Telecommunication Analytics - Know Your Customer

"Submitted towards partial fulfilment of the criteria
for award of PGPBABI by Great Lakes Institute of Management"

Submitted By

Group No. 7

Batch: Jan 2019

Group Members

Nivitha (BACJAN19026)
Selvavinayaagam (BACJAN19042)
Shanmugnathan (BACJAN19044)
Sunandha (BACJAN19048)
Tejaswi (BACJAN19051)

Research Supervisor

Suvajit Mukhopadhyay

Great Lakes Institute of Management

Abstract

The problem of customer churn, which denotes loss of a client to competitors, is a key issue across industries. New customers are difficult to find, especially in saturated industry like telecommunications. It is hard to tap on new customers and the only way to gain new customers is to take away from the competition. Furthermore, it is far less expensive to retain existing customers than to acquire new ones. With attraction offers, customers tend to jump from current service provider to another more easily than ever. This poses a serious challenge for the Telcom industry to reduce churn and retain customers.

Retention is usually a process that identifies customers that are likely to churn, using various predictive modelling techniques, followed by approaching these customers with suitable offers that would persuade the customer into extending the contract. But, can the customer be prevented from even wanting to churn? Can the main churn drivers be mitigated beforehand? This project is focused on a company-wide churn reduction initiative conducted in one of the telecom operators losing business to competition in India.

The primary objective of this project is to help the service provider understand and categorize customers based on the plan, contract, usage, billing and churn data of its subscribers. Along with the analysis on the customer data, the network performance is also analysed in terms of call drops and call quality across its footprint to realize the pains faced by customers. Network Auditing is also done on its networking devices to understand the security risks posed and measures to predict and alarm the Network Operations team monitoring the network and take action proactively.

Acknowledgement

We wish to place on record our deep appreciation for the guidance and support provided to us by our Mentor Mr. Suvajit Mukhopadhyay.

Mr. Suvajit helped us narrow down on the choice of the capstone dataset, as well as the scope definition and implementation of the project. He gave us valuable feedback at every stage to enhance the completeness and the final outcome of the project. His experience, support and thought process guided us to be on the right track towards completion of this project.

We are extremely gifted and fortunate to have Dr PK Viswanathan as our Course Director – BABI Program. His in-depth knowledge coupled with his passion in delivering the subjects to the students has helped us a lot. We are grateful to him for his interesting sessions on Predictive Analytics and Statistics.

We are thankful to Mr. Jenlyn Jude, Program Manager BABI Program for his steady and persistent support extended to us at all times. We also thank all the course faculty of BABI program for providing us a strong foundation in various concepts of analytics & machine learning.

Finally, we would like to thank each and every team member part of the capstone project who shared their views regularly and helped each other to complete the project on time.

Date: 10/Dec/2019

Place: Chennai

Certificate of Completion

Gmail

Nivitha Natarajan <nivithanata@gmail.com>

Project Completion Acknowledgement

1 message

Suvajit Mukhopadhyay <suvajit21@gmail.com> 10 December 2019 at 20:26
To: Jenlyn@greatlearning.in
Cc: Nivitha Natarajan <nivithanata@gmail.com>, Selva vinayagan <selvavinayaganm@gmail.com>, Sunandha Suresh <suna.usha@gmail.com>, Tejaswi Rajalakshmi <tejass.civi.laksh@gmail.com>, Shanmuganathan Thiagarajan <shanmugt@gmail.com>

Dear Jenlyn,

To whom It May Concern,

I hereby certify that the project titled **Telecommunication Analytics - Know Your Customer** for case resolution was undertaken and completed under my supervision by Nivitha N, Selavinayaagam, Shanmuganathan, Sunandha and Tejaswi R of Post Graduate Program in Business Analytics and Business Intelligence (PGPBABI) - Jan 2019.

With Best Regards,

Suvajit Mukhopadhyay

Place: Kolkata

Table of Contents

List of tables and graphs

Abbreviations

ADA	Adaptive Boosting
AUC	Area Under the Curve
CART	Classification and Regression Tree
df/DF	Degrees of Freedom
FN	False Negative
FP	False Positive
GBM	Gradient Boosting Machines
KNN	K Nearest Neighbours
LDA	Linear Discriminant Analysis
NN	Neural Networks
PCA	Principal Component Analysis
RF	Random Forest
ROC	Receiver Operating Characteristics
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machines
TP	True Positive
TN	True Negative
VIF	Variance Inflation Factor

Executive Summary

India is currently the world's second-largest telecommunications market with a subscriber base of 1.20 billion and has registered strong growth over the recent years. Revenues from the telecom sector are expected to grow to US$ 26.38 billion by 2020. The number of internet subscribers in the country is expected to double by 2021 to 829 million and overall IP traffic is expected to grow 4-fold at a CAGR of 30 per cent by 2021. With daily increasing subscriber base, there have been a lot of investments and developments in the sector.
Gone are the days where Telecom customers wait for months to subscribe for a voice or data connectivity. Decades back, the Telecom Operators like BSNL enjoyed a monopoly share of the complete telecom market segment. In today's world, with all major technological advancements, customers have a plethora of choices in hand and they act as a key driver for the success of a Service provider.

The Indian Telecom sector has witnessed exponential growth over the last few years which was due to many factors like reduced tariffs, wider services, evolving consumer consumption patterns and conducive regulations. In 2018, the wireless subscriber base recorded an increase of 13 million subscribers with an overall tele-density of 92%. The year also saw an increase in the rural tele-density from 56% to 59%, while the urban tele-density decreased from 172% to 166%. During the year 2017-18, 98.07 million subscribers have submitted their porting requests to different service providers, the porting requests increased from 272.76 million at the end of March 2017 to 370.83 million at the end of March 2018 which shows subscribers exercising their preference of service providers.

Out of 11 Telecom Operators, only the top 5 players like Reliance Jio, Bharthi Airtel, BSNL, Vodafone and Idea Cellular were able to add new mobile customers in 2018. Loss of customers by Tata Teleservices, Reliance Communications, Telenor, Airvoice, Sistema Shyam and MTNL held back their growth rate to a great extent. The operators have an immense competition to attract new customers to expand their foot print and to stay alive in business. Operators like Loop, Quadrant and Videocon discontinued their services totally or lost to competition. They get into price pressures owing to steep growing competition and at the same time pushed to offer better customer services to stay firm in the market.

Airvoice which dropped from the top 5 list decided to execute a market survey to find out customer needs. Though Tariff rates is a key factor, however, several other major contributing factors came to the forefront as mentioned below:

- Customer Service
- Consumer Usage or **Behaviour** Pattern (Right Products for Right Customers)
- Network Availability and Security
- Call Quality

That turned the attention of management to seek on a transformation plan and the Chief Transformation Officer decided to onboard a team of Data Scientists to use the data to their advantage ,and hence a project was initiated called as "Spark". As the first phase, Airvoice Operator decided to dig deep into its data using analytics to find out key areas relevant for transformation. After lot of brainstorming sessions, the team finalized 3 areas to extract insights as part of phase I.

Key areas identified as part of the scope includes:

- Deeper understanding of the existing customers and identify people at
- Audit Network Performance on Call quality and Call drops
- Predict Network Security events proactively

Key use cases to be derived from the study includes:

- Churn Analysis used to understand the pattern in Customer Attrition
- Network Performance (by call quality, customer rating and area)
- Network Auditing (by analysing Network Traffic and range to see any anomalies)

Chapter 1: Introduction

Problem Statement

Airvoice has been facing increased customer churn rate due to increase in the competitions by other network players and the project aims to analyse the customer usage and network service to address the solution for below two key areas.

1. Reduction of customer churn rate and retention
2. Identification of network performance improvement areas and predict network security events real time.

Prediction of customer churn reason is identified through various machine learning algorithms. Customers are then segmented related to call performance and network types using various clustering algorithms.

Objective

The objective of the study includes exploring, analysing and deriving insights for the organization. These insights will be used by the Strategy Team to formulate subsequent steps and an action Plan. The Data Science Team is mandated:

- To build models which can be used by internal applications of service provider to quickly understand trend and patterns.
- Present top 5 factors related to Customer Churn, Network Performance and Network Security, backed by data.

In Scope

1. The scope of the study includes:
    - Co-ordinating with Management Information Systems' Team to extract relevant data for the study.
    - These data to be ingested within the Analytics Engine of Data Science Team.
    - To analyse, Clean, derive insights from these data
    - Build Machine Learning Models using R and/or Python
2. Quick description for the various modules:
    - **Customer churn:**

      Exploring the various chunks of customer's data that includes the demography and geographical aspects as well to obtain a brief comprehension of the Customer churning.
    - **Call performance:**

      The quality of the voice calls and frequent call drops among the commercial telecom companies are analysed with respect to the location and state of commute.
    - **Network Analysis:**

      A complete network protocol is demonstrated for various network services using the machine learning techniques.

Out of Scope

Scope of the study doesn't include the following:

- Sourcing the data as the Management Information Team will extract relevant data.
- Analysing Social Media Sentiments, Customer Complaints
- Up Sell or Cross Sell Strategy of Various Products to its customers.

Success Criteria

At the end of the study we will be able to extract:

- Customer segmentation based on usage patterns and demography
- Analyse the key reasons for Customer Churn
- Build a model to predict the Customer Churn
- Analyse the key reasons for Network Performance and build a Machine Learning Model to predict the Network Behaviour.
- Analyse on the key patterns behind Network Intrusions and build a Machine Learning Model to predict anomalies on future data.

Data Description

Dataset: Customer Churn Data

The data contains customer details, services, demography and many more details. This data has been collected over a period of last 2 years pulled out from Customer Database of Airvoice. Below is a brief introduction to variables and dataset at hand, which will be used to solve this use case.
Some of the variables in the dataset are

- Customer Location
- Number of Referrals
- Internet Services
- Churn Reason
- Contract
- Offer
- Tenure in Months
- Unlimited Data
- Total Charges

For our project purpose, the data is collected from below mentioned source:
https://community.ibm.com/community/user/businessanalytics/blogs/stevenmacko/2019/07/11/telco-customer-churn-
1113

Churn reason will give us better understanding of the problem faced by customer. Customers have left the network due to better options floated by other competitors in the market.

Figure 1: Customer Churn - Word Graph

Top 5 major reason for Churn are

- Competitor
- Competitor had better devices
- Competitor made better offer
- Attitude
- Dissatisfaction

Dataset: Network Performance

The Data Science team decided to extract data from the Network Performance database. In order to understand the pattern better, they also decided to use their Competitor's data available in Public Domain. The dataset at hand contains Operators name, Indoor/travelling, Network type, Rating, Call drop and Locations over a period of a month. Below is a brief introduction to variables and dataset at hand, which will be used to solve this use case.

For our project purpose, the Data is collected from below mentioned source:
https://data.gov.in/catalog/voice-call-quality-customerexperience?filters%5Bfield_catalog_reference%5D=3995421&format=json&offset=0&limit=6&sort%5Bcreated% 5D=desc

Network performance data of a customer has following fields

- Customer Id
- Operator – Airtel, BSNL, Vodafone, Airvoice and Idea
- Network Type – 2G/3G/4G
- Indoor/Outdoor/Travelling – Variable describing the mode of network usage
- Rating – The variable is in the scale of 1 to 5
- Latitude
- Longitude
- State Name – Data is captured across all Indian States
- Call Drop Category – Call dropped, Poor Network and Satisfactory are the categories.

Call drop is observed almost in all regions of India.

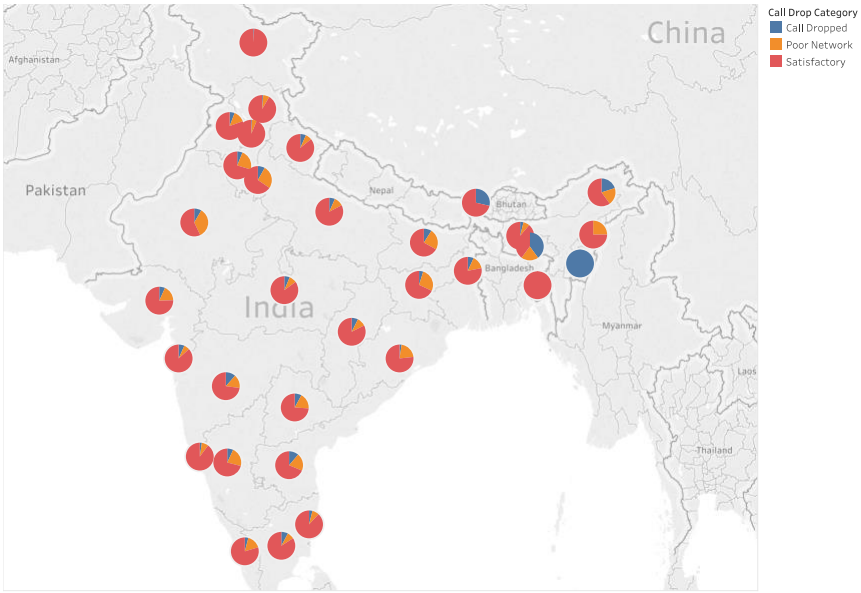


Figure 2: Geo-distribution of Call Drop Category

Maharashtra, Tamil Nadu, UP, Karnataka, Telangana and Gujarat have a greater number of call records comparatively to other states. Around 25% customers did not share the location due to privacy reasons.

Dataset: Network Auditing

The data contains basic features of individual TCP connections pulled out from the Security Operations' Centre database containing suspected attacks on the Organization Network over past 6 months.

For our project purpose, the Data is collected from below mentioned source:
http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

The dataset is classified majorly into Basic, Content and Traffic Features.

Basic Features	*Description*
duration	length (number of seconds) of the connection
protocol_type	type of the protocol, e.g. tcp, udp, etc.
service	network service on the destination, e.g., http, telnet, etc.
src_bytes	number of data bytes from source to destination
dst_bytes	number of data bytes from destination to source
flag	normal or error status of the connection
land	1 if connection is from/to the same host/port; 0 otherwise
wrong_fragment	number of ``wrong'' fragments
urgent	number of urgent packets

Content Features	*Description*
hot	number of ``hot'' indicators
num_failed_logins	number of failed login attempts
logged_in	1 if successfully logged in; 0 otherwise
num_compromised	number of ``compromised'' conditions
root_shell	1 if root shell is obtained; 0 otherwise
su_attempted	1 if ``su root'' command attempted; 0 otherwise
num_root	number of ``root'' accesses
num_file_creations	number of file creation operations
num_shells	number of shell prompts
num_access_files	number of operations on access control files
num_outbound_cmds	number of outbound commands in an ftp session
is_hot_login	1 if the login belongs to the ``hot'' list; 0 otherwise
is_guest_login	1 if the login is a ``guest''login; 0 otherwise

Traffic Features	*Description*
count	number of connections to the same host as the current connection in the past two seconds
	Note: The following features refer to these same-host connections.
serror_rate	% of connections that have ``SYN'' errors
rerror_rate	% of connections that have ``REJ'' errors
same_srv_rate	% of connections to the same service
diff_srv_rate	% of connections to different services
srv_count	number of connections to the same service as the current connection in the past two seconds
	Note: The following features refer to these same-service connections.
srv_serror_rate	% of connections that have ``SYN'' errors
srv_rerror_rate	% of connections that have ``REJ'' errors
srv_diff_host_rate	% of connections to different hosts

46.5 % network connections are interrupted with network intrusions. 53.5%. connections are not interrupted when the connections are made.

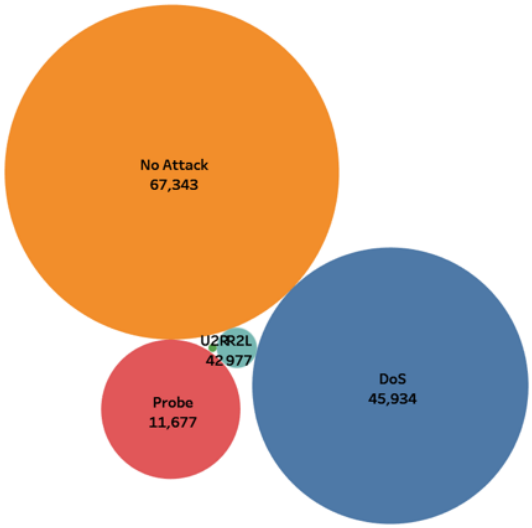


Figure 3: Network Connection Vs Attack Types

The analysis of attacks on the network services is shown below:
http, private, ftp_data,ecr_i ,telnet, finge, ftp service based connections faced lots of DoS or Probe attacks.

