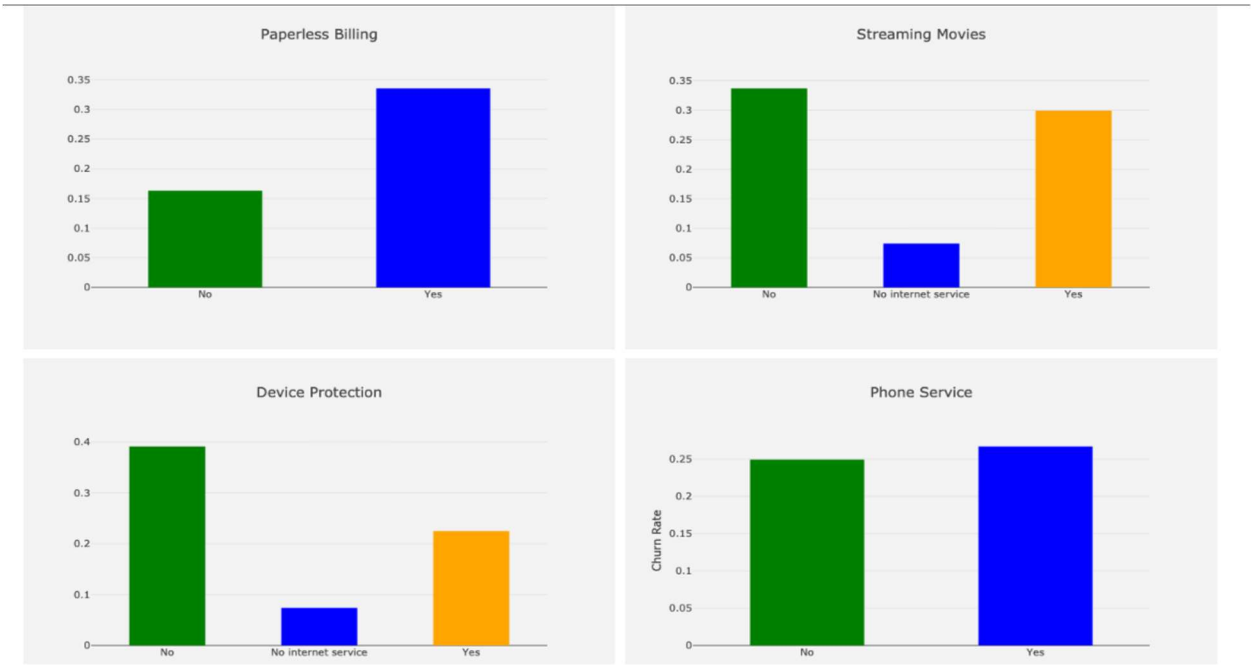

Figure 5.7.5 Churn rate by paperless billing, streaming movies, device protection, phone service

