Telecommunication Customer Churn Analysis

Gunarathne H.M
Computer Science Engineering
University of Moratuwa
Index No: 209328L
Email: gunarathnehm.20@uom.lk

Muthunayake M.N.A
Computer Science Engineering
University of Moratuwa
Index No: 209359G
Email: muthunayakemna.20@uom.lk

Ranasinghe R.K.C
Computer Science Engineering
University of Moratuwa
Index No: 209368H
Email: ranasingherkc.20@uom.lk

***Abstract*—Customer churn is an important problem affecting business and industry, when it comes to the telecommunication industry the basic rule is retain existing customers with the business, and coupled with the high cost associated with acquiring new ones. Machine learning is used to predict and analyze the future behavior of the customers who are most likely to change provided service.The aim of this project is to ascertain customers who want to churn and the major reasons for leaving the existing services. To determine the reasons of the customer churn, logistic regression, random forest and decision trees algorithms are applied.**

I. Introduction

With the high velocity in development of the telco industry, service providers focus more towards the augmentation of the subscriber base. It has been identified that, the cost of acquiring a new customer is significantly higher than retaining an existing customer, this paper will propose a framework to analyze and predict the customer churn and identify factors which affect the customer churn through following hypothesis analysis in order to increase customer retention in the telecommunication domain.

In the telecommunication industry each company provides the customers with huge incentives to attract them to switch to their services, An efficient churn predictive model benefits companies in many ways, It will be easy to limit customer retention campaigns and discounts to selected customers rather than selecting total population, Incorrect predictions could result in a company losing profits because of the counts offered to continuous subscribers. Therefore, right churn prediction is crucial for companies to identify the correct target group.

Telecommunication Customer Churn Data Set is extracted as a csv file format and data set consists of 9491 rows and 14 columns, where most of the attributes represent the customer experience related to the telecommunication industry.

The main objective of this research is to produce a predictive model with better results that access customer churn rate of companies. In addition, this paper will explain factors which have significant relationship with customer churn in the telecom industry through hypothesis validation.

II. METHODOLOGY

Telco data set is obtain by wireless network telecommunication company, the dataset consist of 9491 customers has 12 features and there are no missing data. The best 10 features are listed in below Table 1

TABLE I
Data set schema

Column Name	Description	Type
CustomerId	Unique customer identification	Numeric
Age	Customer age	Numeric
Gender	Customer gender	String
isCustomerSuspended	Whether customer suspended earlier	String
CallDropRate	Call drop rate - customerexperience data	Numeric
NumberOfComplaints	Customer complaints	Numeric
MonthlyBilledAmount - Rs	Avg. monthly bill	Numeric
UnpaidBalance - Rs	Total unpaid balance	Numeric
NumberOfMonthUnpaid	Total unpaid months	Numeric
TotalMinsUsedInLastMonth	Last month voice usage in minutes	Numeric
TotalCallDuration	Total voice usage in minutes	Numeric
AvgCallDuration	Average voice usage in minutes	Numeric
PercentageCallOutsideNetwork	Offnet voice usage % from total usage	Numeric
isChurned	Is customer disconnected?	String

At different levels of business analytics, a massive amount of data is processed and depending on the requirement of the type of analysis, there are 5 main types of analytics: Descriptive, Diagnostic, Predictive, Prescriptive and cognitive analytics. Initially Descriptive, Diagnostic and Predictive methods are selected to describe and recognize the patterns.

III. Hypothesis

1) There is a significant relationship between age and customer churn in Telecom.
2) There is a significant relationship between gender and customer churn in Telecom.
3) There is a significant relationship between customers that was suspended earlier and customer churn in Telecom.
4) There is a significant relationship between call drop ratio and customer churn in Telecom.
5) There is a significant relationship between the number of complaints and customer churn in Telecom.
6) There is a significant relationship between the monthly bill amount of the customer and customer churn in Telecom.
7) There is a significant relationship between unpaid balance and customer churn in Telecom.
8) There is a significant relationship between the number of unpaid months and customer churn in Telecom.
9) There is a significant relationship between total minutes used in the last month and customer churn in Telecom.

10) There is a significant relationship between total call duration and customer churn in Telecom.
11) There is a significant relationship between average call duration and customer churn in Telecom.
12) There is a significant relationship between the percentage of calls outside the network and customer churn in Telecom.