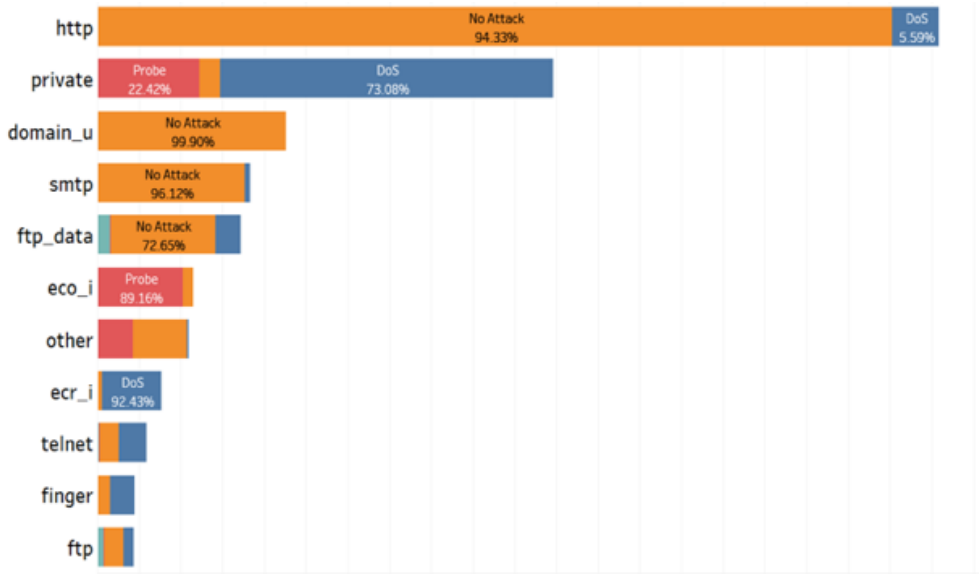


Figure 4: Services and Attacks

Tools and Techniques

Below are some of the tools and techniques used in the capstone.

Tools	Problem	Techniques
Python, R, Orange and Tableau	Customer Churn	Applicable ML algorithms for Classification and Clustering
R, Tableau and Excel	Network Performance	K-means, K-NN
Python, R and Tableau	Network Auditing	Machine Learning algorithms

Limitations

There are certain limitations associated with data collection and the model performance in terms of running time, memory consumption.

- Around 15,000 customers did not wish to share the location and hence it was difficult to impute their location based on other variables.
- Data is huge and clustering algorithm like Agglomerative clustering didn't perform well with the huge dataset of categorical variables.
- Churn dataset had vast number of duplicates and missing data with high degree of class imbalance.
- Some of the models were resource hungry especially when applying cross validation and hyper parameter tuning. So, it was not possible to apply all possible tuning parameters with varied values to improve accuracy.
- All the models were developed for Churn dataset using mlr and caret package in R. There does not exist any support in mlr to plot the derived model for each algorithm like CART, NN, SVM. Due to time limitation the model plots were left unexplored.

Chapter 2: Literature review

Customer Churn

Azeem, Usman & Fong in 2017 used fuzzy logic to predict in telecommunication churn. C4.5, gradient boosting, SVM, linear regression, Artificial Neural Network (ANN), random forest, AdaBoost and neural network were used.
Ismail et al.in 2015 mentioned about the Multilayer Perceptron (MLP) algorithm to predict customer churn. The MLP algorithm was compared between logistic regression and multiple regression. The variables used were customer demographic, customer relationship, billing variables and usage variables.

Vafeiadis et al. in 2015 used data mining algorithms are ANN, decision trees, SVM, Naïve Bayes and logistic regression. These algorithms were then compared with the hyper parameter tuned models. The objective as to find the best the state of data mining algorithms for the customer churn problem. The telecommunications dataset for this research was obtained from UCI Machine Learning Repository.

Customer Call Quality Clustering

Bounsaythip and Rinta-Runsala in 2001 proposed that segmentation as a method to have more targeted communication with the customers; and the process of segmentation explains the characteristics of the customers groups (called segments or clusters) within the data. The diversity of customer needs was influenced by various factor such as lifestyle, income, age, location. The current models for customer profiling are based on customer behaviour based on the transaction records or surveys. Kiang et al. in 2006 surveyed the applications of data mining for customer segmentation purposes that it is the ideal way to obtain customer profitability through careful customer targeting using proper data.

Network Intrusion

Most of the work done on the intrusion anomaly detection systems were based on what machine learning technique to use to improve the accuracy and what are the feature selection technique is used to construct Network intrusion Detection System.

Gisung Kim et.al, presents a new hybrid intrusion detection method that hierarchically combines a misuse detection and anomaly detection. Normal Training dataset is disintegrated into smaller subsets using C4.5, then for each smaller subset the SVM is used to create an anomaly detection model. The characteristic of normal behavior was tracked trough tout his study and splitting subsets. Using C4.5 clustering was not performed as that mat reduce the accuracy of the system.

SVM algorithm was proposed for Intrusion Detection System which provides the best variable selection to each types of attack. This was a multi class classification to identify the number of important features for each of 5 classes (Normal, DoS, Probe, U2R, and R2L). The model output had high performance and the training and testing time decrease for each class. However, the Accuracy decreases for classes DoS, Probe, and R2L, and remains equal for ‘Normal’ and U2R.

In another paper SVM and Discriminant Analysis both combined to identify network intrusion. This study selected only Nine features {2, 12, 23, 24, 29, 31, 32, 36, and 39} are obtained by Discriminant Analysis and modelled and tested using SVM.

Chapter 3: Exploratory Data Analysis

Dataset: Customer Churn
The dataset contains customer details, services, demography and usage, billing information.

Overview of the dataset

- Dataset has 35500 observations and 42 variables
- Churn status is the Dependent Variable and is recorded as "Numeric"
- There are missing values / NAs in variable columns.
    - Referred a Friend
    - Number of Referrals
    - Offer
    - Internet Service
    - Online Backup
    - Device Protection plan
    - Premium Tech Support
    - Streaming TV
    - Streaming Movies
    - Streaming Music
    - Total Refunds
    - Under30
    - Senior Citizen
    - Country
- There are multiple duplicate customer ids and possibly duplicate observations.
- Monthly Charge, Total Charges, Total Extra Data Charges, Total Long-Distance Charges, Total Revenue are closely related and might have strong multi-collinearity.
- Zip Code is recorded as "Numeric" which needs to be converted as "Categorical".
- 15500 observations out of 35500 are reported as Churned observations.

Visual Data Analytics with Exploratory – Independent Categorical Variables

- **Distribution of Referrals and Offer**

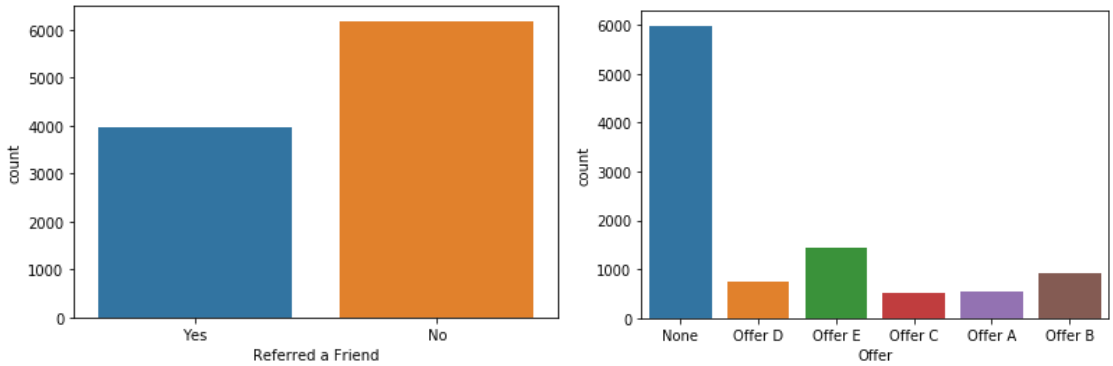


Figure 5: Distribution of Referrals and Offer

Interpretation

- 40% of customers have referred a friend
- Most of the customers do not have any Offers. Offer E and B are the major offers provided to customers among all existing offer possibilities

- **Distribution of Services**

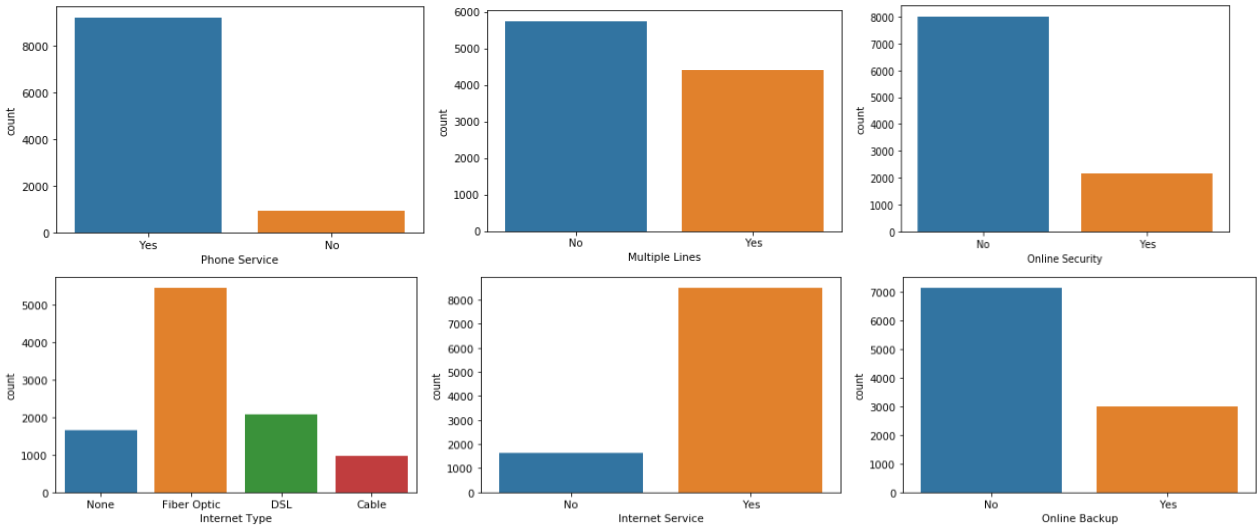


Figure 6: Distribution of Services

Interpretation

- Customers use both Phone and Data Services. Around 10% of customers use either of the services.
- More than 55% of customers have multiple lines
- Fibreoptic and DSL are the top 2 internet types used by subscribers
- Most customers do not prefer online services like security and backup

- **Distribution of add-on services**

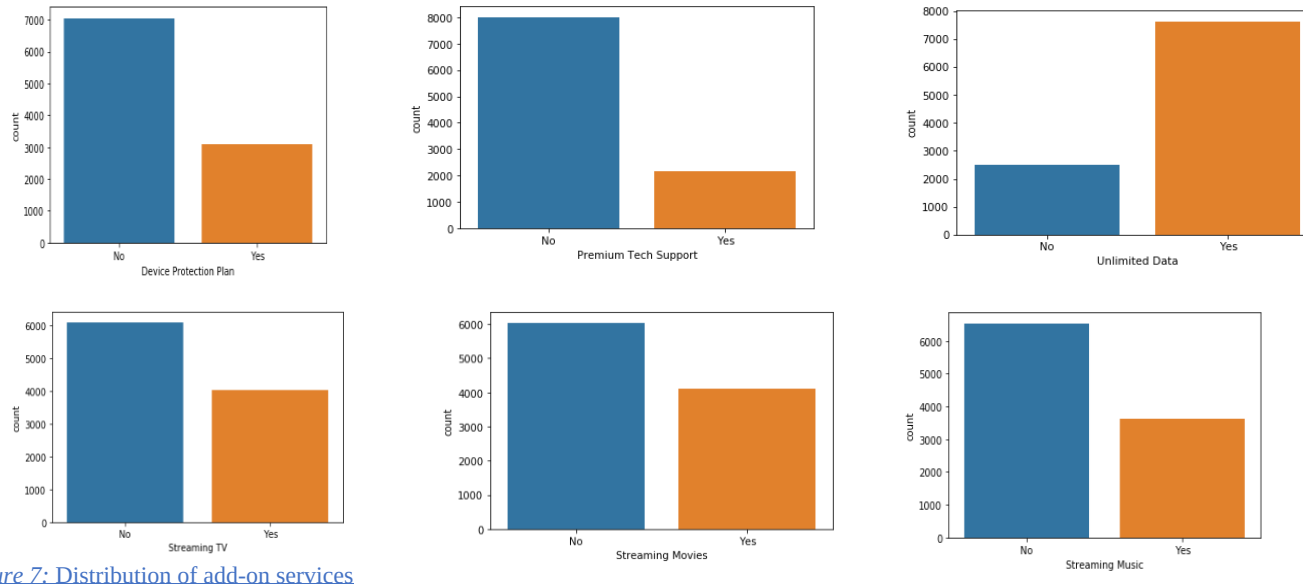


Figure 7: Distribution of add-on services

Interpretation

- Majority of data subscribers prefer unlimited data connectivity option
- Streaming services are least preferred by subscribers
- Device Protection Plan and Premium Tech Support is also not preferred

- **Distribution of Billing details**

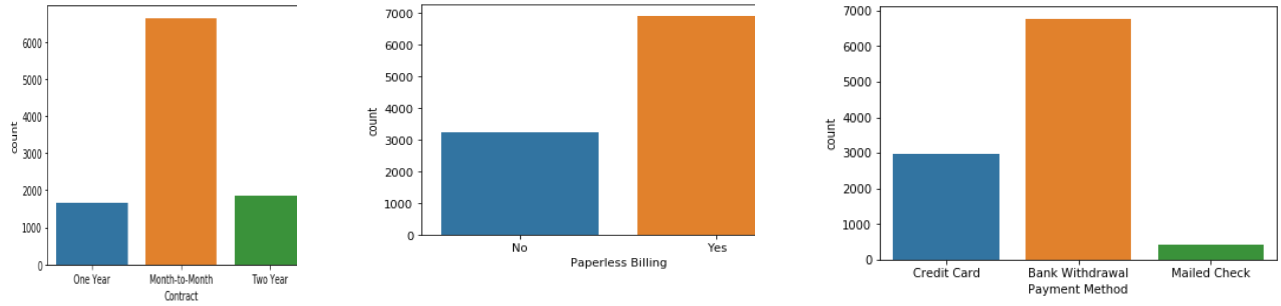


Figure 8: Billing details

Interpretation

- Customers prefer the contract of least risk and term, month on month and one-year contract
- Most of them prefer paperless billing and opt for direct clearance from banks for billing payment

- **Distribution of Customer status**

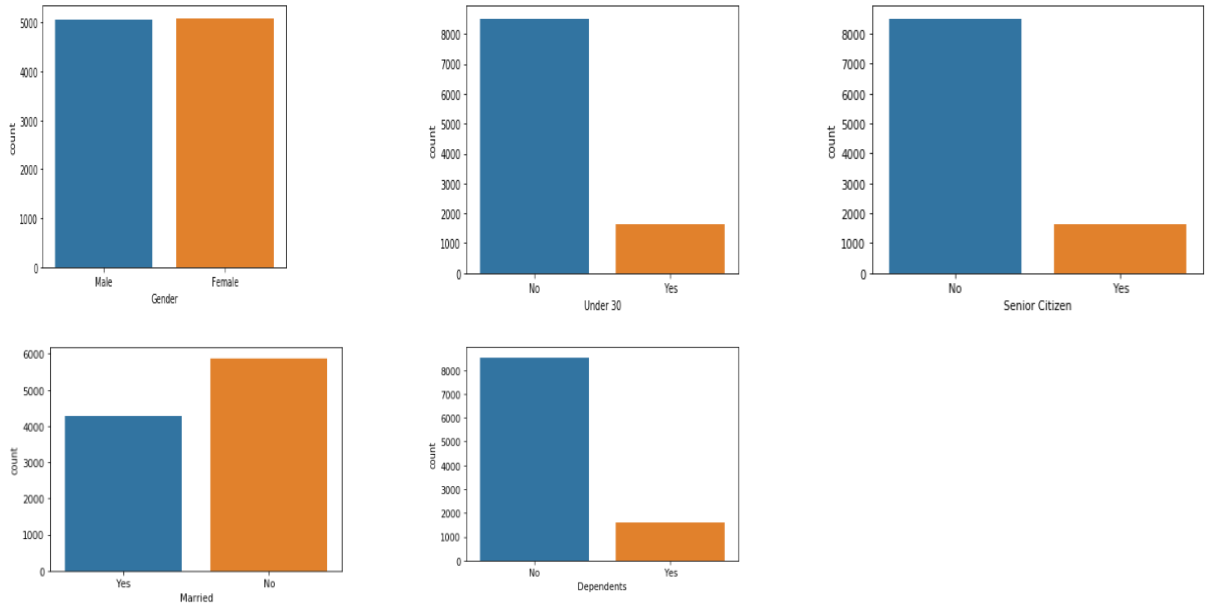


Figure 9: Customer Status

Interpretation

- Customers are equally distributed in terms of sex
- Major proportion of customers are aged between 30 and 65, only 15% of them fall under age group of 30 and senior citizen category
- More than 40% are married and only less than 15% have dependents in their family.