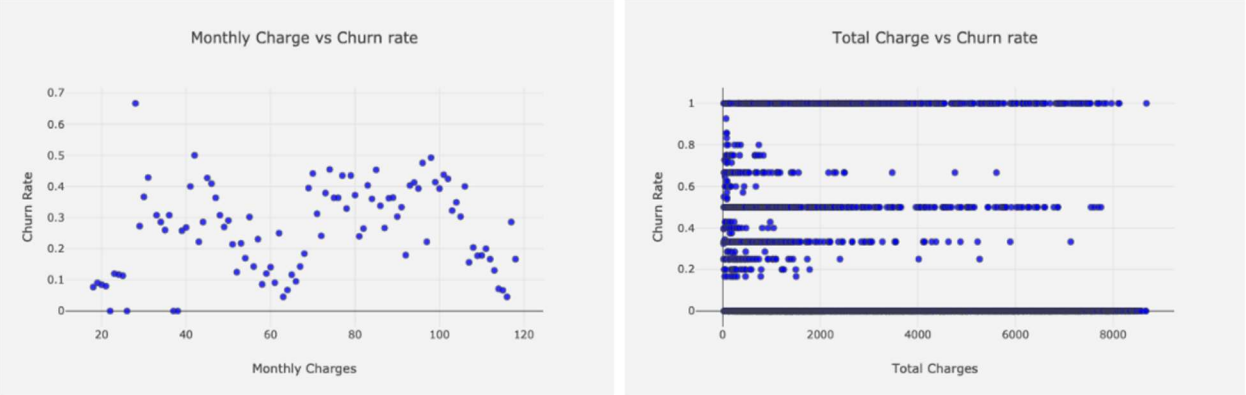

Figure 5.7.6 Monthly charge vs churn rate and Total charge vs churn rate

```
In [14]: kmeans = KMeans(n_clusters=3)
kmeans.fit(df_data[['tenure']])
df_data['TenureCluster'] = kmeans.predict(df_data[['tenure']])
df_data = order_cluster('TenureCluster', 'tenure',df_data,True)
df_data.groupby('TenureCluster').tenure.describe()
```

Out[14]:

TenureCluster	count	mean	std	min	25%	50%	75%	max
0	2941.0	7.801428	6.227163	0.0	2.0	6.0	13.0	21.0
1	1929.0	34.792120	8.297679	22.0	27.0	35.0	42.0	49.0
2	2173.0	63.475380	7.172433	50.0	57.0	65.0	70.0	72.0

Figure 5.7.7 Tenure cluster

Figure 5.7.8 Tenure cluster vs churn rate

In [17]:
```
kmeans = KMeans(n_clusters=3)
kmeans.fit(df_data[['MonthlyCharges']])
df_data['MonthlyChargeCluster'] = kmeans.predict(df_data[['MonthlyCharges']])
df_data = order_cluster('MonthlyChargeCluster', 'MonthlyCharges',df_data,True)
df_data.groupby('MonthlyChargeCluster').MonthlyCharges.describe()
```

Out[17]:

MonthlyChargeCluster	count	mean	std	min	25%	50%	75%	max
0	1892.0	23.384619	5.660437	18.25	19.80	20.40	25.0500	42.40
1	2239.0	61.628808	10.441432	42.60	51.80	61.55	70.7000	77.80
2	2912.0	94.054258	10.343944	77.85	85.05	93.90	101.9125	118.75