Following sections depict the Descriptive, Diagnostic and Predictive methods which were used to describe and recognize the patterns of customer churn.

IV. Descriptive Analytic

As a data analyst, when searching for an answer for ”What is happening in my business?” Descriptive analytics comes into the picture.

The selected dataset contains the telecom customer churn data with 9490 total records. It contains 14 columns with telecom user’s demographic data, behavioral data and transactional data including customer ID and customer churn status. A descriptive analysis was done to explore the selected telecom churn dataset using telecom user’s demographic data variables such as gender and age. Further the dependent variable of dataset, churn status was analyzed to identify its distribution. **Gender analysis:** The gender of majority users in selected telco dataset are females. However, the following figure 1 shows that gender wise distribution of population is similar to each other.

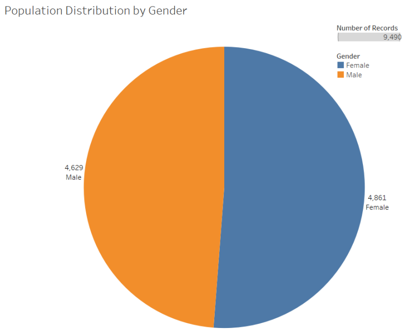

Fig. 1. Population Distribution by Gender

Churn distribution: According to the following figure 2, there are 858 churned users and 8,632 active (not churned) users from the total of 9,490 users in the selected telco dataset. The percentage of churn is 9.041

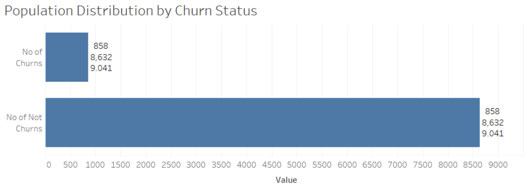

Fig. 2. Population Distribution by Churn

Customer Age analysis: Most of the users in the selected dataset are above 60 years old and there are few users less than 18 years of age. However, the following figure 3 shows that age wise distribution of population is similar to each other.

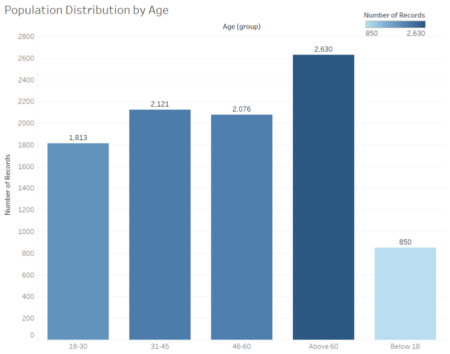

Fig. 3. Population Distribution by Age

V. Diagnostic Analysis

At this stage, historical data can be measured against other data to answer the question of *“why something happened”*.

A diagnostic analysis was done for identifying the relationship between customer churn and independent variables such as telco user’s demographic data, behavioral data and transnational data. These variables contain customer age, gender, customer was suspended earlier, call drop rate, number of complaints, monthly bill amount, number of unpaid months, unpaid bill value, average call duration, off-net call proportion and last month total usage. The analysis results are summarized below.

Relationship between gender and customer churn: According to the following figure 4, 8.7% has churned from the total male population of telco users whereas 9.2% has churned from total female users. Therefore, the churn percentage of females is 0.5% higher than the churn percentage of males in the selected telco dataset.

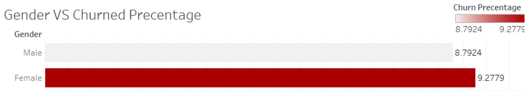

Fig. 4. Gender Vs Churned Percentage

Relationship between age and customer churn: According to the following figure 5, 14% has churned from users age below 18, 13.7% has churned from users age between 18 - 30, 11.8% has churned from users age between 31 - 45, 5% has churned from users age between 46 - 60, and 5% has churned from users age above 60. Therefore, highest churn percentages are recorded from age groups below 45 in the selected telco dataset.

