Churn Prediction of Fiber to Home Users

A Thesis Submitted for the Degree of Master of Business Analytics

W.M.S.M.M Wijayarathne

University of Colombo School of Computing

2020

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis. This thesis has also not been submitted for any degree in any university previously.

Student Name: W.M.S.M.M. Wijayarathne

Registration Number:18880392

Index Number:2018/BA/039

_____

Signature of the Student & Date

This is to certify that this thesis is based on the work of Mr.Menuka Wijayarathne under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by,

Supervisor Name:

Manjusri Wickramasinghe

Signature of the Supervisor & Date

Manjusri Wickramasinghe

Digitally signed by Manjusri Wickramasinghe
DN: cn=Manjusri Wickramasinghe, o=University of Colombo, ou, email=mie@ucsc.cmb.ac.lk, c=LK
Date: 2021.09.14 22:31:00 +05'30'

I would like to dedicate this thesis to Network Engineers who dedicated their lives to keep the world connected.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor Dr. Manjusri Wickramasinghe for his guidance and encouragement throughout this project. I would also like to extend my gratitude to Dialog Axiata PLC for providing me the resources.

ABSTRACT

Due to the competitive strategies of service providers, Customers actively swap from one service provider to another to satisfies their needs which triggers them to ‘churn’. Hence, It’s time for the telecommunication industry to make necessary predictive decisions to increase the **‘happiness factor of customers’ and** ‘**Retaining the customers with stabilizing their market value.’**This study discusses the strategies that can predict the Churn and the remedies of minimizing the churn factor of the fiber to home users. That could help relevant stakeholders to take a suitable decision at the right time and mitigate customers from churning, thereby reducing the churn rate. 14,461 instances related to a telecom sector are used in this study with fourteen network-related feature parameters for developing a churn prediction model for fiber to the home users, in which some of the features are numerical and some of the features are categorical. Among these features, eight significant features are selected using the Recursive Feature Elimination Technique. These variables are used as input variables for Logistic Regression classifier to build a churn prediction model. Model performance is measured using Recall, Precision, F1-score, Support, confusion Matrix, and ROC curve. It was able to achieve significant accuracy from the model with 0.86 area under the curve. Market research survey was carried out to observe why people are unhappy with fiber to home services. These results are used to develop business and technical strategies to retain the fiber to the home customers.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Motivation

Telecommunication companies invest millions in new technologies expecting a higher level of return on investment. Specifically, when offering fiber-to-home solutions, millions of dollars are spent on making the required infrastructure. Despite the cost of infrastructure, the customer would leave the service provider if they wouldn't meet their expectations including speed and coverage. This situation will be worsened due to the increasing number of competitors who try to offer high-quality services at competitive rates.

Telecommunication companies should carefully observe the strategies to mitigate the deliberate churns of fiber-to-home customers. Moreover, there would be many reasons for churn such as competitor's advertising strategies, promotional packages, seasonal packages, and poor customer care service and poor network quality, etc. Particularly as more the network quality gets poor, the more customer tends to churn. Hence, it's important to maintain a seamless and quality service provider infrastructure that values what customers paid for.

1.2 Statement of the problem

The problem is to predict the potential churners and potential non-churners who subscribed with fiber to home telecommunication services.

1.3 Research Aims and Objectives

1.3.1 **Aim**

This study aims to build a model for the early prediction of churn of fiber-to-home customers by using the network parameters of the Optical line terminator(OLT) and Optical Network Terminator(ONT) and to discuss the strategies in minimizing the churning factor of fiber to home(FTTH) customers. Through this process, it is intended to proactively identify the potential customer churns and apply appropriate strategies to retain these customers before they leave the service provider. Below depicted network parameters are used for the research.

- Optical Line Terminator Model (ONT)
- Distance between OLT and ONT(km)
- Receiving power(dbmv) of ONT

- Transmitting Power(dbmv) of ONT
- Total ONT count per OLT port
- Total Third-party ONT count per port-
- Total Original ONT count per port
- Television over IP service count per individual ONT
- Voice service count per individual ONT
- High-speed internet service count per individual ONT
- Total Service count per individual ONT

1.3.2 Objectives

The objective of the study is to predict the potential churners who subscribed to fiber to home services and to develop proactive business and technical strategies to retain these customers.

1.4 Scope

Table 3. 2:Project Scope

Project Task	**Scope of the tasks**
Data collection	- Extraction and preprocessing of required data
Build a supervised learning model for churn prediction	- Implementing a binary classification model using logistic regression and the model accuracy is observed. - Observe the methods to increase the model accuracy.
Find hidden relationships and hidden behavioral patterns of data	- Using the K means clustering algorithm, partitioning the customer data into groups.
Market research survey	- Factors contributing to the dissatisfaction of FTTH customers are observed.
Build up a service model for churn mitigation	- Finding business strategies to retain customers based on the collected data. - Finding proactive strategies to enhance degraded network parameters to retain customers.

1.5 Background of the study

Data for this study is obtained from dialog Axiata PLC. Dialog Offers three types of Fiber to home service variants for customers including IPTV (Television Service over IP address including Video on Demand and rewind TV), VOICE, and HSI (High-Speed Internet). Customers can purchase either IPTV, VOICE, or HSI using a single fiber optic connection based on their preference. The monthly subscription fee of FTTH services is based as below.

- HSI – Bandwidth of the data package
- IPTV – Channels per package
- VOICE – Number of voice connections

Fiber to the home solution (FTTH) is a low latency solution, with high performance and seamless experience where customers can subscribe to higher bandwidth packages. FTTH services are for domestic customers who don't require guaranteed bandwidths. These services can be impacted by network congestion. Hence dedicated bandwidth is not guaranteed. Several data sources are used throughout this study including FTTH related live network elements and billing details to analyze and build an accurate predictive model to identify potential churns.

1.6 Feasibility Study

Required data are retrieved by the below data sources which are governed by Dialog Axiata PLC. Permission for accessing the data was granted by the group chief technology officer at Dialog Axiata PLC [Appendix A]. The sole purpose of collecting the below mentioned data is to implement a churn predictive learning model for FTTH users. Moreover, no financial cost is involved in carrying out this study.

1. Work Order Management System(WOM)

Work order details of the new installation are located on WOM. Data with the below attributes are considered for this study.

- Date of the New installations.
- Optical performance details of Optical Network Terminator (ONT) at the installation stage. (signal to noise ratio, Downstream power levels)

2. Customer Relationship Management Portal(CRM)

The connection status and subscription details of the customers are located on CRM. Below attributes of CRM are considered in this study.

- Billing status of the customer
- Circuit number of the customer - unique number to index the customer
- Monthly bandwidth consumption of customers - This relates to customers with internet packages. The monthly data consumption of a particular user can be exported and analyzed.
- IPTV subscription details of the customer - This reveals, subscribed IPTV package of the user (Low cost, Medium Cost, high cost).
- Voice subscription details of the customer - This reveals the amount of subscribed voice connections per user.
- Provisioned services for individual users-This shows how many services are commissioned per each user

3. Service commissioning details/Configuration details of Optical Line Terminator

Below data are retrieved by logging into the respective network element called Optical Line Terminator(OLT) and using CLI scraping.

- Optical Line Terminator Model (ONT)
- Distance between OLT and ONT(km)
- Receiving power(dbmv) of ONT
- Transmitting Power(dbmv) of ONT
- Total ONT count per OLT port
- Total Third-party ONT count per port-
- Total Original ONT count per port
- Television over IP service count per individual ONT
- Voice service count per individual ONT
- High-speed internet service count per individual ONT
- Total Service count per individual ONT
- Multiple Service /Single Service

4. JIRA Platform

Complaints Related to FTTH services are lodged on this platform. Below attributes of the JIRA platform are considered in this study.

- The root cause of the Fault – Examples: IPTV freezing, Internet connection lost, internet speed is low, voice call dropping, etc
- Number of Optical Network Termination units per passive optical port when the fault

is handled by the technical agent.

- Number of service request per subscriber (Number of fault escalations per subscriber)

1.7 Structure of the thesis

The rest of the paper is presented as follows. Related works are presented in Chapter 2.In Chapter 3, the architecture of the churn prediction and mitigation model is introduced in detail. The experimental setting and results are exhibited in Chapter 4. Conclusions and directions are discussed in chapter five for further research.

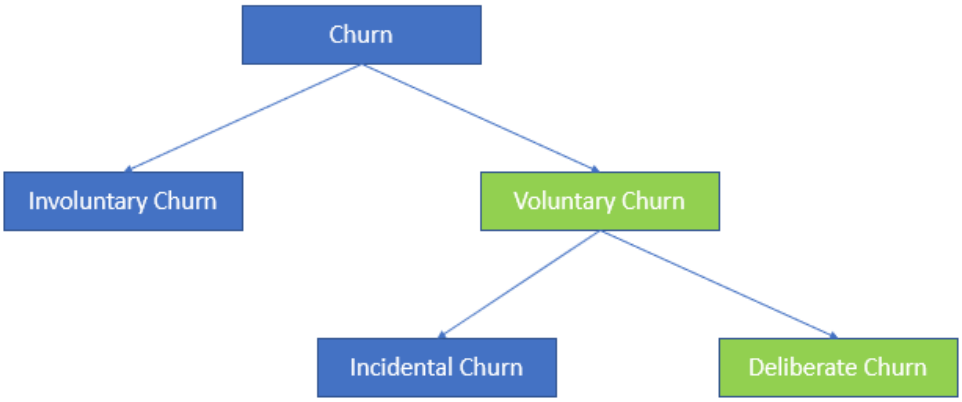
CHAPTER 2

LITERATURE REVIEW

2.1 A Literature Review

Churns can happen in two ways. Sometimes customer voluntarily cancels their subscription by contacting service provider or either company would remove the subscription of a customer without any action by the customer (Rahman et al. 2011). When a Service provider disconnects a customer due to non-payment or fraud, it is called involuntary churn – Figure 2.1. Voluntary churn or the scenario of canceling a subscription by the customer himself is diverse. Voluntary churn can be classified as incidental churn and deliberate churn as in the below figure2.1. Incidental churn represents customer's financial problems and customer's relocation to a new location where the existing telecommunication services are not available.

FIGURE 2.1:Types of Churn

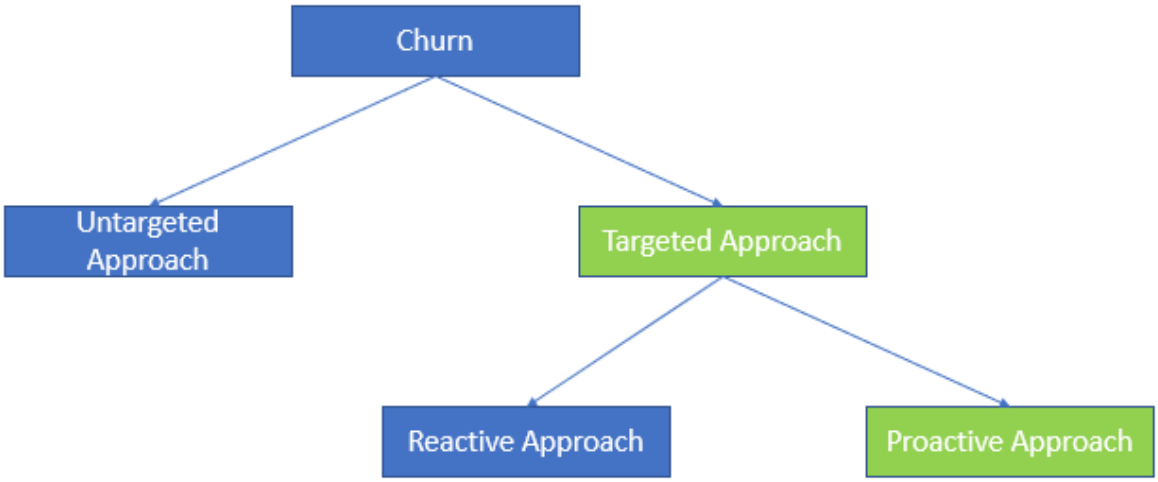


Telecommunication companies should observe strategies to mitigate the deliberate churns since there are many causes for deliberate churn to occur such as competitor's advertising strategies, promotional packages, seasonal packages, and poor customer care service of the service provider, poor service level agreements, and bad network quality, etc. The focus of this proposal is to overcome the deliberate voluntary churn of fiber to home customers.