- **Distribution of Churn**

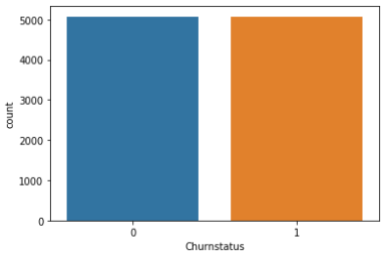


Figure 10: Customer Churn Status

Interpretation

- 50% of customers churn

Visual Data Analytics with Exploratory – Independent Continuous Variables

Correlation Plot

	Number of Referrals	Tenure in Months	Avg Monthly Long Distance Charges	Avg Monthly GB Download	Monthly Charge	Total Charges	Total Refunds	Total Extra Data Charges	Total Long Distance Charges	Total Revenue	Age	Number of Dependents
Number of Referrals	1.0	0.36	0.0066	0.03	-0.0048	0.27	0.038	0.0095	0.24	0.28	-0.046	0.3
Tenure in Months	0.36	1.0	0.034	0.038	0.23	0.86	0.075	0.12	0.71	0.88	-0.0063	0.13
Avg Monthly Long Distance Charges	0.0066	0.034	1.0	-0.025	0.17	0.09	-0.0087	0.0039	0.56	0.23	-0.0069	-0.022
Avg Monthly GB Download	0.03	0.038	-0.025	1.0	0.34	0.18	0.011	0.079	0.0083	0.14	-0.4	0.12
Monthly Charge	-0.0048	0.23	0.17	0.34	1.0	0.59	0.042	0.12	0.25	0.53	0.15	-0.16
Total Charges	0.27	0.86	0.09	0.18	0.59	1.0	0.056	0.14	0.67	0.98	0.042	0.037
Total Refunds	0.038	0.075	-0.0087	0.011	0.042	0.056	1.0	0.02	0.044	0.054	0.016	0.028
Total Extra Data Charges	0.0095	0.12	0.0039	0.079	0.12	0.14	0.02	1.0	0.085	0.14	0.029	-0.012
Total Long Distance Charges	0.24	0.71	0.56	0.0083	0.25	0.67	0.044	0.085	1.0	0.81	-0.0061	0.079
Total Revenue	0.28	0.88	0.23	0.14	0.53	0.98	0.054	0.14	0.81	1.0	0.031	0.052
Age	-0.046	-0.0063	-0.0069	-0.4	0.15	0.042	0.016	0.029	-0.0061	0.031	1.0	-0.13
Number of Dependents	0.3	0.13	-0.022	0.12	-0.16	0.037	0.028	-0.012	0.079	0.052	-0.13	1.0

Figure 11: Customer Churn - Correlation Plot

Interpretation

- Tenure in Months is closely correlated with
  - Total Charges
  - Total Long-Distance Charges
  - Total Revenue
  - Number of Referrals
- Average Monthly GB Download has a negative correlation with Age

Histogram Representation

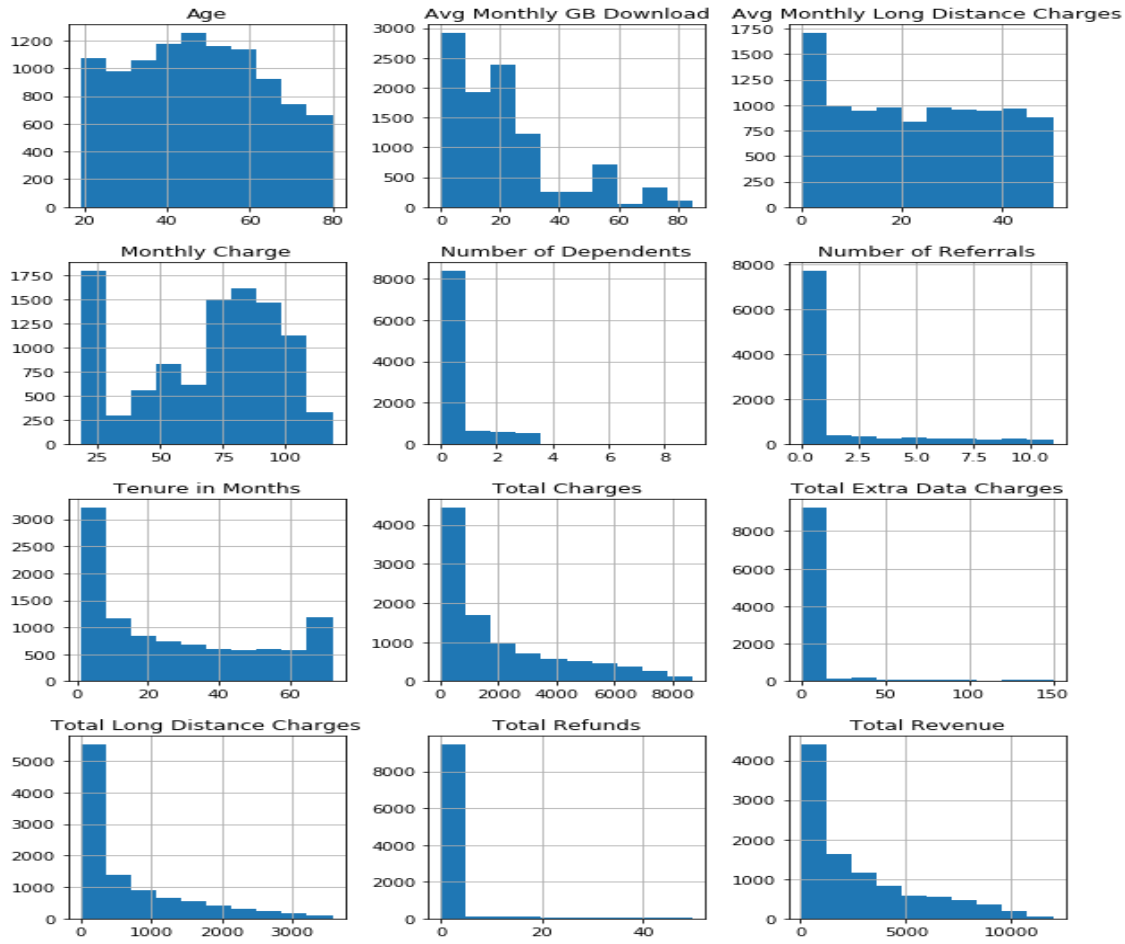
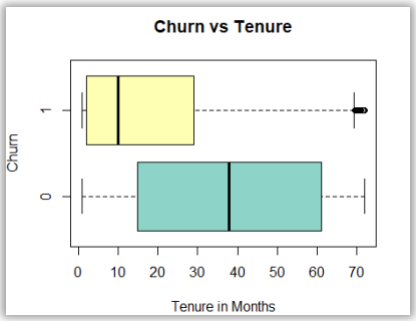


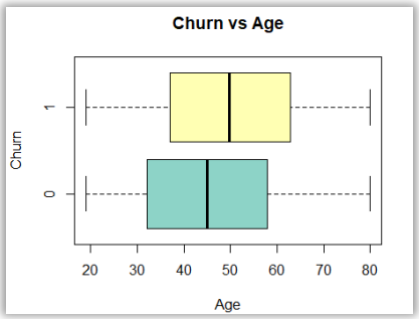
Figure 12 : Continuous Variables – Histogram Representation

Interpretation

- Age is normally distributed, major age group between 30 and 65
- Average monthly download is high for under 30 age group
- Monthly charge is highly distributed between 60 to 90.17% of customers have a monthly charge reached 1000
- More than 40% of customers have tenure less than 20 months

ChurnRelation with Continuous Variables





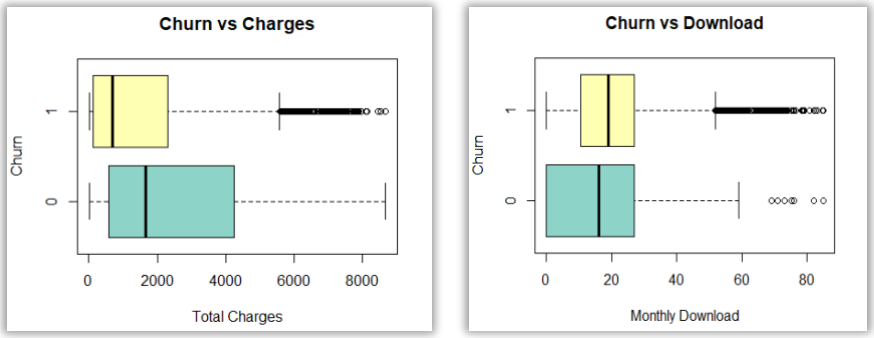


Figure 13 : Churn vs Independent variables

Interpretation

- High number of churn customers have a tenue less than 30 months. % is high in the initial 2 years of customer onboarding, this is closely linked with the type of contract.
- Age group between 35 – 65 contribute to high churn
- Churn increases with Total Charges reaching 2000 threshold on the lower end. At the higher end, there are customer churning with total charges greater than 6000
- Customer who do not have unlimited data service have a high chance of churn when the download usage is beyond 30GB

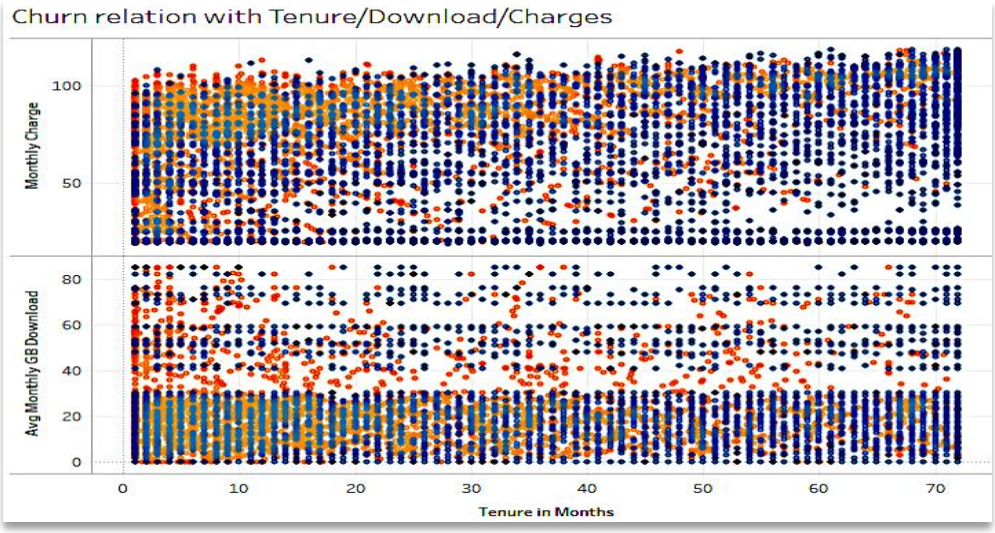
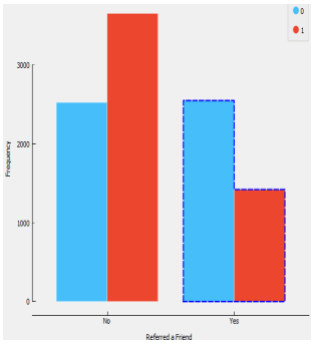


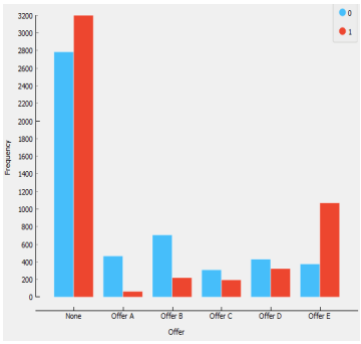
Figure 14: Monthly Charge, Download vs Tenure

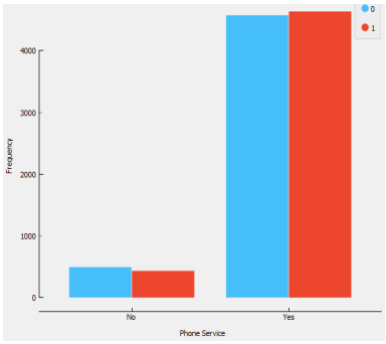
Interpretation

- Churn is spread and concentrated for download below 30 GB across all tenures
- Higher churn when tenure is low and monthly charge is high.

Churn Relation with Categorical Variables







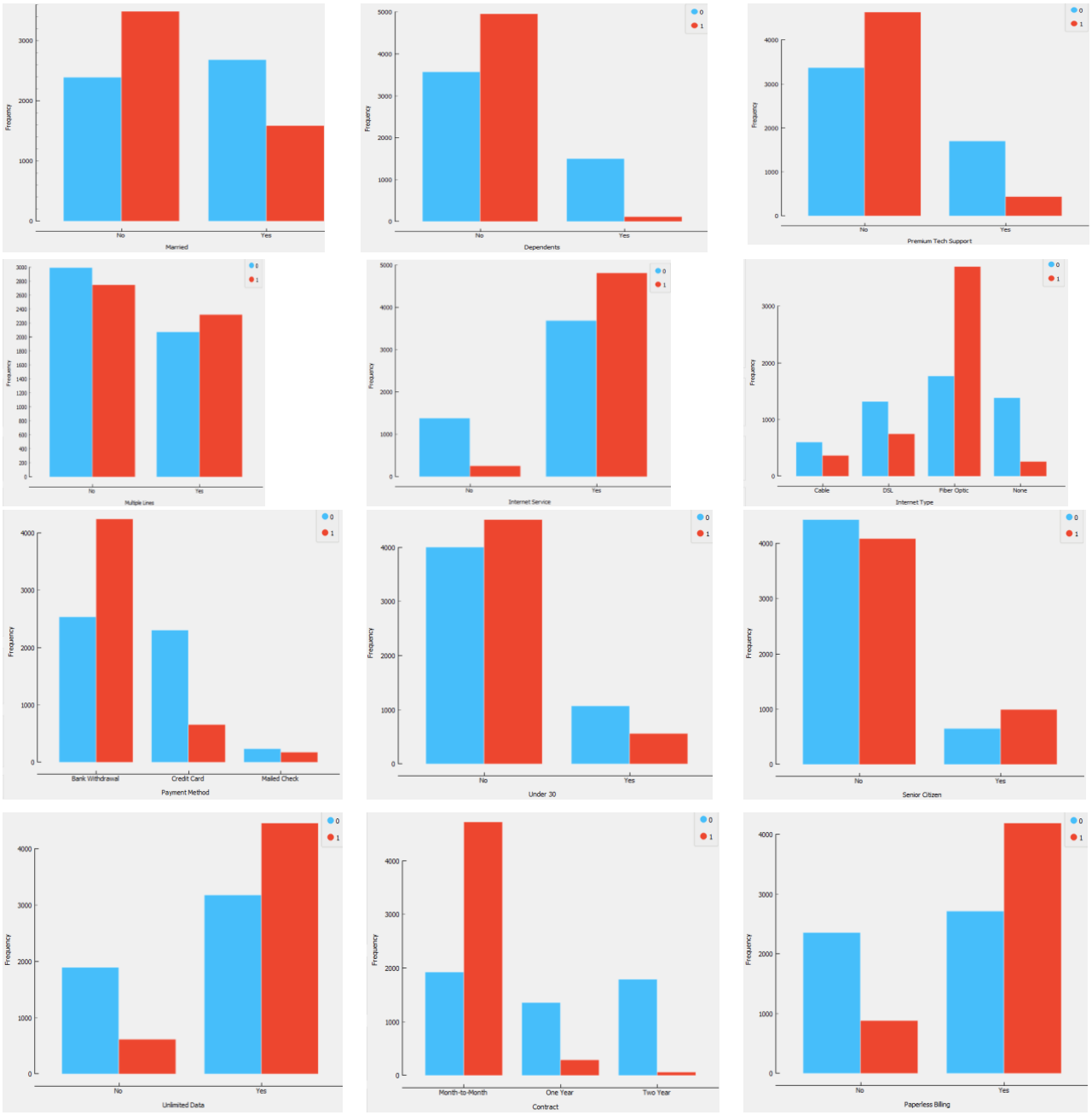


Figure 15: Churn vs Categorical Variable

Interpretation

- Churn is high in offer “None” and “Offer E”
- Churn happens in both Phone and Internet service
- Fibre Optic internet type has highest churn compared with other types
- High even with customers having multiple lines
- Churn high with customers not having Premium tech support
- Subscribers with unlimited data also have high churn
- Streaming and Online Services are not significant for Churn
- Month on Month has the highest proportion of churn
- Churn is major in age group 35 to 65