Figure 5.7.9 Monthly charge cluster

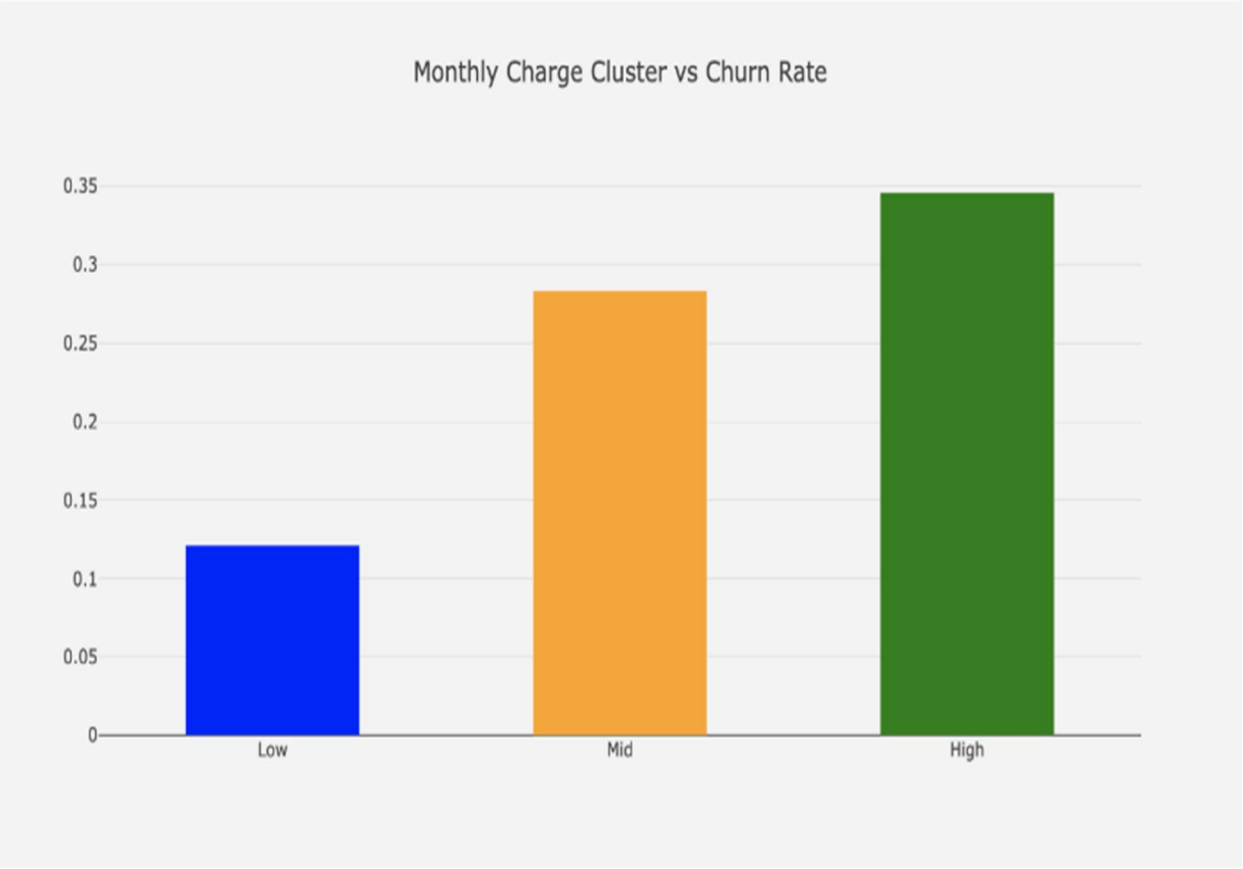

Figure 5.7.10 Monthly charge cluster vs churn rate

```
In [23]: kmeans = KMeans(n_clusters=3)
         kmeans.fit(df_data[['TotalCharges']])
         df_data['TotalChargeCluster'] = kmeans.predict(df_data[['TotalCharges']])
         df_data = order_cluster('TotalChargeCluster', 'TotalCharges',df_data,True)
         df_data.groupby('TotalChargeCluster').TotalCharges.describe()
```

Out[23]:

TotalChargeCluster	count	mean	std	min	25%	50%	75%	max
0	4160.0	686.204087	572.025502	18.80	161.4875	535.95	1139.4125	1975.85
1	1613.0	3272.602139	814.144302	1978.65	2548.6500	3211.20	3970.4000	4779.45
2	1259.0	6292.972558	1003.372938	4783.50	5468.7000	6145.85	7040.1500	8684.80