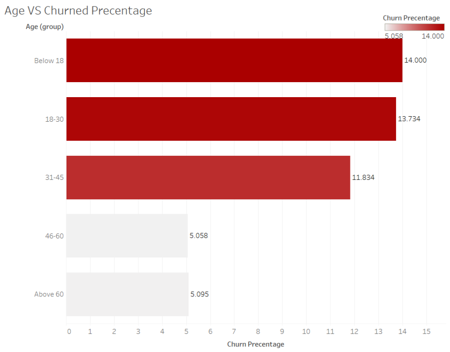

Fig. 5. Age Vs Churned Percentage

Relationship between previously suspended customer and churn: According to the following figure 6, 10.7% has churned from total users who have not suspended previously whereas 9.0% has churned from total users who have suspended at least one time previously. Therefore, the churn percentage of never suspended users is 1.7% higher than at least one time suspended users in the selected telco dataset.

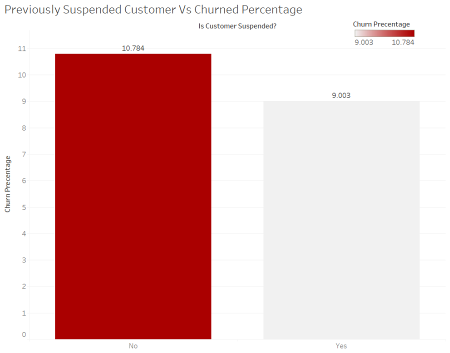

Fig. 6. Previous Suspended Customers Vs Churned Percentage

Relationship between call drop rate and customer churn: According to the following figure 7, customer churn percentage has increased with the call drop rate in the selected telco dataset.

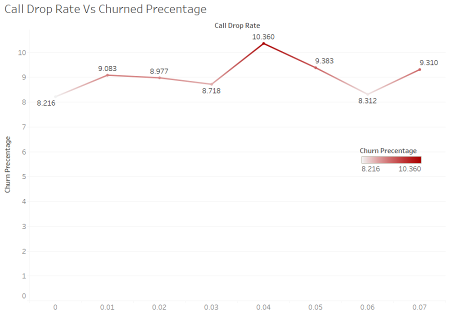

Fig. 7. Call Drop Rate Vs Churned Percentage

Relationship between monthly bill amount and customer churn: According to the following figure 8, customer churn percentage has increased with the monthly bill amount in the selected telco dataset.

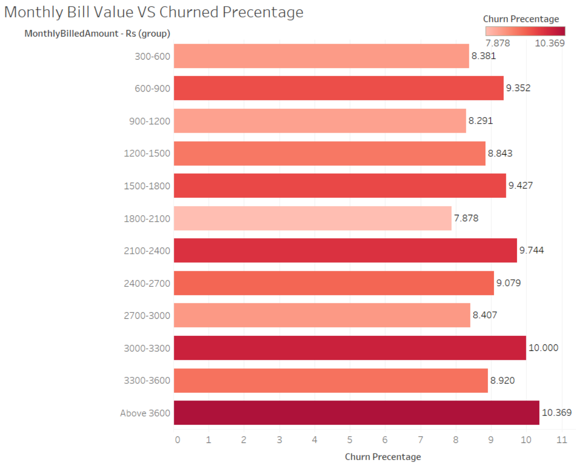

Fig. 8. Monthly Bill Amount Vs Churned Percentage

Relationship between no of unpaid months and customer churn: According to the following figure, customer churn percentage has slightly increased with the no of unpaid months in the selected telco dataset.

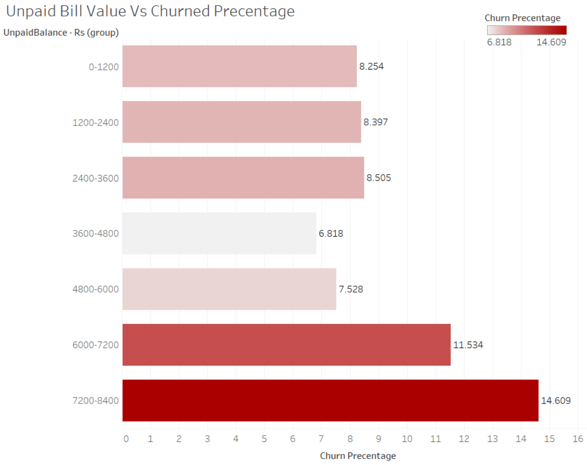

Fig. 9. Unpaid Bill Value Vs Churned Percentage

Relationship between average call duration - minutes and customer churn: According to the following figure **??**, customer churn percentage has decreased with the average call duration in the selected telco dataset.