Churn Demography

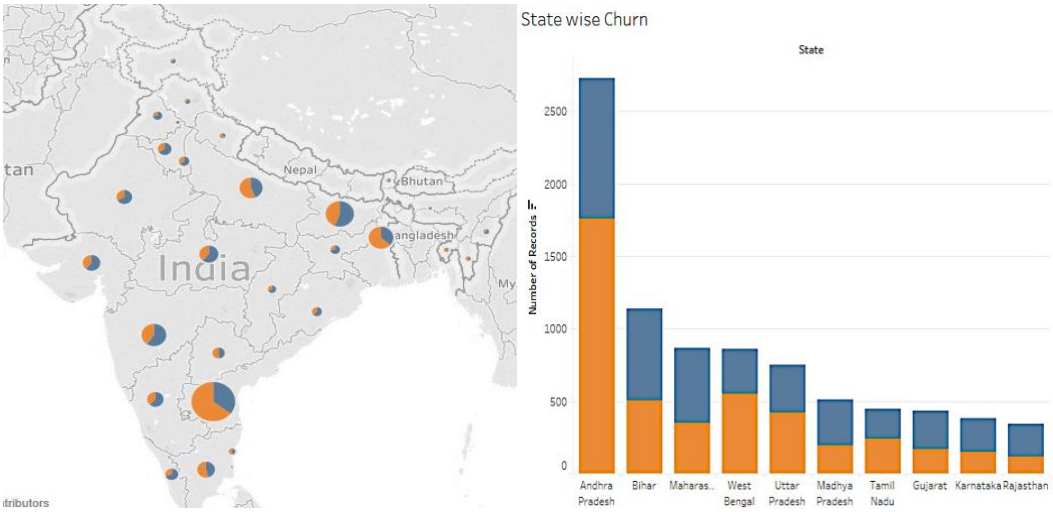
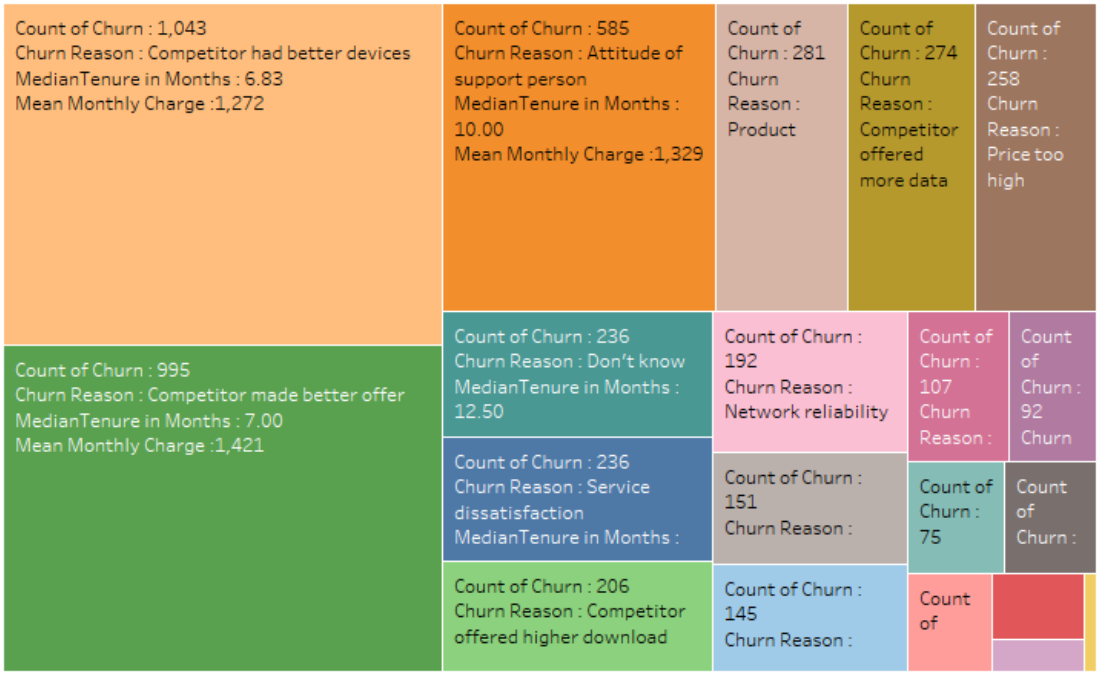


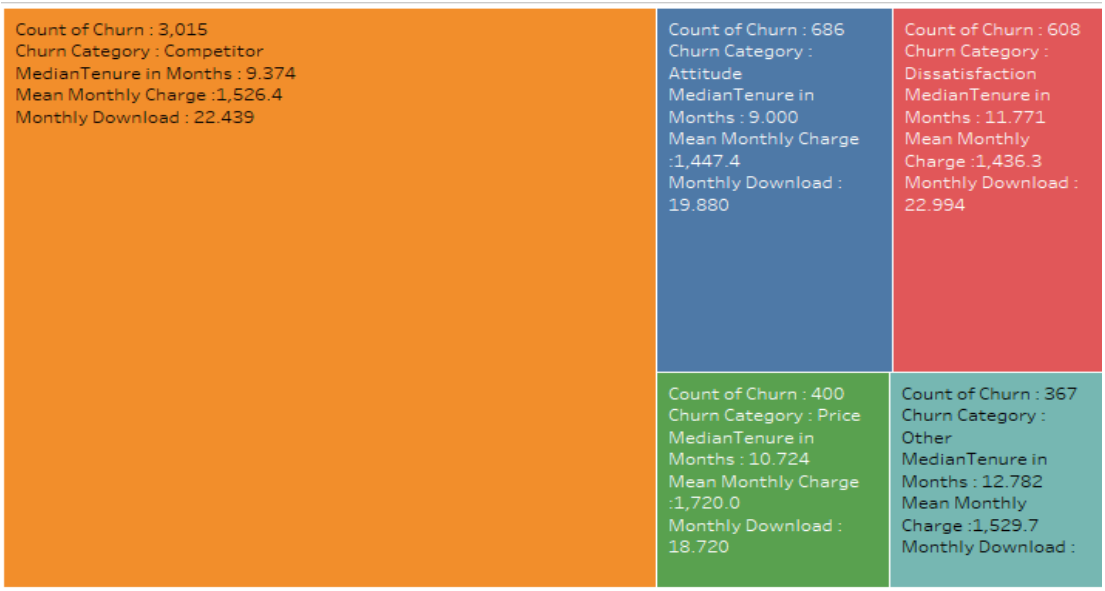
Figure 16: Demography

Interpretation

- Andhra Pradesh, Bihar and Maharashtra have the top 3 customer base
- High churns in Andhra Pradesh followed by West Bengal and Bihar
- Chittoor and Visakhapatnam have the high churn % across all cities

Churn Reasons





Interpretation

- High Churn rate due to competition
- Competitors provide better offers and data download limits
- Poor Customer support
- Dis-satisfaction due to call / network quality, limited services

Business Insights

- Around 40% of customers have referred a friend. Many new customers through referral indicates high advocacy rate or referral offers
- Around 59% of subscribers do not have any offers. It is recommended to revisit the Offer plans
- Around 67% of customers opt for month on month contract. The Service Provider has to think on the various contract options and plans to reduce attrition as this is closely correlated to Churn
- Average tenure of customers is less than 24 months which is a concern to work on to retain customers
- Premium Tech Support is engaged by 21% of customers indicating concerns in standard support service offered to customers
- Streaming services are least preferred
- Majority of customers are of age group in between 35 to 65 years
- Andhra Pradesh and Bihar have a high customer base and churn %
- Competition and Dissatisfaction, Attitude of support person are the major churn reasons
- Very High Churn Rate of 50%
- Service provider is losing customers heavily to competition primarily due to price, choice of products, data limits

Dataset: Call Performance

The data contains information on customer details like location details, network details, ratings and reason for call drop. Below is a brief introduction to variables considered for the problem statement

Overview of the Dataset

RJio is used widely across India with more people using 4G network type and more customers are satisfied with the network performance. Average customer ratings are above 3 where the ratings are in the range of 1 to 5.

Interpretation of summary of dataset

- The dataset has customer details like State of living, Latitude and Longitude of the location, Operator, Network Type, Travelling Type, Ratings given by the customer and reason for the call drop.
- No. of rows: 63336, No. of columns: 8
- State Name column has 15966 null values.
- Call Drop Category is the independent(target) variable.
- Latitude and longitude are **-1** and **0** for customers who turned off the location sharing.
- **Operator**: RJio is widely used across India
- **Network Type**: 4G network is used predominantly by customers.
- **Ratings**: On average, customers have given 3 ratings to the service.
- **Call Drop Category:** Majority of customers are satisfactory.

Study of target variable and relationship between independent variable

Call Drop Category (Target Variable)

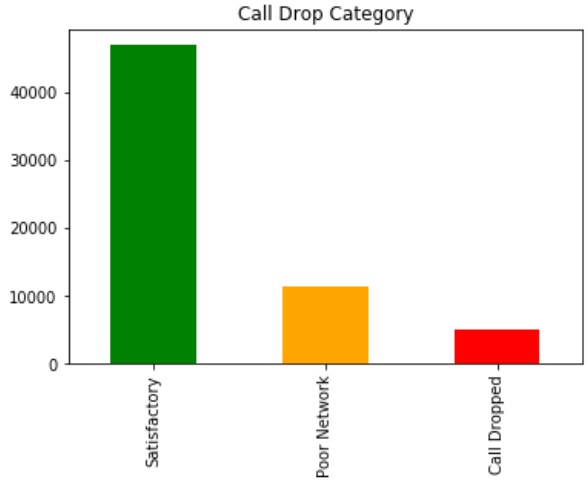


Figure 17: Call Drop Category

Interpretation

Most of the customers are happy and satisfied (74 %) with the customers and there is more feedback on poor network (17.7%) followed by call dropped feedback (8.06 %).

Relationship of Rating variable with Call Drop Category variable

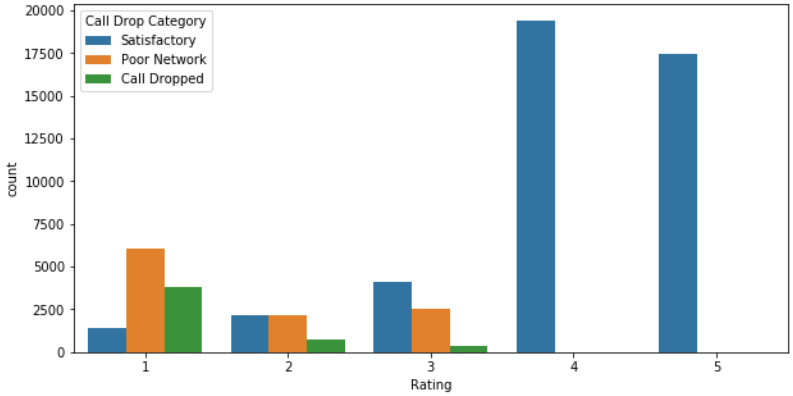


Figure 18: Rating Vs Call Drop Category

Interpretation

- Most of the satisfied customers have given good ratings of 4 and 5 and customer who had poor network and call dropped have given lesser ratings.
- There are customers who are satisfied and have given lesser ratings.

Latitude & Longitude co-ordinates

Interpretation

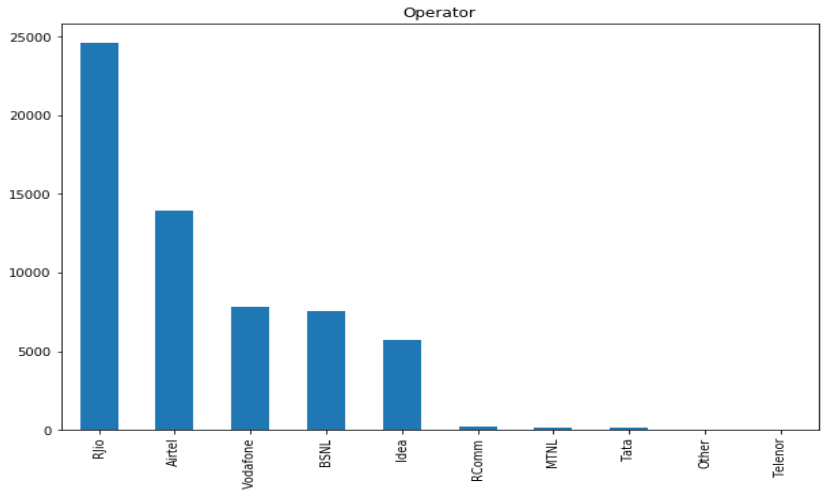
Call drops and poor network are observed cross India and it is not specific to any particular state.

Operator

Table 1 : Operator Distribution

Airtel	14864
Airvoice	8468
BSNL	8084
Idea	5734
Rcomm	217
Tata	197
MTNL	142
Other	36
Telenor	2

Figure 19: Operator



Interpretation

RJio, Airtel, Vodafone and BSNL are leading the market.

Relationship of Operator variable with Call Drop Category variable

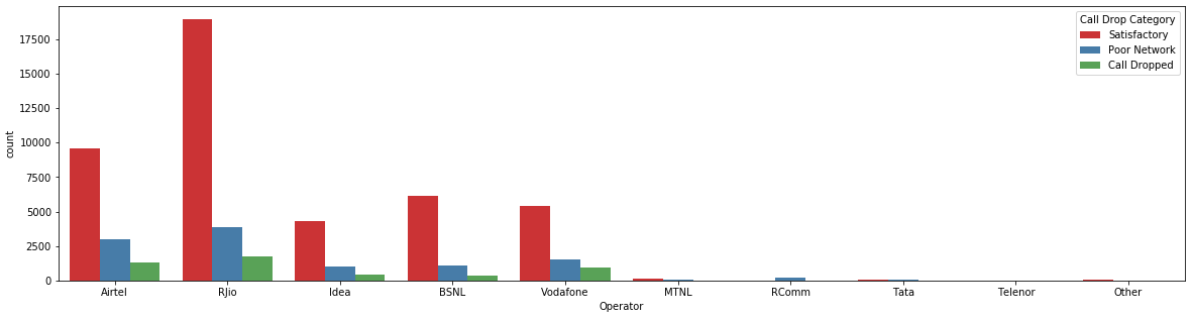


Figure 20: Operator Vs Call Drop Category

Interpretation

The below graph clearly says that most of the network had call drops and signs of poor network.