Figure 5.7.11 Total charge cluster

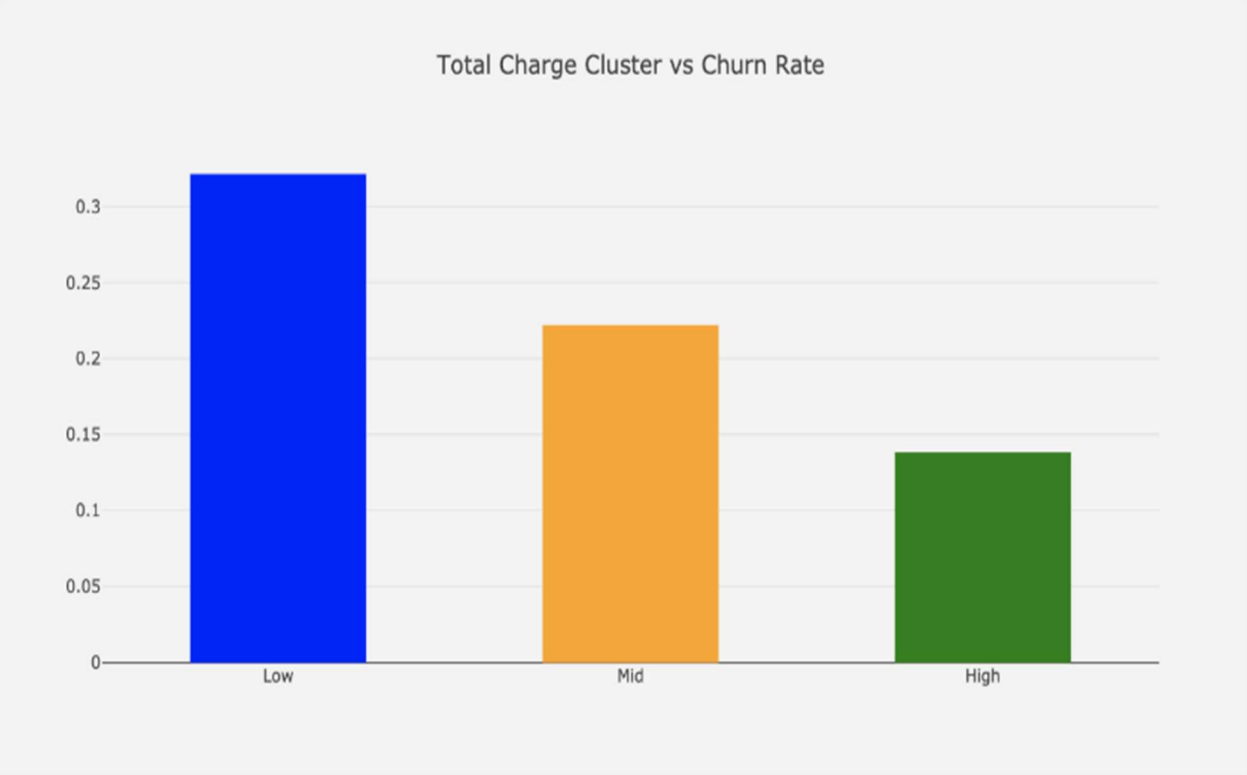

Figure 5.7.12 Total charge cluster vs churn rate

In [29]:
```
import statsmodels.api as sm
import statsmodels.formula.api as smf


glm_model = smf.glm(formula='Churn ~ {}'.format(glm_columns), data=df_data, family=sm.families.Binomial())
res = glm_model.fit()
print(res.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                  Churn   No. Observations:                 7032
Model:                            GLM   Df Residuals:                     7002
Model Family:                Binomial   Df Model:                           29
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2902.2
Date:                Wed, 01 Apr 2020   Deviance:                       5804.5
Time:                        22:13:08   Pearson chi2:                 7.60e+03
No. Iterations:                   100   Covariance Type:             nonrobust
------------------------------------------------------------------------------
```