Figure 2.2 depicts that there are two ways of mitigating customer churn namely untargeted and targeted approach. Relying on superior products and mass advertisement to increase brand loyalty and retain customers are categorized as an untargeted approach, while in a targeted approach, firms will identify customers who are likely to churn and offer direct incentives to avoid that to happen (Nafis et al.2017).

The targeted approach can be divided into two categories as reactive approach and proactive approach. In the reactive approach, the company waits until customers decide to disconnect the subscription. Accordingly, the company offers incentives to the specific customer. As an example, introducing cheaper packages, to retain the customer with the current service provider. For a proactive approach, the company deliberately observes and identifies the subscribers who would likely churn very soon. And then targets these bulk of customers with special packages or incentives to refrain them from churning.

FIGURE 2.2:Churn Mitigation Techniques



The proactive approach will not be effective in churn mitigation if the customers are not accurately classified, because a financial loss would occur when allocating budget for wrongly targeted customers. Thus, customer churn classification should be as accurate as possible. Strategies of the proactive approach and reactive approach are taken into the consideration throughout this research proposal.

Multiple service requirements over a single access system might be a requirement of many customers. As an example, Sri Lankan customers usually purchase high-speed internet and IPTV over the same fiber connection without any preference for separate access systems for additional services. Thus, it is vital to observe the association between services which can be offered to customers using a single access system to mitigate the churn factor. (Such as people that buy X service type also tends to buy Y Service type).

Customer churn is affected due to various reasons in the service provider environment such as service quality, network coverage, congestion, billing issues, costs, infrastructure-related

issues, and limitation of technologies. Customer churn is a binary classification problem since there are two possibilities for dependant variables either customer churning or not. Various classification techniques can be used to address this binary classification problem. Höppner's 'Profit-driven decision trees for churn prediction' reveals how decision trees can be used for binary classification problems (Höppner et al. 2020). Here they have built up a profit-based model where the profit concerns are directly integrated with the model classifier. Adnan. et al used various data transformation techniques in analyzing data related to the telecommunication sector and evaluated the performance of the underlying classifier (Amin et al. 2019). They have used Naive Bayes (NB), K-Nearest Neighbour (KNN), Gradient Boosted Tree (GBT), Single Rule Induction (SRI), and Deep learner Neural net (DP) as underlying classifiers. AdaBoost is another classification technique, which can be used with particle swarm optimization-based undersampling method to enhance the model accuracy (Amin et al. 2019). Association rule is an unsupervised learning Model which does not contain a direct prediction model and it detects associations between discrete events, products, or attributes (Mitkees et al. 2017). Customer churn is also predicted by Srivastava in his paper 'Performance Analysis of Hoeffding and Logistic Algorithm for Churn Prediction in Telecom Sector' (Bhatnagar et al. 2020). In this paper logistic algorithm is used to address the binary classification problem of churning.

In constructing a binary classification model imbalance of the dependent variable should be discussed. Class imbalance presents significant challenges to customer churn prediction such that they may classify all instances into the majority class, resulting in overall high accuracy but unacceptably low precision to the minority class of interest (Zhu et al. 2018). Particle swarm optimization(PSO) based under-sampling method is one of the solutions for data imbalance in data mining (Adris et al. 2017). Telecom datasets generally have fewer instances of churner class and most of the dataset comprises of non-churner instances [7]. The imbalance class distribution present in telecom datasets mostly results in low prediction performance of classification algorithms. The notion that only duplicating minority class through random oversampling or discarding majority class using random under-sampling may not improve prediction results (Adris et al. 2017). Adris used telecommunication data set which comprises 50,000 instances where there are only 3276(0.06% portion from whole instances amount) churner instances. After applying PSO-based under-sampling, the training dataset has a balanced class distribution with an equal number of churners and non-churners (Adris et al. 2017). Mitigation of the data imbalance using over-sampling (ROS), random under-sampling (RUS), and synthetic minority

oversampling technique (SMOTE), and increasing the model accuracy is discussed in Zhu's 'Investigating Decision Tree in Churn Prediction with Class Imbalance' paper (Zhu et al. 2018). Random resampling techniques including random oversampling and random undersampling, are easy to understand and use, but they also produce some side effects such as overfitting or information loss by duplicating or deleting examples from the training sets (Pan et al. 2018). Several New techniques such as the synthetic minority oversampling technique (SMOTE) are introduced to overcome these sampling issues which creates new examples for the minority class inferred from existing examples (Pan et al. 2018). Hence when addressing the class imbalance in churn prediction analysis we should be careful in selecting appropriate sampling techniques such that they may not overfit, underfit, or causes information loss.

These researches have mainly focused on the dependant variable of binary classification problems and improving the accuracy of the binary classification model. Proactive approaches to mitigate potential churnings using the churn classification model are not discussed in these researches. Telecommunication companies can retain their customers if they can proactively monitor the potential churners and fix the parameters of network-level to uplift the customer's user experience. Thus, in this work, we propose a churn prediction model where independent variables are the network level parameters and the dependant variable is the binary classification problem of churning. Using this framework we can predict potential churners and proactively enhance degraded network parameters. Furthermore, the customer's buying pattern is observed to build business strategies to retain potential churners.

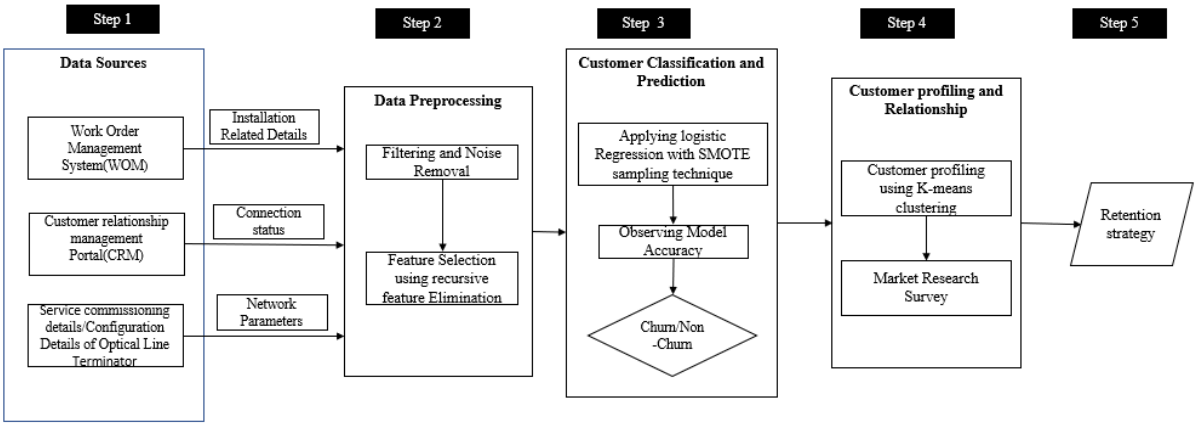
CHAPTER 3

METHODOLOGY

3.1 Proposed Architecture

The following figure3.1 depicts the architecture of the proposed churn prediction model.

FIGURE 0.1: Proposed Model for Churn Prediction and Churn Mitigation



The proposed model comprises five main steps. These steps are explained in detail below.

3.1.1 Step 1

In the first step, required data is retrieved from data sources. Retrieval of data is explained in detail in the feasibility study in section 1.6.

3.1.2 Step 2

In the second step, extracted data are preprocessed. Processed dataset splits to Training and Test data set. Recursive Feature Elimination Technique is applied to Training data set to identify best and worst-performing features. These features are selected as input variables for the model.

3.1.3 Step 3

In the third step, the logistic regression classification algorithm is used for the binary classification. Class imbalance of predictive variable is addressed with Synthetic Minority Oversampling Technique (SMOT) by oversampling the minority class. Model performance is evaluated using the confusion matrix, recall,f-measure, support, and ROC curve.

3.1.4 Step 4

In the fourth step, customer profiling is performed using k-means clustering techniques. Cluster analysis is based on the patterns of network parameters captured from the data. Further, a market research survey is carried out to observe why people are unhappy with fiber to home services.

3.1.5 Step 5

In the final step, the model recommends retention strategies for churners. Results of step 4 are used to develop a retention strategy in Step 5. These retention strategies contain proactive approaches which can be applied on the network infrastructure layer as well as business strategies.

3.2 Rationales to select algorithms for the proposed model

3.2.1 Step -1: The logistic regression classification algorithm

This research addresses a binary classification problem. Here we need to explain the relationship between the dependent binary variable with categorical form (Churn or non-Churn) and multiple nominal independent variables(Network Parameters). Logistic regression is a machine learning algorithm that is used as a supervised learning technique for classifying binary classification problems. The proposed system uses logistic regression model in classifying churners and non-churners.

3.2.2 Step -1: Feature selection using Recursive Feature Elimination Method

It is important to find out significant features of the Training data set to reduce the complexity of the model and to reduce the time complexity for a model to get trained and finally to result in a more accurate model without overfitting or underfitting. Recursive Feature Elimination (RFE) is a feature selection technique. A ranking of features with the corresponding accuracy is produced through RFE. RFE is an efficient technique for eliminating features from a training dataset for feature selection. thus, RFE is used in this research.

3.2.3 Step -3: Synthetic Minority Oversampling Technique (SMOT)

Class imbalance is a major concern for churn predictive models since the number of instances of churners is usually low. For this study, it is proposed an over-sampling approach in which

the minority class is over-sampled by creating 'synthetic samples'. This is a data augmentation for the minority class and it's called Synthetic Minority Oversampling Technique.

3.2.4 Step 4:K-means algorithm

K-means clustering is an unsupervised learning algorithm. The proposed system uses K- mean algorithm to profile the churn customer data into groups based on available transformed data.

3.3 Dataset description

In this study, the dataset is obtained from Dialog Axiata PLC. We have used 14,461 instances with fourteen network-related feature parameters for developing a churn prediction model, in which some of the features are numerical and some of the features are categorical. The data are extracted from four data sources, namely network elements(optical line terminator-OLT, ONT-Optical Network Terminator), Work order management system(WOM), customer relationship management system(CRM), and customer fault management system(JIRA). The below table gives an insight into extracted features from preceding data sources. It contains labeled data with two classes where 90.3% data are labeled as billing connected customers that represent non-churners and 9.6% data are labeled as billing disconnected customers that represent churners.

Table 3. 3:Feature List

Work order management system(WOM)	**Customer fault management system(JIRA)**	**Network Element Parameters**	**customer relationship management system(CRM)**
Date of the New installations	The root cause of the Fault	Optical Line Terminator Model	billing Status of the Customer- This shows whether a customer is billing connected/Temporary disconnected/Permanently disconnected.
Optical performance details of Optical Network Terminator(ONT) at the installation stage(signal to noise ratio, downstream power levels)	Number of optical network termination units per passive optical port when the fault is handled by the agent	Distance between OLT and ONT(km)	Circuit Number of the customer - unique number to index the customer

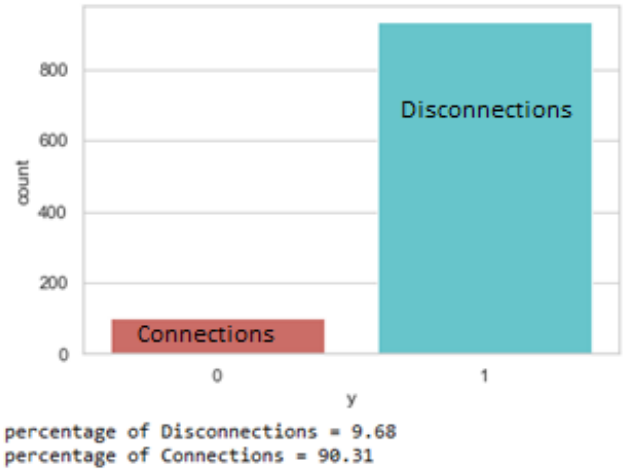
	Number of Service request per User (Number of Fault Escalations)	Receiving power(dbmv) of ONT	Monthly bandwidth consumption of customer-This relates to customers with internet packages. The monthly data consumption of a particular user can be exported and analyzed.
		Transmitting power(dbmv) of ONT	IPTV subscription details of the customer -This reveals the type of IPTV package the user has subscribed to(Low cost, Medium Cost, high cost)
		Total service count per individual ONT	Voice subscription details of the customer-This reveal the amount of subscribed voice connections per user
		High-speed internet service count per individual ONT	Services provisioned for individual users-This shows the number of services commissioned per subscriber.
		Voice service count per individual ONT Total Original Ont count per port	
		Television over IP service count per individual ONT	
		Total Third-party ont count per port-	
		Total ONT count per OLT port	