Network Type

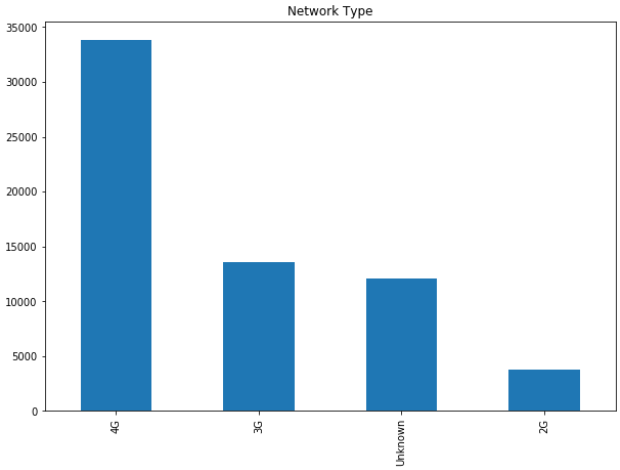


Figure 21: Network Type

Interpretation

Most of the customers are using 4G, followed by 3G and 2G. Network types are not identified for 12k customers.

Relationship between Network Type and Call Drop Category

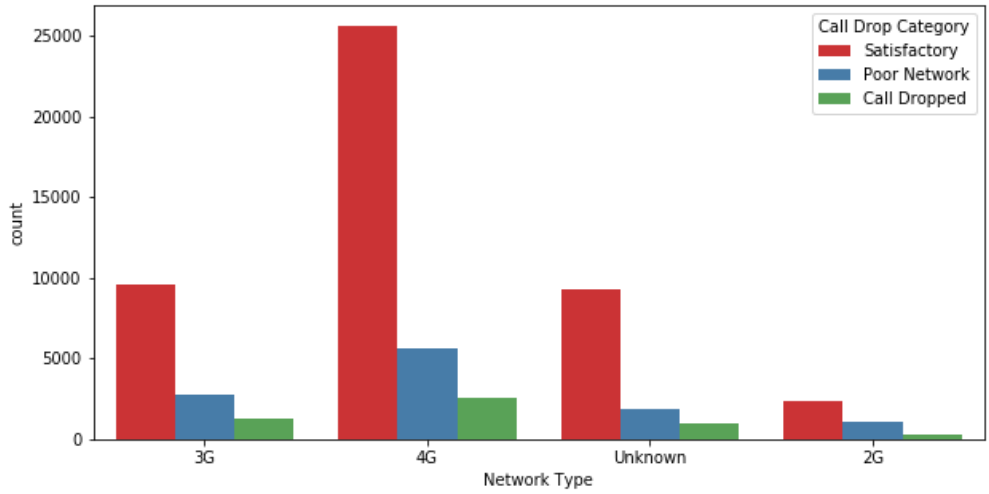


Figure 22: Network Type Vs Call Drop Category

Interpretation

- Call drops, poor network is again distributed across different network type.
- But the satisfaction is more with 4G network.

Indoor Outdoor Travelling

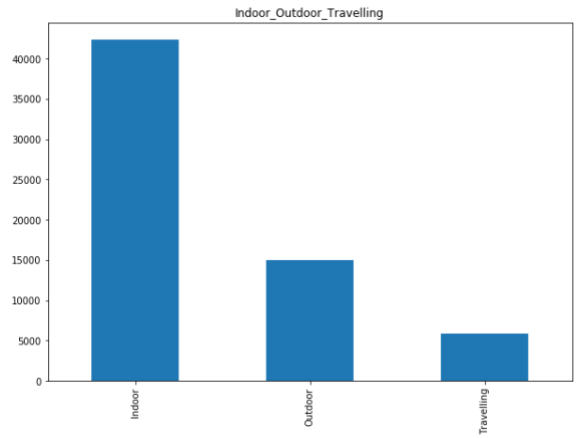


Figure 23: Travel Type

Interpretation

Indoor usage was extremely high comparatively to Outdoor and Travelling.

Relationship between Network Type and Call Drop Category

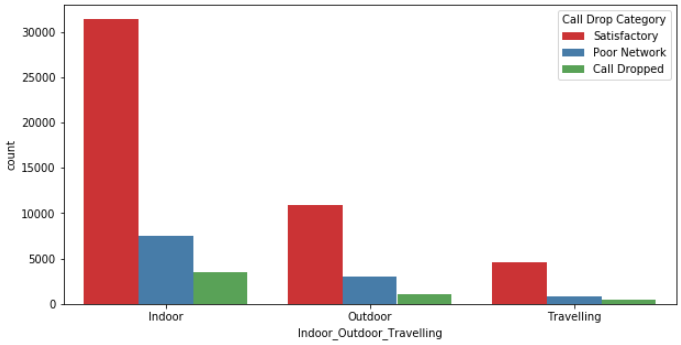


Figure 24: Travel Type vs Call Drop Category

Interpretation

Call drops are occurring irrespective of where the customer is indoor, outdoor or travelling.

Relationship between State and Call Drop Category

State Name	Call Dropped	Poor Network	Satisfactory
Andhra Pradesh	183	326	1114
Arunachal Pradesh	1	1	3
Assam	25	49	594
Bihar	78	207	575
Chandigarh	1	6	103
Chhattisgarh	39	54	439
Dadra and Nagar Haveli	1	1	13
Goa	3	10	114
Gujarat	204	536	2275
Haryana	104	401	1200
Himachal Pradesh	3	19	242
Jharkhand	23	132	331
Karnataka	251	803	2575
Kashmir	1	0	101
Kerala	59	234	1146
Madhya Pradesh	107	148	1458
Maharashtra	1081	1449	6799
Manipur	1	0	0
Meghalaya	2	1	2
NCT	217	710	1755
Nagaland	0	3	9
Odisha	48	419	1541
Pondicherry	1	2	22
Punjab	27	65	382
Rajasthan	97	404	666
Sikkim	2	0	5
Tamil Nadu	292	304	3158
Telangana	220	510	2064
Tripura	0	0	4
Uttar Pradesh	272	443	3413
Uttarakhand	20	22	256
West Bengal	293	656	3391

Figure 25: States Vs Call Drop Category

Interpretation

As Maharashtra, West Bengal, UP, Tamil Nadu and Karnataka are the leading states, more call drops or poor network is also noticed in all top states.

Business Insights

- Most of the customers have used **4G network** with more network connections coming from **Maharashtra.**
- Call Drops and Poor Network has been observed across all the networks including Rjio, Airtel, Vodafone and BSNL.
- **Satisfied** customers have given 4 or 5 ratings, while few customers have given **lesser** ratings.
- Customers have tried to use the network mostly indoors but call drops and poor network are observed even if the call is placed while the customer is indoor or travelling and outdoor.
- **RJio, Airtel and Vodafone** are the leading network operators.
- Call drops and poor network are observed across India.

Dataset: Network Intrusion

Overview of the Data

The telecom networks are growing larger in recent years. As a great variety of people all over the world are connecting, they are unconsciously encountering the number of security threats such as viruses, worms and attacks from hackers. Now firewalls, anti-virus software, message encryption, secured network protocols, password protection and so on are not sufficient to assure the security in networks, when some intrusions take advantages of weaknesses in systems to threaten. The project is focused on developing machine learning (ML) algorithms to classify various network attacks.

The dataset is used to build a model that can distinguish between and classify good connections and bad connections. The attacks fall into four main classes:

DoS	Back, Land, Neptune, Pod, Smurf, Teardrop
R2L	ftp_write, guess_passed, imap, multihop, phfspy, warezclient, warezmaster
U2R	Buffer_overfloe, loadmodule, perl, rootkit
PROBE	ipsweep,nmap,portsweep,satan

Table 2: Network Attack Types

Introduction to Attack Types

- **DoS**- Denial of DOS: denial-of-service, e.g. syn flood;
- **R2L**- Unauthorized access from a remote machine, e.g. guessing password;
- **U2R**- Unauthorized access to local superuser (root) privileges, e.g., various ``buffer overflow'' attacks;
- **Probing**- Surveillance and other probing, e.g., port scanning.