Figure 5.7.13 Linear model regression results

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.2555	0.276	0.924	0.355	-0.286	0.797
gender	-0.0233	0.065	-0.358	0.721	-0.151	0.104
SeniorCitizen	0.2212	0.085	2.610	0.009	0.055	0.387
Partner	0.0011	0.078	0.014	0.988	-0.152	0.154
Dependents	-0.1309	0.090	-1.453	0.146	-0.307	0.046
tenure	-0.0620	0.008	-7.440	0.000	-0.078	-0.046
PhoneService	0.2323	0.403	0.576	0.564	-0.558	1.023
PaperlessBilling	0.3451	0.075	4.614	0.000	0.198	0.492
MonthlyCharges	-0.0345	0.032	-1.082	0.279	-0.097	0.028
TotalCharges	0.0001	9.97e-05	1.150	0.250	-8.08e-05	0.000
MultipleLines_No	-0.1138	0.130	-0.876	0.381	-0.368	0.141
MultipleLines_No_phone_service	0.0231	0.160	0.144	0.885	-0.291	0.337
MultipleLines_Yes	0.3461	0.283	1.221	0.222	-0.209	0.902
InternetService_DSL	-0.6036	0.226	-2.669	0.008	-1.047	-0.160
InternetService_Fiber_optic	1.0556	0.578	1.826	0.068	-0.077	2.188
InternetService_No	-0.1964	0.091	-2.152	0.031	-0.375	-0.018
OnlineSecurity_No	0.3288	0.108	3.040	0.002	0.117	0.541
OnlineSecurity_No_internet_service	-0.1964	0.091	-2.152	0.031	-0.375	-0.018
OnlineSecurity_Yes	0.1231	0.261	0.472	0.637	-0.389	0.635
OnlineBackup_No	0.2224	0.107	2.082	0.037	0.013	0.432
OnlineBackup_No_internet_service	-0.1964	0.091	-2.152	0.031	-0.375	-0.018
OnlineBackup_Yes	0.2295	0.261	0.881	0.379	-0.281	0.740
DeviceProtection_No	0.1511	0.107	1.408	0.159	-0.059	0.361
DeviceProtection_No_internet_service	-0.1964	0.091	-2.152	0.031	-0.375	-0.018
DeviceProtection_Yes	0.3008	0.261	1.154	0.249	-0.210	0.812
TechSupport_No	0.3161	0.108	2.929	0.003	0.105	0.528
TechSupport_No_internet_service	-0.1964	0.091	-2.152	0.031	-0.375	-0.018

Figure 5.7.14 Linear model regression results

```
In [30]: np.exp(res.params)
```

```
Out[30]: Intercept                                1.291094
         gender                                   0.976999
         SeniorCitizen                            1.247633
         Partner                                  1.001130
         Dependents                               0.877318
         tenure                                   0.939874
         PhoneService                             1.261550
         PaperlessBilling                         1.412110
         MonthlyCharges                           0.966097
         TotalCharges                             1.000115
         MultipleLines_No                         0.892446
         MultipleLines_No_phone_service           1.023419
         MultipleLines_Yes                        1.413587
         InternetService_DSL                      0.546815
         InternetService_Fiber_optic              2.873624
         InternetService_No                       0.821650
         OnlineSecurity_No                        1.389280
         OnlineSecurity_No_internet_service       0.821650
         OnlineSecurity_Yes                       1.131047
         OnlineBackup_No                          1.249079
         OnlineBackup_No_internet_service         0.821650
         OnlineBackup_Yes                         1.258001
         DeviceProtection_No                      1.163101
         DeviceProtection_No_internet_service     0.821650
```

Figure 5.7.14 Linear model regression results

```
TechSupport_Yes                          1.145503
StreamingTV_No                           0.943367
StreamingTV_No_internet_service          0.821650
StreamingTV_Yes                          1.665674
StreamingMovies_No                       0.943602
StreamingMovies_No_internet_service      0.821650
StreamingMovies_Yes                      1.665259
Contract_Month_to_month                  2.160961
Contract_One_year                        1.095171
Contract_Two_year                        0.545543
PaymentMethod_Bank_transfer__automatic_  1.033994
PaymentMethod_Credit_card__automatic_    0.947467
PaymentMethod_Electronic_check           1.379340
PaymentMethod_Mailed_check               0.955442
TenureCluster_High                       1.597177
TenureCluster_Low                        0.899531
TenureCluster_Mid                        0.898646
MonthlyChargeCluster_High                1.068921
MonthlyChargeCluster_Low                 1.111221
MonthlyChargeCluster_Mid                 1.086954
TotalChargeCluster_High                  1.538368
TotalChargeCluster_Low                   0.729169
TotalChargeCluster_Mid                   1.150983
dtype: float64
```