CHAPTER 4

EVALUATION AND RESULTS

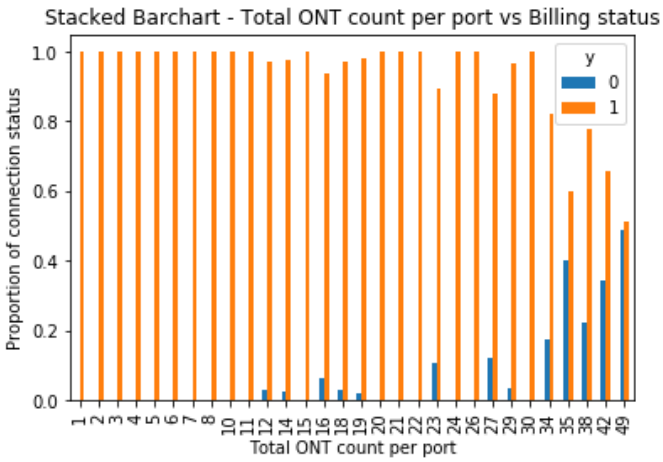
4.1 Data visualization and descriptive statistics

FIGURE 4.1:Class imbalance of predictive Variables



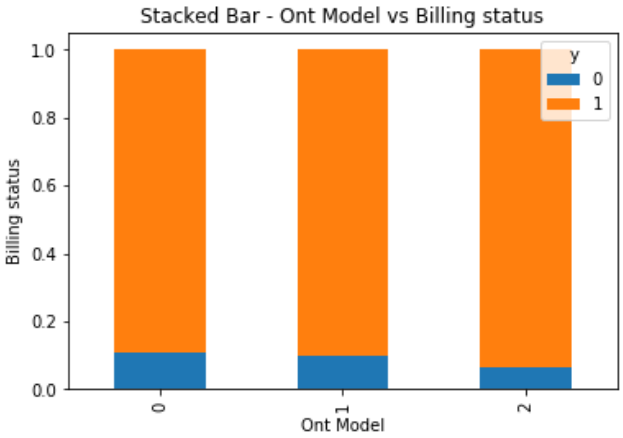
As it is shown in figure 4, classes are imbalanced as the percentage of disconnections is 9.69% and the percentage of connections is 90.31%.

FIGURE 4.2: Total ONT count per Port



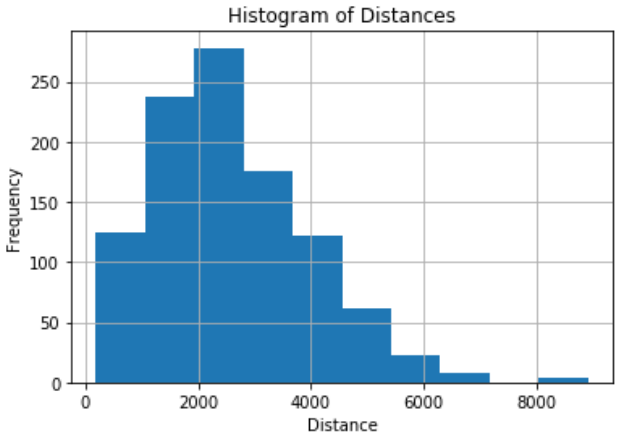
According to the bar chart presented in Figure 4.2, there is a gradual increment of the billing disconnected proportion as the total optical network terminator count per port get increases. Thus, the total optical network terminator count per port can be a good predictor of the outcome variable.

FIGURE 4.3: Ont Model vs Billing Status



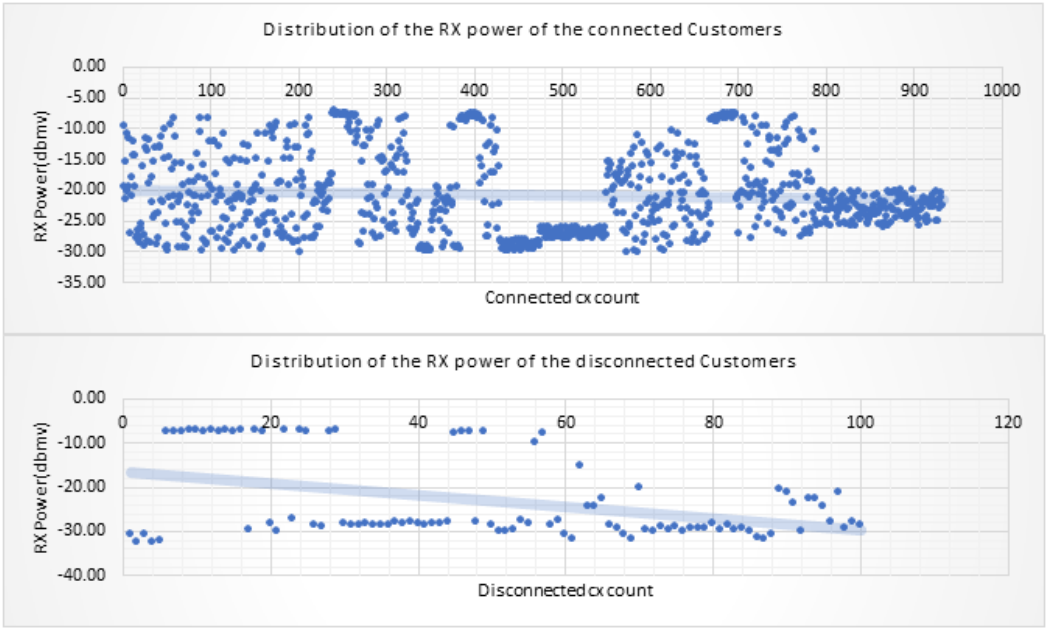
As shown in figure 4.3, slightly higher disconnections are reported for optical network terminator model 0 and optical network terminator model 1 compared to network terminator model 2. There would be a slight association between billing status and the optical network terminator model.

FIGURE 4.4: Histogram of Distance between OLT and ONT



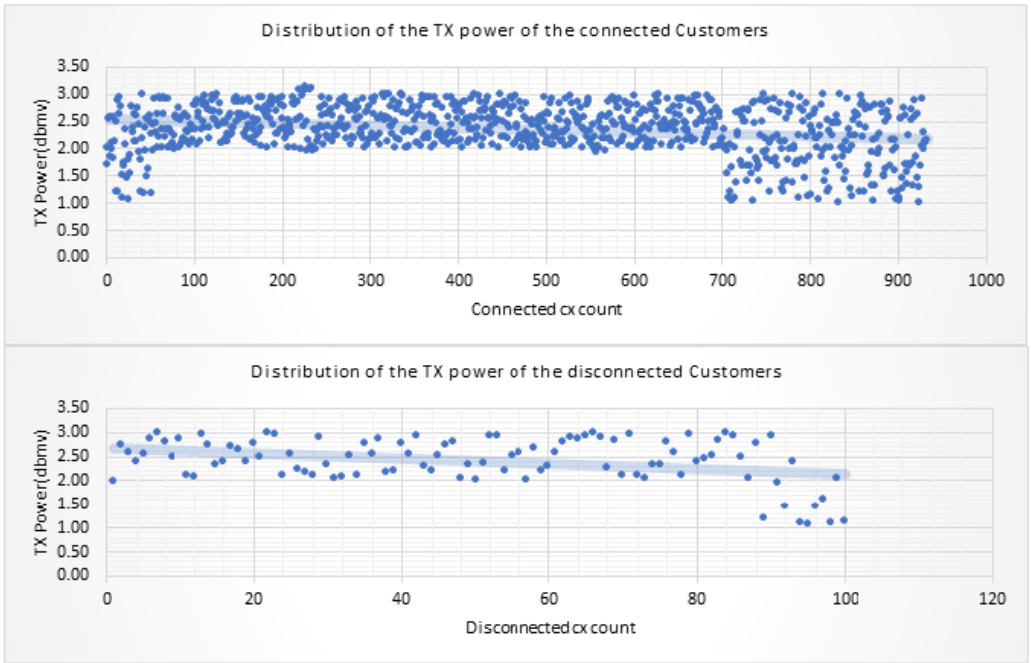
According to figure 4.4, the distance between OLTs and ONTs mostly lies between 2000m and 3000m.

FIGURE 4.5: Distribution of the ONT's Receiving Power vs Billing State



The distribution of the customer edge device's(Optical Network Terminator-ONT) receiving power is shown with respect to billing status in Figure 4.5. Billing connected customer's receiving power has scattered between -29dbmv and -8dbmv while disconnected customer's receiving power tends to scatter only around -8dbmv and -30dbmv. hence receiving power attribute would be a good predictor of the outcome variable.

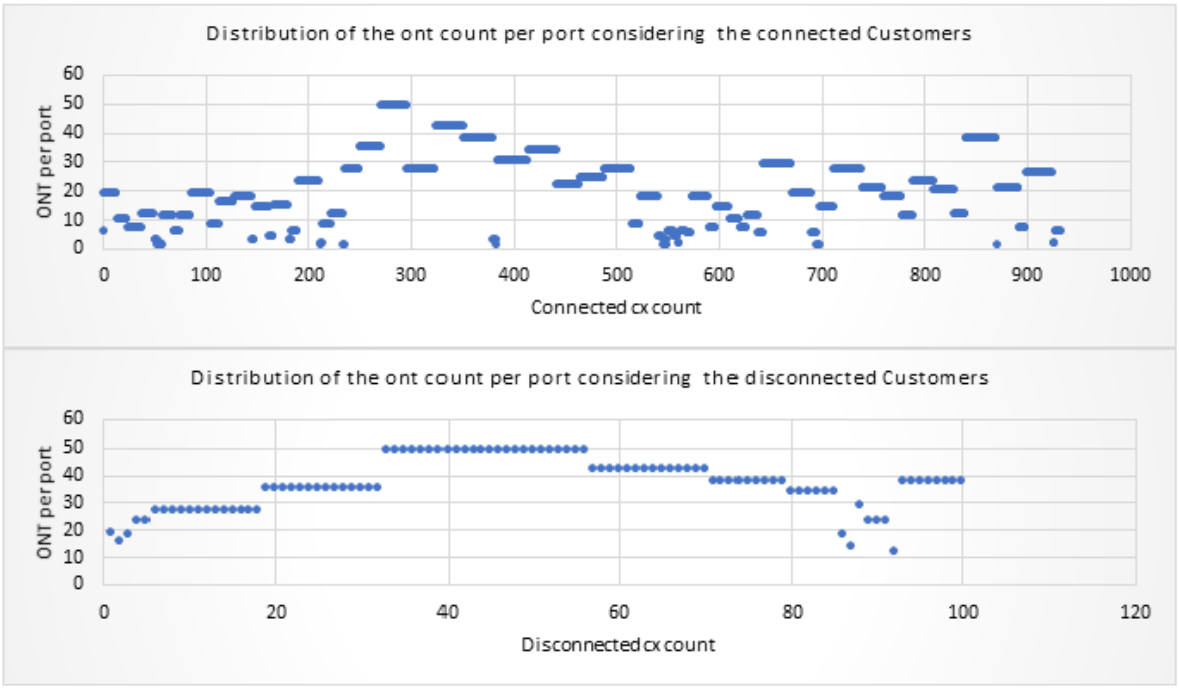
FIGURE 4.6: Distribution of the transmitting power of ONT vs Billing Status



As shown in Figure 4.6 transmitting power of the Optical network terminators of disconnected and connected customers scattered between 2dbmv and 3 dbmv in the same

manner, but some outliers are scattered till 1dbmv.it is a bit hard to observe an association between transmitting power attribute with billing status.

FIGURE 4.7: Distribution of the Total ONT count Per-port vs Billing Status



According to Figure 4.7, scatters of disconnected cx's scatter plot get sparse to dense when the Optical network terminator count per port grows from 10 onwards. And the highest scatter density is shown when ONT per port count is between 50 and 30.and the connected customer's scatter density is high between 1 to 35. hence this attribute might have an association with the outcome variable. Descriptive Statistics of the Distance between OLT and ONT of disconnected and connected subscribers are shown in Figure 4.8.