Distribution of Attacks

U2R	**0.03%**
R2L	**0.78%**
Probe	**9.27%**
No Attack	**53.46%**
DoS	**36.46%**

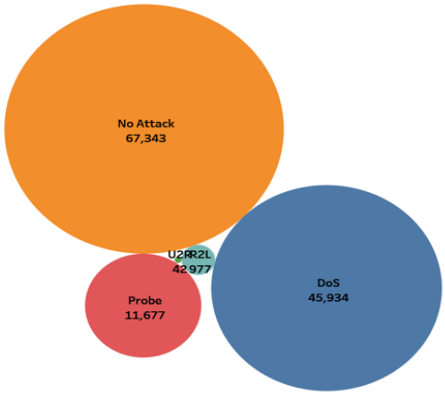


Figure 26: Distribution of Attacks

Study of Attack with Independent variables

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 125973 entries, 0 to 125972
Data columns (total 45 columns):
duration                        125973 non-null int64
protocol_type_icmp              125973 non-null int64
protocol_type_tcp               125973 non-null int64
protocol_type_udp               125973 non-null int64
src_bytes                       125973 non-null int64
dst_bytes                       125973 non-null int64
land                            125973 non-null int64
wrong_fragment                  125973 non-null int64
urgent                          125973 non-null int64
hot                             125973 non-null int64
num_failed_logins               125973 non-null int64
logged_in                       125973 non-null int64
num_compromised                 125973 non-null int64
root_shell                      125973 non-null int64
su_attempted                    125973 non-null int64
num_root                        125973 non-null int64
num_file_creations              125973 non-null int64
num_shells                      125973 non-null int64
num_access_files                125973 non-null int64
num_outbound_cmds               125973 non-null int64
is_host_login                   125973 non-null int64
is_guest_login                  125973 non-null int64
count                           125973 non-null int64
srv_count                       125973 non-null int64
serror_rate                     125973 non-null float64
srv_serror_rate                 125973 non-null float64
rerror_rate                     125973 non-null float64
srv_rerror_rate                 125973 non-null float64
same_srv_rate                   125973 non-null float64
diff_srv_rate                   125973 non-null float64
srv_diff_host_rate              125973 non-null float64
dst_host_count                  125973 non-null int64
dst_host_srv_count              125973 non-null int64
dst_host_same_srv_rate          125973 non-null float64
dst_host_diff_srv_rate          125973 non-null float64
dst_host_same_src_port_rate     125973 non-null float64
dst_host_srv_diff_host_rate     125973 non-null float64
dst_host_serror_rate            125973 non-null float64
dst_host_srv_serror_rate        125973 non-null float64
dst_host_rerror_rate            125973 non-null float64
dst_host_srv_rerror_rate        125973 non-null float64
class_types                     125973 non-null object
class                           125973 non-null object
target class                    125973 non-null int64
diff_level                      125973 non-null int64
dtypes: float64(15), int64(28), object(2)
memory usage: 43.2+ MB
```

Interpretation

- Dataset has 125972 records and 45 variables.
- The variable target class is the target categorical and binary variable.
- There are no missing values.
- There are 6 binary features,3 nominal / character and 32 continuous.
- There are 125972 records out of which 67343 are No attack connects and 58629 are Attack Connections.

Sample Data

	duration	protocol_type_icmp	protocol_type_tcp	protocol_type_udp	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	...	dst_host_same_src_port_r
0	0	0	1	0	491	0	0	0	0	0	...	0
1	0	0	0	1	146	0	0	0	0	0	...	0
2	0	0	1	0	0	0	0	0	0	0	...	0
3	0	0	1	0	232	8153	0	0	0	0	...	0
4	0	0	1	0	199	420	0	0	0	0	...	0

5 rows × 45 columns

Five Point Summary

	duration	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	num_failed_logins	logged_in
count	125973.00000	1.259730e+05	1.259730e+05	125973.000000	125973.000000	125973.000000	125973.000000	125973.000000	125973.000000
mean	287.14465	4.556674e+04	1.977911e+04	0.000198	0.022687	0.000111	0.204409	0.001222	0.395736
std	2604.51531	5.870331e+06	4.021269e+06	0.014086	0.253530	0.014366	2.149968	0.045239	0.489010
min	0.00000	0.000000e+00	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.00000	0.000000e+00	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.00000	4.400000e+01	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	0.00000	2.760000e+02	5.160000e+02	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
max	42908.00000	1.379964e+09	1.309937e+09	1.000000	3.000000	3.000000	77.000000	5.000000	1.000000

	num_failed_logins	logged_in	num_compromised	...	dst_host_srv_count	dst_host_same_srv_rate	dst_host_diff_srv_rate	dst_host_same_src_port_rate
count	125973.000000	125973.000000	125973.000000	...	125973.000000	125973.000000	125973.000000	125973.000000
mean	0.001222	0.395736	0.279250	...	115.653005	0.521242	0.082951	0.148379
std	0.045239	0.489010	23.942042	...	110.702741	0.448949	0.188922	0.308997
min	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	...	10.000000	0.050000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	...	63.000000	0.510000	0.020000	0.000000
75%	0.000000	1.000000	0.000000	...	255.000000	1.000000	0.070000	0.060000
max	5.000000	1.000000	7479.000000	...	255.000000	1.000000	1.000000	1.000000

	dst_host_srv_diff_host_rate	dst_host_serror_rate	dst_host_srv_serror_rate	dst_host_rerror_rate	dst_host_srv_rerror_rate	diff_level
count	125973.000000	125973.000000	125973.000000	125973.000000	125973.000000	125973.000000
mean	0.032542	0.284452	0.278485	0.118832	0.120240	19.504060
std	0.112564	0.444784	0.445669	0.306557	0.319459	2.291503
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	18.000000
50%	0.000000	0.000000	0.000000	0.000000	0.000000	20.000000
75%	0.020000	1.000000	1.000000	0.000000	0.000000	21.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	21.000000

Figure 27: Summary

Checking Missing Values

There are no null values in the data.

Basic Features

feature name	*description*	*type*
duration	length (number of seconds) of the connection	continuous
protocol_type	type of the protocol, e.g. tcp, udp, etc.	discrete
service	network service on the destination, e.g., http, telnet, etc.	discrete
src_bytes	number of data bytes from source to destination	continuous
dst_bytes	number of data bytes from destination to source	continuous
flag	normal or error status of the connection	discrete
land	1 if connection is from/to the same host/port; 0 otherwise	discrete
wrong_fragment	number of ``wrong'' fragments	continuous
urgent	number of urgent packets	continuous

Duration Vs Attacks

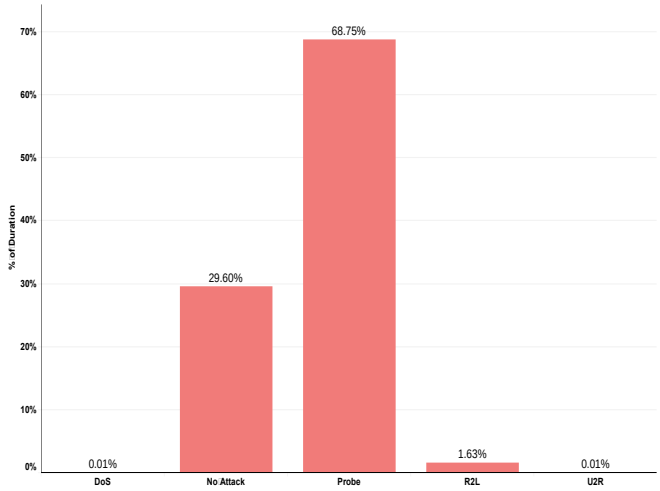


Figure 28: Duration of the Attacks

Interpretation

- Duration of Probe attack connections are higher compared to other connection.
- 31.39 % of total duration is accounted by good Connections

Protocol Vs Attacks

Protocol variable has 3 categories: TCP, CMP & UDP.

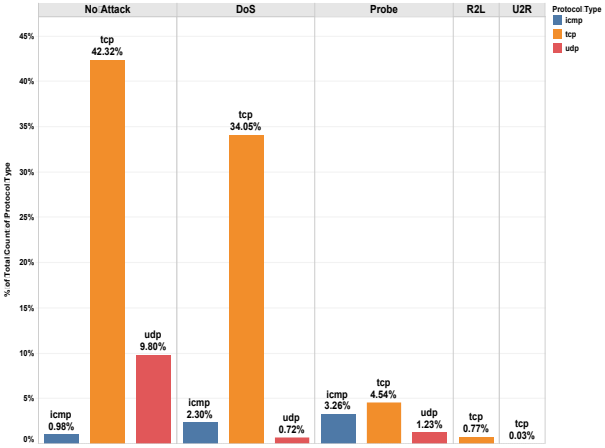


Figure 29: Protocol Vs Attacks

Interpretation

- 79.59% of no attack connections used TCP.
- 91.8 % of DoS attack used TCP protocol.
- 50% of Probe attack are via TCP followed by ICMP with 34.5%.
- TCP protocol is used for all the R2L and U2R attack connections.

Service Vs Attacks

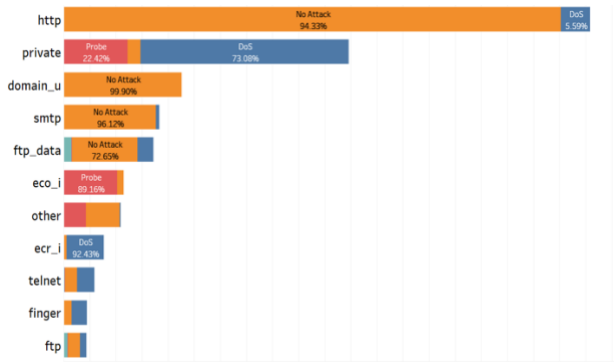


Figure 30: Service Vs Attacks

Interpretation

- http service is used widely.
- 94% of http is used by No Attack.
- DoS attack used 73% private services.
- % 99.9 of Domain_u service is used by No attack.
- 80% Eco_i service is used by Probe attack.
- 92% of ecr_i service is used by DoS.

Land vs Attacks

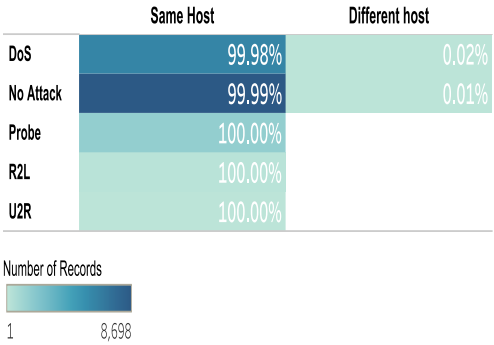


Figure 31: Land vs Attacks

Interpretation

- Most of the connections are from or to the same host.

Content Features

feature name	*description*	*type*
hot	number of ``hot'' indicators	continuous
num_failed_logins	number of failed login attempts	continuous
logged_in	1 if successfully logged in; 0 otherwise	discrete
num_compromised	number of ``compromised'' conditions	continuous
root_shell	1 if root shell is obtained; 0 otherwise	discrete
su_attempted	1 if ``su root'' command attempted; 0 otherwise	discrete
num_root	number of ``root'' accesses	continuous
num_file_creations	number of file creation operations	continuous
num_shells	number of shell prompts	continuous
num_access_files	number of operations on access control files	continuous
num_outbound_cmds	number of outbound commands in an ftp session	continuous
is_hot_login	1 if the login belongs to the ``hot'' list; 0 otherwise	discrete
is_guest_login	1 if the login is a ``guest''login; 0 otherwise	discrete

Figure 32: Content Features

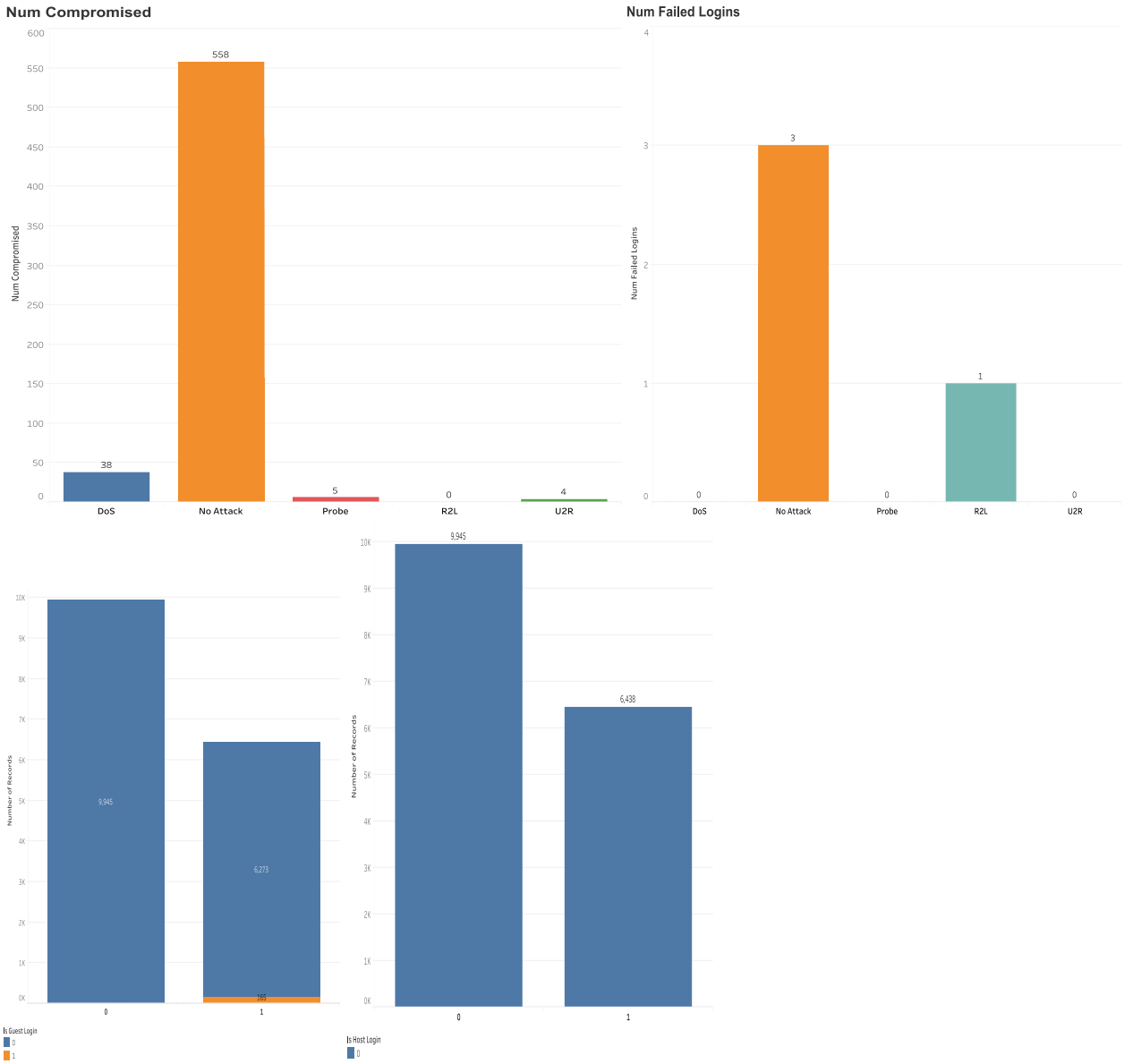


Figure 33: Login Types Vs Attacks

Interpretation

- 6,438 logins were observed. 165 of those logins seems to be Guest login.
- Total number of compromised was 49.
- There was only one login failed during the attack.

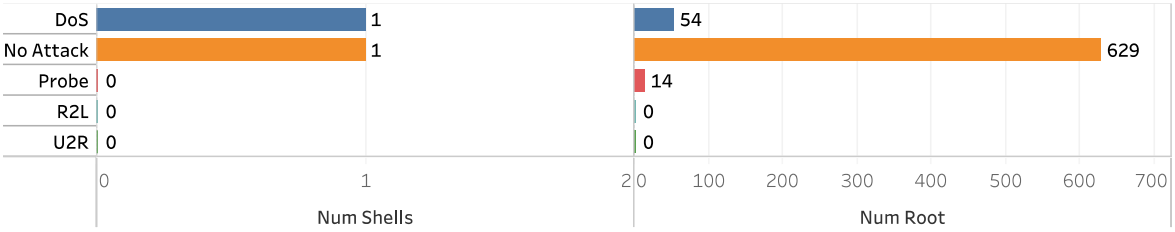


Figure 34: Root Access vs Attacks

Interpretation

- Number of root access has been more.
- DoS & Probing attack had 54 and 14 times root accesses.

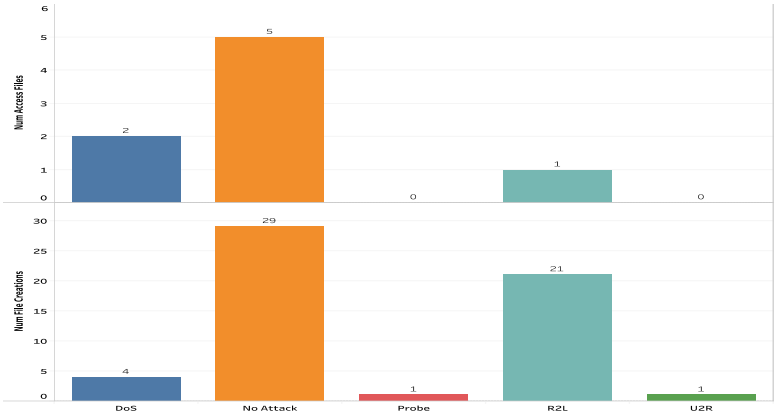


Figure 35: File Access vs Attacks

Interpretation

- 3 times the operation was performed on access control files and 27 file related operations were performed during the attack.

Traffic Features

feature name	*description*	*type*
count	number of connections to the same host as the current connection in the past two seconds	continuous
	Note: The following features refer to these same-host connections.	
serror_rate	% of connections that have ``SYN'' errors	continuous
rerror_rate	% of connections that have ``REJ'' errors	continuous
same_srv_rate	% of connections to the same service	continuous
diff_srv_rate	% of connections to different services	continuous
srv_count	number of connections to the same service as the current connection in the past two seconds	continuous
	Note: The following features refer to these same-service connections.	
srv_serror_rate	% of connections that have ``SYN'' errors	continuous
srv_rerror_rate	% of connections that have ``REJ'' errors	continuous
srv_diff_host_rate	% of connections to different hosts	continuous

Figure 36: Traffic Features

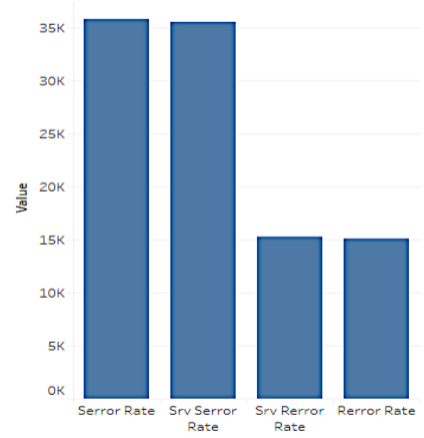


Figure 37: Error Rate

Interpretation

- 35k connections of SYN & REJ errors have been observed in same host connections.
- 15k connections of SYN & REJ errors have been observed in same service connections.

Visualizing the Continuous Variables

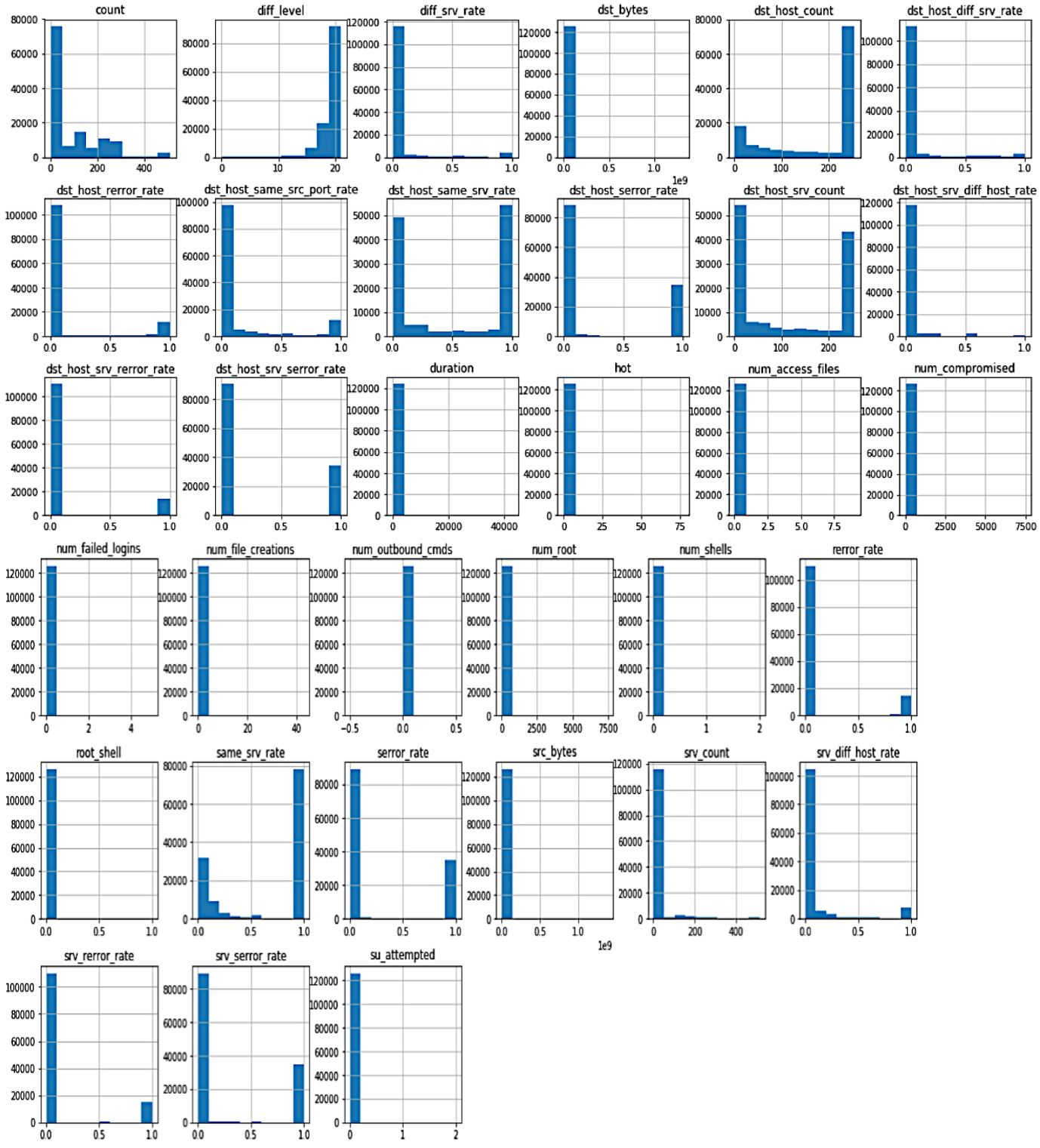


Figure 38: Continuous Variables

Interpretation

- From the above histogram the variables dst_host_rerror_rate, dst_host_serror_rate, dst_host_srv_serror_rate,hot,num_acces_files,num_compromised_num_failed_logins,num_outbound_cmds,num_file_creations,num_root,num_shell,root_shell,su_attempted most of the values re 0's.
- The variable count ranges from 0 to 600.
- Dst_hot_count has min value of 0 and more than 70000 values are above 200.
- Srv_serror_rate had 80% of the values are 0's and the remaining are 1's

- Thus the distribution of data is not normal.

Multi-collinearity

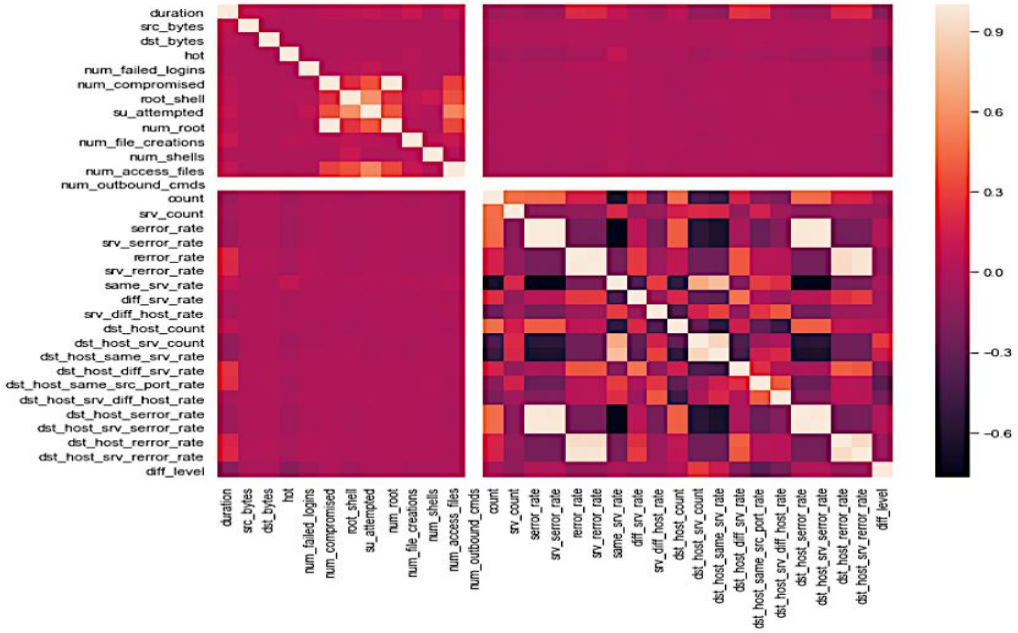


Figure 39: Multicollinearity

Interpretation

- The variables num_root and num_compromised are highly positively correlated.
- Serror_rate, srv_serror ,dst_host_serror_rate and dst_host_srv_serror_rate are highly positively correlated.
- rerror_rate, srv_rerror ,dst_host_rerror_rate and dst_host_srv_rerror_rate are highly positively correlated.

Chapter 4: Model Building

Feature engineering is performed on the dataset before applying to the machine learning algorithm to generate better model with best accuracy.

Dataset: Customer Churn

Feature Engineering

Removal of Duplicate Observations

- Duplicate records are removed from the dataset and the number of unique observations is 6945.

Missing Value Imputation

- Missing and Null values in the variables has also dropped after removal of duplicates which skewed the data.