Figure 5.7.15 Linear model regression results

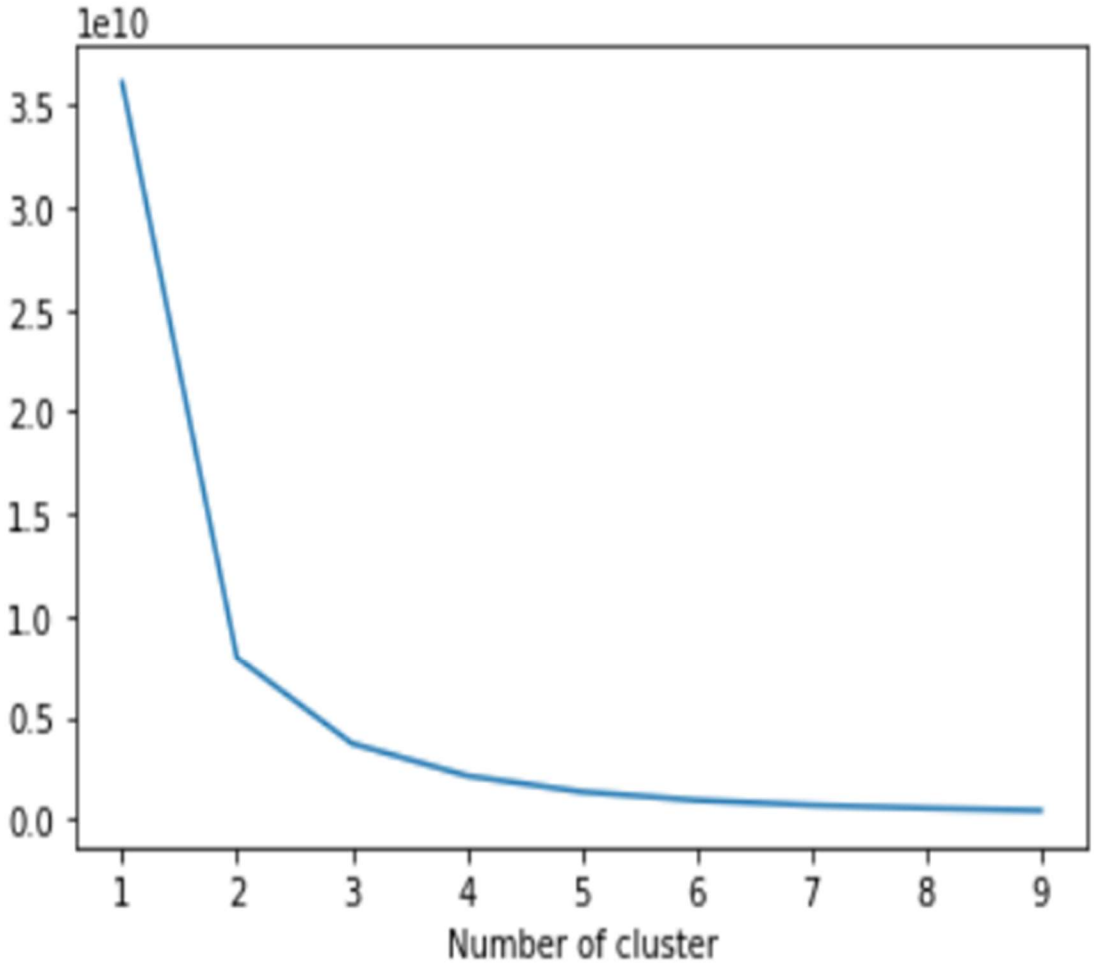

Figure 5.7.16 ROC curve

```
In [31]: #create feature set and labels
X = df_data.drop(['Churn','customerID'],axis=1)
y = df_data.Churn
#train and test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.05, random_state=56)
#building the model
xgb_model = xgb.XGBClassifier(max_depth=5, learning_rate=0.08, objective= 'binary:logistic',n_jobs=-1).fit(X_train, y_train)

print('Accuracy of XGB classifier on training set: {:.2f}'
      .format(xgb_model.score(X_train, y_train)))
print('Accuracy of XGB classifier on test set: {:.2f}'
      .format(xgb_model.score(X_test[X_train.columns], y_test)))
```

```
Accuracy of XGB classifier on training set: 0.84
Accuracy of XGB classifier on test set: 0.81
```