FIGURE 4.8: Descriptive Statistics of Distance between OLT and ONT considering customer's billing status

Descriptive Statistics of Distance for connected customers	
Mean	**2580.957128**
Standard Error	45.33980535
Median	2458
Mode	2496
Standard Deviation	1384.906563
Sample Variance	1917966.187
Kurtosis	1.280134542
Skewness	0.823466217
Range	8734
Minimum	186
Maximum	8920
Sum	2408033
Count	**933**

Descriptive Statistics of Distance for disconnected customers	
Mean	**2881.97**
Standard Error	126.6974639
Median	3027.5
Mode	2394
Standard Deviation	1266.974639
Sample Variance	1605224.736
Kurtosis	-0.944757723
Skewness	-0.283608938
Range	4579
Minimum	304
Maximum	4883
Sum	288197
Count	**100**

4.2 Noise removal

It is vital for making data useful because noisy data can ultimately lead to a less accurate model. For the proposed research, it is mainly used network-related parameters which are directly fetched from the network element called 'optical line terminator'. The main challenge is to retrieve the network-related parameters of the disconnected customers where the edge device is already removed. As a solution for this, network-related information on user acceptance test reports and escalated faults are referred. Using those references, it was prepared a disconnected customer database without any null values. Data are collected and processed using python web scraping techniques and excel software.

4.3 Attribute Information (Transformation of categorical data to numerical data)

The information of retrieved attributes is shown in the following Table

Table 4.1:Information of Attributes

Data Set Characteristics:	Multivariate	Number of instances	14461
Attribute Characteristics:	Categorical, Integer	Number of Attributes:	14
Associate tasks	Classification/Clustering	Missing Values?	No
Dependent variable	Billing Status(Y)	Independent variable count	13

The feature list is as follows.

1. Customer Reference - **Numerical**
2. ONT Model (Huawei 045H=**0**,Telpo MONUH116=**1**,Huawei EG8145V=**2**)- **Categorical**
3. Distance between OLT and ONT(km)-**Numerical**
4. Receiving power(dbmv) - **Numerical**
5. Transmitting Power(dbmv)- **Numerical**
6. Total ONT count per OLT port- **Numerical**
7. Total Third party ont count per port- **Numerical**
8. Total Original Ont count per port - **Numerical**
9. Television over IP service count -**Numerical**
10. Voice service count - **Numerical**

11. High speed internet service count - **Numerical**
12. Total Service count - **Numerical**
13. Multiple Service /Single Service (Multiple service=1, Single service=0)- **Categorical**
14. Billing Status(connected=1/Disconnected=0) -**Categorical**

4.4 Accomplishments

4.4.1 Over-sampling using Synthetic Minority Oversampling Technique

FIGURE 4.9: Oversampled Data Set

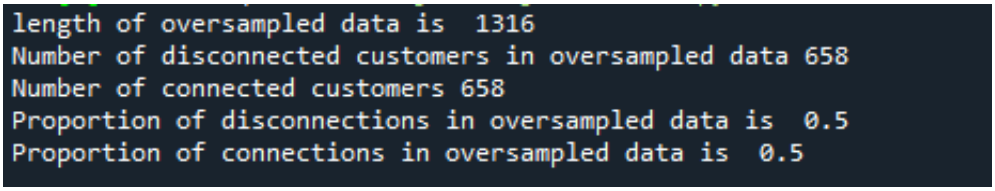
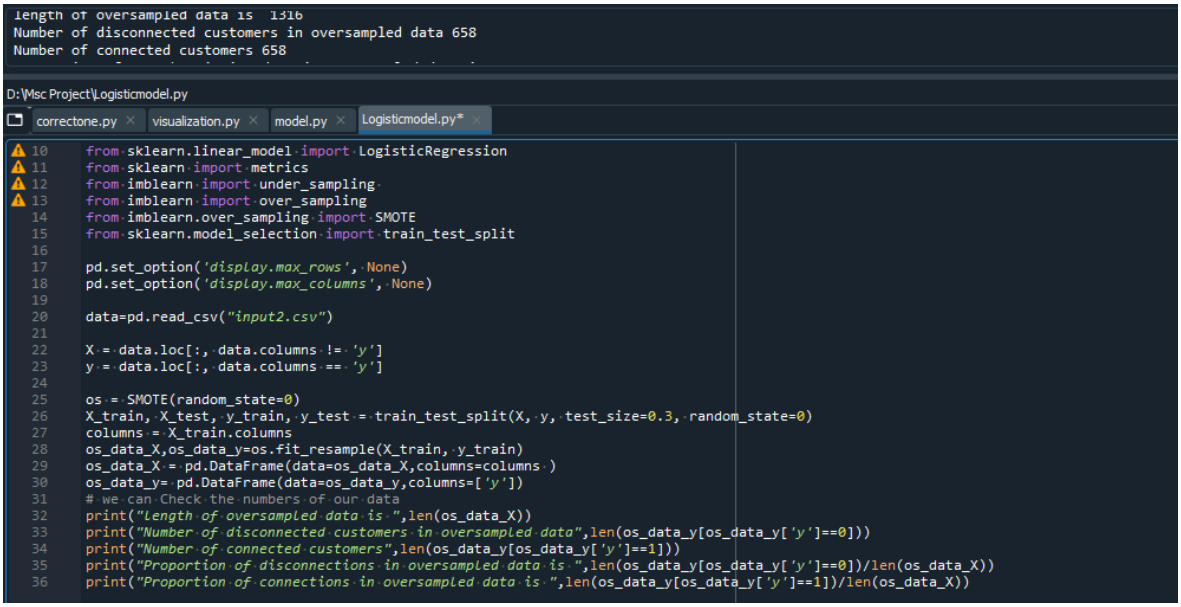
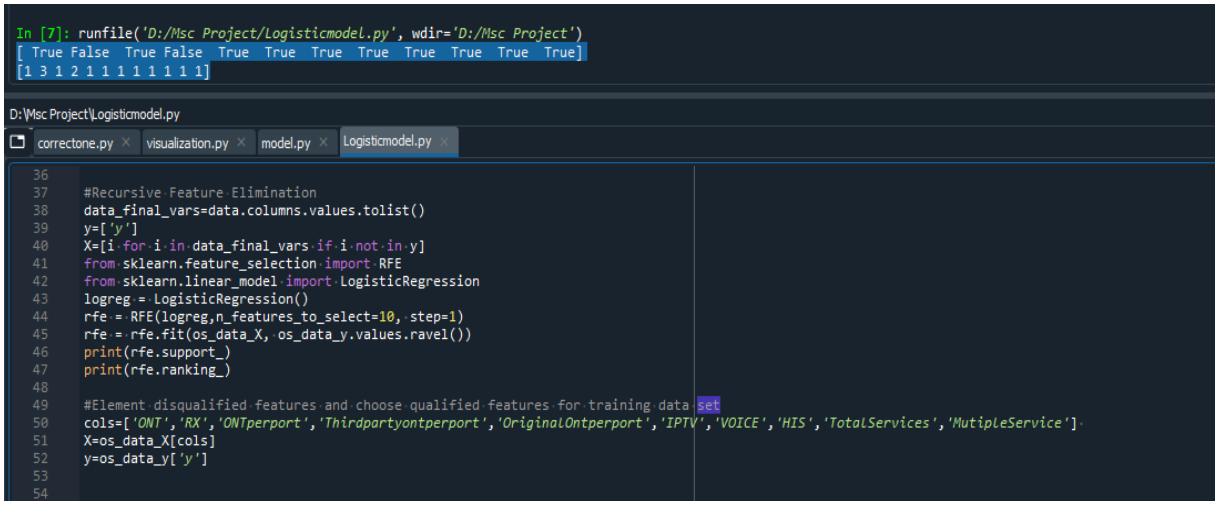


FIGURE 4.10: Python Code for SMOTE



4.4.2 Recursive Feature Elimination

FIGURE 4.11: Feature Selection and Related Python Codes



```
In [7]: runfile('D:/Msc Project/Logisticmodel.py', wdir='D:/Msc Project')
[ True False  True False  True  True  True  True  True  True  True  True]
[1 3 1 2 1 1 1 1 1 1 1 1]

D:\Msc Project\Logisticmodel.py
```

correctone.py | visualization.py | model.py | Logisticmodel.py

```
36
37  #Recursive Feature Elimination
38  data_final_vars=data.columns.values.tolist()
39  y=['y']
40  X=[i for i in data_final_vars if i not in y]
41  from sklearn.feature_selection import RFE
42  from sklearn.linear_model import LogisticRegression
43  logreg = LogisticRegression()
44  rfe = RFE(logreg,n_features_to_select=10, step=1)
45  rfe = rfe.fit(os_data_X, os_data_y.values.ravel())
46  print(rfe.support_)
47  print(rfe.ranking_)
48
49  #Element disqualified features and choose qualified features for training data set
50  cols=['ONT','RX','ONTperport','Thirdpartyontperport','OriginalOntperport','IPTV','VOICE','HIS','TotalServices','MutipleService']
51  X=os_data_X[cols]
52  y=os_data_y['y']
53
54
```

4.4.3 Implementing the model

FIGURE 4.12: Summary Statistics of the Model

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.324352
         Iterations: 35
                                Results: Logit
================================================================================
Model:                    Logit                 Pseudo R-squared:       0.532
Dependent Variable:       y                     AIC:                    873.6940
Date:                     2021-02-28 21:13      BIC:                    925.5175
No. Observations:         1316                  Log-Likelihood:         -426.85
Df Model:                 9                     LL-Null:                -912.18
Df Residuals:             1306                  LLR p-value:            3.6346e-203
Converged:                0.0000                Scale:                  1.0000
No. Iterations:           35.0000
--------------------------------------------------------------------------------
                      Coef.     Std.Err.      z    P>|z|      [0.025       0.975]
--------------------------------------------------------------------------------
ONT                    1.4848       0.1633  9.0940 0.0000        1.1648       1.8048
RX                     0.0329       0.0111  2.9690 0.0030        0.0112       0.0546
ONTperport            -6.2411       1.0716 -5.8241 0.0000       -8.3414      -4.1408
Thirdpartyontperport   6.0918       1.0703  5.6917 0.0000        3.9940       8.1895
OriginalOntperport     6.1170       1.0707  5.7130 0.0000        4.0184       8.2156
IPTV                  35.9041 3098429.7553  0.0000 1.0000 -6072774.8248 6072846.6330
VOICE                 33.0791 3098429.7553  0.0000 1.0000 -6072777.6498 6072843.8080
HIS                   35.4388 3098429.7553  0.0000 1.0000 -6072775.2902 6072846.1677
TotalServices        -31.7511 3098429.7553 -0.0000 1.0000 -6072842.4800 6072778.9778
MutipleService        -3.1219       0.4206 -7.4232 0.0000       -3.9462      -2.2976
================================================================================
```

The p-values for most of the variables are smaller than 0.05, except for four variables. therefore, we will remove them.

FIGURE 4.13: Summary Statistics of the Model(2nd cycle)

```
Optimization terminated successfully.
         Current function value: 0.394467
         Iterations 10
                         Results: Logit
==================================================================
Model:                Logit            Pseudo R-squared: 0.431
Dependent Variable:   y                AIC:              1050.2368
Date:                 2021-02-28 21:13 BIC:              1081.3309
No. Observations:     1316             Log-Likelihood:   -519.12
Df Model:             5                LL-Null:          -912.18
Df Residuals:         1310             LLR p-value:      1.1601e-167
Converged:            1.0000           Scale:            1.0000
No. Iterations:       10.0000
------------------------------------------------------------------
                      Coef.  Std.Err.    z    P>|z|   [0.025  0.975]
------------------------------------------------------------------
ONT                   1.6463   0.1455 11.3161 0.0000  1.3611  1.9314
RX                   -0.0469   0.0079 -5.9586 0.0000 -0.0623 -0.0314
ONTperport           -6.0164   1.0800 -5.5707 0.0000 -8.1331 -3.8996
Thirdpartyontperport  5.8847   1.0791  5.4535 0.0000  3.7698  7.9996
OriginalOntperport    5.9658   1.0796  5.5260 0.0000  3.8499  8.0818
MutipleService        1.4421   0.1619  8.9084 0.0000  1.1249  1.7594
==================================================================
```

4.4.4 Logistic Regression Model Fitting

FIGURE 4.14: Python Code for Logistic Regression Model Fitting