- All Categorical Variables with NAs are replaced by extracting the information from related variables or based on the frequency of categorical values.
- For Continuous variables like "Total Refund" and "Number of Dependents", instead of mean or median, the value is replaced with "0" based on density function.
- The overall number of observations after cleaning is 6945 without losing information due to NA and 1869 / 6945 customer have churned which is around 27%.

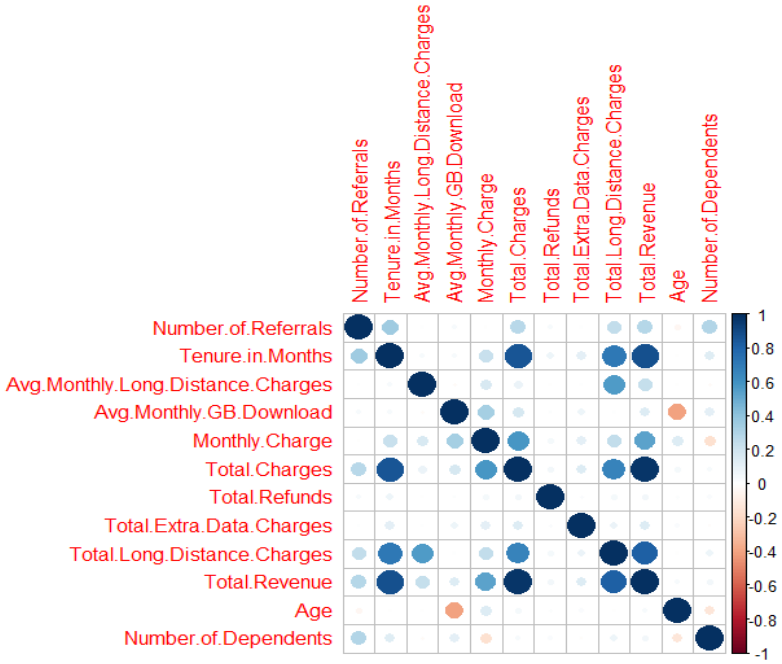
Data Balancing

- The dataset which is arrived after wrangling has a total of 6945 observations with 27% of churn.
- This is imbalanced and we shall do Synthetic oversampling of minority class to balance the dataset.
- **SMOTEN** is used to balance the data, as the dataset contains both categorical and continuous
- variables. This balanced dataset after SMOTEN has 10152 observations with equal proportions of
- churn customers. This data set will be used for further exploratory analysis.

Feature Selection

In order to obtain an accurate model, it is necessary to refine the data that adds more contribution to the model through feature selection techniques – Correlation plot, Variable significance based on regression, VIF, Information Gain.

Correlation Plot



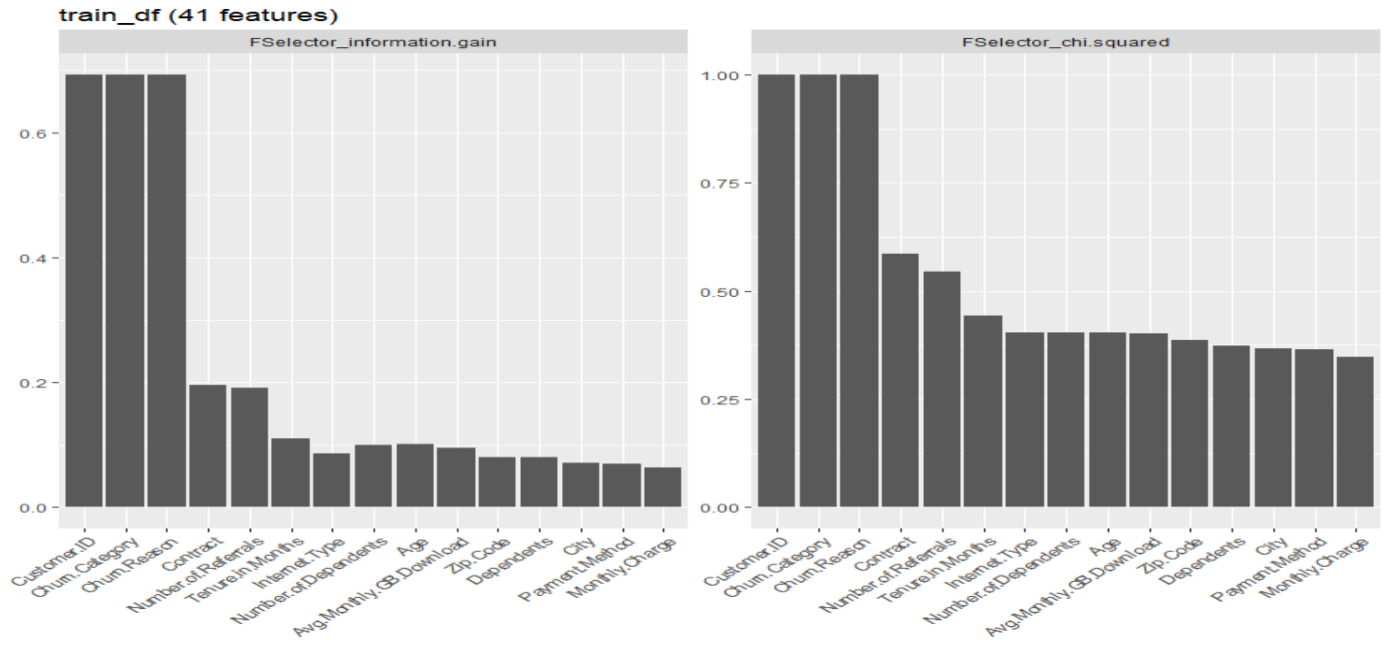
Variance Inflation Factor &Multi-collinearity

High intercorrelation amongst the independent variables leads to multicollinearity, which may lead to erroneous/imprecise estimation of regression coefficients. Using Variance Inflation Factor (VIF) is a way of treating for multicollinearity.

The optimal cut-off for VIF is set at 5. City and residing city, followed by tier flag and gender show a VIF >5.

Number.of.Referrals	1.520055	1	1.232905
Tenure.in.Months	13.244176	1	3.639255
Offer	2.186023	5	1.081348
Phone.Service	1.484967	1	1.218592
Multiple.Lines	1.627507	1	1.275738
Internet.Type	5.403116	3	1.324669
Avg.Monthly.GB.Download	2.369272	1	1.539244
Online.Security	1.152125	1	1.073371
Online.Backup	1.395772	1	1.181428
Device.Protection.Plan	1.424875	1	1.193681
Premium.Tech.Support	1.211084	1	1.100493
Streaming.TV	1.691215	1	1.300467
Streaming.Movies	3.215653	1	1.793224
Streaming.Music	2.973325	1	1.724333
Unlimited.Data	1.986362	1	1.409384
Contract	1.717019	2	1.144705
Paperless.Billing	1.189814	1	1.090786
Payment.Method	1.205567	2	1.047847
Total.Charges	17.629418	1	4.198740
Gender	1.016829	1	1.008379
Age	2.579729	1	1.606153
Under.30	2.153914	1	1.467622
Senior.Citizen	1.841046	1	1.356852
Married	1.845446	1	1.358472
Dependents	3.502118	1	1.871395
Number.of.Dependents	3.500228	1	1.870890

Variable Information Gain



Logistic Regression

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	819	196
1	73	301

Table 3: Logistic Regression - Confusion Matrix

	Accuracy	Precision	F1
Logistic Regression	0.8063	0.8048	0.6912

Table 4: Scores - Logistic Regression

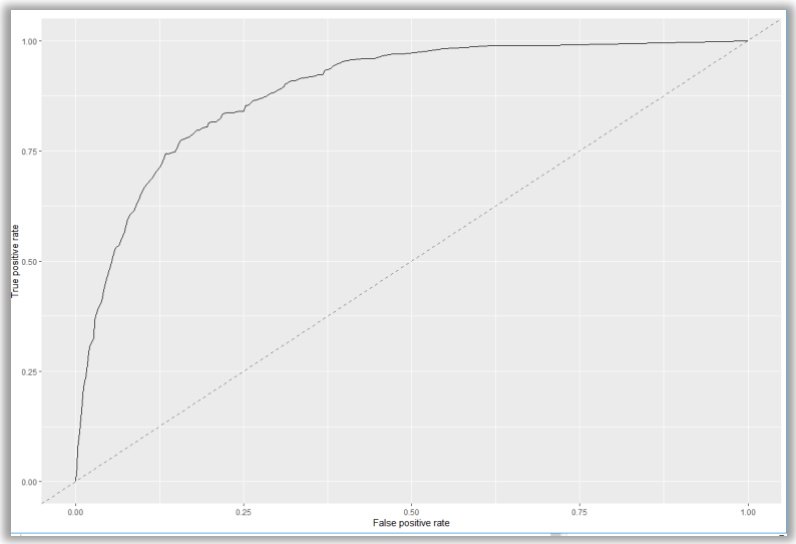


Figure 40: ROC Curve

AUC	**0.8882695**

Table 5: AUC - LR

There is no tuning parameter in logistic Regression.

Naive Bayes

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	756	259
1	72	302

Table 6: Naive Bayes -Confusion Matrix

	Accuracy	Precision	F1
Naive Bayes	0.7617	0.8075	0.6460

Table 7: Naïve Bayes - Scores

AUC	**0.8446801**

Applying Cross Validation to check accuracy of the model.

Hyperparameter Tuning

Confusion Matrix and Statistics

Prediction	Reference	
	0	1

0	755	260
1	72	302

Table 8: Naive Bayes - HPT Confusion Matrix

	Accuracy	Precision	F1
Naive Bayes	0.761	0.8075	0.6453

Table 9: Naïve Bayes–HPT Scores

LASSO Regression

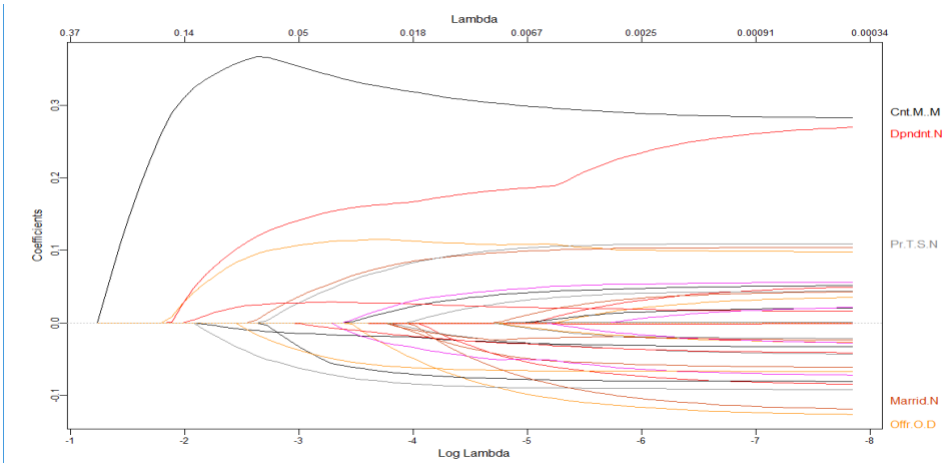


Figure 41: LASSO Regression

Observation

Contract, Dependents, Premium tech support, Married, Offers are the top 5 variables identified using LASSO regression.

RIDGE Regression

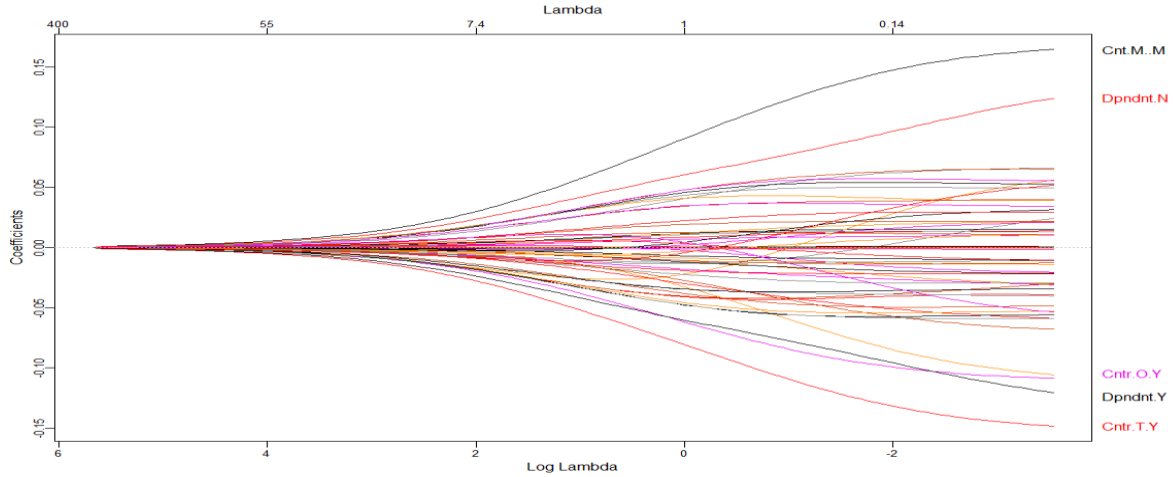


Figure 42: RIDGE Regression

Observation

Contract and Dependents are the two variables of high importance identified in RIDGE regression.

LDA - Linear Discriminant Analysis

Confusion Matrix and Statistics

Prediction	Reference	
	0	**1**
0	796	219
1	69	305

Table 10: LDA- Confusion Matrix

	Accuracy	Precision	F1
LDA	0.7927	0.8155	0.6793

Table 11: LDA - Scores

AUC	**0.8819841**

Table 12: LDA -AUC

Hyperparameter Tuning

Confusion Matrix and Statistics

Prediction	Reference	
	0	**1**
0	796	219
1	69	305

Table 13: LDA- HPT Confusion Matrix

	Accuracy	Precision	F1
LDA	0.7927	0.8155	0.6793

Table 14: LDA –HPT Scores

CART – Decision Trees

Confusion Matrix and Statistics

Prediction	Reference	
	0	**1**
0	860	155
1	110	264

Table 15: CART -Confusion Matrix

	Accuracy	Precision	F1
CART	0.8092	0.7059	0.6658

Table 16: CART - Scores

AUC	**0.8319064**

Table 17: CART - AUC

Hyperparameter Tuning

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	853	162
1	115	259

Table 18: CART –HPT Confusion Matrix

	Accuracy	Precision	F1
CART	0.8006	0.6925	0.6516

Table 19: CART-HPT Scores

Random Forest - Bagging

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	1006	9
1	6	368

Table 20: RF-Confusion Matrix

	Accuracy	Precision	F1
Random Forest	0.9892	0.9840	0.9800

Table 21: RF - Scores

AUC	0.9993757

Table 22: RF - AUC

Hyperparameter Tuning

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	958	57
1	49	325

Table 23: RF- HPT Confusion Matrix

	Accuracy	Precision	F1
Random Forest	0.9237	0.8690	0.8598

Table 24: RF–HPT Scores

Support Vector Machine

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	882	133
1	79	295

Table 25: SVM- Confusion Matrix

	Accuracy	Precision	F1
SVM	0.9892	0.9840	0.9800

Table 26: SVM - Scores

AUC	0.9166829

Table 27: SVM - AUC

Hyperparameter Tuning

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	931	84
1	50	324

Table 28: SVM- HPT Confusion Matrix

	Accuracy	Precision	F1
SVM	0.9035	0.8663	0.8286

Table 29: SVM –HPT Scores

GBM – Gradient Boosting

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	806	209
1	71	303

Table 30: GBM- Confusion Matrix

	Accuracy	Precision	F1
GBM	0.7984	0.8102	0.6840

Table 31: GBM- Scores

AUC	0.8870788

Table 32: GBM - AUC

Hyperparameter Tuning

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	914	101
1	77	297

Table 33: GBM- HPT Confusion Matrix

	Accuracy	Precision	F1
GBM	0.8719	0.7941	0.7694

Table 34: GBM–HPT Scores

ADA- Adaptive Boosting

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	872	143
1	94	280

Table 35: ADA- Confusion Matrix

	Accuracy	Precision	F1
ADA	0.8294	0.7487	0.7026

Table 36: ADA - Scores

AUC	0.9073957

Table 37: ADA - AUC

Hyperparameter Tuning

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	898	117
1	78	296

Table 38: ADA- HPT Confusion Matrix

	Accuracy	Precision	F1
ADA	0.8596	0.7914	0.7522

Table 39: ADA– HPT Scores

KKNN – K Neatest Neighbours

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	951	64
1	31	343

Table 40: KKNN–Confusion Matrix

	Accuracy	Precision	F1
KKNN	0.9316	0.9171	0.8784

Table 41: KKNN– Scores

AUC	0.9866942

Table 42: KKNN– AUC

Hyperparameter Tuning

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	936	79
1	65	309

Table 43: KKNN– HPT Confusion Matrix

	Accuracy	Precision	F1
KKNN	0.8963	0.8262	0.8110

Table 44: KKNN– HPT Scores

Neural Networks

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	1193	330
1	92	469

Table 45: NNET- Confusion Matrix

	Accuracy	Precision	F1
NNET	0.7975	0.8360	0.6897

Table 46: NNET - Scores

AUC	0.8136008

Table 47: NNET - AUC

Hyperparameter Tuning

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	948	575
1	70	491

Table 48: NN -HPT Confusion Matrix

	Accuracy	Precision	F1
NNET	0.8905	0.8752	0.6036

Table 49: NN–HPT Scores

Comparison of models with default parameters

	Accuracy	Precision	F1
Logistic Regression	81%	80%	69%
Naive Bayes	76%	81%	65%
LDA	79%	82%	68%
CART	81%	71%	67%
Random Forest	99%	98%	98%
SVM	85%	79%	74%
GBM	80%	81%	68%
ADA	83%	75%	70%
KKNN	93%	92%	88%
NN	80%	84%	69%

Table 50: Model Score Comparison

Random Forest is giving highest accuracy with 98.92 % with default parameters.

Comparison of models with hyper parameters

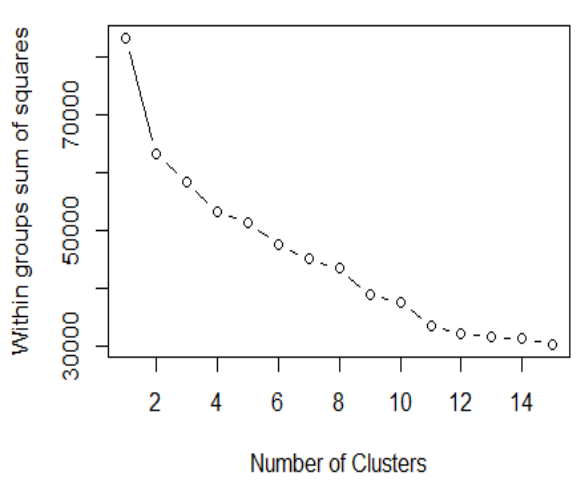
	Accuracy	Precision	F1
Naive Bayes	76%	81%	65%
LDA	79%	82%	68%
CART	80%	69%	65%
Random Forest	92%	87%	86%
SVM	90%	87%	83%
GBM	87%	79%	77%
ADA	86%	79%	75%
KKNN	90%	83%	81%
NN	69%	87%	60%

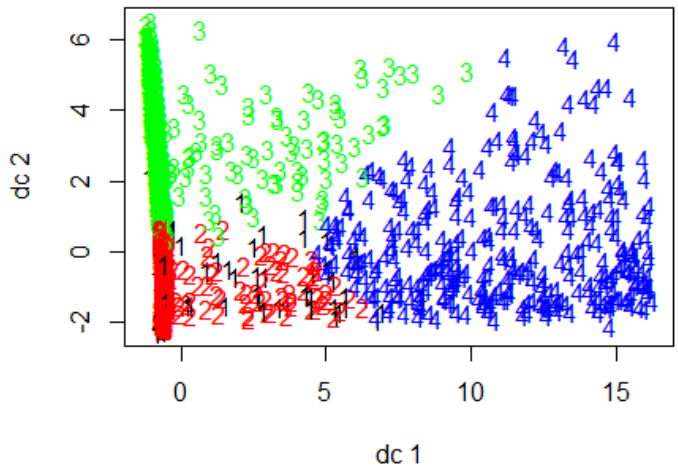
Even after performing cross validation and tuning the hyper parameters, overfitting is removed, Random forest still gives the higher accuracy comparatively. For the subject dataset, we intend to reduce the number of misclassifications, as False Positives and False Negatives which usually is related to the business costs incurred to retain customers.

Misclassification is the least in Random Forest model.

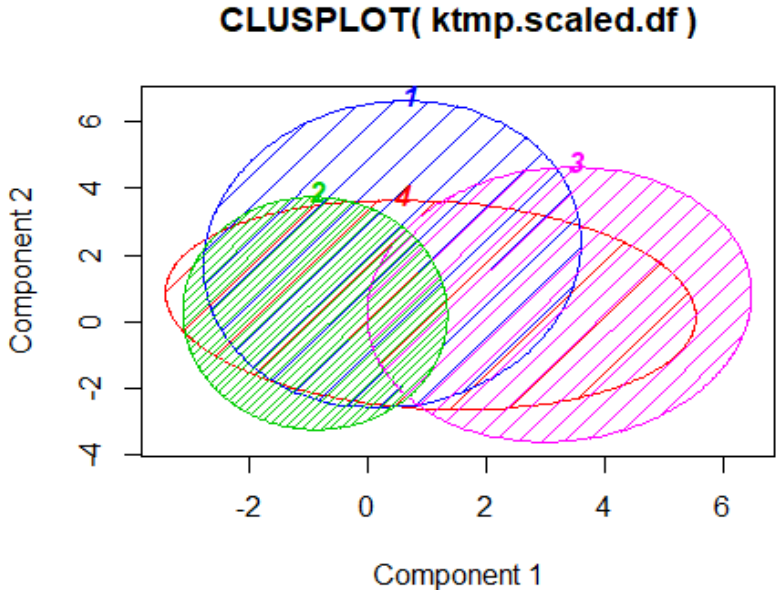
K Means Clustering

Customer segmentation is done on the Churn dataset to understand the grouping of customers based on the attributes. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.





Based on the wssplot, using the elbow method, the optimum number of K value is found to be "4". This is also verified with "Nbclust" package in R.



Based on the results, the broad grouping of customers is classified as –

Cluster 1 **Value Customers**	Low/No Referrals, Low Tenure ~ 1-year, Low Voice/Data usage, Low Total Charges, Low Revenue, Age group ~ 45, Low dependents **High Customer Base – 50%**
Cluster 2 **Pragmatists**	Good Referrals, Highest Tenure (loyal)~ 5 years, High Voice/Data usage, High Monthly/Total charges ~ 60, High Revenue, Low Age group, Extra Data/Long Distance Charges, Medium No. of Dependents **Second High Customer Base – 28%**
Cluster 3 **Promoters**	Highest Referrals, Medium Tenure ~ 3 years, Low Monthly charges, Medium Long-Distance Charges, Low Data Usage, Medium Revenue, High dependents, Medium % of customers group, Age group ~ 45 **Customer Base – 17%**
Cluster 4 **Sound Seniors**	Good Referrals, Tenure ~ 3 years, Second High Monthly/Total Charges, Highest Refunds, Second High Long-Distance Charges, Second High Revenue, Old age group – 53, Medium No. of Dependents **Customer Base – 5%**

Dataset: Network Performance

Feature Engineering

Missing Value Imputation
There are 15966 records with missing values for Latitude, longitude and State Name. These are records of customers who wished, not to share their location. With the other variables like operator, network type, etc. it is impossible to impute the location and state names. Hence those records are dropped for the model building.

Outliers
Data has been collected across geographical regions and the histogram on distribution of rating, latitude and longitude doesn't show any outliers.

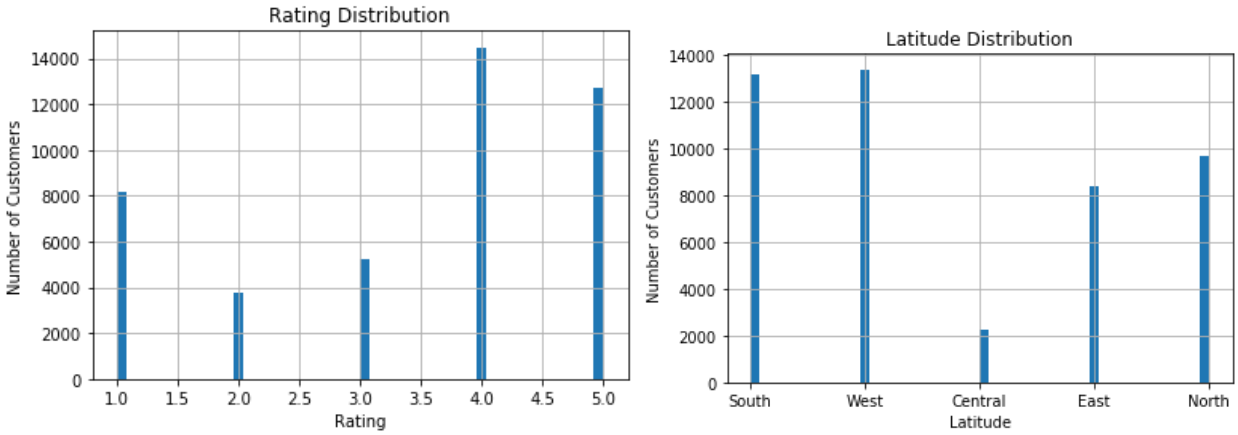


Figure 43: Outliers

Model Building

Clustering is performed on the Network performance data to segment customers based on network type, call performance, Operator and access location.

There are two types of clustering used for segmentation of data:

1. K-means Clustering
2. K-Nearest Neighbors Algorithm

K-means Clustering

The objective of K-means clustering is to group customers based on the other information available.

- For clustering purpose, all variable is converted into numeric.
- Call Drop Category is in ordinal variable and hence it a numeric with Satisfactory of 1, Call Dropped of 2 and Poor Network of 3.
- State Name field has 32 values and for clustering purpose, state values grouped into Central, South, West and East region.
- Latitude and Longitude column is dropped as it is no longer to needed as the state values are grouped into Regions.
- Other variables like Operator, Network Type, Indoor_Outdoor_Travelling variable are converted into numeric variable by creation of dummy variables of n-1 orders.
- Call Drop Category variable is highly correlated with Rating variable and so we would be removing Rating variable for the model building.

The Final Dataset has 14 columns