Figure 5.7.17 Accuracy of the model

```
In [32]: y_pred = xgb_model.predict(X_test)
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.84      0.93      0.88       266
           1       0.67      0.45      0.54        86

   micro avg       0.81      0.81      0.81       352
   macro avg       0.76      0.69      0.71       352
weighted avg       0.80      0.81      0.80       352
```

Figure 5.7.18 Detailed analysis of accuracy of the model

```
In [34]: df_data['proba'] = xgb_model.predict_proba(df_data[X_train.columns])[:,1]
         df_data[['customerID', 'proba']].head()
```

Out[34]:

	customerID	proba
0	7590-VHVEG	0.514636
1	6713-OKOMC	0.144643
2	7469-LKBCI	0.017225
3	8779-QRDMV	0.913855
4	1680-VDCWW	0.044838

Figure 5.7.19 Result customer churn probability of each customer in the data-set

Chapter-6 Advantages / Disadvantage

Advantages

Adopting Machine Learning for Churn Prediction has several advantages over traditional business rules:

1. Machine Learning relies on finding patterns and relationships in large amounts of data, the rules discovered by the Machine Learning model are guaranteed to be supported by evidence instead of intuition/hunches.

2. Unlike humans, who are limited by the number of variables/factors they can account for when crafting their business rules, Machine Learning algorithms can process and extract patterns from many variables, which results in (typically) more complex and comprehensive rules.

3. Assuming there is good quality data available, Machine Learning is able to learn highly accurate rules in a much shorter time when compared to a human which usually needs a significant amount of experience and domain knowledge to devise accurate rules, i.e., the ROI tends to be higher for Machine Learning.

4. A fourth advantage is that Machine Learning is able to timely detect concept drift and adapt the rules accordingly, thus being more adaptive to changes, i.e., if the accuracy of the churning predictions start to degrade over time, Machine Learning quickly detects this and adapts the rules to the new scenario, ensuring the prediction of churners is reliable for the business.

Disadvantages

1. Modeling churn is difficult because there is inherent uncertainty when measuring churn. This uncertainty does not change churn's status as an essential SaaS metric. What this uncertainty does change is how we utilize churn metrics.

2. For any metric you use to calculate churn, make sure you understand its limitations. The smaller the number of customers you have, the more likely month-to-month churn may appear to move up or down based solely on chance.

3. If you have a large number of customers, measuring quarterly and annual churn rates can give very different results depending on how churn is calculated and how much the true rate of churn changes during the longer period. Despite answering a simple question, churn is a complicated metric.

Chapter-7 Applications

The ability to predict that a particular customer is at a high risk of churning, while there is still time to do something about it, represents a huge additional potential revenue source for every online business. Besides the direct loss of revenue that results from a customer abandoning the business, the costs of initially acquiring that customer may not have already been covered by the customer's spending to date. (In other words, acquiring that customer may have actually been a losing investment.) Furthermore, it is always more difficult and expensive to acquire a new customer than it is to retain a current paying customer.

In order to succeed at retaining customers who would otherwise abandon the business, marketers and retention experts must be able to (a) predict in advance which customers are going to churn through churn analysis and (b) know which marketing actions will have the greatest retention impact on each particular customer. Armed with this knowledge, a large proportion of customer churn can be eliminated.

Churn prediction modeling techniques attempt to understand the precise customer behaviors and attributes which signal the risk and timing of customer churn. The accuracy of the technique used is obviously critical to the success of any proactive retention efforts.