```
Accuracy of logistic regression classifier on test set: 0.86
```

D:\Msc Project\Logisticmodel.py

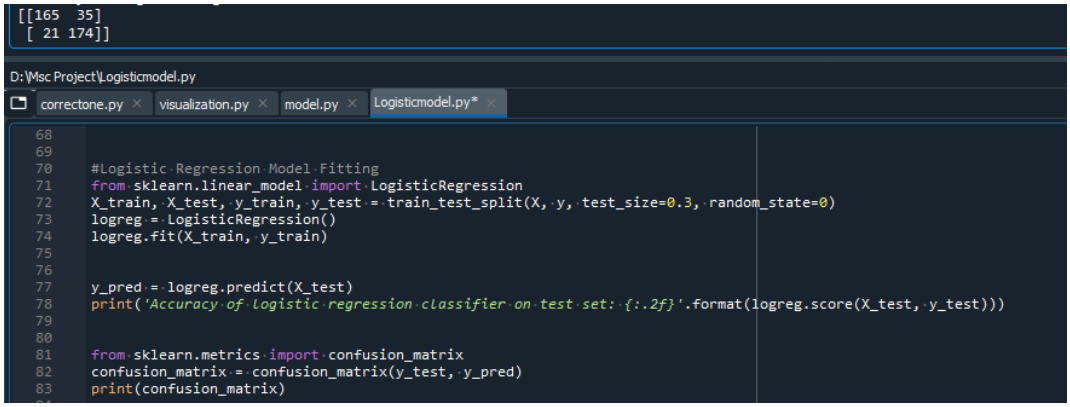
correctone.py | visualization.py | model.py | Logisticmodel.py

```
53
54  #observing the p value of the features
55  import statsmodels.api as sm
56  # logit_model=sm.Logit(y,X)
57  # result=logit_model.fit()
58  # print(result.summary2())
59
60  #observing the p value of the features after removing the feature where p values >0.05
61
62  cols1=['ONT','RX','ONTperport','Thirdpartyontperport','OriginalOntperport','MutipleService']
63  X=os_data_X[cols1]
64  y=os_data_y['y']
65  logit_model=sm.Logit(y,X)
66  result=logit_model.fit()
67  #print(result.summary2())
68
69
70  #Logistic Regression Model Fitting
71  from sklearn.linear_model import LogisticRegression
72  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
73  logreg = LogisticRegression()
74  logreg.fit(X_train, y_train)
75
76
77  y_pred = logreg.predict(X_test)
78  print('Accuracy of logistic regression classifier on test set: {:.2f}'.format(logreg.score(X_test, y_test)))
```

4.4.5 Confusion matrix

The result is telling us that we have ***165+174*** correct predictions and ***21+35*** incorrect predictions.

FIGURE 4.15: Evaluating Model Accuracy-Confusion Matrix



```
[[165  35]
 [ 21 174]]
```

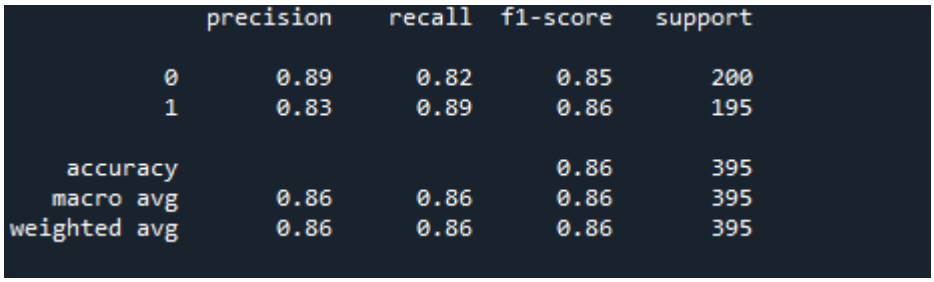
D:\Msc Project\Logisticmodel.py

correctone.py | visualization.py | model.py | Logisticmodel.py*

```
68
69
70  #Logistic Regression Model Fitting
71  from sklearn.linear_model import LogisticRegression
72  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
73  logreg = LogisticRegression()
74  logreg.fit(X_train, y_train)
75
76
77  y_pred = logreg.predict(X_test)
78  print('Accuracy of logistic regression classifier on test set: {:.2f}'.format(logreg.score(X_test, y_test)))
79
80
81  from sklearn.metrics import confusion_matrix
82  confusion_matrix = confusion_matrix(y_test, y_pred)
83  print(confusion_matrix)
```

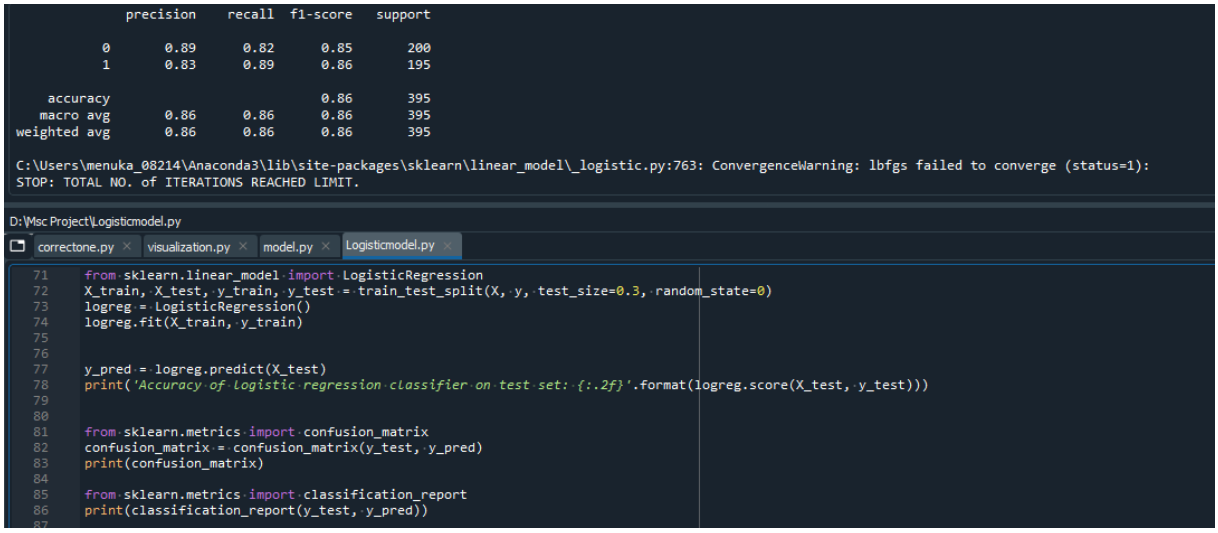
4.4.6 Compute precision, recall, F-measure, and support

FIGURE 4.16: Evaluating Model Accuracy - Precision,Recall,F1-Score,Support



	precision	recall	f1-score	support
0	0.89	0.82	0.85	200
1	0.83	0.89	0.86	195
accuracy			0.86	395
macro avg	0.86	0.86	0.86	395
weighted avg	0.86	0.86	0.86	395

FIGURE 4.17: Evaluating Model Accuracy - Precision,Recall,F1-Score,Support(Python Code)



```
              precision    recall  f1-score   support

           0       0.89      0.82      0.85       200
           1       0.83      0.89      0.86       195

    accuracy                           0.86       395
   macro avg       0.86      0.86      0.86       395
weighted avg       0.86      0.86      0.86       395

C:\Users\menuka_08214\Anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:763: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
```

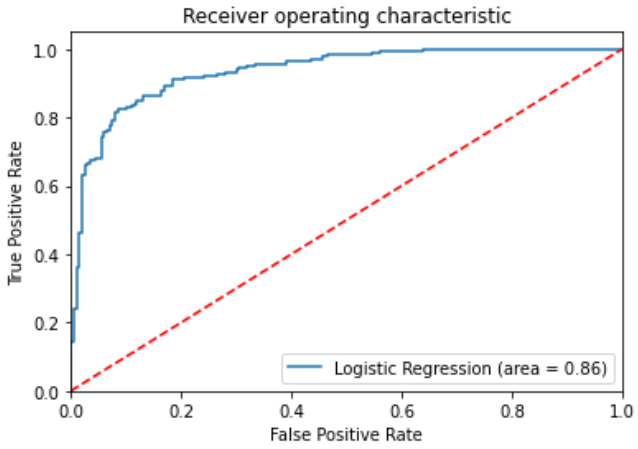
D:\Msc Project\Logisticmodel.py

correctone.py | visualization.py | model.py | Logisticmodel.py