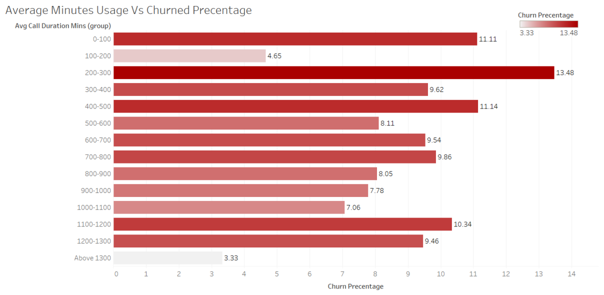

Fig. 10. f9 Bill Value Vs Churned Percentage

Relationship between off-net call proportion and customer churn: According to the following figure 11, customer churn percentage has not changed with the proportion of calls to other networks in the selected telco dataset. Therefore, there is no significant relationship between off-net calls and churn in selected dataset.

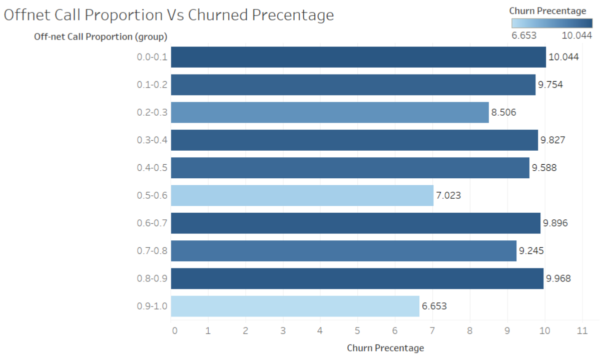

Fig. 11. Offset Call Proportion Vs Churned Percentage

Relationship between last month total usage – minutes and customer churn: According to the following figure 12, customer churn percentage has not changed with the last month usage in the selected telco dataset. Therefore, there is no significant relationship between last month usage and churn in the selected dataset.

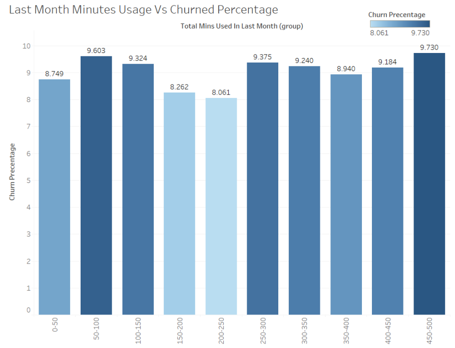

Fig. 12. Last Month Minute Usage Vs Churned Percentage

VI. Predictive Analysis

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Logistic Regression, Random Forest and Decision Trees are used as learning algorithms to train the initial machine learning models.

A. Logistic Regression

Classification is a very important field of supervised machine learning. A large number of important machine learning problems fall within this area. Logistic Regression is one of the fundamental classification techniques. It belongs to the group of linear classifiers. In general logistic regression is fast and relatively uncomplicated and also it's convenient to interpret the result.

There are several packages needed for logistic regression in Python implementation. All of them are free and open-source, with lots of available resources. **NumPy**, which is a fundamental package for scientific and numerical computing in Python. Pandas for data manipulation and analysis. Specially, it renders data structures and operations for manipulating numerical frames and time series. **Matplotlib** to visualize the results of classification, **Scikit-learn** It features various classification, regression and clustering algorithms including support vector machines.K-means, random forest etc. Individual variables can be plotted for numeric variables to clearly identify the distribution of the input attributes. It looks like perhaps few of the input variables have a Gaussian distribution and a lot of input variables show even distribution. It is very common to see categorical features in a dataset. Here mainly **Gender**, **isCustomerSuspended** and **isChurn** are categorical. However, machine learning algorithms can only read numerical values. It is essential to encoding categorical features into numerical values.