The probability of churn can be predicted using various statistical or machine learning techniques. These methods process historical purchase and behavior data in order to predict the probability of cancellation per customer.

1. A well-constructed model can inform a wide range of decisions and flow into numerous internal tools or applications. For example, some common use cases for a churn model are:

2. Measuring feature impacts on the likelihood of churn in order to understand why customers choose to leave, which can inform long-term retention initiatives

3. Creating churn risk scores that can indicate who is likely to leave, and using that information to drive retention campaigns

4. Predicting the probability of churn and using it to flag customers for upcoming email campaigns
Integrating outputs with internal apps, such as a customer call center, to provide relevant real-time churn risk information

5. Discounting strategically with promotion campaigns to customers with a high cancellation risk

Chapter-8 Conclusion and future work

The importance of customer churn prediction in the telecom and other service based corporations market is to help companies make more profit. It has become known that predicting churn is one of the most important sources of income to telecom companies. Hence, this model aimed to build a system that predicts the churn of customers based on the Telco Customer Churn data-set that was available on the Kaggle website. These prediction models need to achieve high AUC values. To test and train the model, the sample data is divided into 80% for training and 20% for testing. We chose to perform cross-validation with 10-folds for validation and hyperparameter optimization. We have applied feature engineering, effective feature transformation and selection approach to make the features ready for machine learning algorithms. The method of preparation and selection of features and entering the mobile social network features had the biggest impact on the success of this model, since the value of AUC in data-set reached 87%. XGBOOST binary classifier model achieved the best results in all measurements. The AUC value was 93.301%. and the accuracy achieved by our model was about 87%.

We identified the main churn factors from the dataset and performed cluster profiling according to their risk of churning. Finally, we provided guidelines on customer retention for decision-makers of the telecom companies. In future, we will further investigate eager leaning and lazy learning approaches for better churn prediction. The study can be further extended to explore the changing behavior patterns of churn customers by applying Artificial Intelligence techniques for predictions and trend analysis.

Abbreviation

ML: Machine learning

CDR: call detail record

CRM: customer relationship management

SMS: short message service

XG Boost: Extreme Gradient Boosting

ANN: Artificial Neural Network

SVM: Support Vector Machines

NB: Naive Bayes

LR: Logistic Regression

DT: Decision Tree

AUC: Area Under the Curve

SNA: Social Network Analysis

GBM: Gradient Boosted Machine

CSV: comma-separated values

REFERENCES

1) Thanasis Vafeiadis, Konstantinos I. Diamantaras, G. Sarigiannidis, K. Ch Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction", *Simulation Modelling Practice and Theory*, vol. 55, pp. 1-9, 2015.

2) Abbas Keramati, Ruholla-Jafari-Marandi Mohammed, Aliannejadi ImanAhmadian, MahdiehMozaffari UldozAbbasi, "Improved churn prediction in telecommunication industry using data mining techniques", *Applied Soft Computing*, vol. 24, pp. 994-1012, 2014.

3) Veronikha Effendy, ZK Abdurahman Baizal, "Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest", 2014 2nd International Conference on Information and Communication Technology (ICoICT), pp. 325-330, 2014.

4) V. Umayaparvathi, K. Iyakutti, "Applications of Data Mining Techniques in Telecom Churn Prediction", *International Journal of Computer Applications (0975-8887)*, vol. 42, no. 20, March 2012.

5) Preeti. Dalvi, Siddhi K. Khandge, Ashish Deomore, Aditya Bankar, V. A. Kanade, "Analysis of customer churn prediction in telecom industry using decision trees and logistic regression", *Colossal Data Analysis and Networking (CDAN) Symposium on*, pp. 1-4, 2016.

6) Duyen Do, Phuc Huynh, Phuong Vo, Tu Vu, "Customer churn prediction in an internet service provider", *Big Data(Big Data) 2017 IEEE International Conference on*, pp. 3928-3933, 2017.