```
71  from sklearn.linear_model import LogisticRegression
72  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
73  logreg = LogisticRegression()
74  logreg.fit(X_train, y_train)
75
76
77  y_pred = logreg.predict(X_test)
78  print('Accuracy of logistic regression classifier on test set: {:.2f}'.format(logreg.score(X_test, y_test)))
79
80
81  from sklearn.metrics import confusion_matrix
82  confusion_matrix = confusion_matrix(y_test, y_pred)
83  print(confusion_matrix)
84
85  from sklearn.metrics import classification_report
86  print(classification_report(y_test, y_pred))
```

4.4.7 ROC Curve

FIGURE 4.18: ROC Curve



4.4.8 Performing customer profiling using k-means based on network parameters

Based on the elbow plot shown in Figure 4.19, it is chosen six as an optimal number of clusters to initiate the Kmeans clustering for Customer segmenting using Network parameters.

FIGURE 4.19: Optimal Number of Cluster selection using elbow method

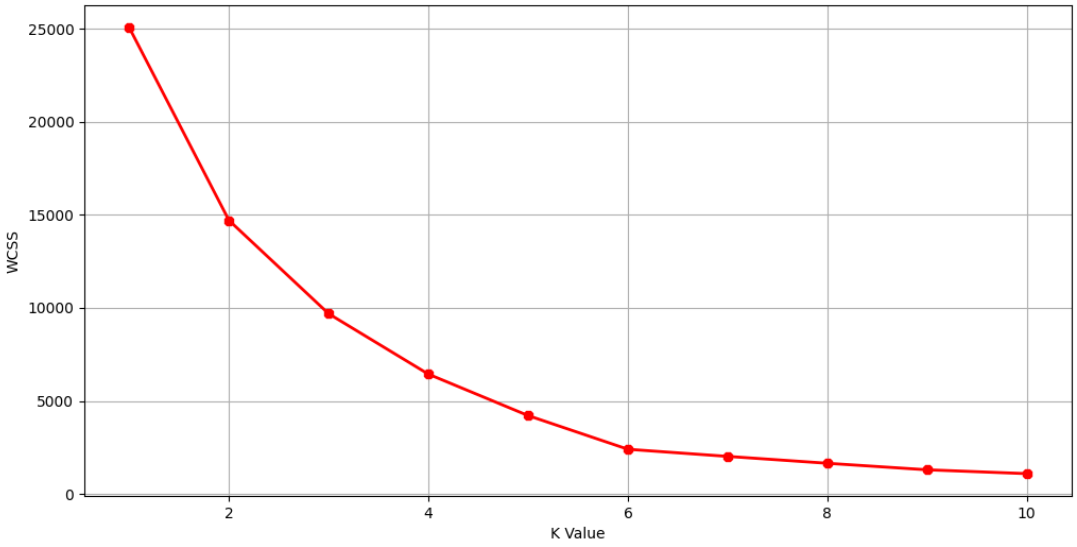


FIGURE 4.20: Customer segmentation based on Receiving Power and Transmitting power of ONT

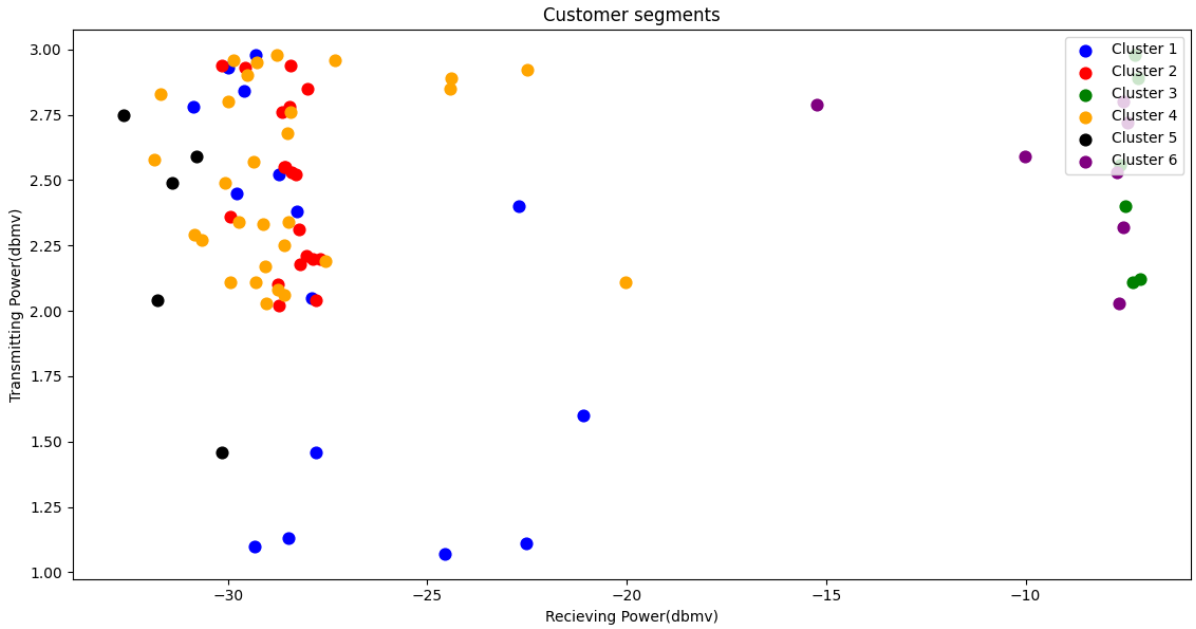


Figure 4.20 shows the distribution of the six clusters. it is quite hard to segregate these six clusters by giving meaningful interpretation individually. Nevertheless, cluster density gets higher as transmitting power increases from 1.75dbmv and receiving power decreases from -28 dbmv. Hence it could interpret cx segmentation as below.

Cluster1: Customers with low receiving power(< - 28dbmv) and High transmitting Power(> +2 dbmv)

FIGURE 4.21: Customer segmentation based on third party ont count per port and Distance between OLT and ONT

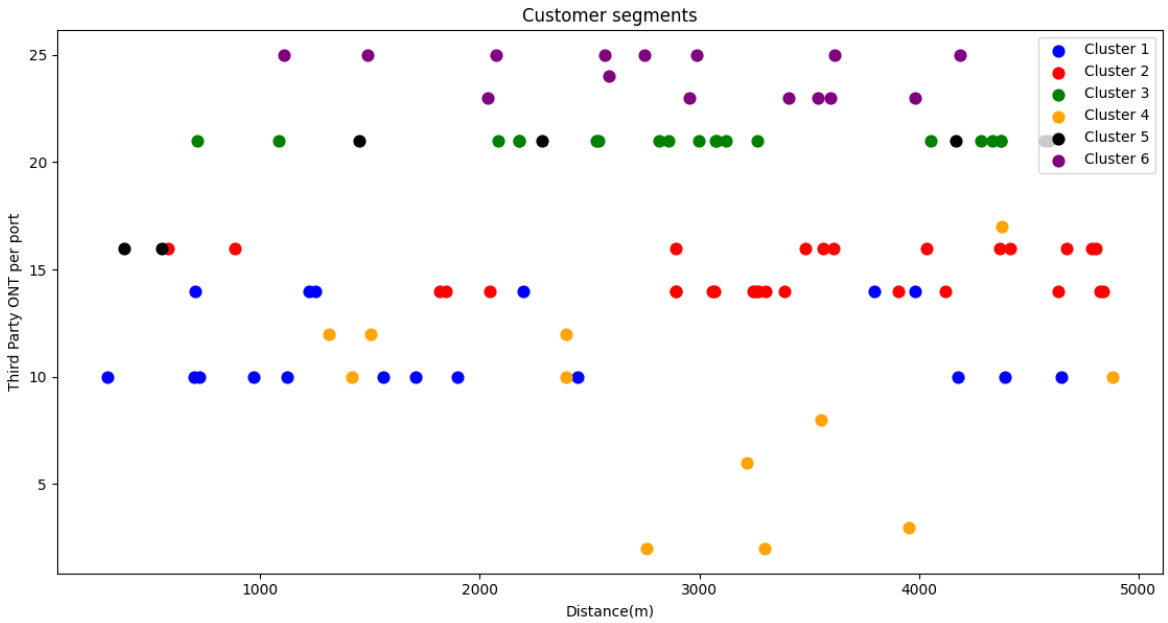


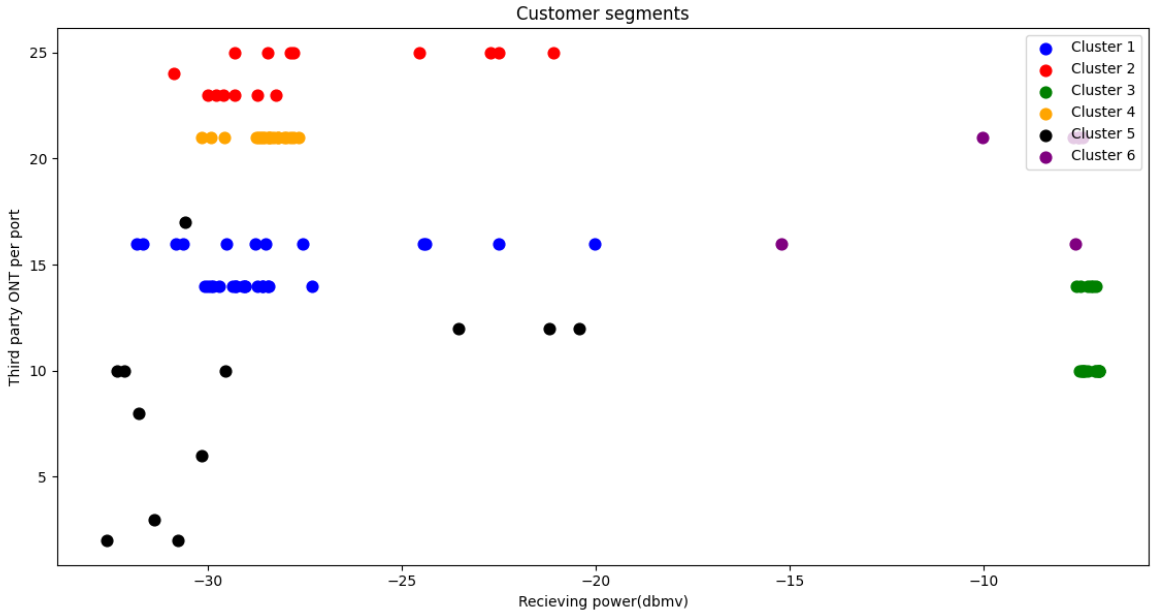
Figure 4.21 shows a distribution of 6 clusters, where we can interpret only cluster six, cluster two, and cluster three in a meaningful way.

Cluster 2: low number of third-party ONTs per port with distance from 3000m to 5000m

Cluster 3: average number of third-party ONTs per port with distance from 3000m to 5000m

Cluster 6: high number of third-party ONTs per port with distance from 3000m to 5000m

FIGURE 4.22: Customer segmentation based on third party ont count per port and Receiving power of ONT



According to figure 4.22, clusters can be interpreted as below.

Cluster 1:Average third party ont count per port and receiving power (< -28 dbmv)

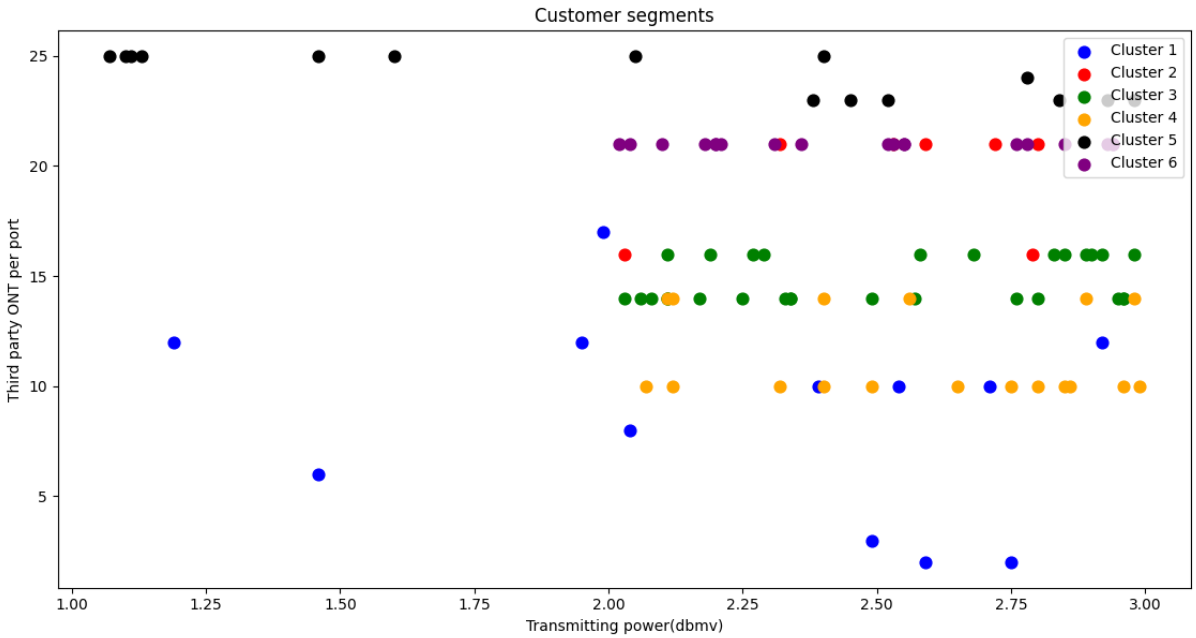
Cluster 2, Cluster 4:High third party ont count per port and receiving power (< -28 dbmv)

Cluster 3: Average third party ont count per port and receiving power (>-7 dbmv)

Cluster 5: Low third party ont count per port and receiving power (< -28 dbmv)

Cluster 6: Cannot be interpreted.

FIGURE 4.23: Customer segmentation based on third party ont count per port and transmitting power of ONT



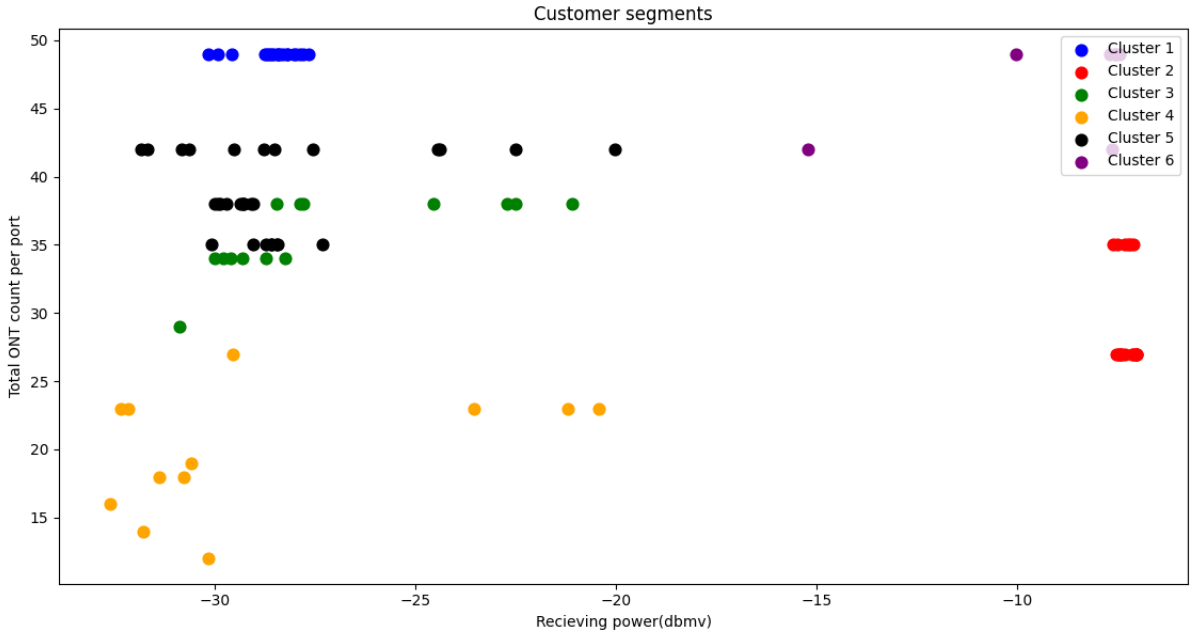
According to figure 4.23, clusters can be interpreted as below.

Cluster 1, Cluster 4: Low third party ont count per port and high transmitting power (>2dbmv)

Cluster 2, Cluster 6, Cluster 5: High third party ont count per port and high transmitting power (> 2 dbmv)

Cluster 3: Average third party ont count per port and high transmitting power (> 2 dbmv)

FIGURE 4.24: Customer segmentation based on total ont count per port and receiving power of ONT



According to figure 4.24, clusters can be interpreted as below.

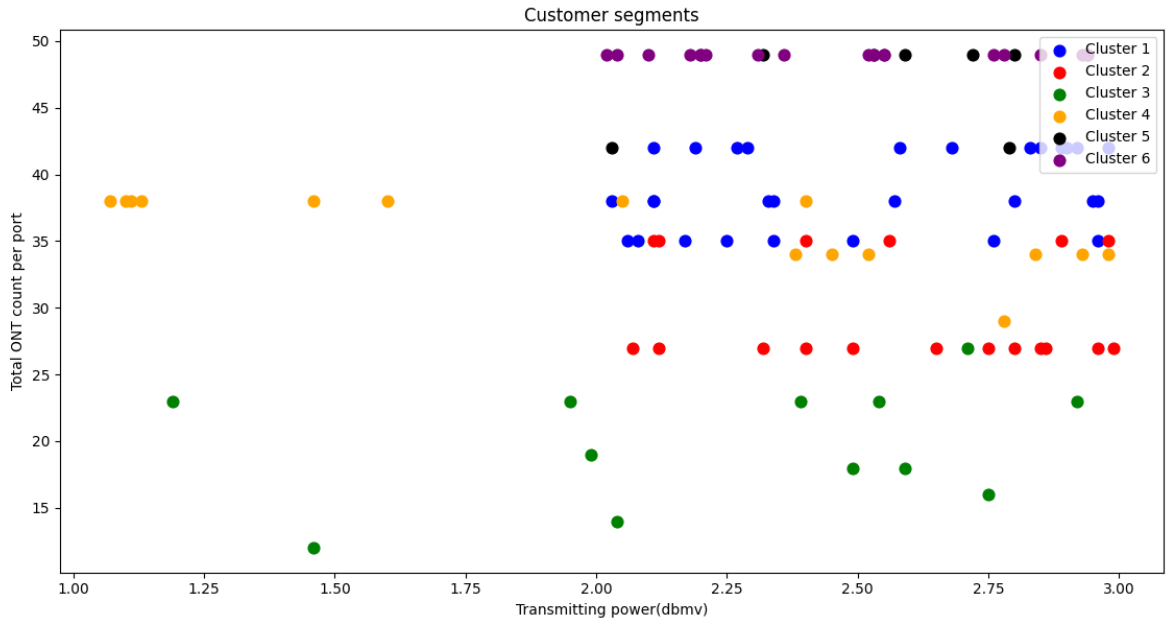
Cluster 1: High Total ont count per port and receiving power (< -28 dbmv)

Cluster 2: Average Total ont count per port and receiving power (> -8 dbmv)

Cluster 3, Cluster 5: Average Total ont count per port and receiving power (< -28 dbmv)

Cluster 4: Low total ont count per port and receiving power(< -28 dbmv)

FIGURE 4.25: Customer segmentation based on total ont count per port and transmitting power of ONT



According to figure 4.25, clusters can be interpreted as below.

Cluster 5, Cluster 6: High Total ont count per port and high transmitting power (> 2 dbmv)

Cluster 1, Cluster 4: Average Total ont count per port and high transmitting power (> 2 dbmv)

Cluster 2, Cluster 3: Low Total ont count per port and high transmitting power (> 2dbmv)

4.4.9 Market research survey

A Market Research was carried out for disconnected customers. Table 2.2 shows the questionnaire which was raised at the disconnected customers. This questionnaire is created by analyzing the fault escalations of disconnected customers. These Survey results are used to develop business strategies for churn mitigation.

Table 4.2: Questions for Market Research Survey

Satisfy with the Speed of the connection?	Excellent=10 Average=5 poor=1 (Customer can choose desired weight from 1 to 10)
Is there are any sudden drops with the internet connection?	Often=10 Average=5 seldom=1 (Customer can choose desired weight from 1 to 10)

how often TV Related issues occur?	Often=10 Average=5 seldom=1 (Customer can choose desired weight from 1 to 10)
How often Voice Dropping issues occur?	Often=10 Average=5 seldom=1 (Customer can choose desired weight from 1 to 10)

4.4.10 Performing customer profiling using k-means based on survey results

Based on the elbow plot shown in Figure 4.26, it is chosen two as an optimal number of clusters to initiate the Kmeans clustering for Customer segmenting using Network parameters.

FIGURE 4.26: Optimal Number of Cluster selection using elbow method

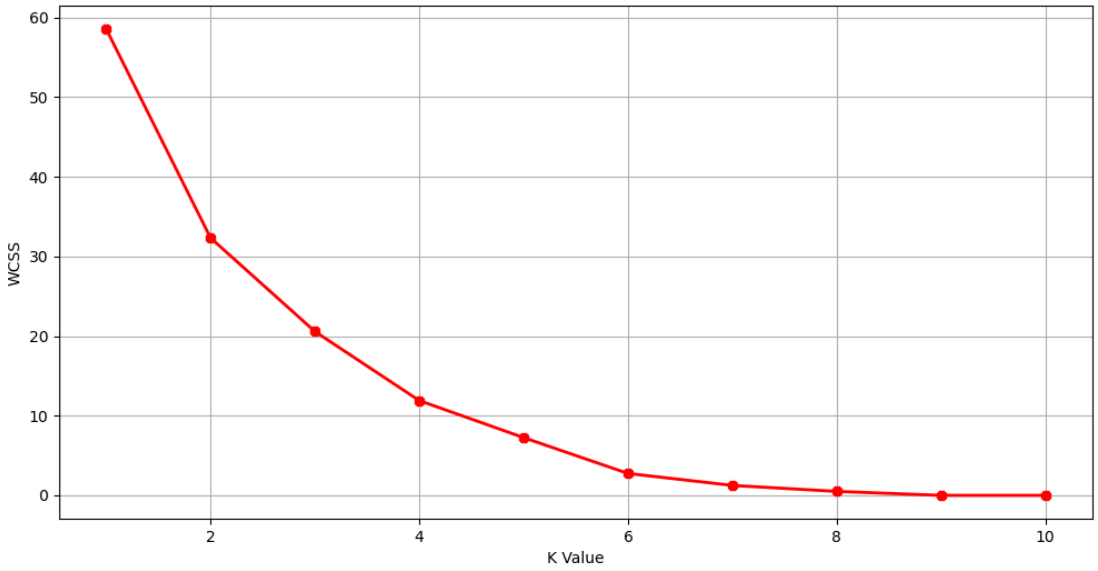
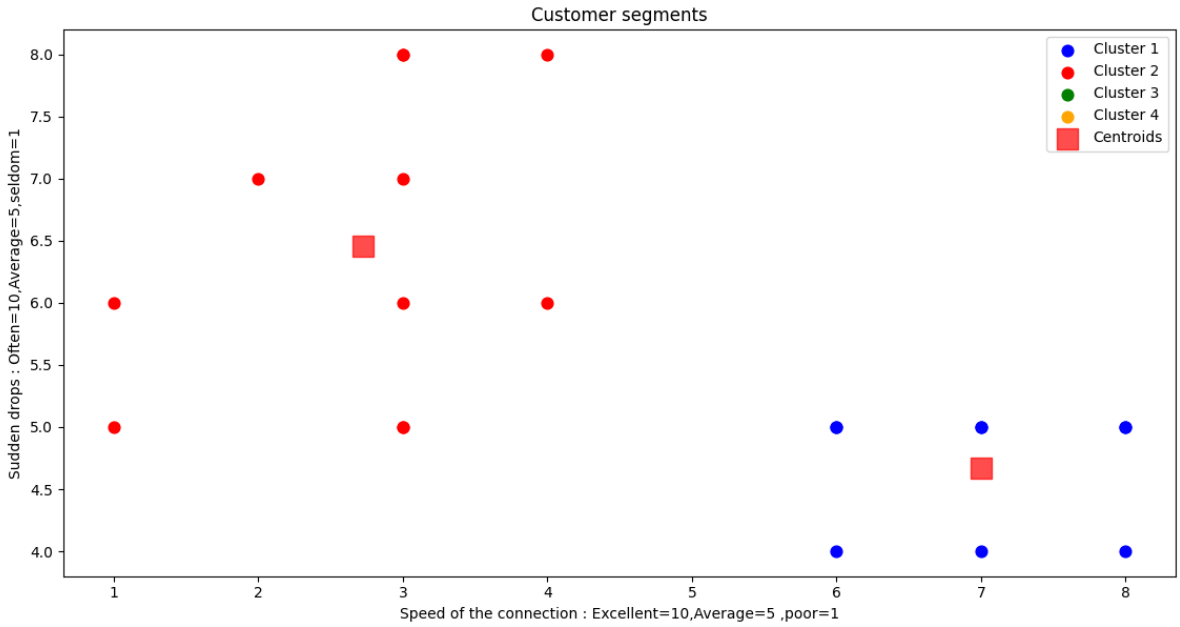


FIGURE 4.27: Customer segmentation based on survey results