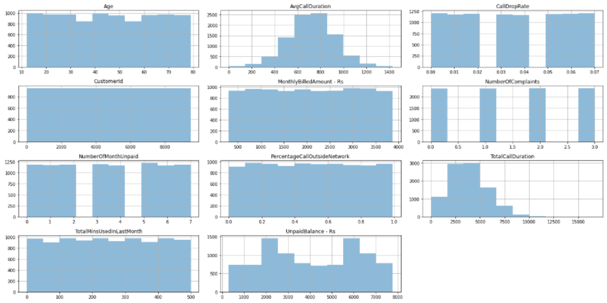

Fig. 13. Plotting histograms for input variables

Map function is used to substituting each value in a series with another value (0 and 1).

Machine learning models should be simple as much as possible to get their performance out. Therefore removed **AvgCallDuration** from the initial data set since **AvgCallDuration** is correlated with **TotalCallDuration**.The data set will be divided into two sets to train the model,0.7 is used to train the model and 0.3 is used as a test data set

We can evaluate the predictions by comparing them to the expected results in the validation set, then calculate classification accuracy is 91% for the logistic regression.

B. Random Forest

The random overall structure consists of a large number of individual decision trees that work as a whole. Every single tree in the random forest splits out a class prediction and the class with the most votes becomes the prediction of our model. The basic concept behind the random forest is simple but powerful that is why the random forest model works so well in data science.

1) Feature Selection: During a random selection of entities in a normal decision tree, when it is time to split a node, we consider all the possible entities and choose the one that causes the greatest separation between the observations of the left node and the right node. However, each tree in a random forest can only choose from a random subset of entities. This forces even greater variation between the trees in the model and ultimately leads to a weak correlation between the trees and greater diversification.

2) Correlation: The low correlation among the features is the key point in random forest. Correlation can be computed in numerical values, so that categorical data need to be mapped with the numerical values. Following figure will visualize the Pearson correlation of each feature.

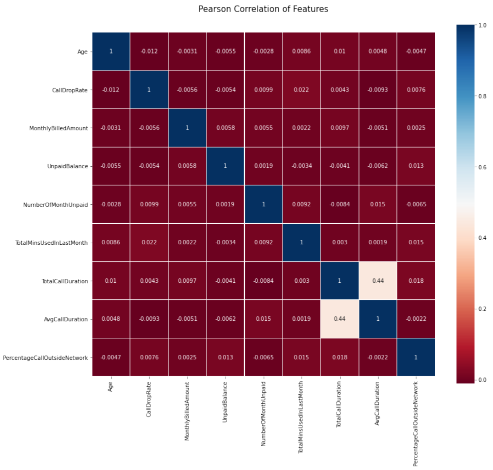

Fig. 14. Pearson Correlation of Features

Strong correlation can be removed since it has the same impact when it comes to the model being trained. Above visualization shows that there is no strong correlation between each feature.

3) Correcting Class Imbalance: According to the figure 15, class distribution is imbalanced. While working with an imbalanced classification problem, minority class is the most important as it has less examples in the dataset. It is hard to create a prediction model. So balancing is important before a model is trained.

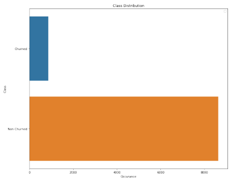

Fig. 15. Class Distribution

4) Feature Importance: Figuring out the most importance out of the features is necessary when a tree is being formed. Each feature can be averaged and the features are classified according to this measurement.

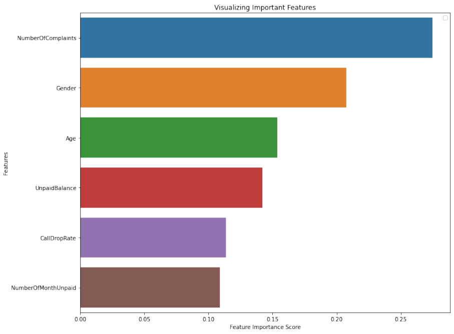

Fig. 16. Feature Importance

5) Training the Model: The data set will be divided into two sets to train the model, 90% is used to train the model and 10% is used as a test data set. As a data cleaning step, correlated features need to be removed as well as identified important features will be used to train the data model. Final outcomes of the accuracy which is produced by the random forest classifier are shown below table.

	precision	**recall**	**f1-score**	**support**
No	0.90	0.69	0.78	425
Yes	0.12	0.35	0.17	49
accuracy			0.66	474
macro avg	0.51	0.52	0.48	474
weighted avg	0.82	0.66	0.72	474

C. Decision Tree

Decision trees are a powerful and popular tool for classification and prediction. It is a supervised machine learning technique for inducing a decision tree from training data. A decision tree is a mapping from observations about an item to conclusions about its target value. In the tree structures, each non leaf node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. To predict the telco customer churn using Decision trees, Apache Spark, Spark ML and Pandas were used.