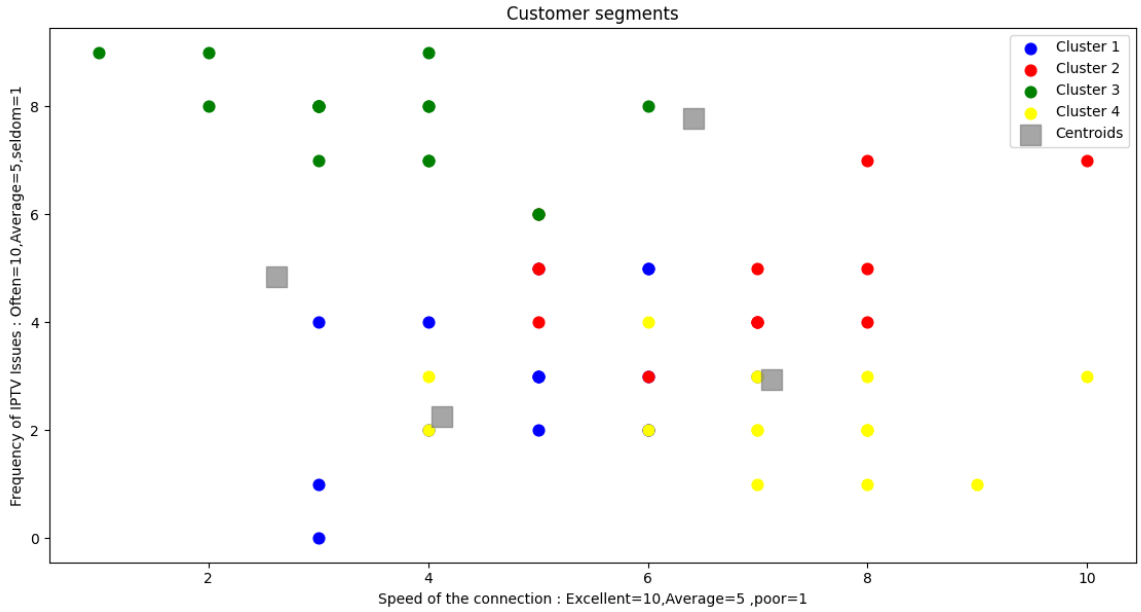


According to Figure 4.27 clusters are identified as below.

Cluster 2: Customer experiences severe drops in internet connection with low internet speed.

Cluster 1: Customer experiences drops in low rate with high internet speed.

FIGURE 4.28: Customer segmentation based on survey results

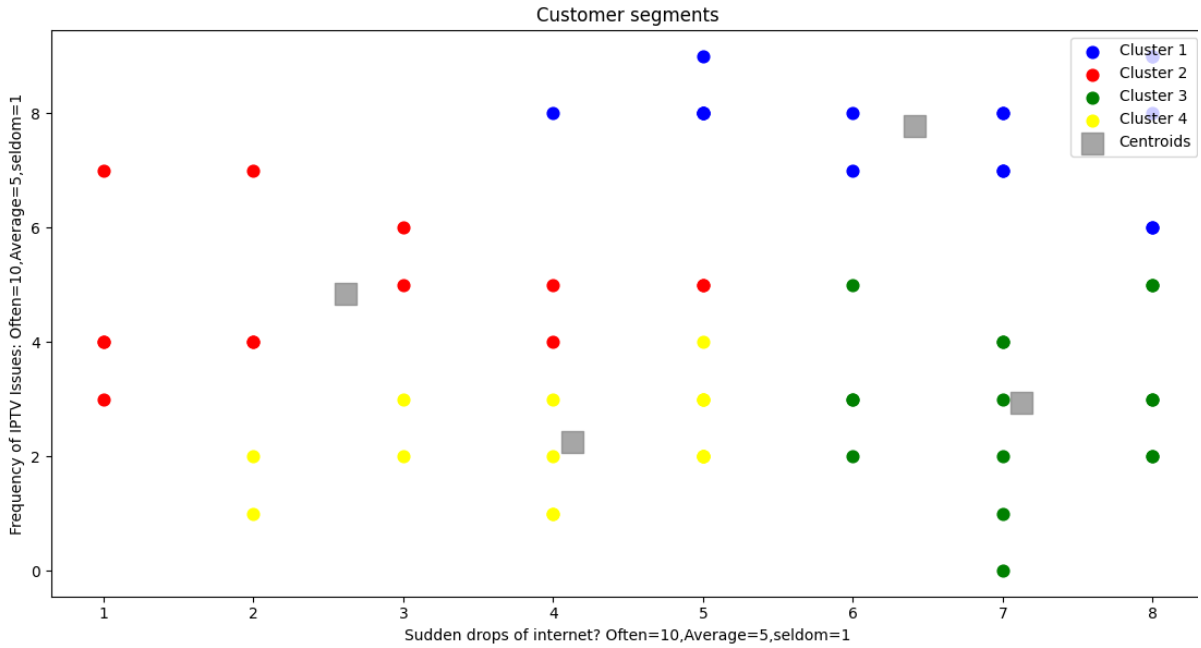


According to Figure 4.28 clusters are identified as below.

Cluster 3: Customer experiences frequent issues with IPTV connection, with low internet speed

Cluster 4: Customer experiences issues with IPTV connection less frequently, with low internet speed

FIGURE 4.29: Customer segmentation based on survey results



According to Figure 4.28 clusters are identified as below.

Cluster 1: Customer experiences iptv issues frequently and drops in internet connection frequently

Cluster 2: Customer experiences iptv issues frequently and drops in internet connection less frequently

Cluster 3: Customer experiences iptv issues less often and drops in internet connection frequently

Cluster 4: Customer experiences iptv issues less often and drops in internet connection less often

4.4.11 Retention Strategy

Clusters that are identified in section 4.4.8 and section 4.4.10 are listed in the below table. Retention strategies are proposed based on the Nature of these clusters.

Table 4.3: Identified Clusters

Identified Clusters based on Network parameters	Identified Clusters based on Survey Results
low number of third-party ONTs per port with distance from 3000m to 5000m	Customer experiences severe drops in internet connection with low internet speed.
average number of third-party ONTs per port with distance from 3000m to 5000m	Customer experiences internet drops at low rate with high internet speed.
high number of third-party ONTs per port with distance from 3000m to 5000	Customer experiences frequent issues with IPTV connection, with low internet speed
Average third party ont count per port and receiving power (< -28 dbmv)	Customer experiences issues with IPTV connection less frequently, with low internet speed

Cluster 4:High third party ont count per port and receiving power (< -28 dbmv)	Customer experiences iptv issues and drops in internet connection frequently
Average third party ont count per port and receiving power (>-7 dbmv)	Customer experiences iptv issues frequently and drops in internet connection less frequently
Low third party ont count per port and receiving power (< -28 dbmv)	Customer experiences iptv issues less often and drops in internet connection frequently
Low third party ont count per port and high transmitting power (> 2dbmv)	Customer experiences iptv issues and drops in internet connection less often
High third party ont count per port and high transmitting power (> 2 dbmv)	
Average third party ont count per port and high transmitting power (> 2 dbmv)	
High Total ont count per port and receiving power (< -28 dbmv)	
Average Total ont count per port and receiving power (> -8 dbmv)	
Average Total ont count per port and receiving power (< -28 dbmv)	
Low total ont count per port and receiving power(< -28 dbmv)	
High Total ont count per port and high transmitting power (> 2 dbmv)	
High Total ont count per port and high transmitting power (> 2 dbmv)	
Low Total ont count per port and high transmitting power (> 2dbmv)	

Technical retention Strategies for the customer segments,

1. Reduction of Distance between Optical line terminator and Optical network Terminator
2. Maintain Receiving power threshold of Optical Network Terminator between -8 dbmv and -27 dbmv
3. Maintain Transmitting power of optical Network Terminator below 2 dbmv

Marketing retention strategies for the customer segments are shown in Table 4.4

Table 4.4: Marketing Strategies

Identified Clusters based on Survey Results	Marketing Strategies
Customer experiences severe drops in internet connection with low internet speed.	• Replace Customer End equipment with New units. • Offer a Waiver for the customer (for internet package) • Offer an alternative Access system with a free data quota for a certain period.
Customer experiences internet drops at low rate with high internet speed.	• Replace Customer End equipment with New units. • Offer Access point for better wifi

	coverage
Customer experiences frequent issues with IPTV connection, with low internet speed	• Replace Customer End equipment with New units. • Offer a Waiver for the customer (for internet package)
Customer experiences issues with IPTV connection less frequently, with low internet speed	• Offer a Waiver for the customer (for internet package) • Offer an alternative Access system for the internet with a free data quota for a certain period.
Customer experiences iptv issues frequently and drops in internet connection frequently	• Replace Customer End equipment with New units. • Offer Access point for better wifi coverage
Customer experiences iptv issues frequently and drops in internet connection less frequently	• Replace Customer End equipment with New units.
Customer experiences iptv issues less often and drops internet connection frequently	• Offer Access point for better wifi coverage

CHAPTER 5

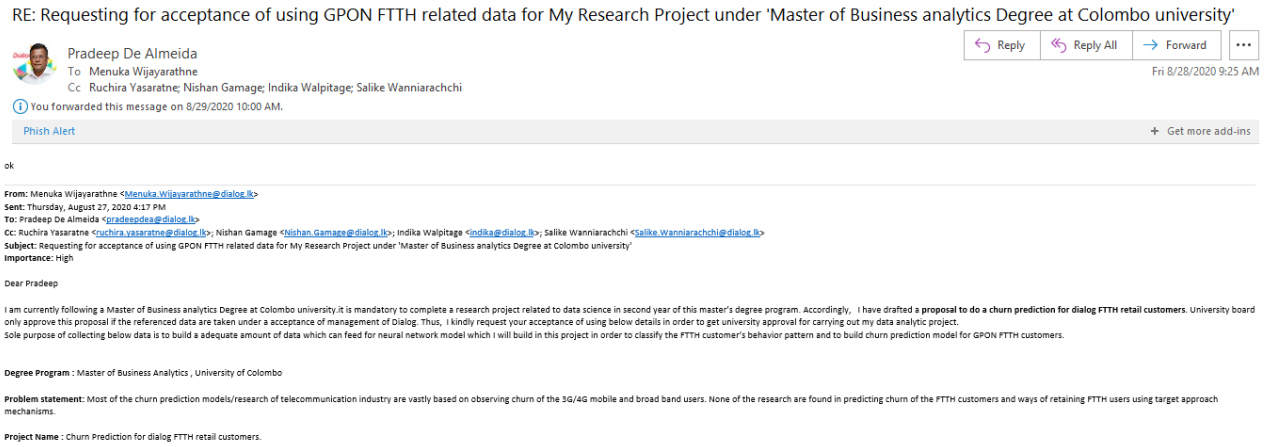
CONCLUSION AND FUTURE WORK