1) Data Preparation: Telecom data set was loaded from a CSV file which contains 9490 total records. It contains 14 columns with telecom user's demographic data, behavioral data and transactional data including customer ID and customer churn status. It was visualized and analyzed the data using pandas and removed the correlated fields to improve model accuracy as follows. A randomly sampled portion of the data (10%) used for analysis using pandas and it was identified that TotalCallDuration and AvgCallDuration are correlated. Therefore, TotalCallDuration was removed, which is a one column of a pair of correlated fields. There were about 10 times as many False churn samples as True churn samples. Therefore, stratified sampling was used to put the two sample types on the same footing. So, all instances of the Churn=True class were kept, but downsampled the Churn=False class to a fraction of 858/8632.

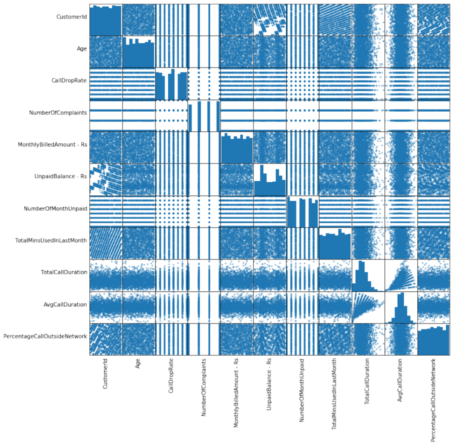

Fig. 17. Correlations between the numeric columns

2) Training the model: Predictive modelling was done using the Spark ML package. Decision Tree model was generated using the training set and evaluated with the testing set. The data set was divided into two sets using a random split function, 80% was used to train the model and 20% was used as a test data set. The cross validator used the ParamGridBuilder to iterate through the maxDepth parameter of the decision tree and evaluated the models using the F1-score, repeating 3 times per parameter value for reliable results. The Decision tree model produced using the cross-validation process was one with a depth of 7 as follows. DecisionTreeClassificationModel $uid = DecisionTreeClassifier_0c5a732f2dc2$ of depth 7 with 115 nodes Therefore it was assumed that a tree depth of 7 would perform well. Finally, the predictions and evaluations were done for the test data set with the f1 accuracy of: 0.6849696758911609. Follows the sample of prediction results:

	Customer Id	Description
0	4	1
1	5	1
2	7	1
3	8	0
4	15	0

VII. Recommendations and Conclusion

A. Result

As per the descriptive and diagnostic analysis, key features of telco dataset have positive relationships with customer churn. However, there was a weak relationship between off-net call proportion and last month total usage – minutes with customer churn. The Logistic regression model was trained using 11 features such as Age, Gender, isCustomerSuspended, CallDropRate, NumberOfComplaints, MonthlyBilledAmount - Rs, UnpaidBalance - Rs, NumberOfMonthUnpaid, TotalMinsUsedInLastMonth, AvgCallDuration, PercentageCallOutsideNetwork. The classification accuracy of the model is 0.91 on the test dataset. The Random Forest model was trained using all the features in telecom dataset. The classification accuracy of the model is 0.89 on the test dataset. Finally, the Decision tree model was trained using 11 features in telecom dataset except TotalCallDuration. The classification accuracy of the model is 0.68 on the test dataset.

B. Future Work

This study can be further extended and used to support future studies focusing on Customer Churn prediction in Telecom.
Some of the future directives are mentioned below:

- Identify more features and their effectiveness for customer churn.
- Prescriptive analytics to identify solutions to prevent customer churn by providing targeted campaigns.

C. Assumptions or Disclosure

- The sample telecom churn data-set is used as an estimate of the larger population to generalize the findings.
- Only the customer age, gender, customer was suspended earlier, call drop ratio, number of complaints, monthly bill amount, unpaid balance, number of unpaid months, total minutes used in the last month, total call duration, average call duration and percentage of calls outside the network.