In conclusion, the present results of the study show that the use of machine learning techniques to construct a binary classifier in predicting churn of fiber to home customers based on network parameters. The obtained results show that the proposed churn model performed better by using logistic regression technique, resulted with better F-measure of 86%. Further, it is used clustering techniques to segment the disconnected customers based on network parameters and results of a market research survey. Finally, retention strategies are provided for the telecommunication company so that they can proactively retain the potential churners.

Concerning future research, it is intended to use this concept in predicting faults in dynamic telecommunication infrastructures proactively and to enhance the end user experience effectively with maximizing the return on investment.

APPENDICES

Appendix A - Permission for Accessing the data was granted by the group chief technology officer at dialog Technology.



RE: Requesting for acceptance of using GPON FTTH related data for My Research Project under 'Master of Business analytics Degree at Colombo university'

Pradeep De Almeida
To Menuka Wijayarathne
Cc Ruchira Yasaratne; Nishan Gamage; Indika Walpitage; Salike Wanniarachchi

Reply | Reply All | Forward | ...
Fri 8/28/2020 9:25 AM

You forwarded this message on 8/29/2020 10:00 AM.

Phish Alert | + Get more add-ins

ok

From: Menuka Wijayarathne <Menuka.Wijayarathne@dialog.lk>
Sent: Thursday, August 27, 2020 4:17 PM
To: Pradeep De Almeida <pradeepdea@dialog.lk>
Cc: Ruchira Yasaratne <ruchira.yasaratne@dialog.lk>; Nishan Gamage <Nishan.Gamage@dialog.lk>; Indika Walpitage <indika@dialog.lk>; Salike Wanniarachchi <Salike.Wanniarachchi@dialog.lk>
Subject: Requesting for acceptance of using GPON FTTH related data for My Research Project under 'Master of Business analytics Degree at Colombo university'
Importance: High

Dear Pradeep

I am currently following a Master of Business analytics Degree at Colombo university.it is mandatory to complete a research project related to data science in second year of this master's degree program. Accordingly, I have drafted a **proposal to do a churn prediction for dialog FTTH retail customers**. University board only approve this proposal if the referenced data are taken under a acceptance of management of Dialog. Thus, I kindly request your acceptance of using below details in order to get university approval for carrying out my data analytic project.
Sole purpose of collecting below data is to build a adequate amount of data which can feed for neural network model which I will build in this project in order to classify the FTTH customer's behavior pattern and to build churn prediction model for GPON FTTH customers.

Degree Program : Master of Business Analytics , University of Colombo

Problem statement: Most of the churn prediction models/research of telecommunication industry are vastly based on observing churn of the 3G/4G mobile and broad band users. None of the research are found in predicting churn of the FTTH customers and ways of retaining FTTH users using target approach mechanisms.

Project Name : Churn Prediction for dialog FTTH retail customers.

REFERENCES

[1] Aziz, A., Ismail, N., Ahmad, F. (2013). Mining students' academic performance. Journal of Theoretical and Applied Information Technology,53(3):485-495.

[2] Rahman, M., Lazim, Y., Mohamed, F. (2011). Applying rough set theory in multimedia data classification. International Journal of New Computer Architectures and their Applications, 1(3):683-693.

[3] Nafis, N., Makhtar, M., Awang, M., Rahman1, M., and Deris, M. (2017). Churn classification model for local telecommunication Company based on rough set theory. ARPN Journal of Engineering and Applied Sciences.

[4] Ullah, I., Raza, B., Malik, A., Imran, M., Islam, S., and Kim, S. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector, in IEEE Access, vol. 7, pp. 60134-60149 , doi: 10.1109/ACCESS.2019.2914999.

[5] Höppner, S., Stripling, E., Baesens, B., vanden, S., and Verdonck,T. (2020). Profit driven decision trees for churn prediction,European Journal of Operational Research,Volume 284, Issue 3.

[6] Amin, A., Shah, B., Masood, A., Joaquim, F., Ali G., Rocha, A., and Anwar, S. (2019). Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods,International Journal of Information Management,Volume 46,Pages 304-319,ISSN 0268-4012.

[7] Adris, A., Iftikhar, A., ur Rehman, Z. (2017). Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling. Clust. Comput., 1–15

[8] Mitkees, I. M., Badr, S. M., & ElSeddawy, A. I. B. (2017, December). Customer churn prediction model using data mining techniques. InComputer Engineering Conference (ICENCO), 2017 13th International (pp. 262-268). IEEE.

[9] Bhatnagar, A., and Srivastava, S. (2020). Performance Analysis of Hoeffding and Logisitic Algorithm for Churn Prediction in Telecom Sector, 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM), Dubai, United Arab Emirates, pp. 377-380, doi: 10.1109/ICCAKM46823.2020.9051503.

[10] Zhu, B., Xie, G., Yuan, Y., & Duan, Y. (2018). Investigating Decision Tree in Churn Prediction with Class Imbalance. Proceedings of the International Conference on Data Processing and Applications - ICDPA 2018. doi:10.1145/3224207.3224217.

[11] Zhu, B., Pan, Y., & Gao, Z. (2018). Application of Active Learning for Churn Prediction with Class Imbalance. Proceedings of the 2018 International Conference on Machine Learning Technologies - ICMLT '18. doi:10.1145/3231884.3231900.

[13]Mostert, W., Malan, K., & Engelbrecht, A. (2018). Filter versus wrapper feature selection based on problem landscape features. Proceedings of the Genetic and Evolutionary Computation Conference Companion on - GECCO '18. doi:10.1145/3205651.3208305.