```
[1] "Operator.Airtel"                  "Operator.Airvoice"
 [3] "Operator.Idea"                   "Operator.RJio"
 [5] "Indoor_Outdoor_Travelling.Indoor"  "Indoor_Outdoor_Travelling.Outdoor"
 [7] "Network.Type.2G"                 "Network.Type.3G"
 [9] "Network.Type.4G"                 "Call.Drop.Category"
[11] "State.Region.Central"            "State.Region.East"
[13] "State.Region.North"              "State.Region.South"
```

Determination of cluster by WSS plot

WSS plot will help us to determine number of clusters to be selected.

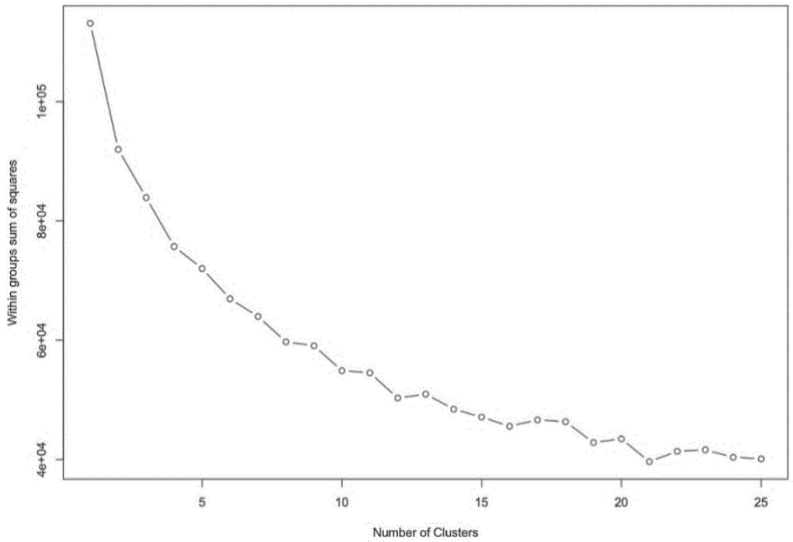


Figure 44: WSS plot – Determination of Clusters

At 5thpoint, there is an elbow joint from where the value stops to decrease significantly and hence, we will proceed with 5 clusters.

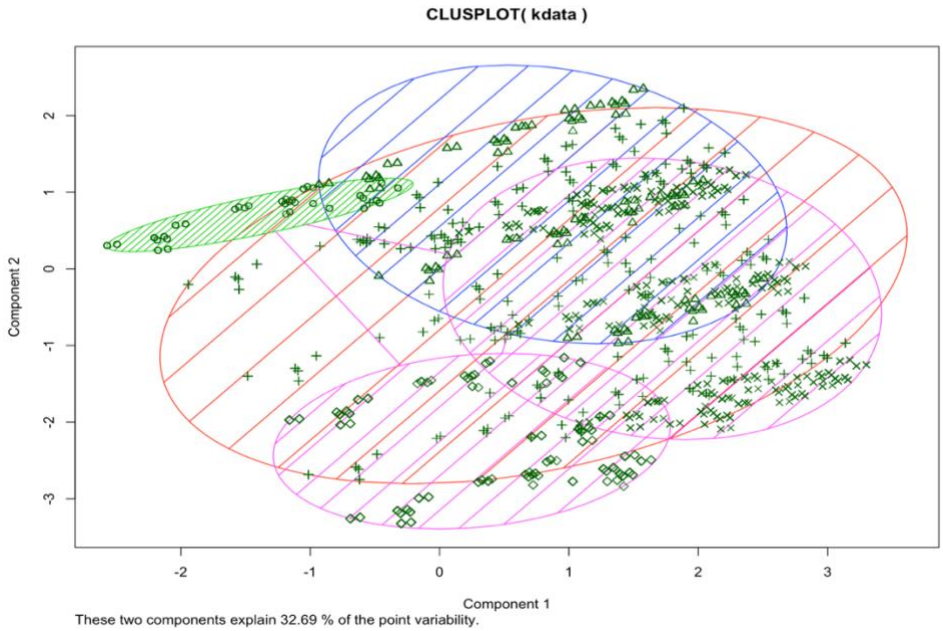


Figure 45: Kmeans Cluster Plot

Above image is showing the picturization of 5 different clusters with each cluster differentiated in different colors.The result from the kmeans cluster is analyzed with the source data to understand on customer segmentation.

	Cluster				
	1	2	3	4	5
Central	549	497	150	546	502
East	3,220	1,139	497	1,884	1,662
North	2,761	2,453	705	2,098	1,682
South	3,135	3,327	1,066	3,690	1,940
West	3,747	1,772	1,494	4,206	2,148

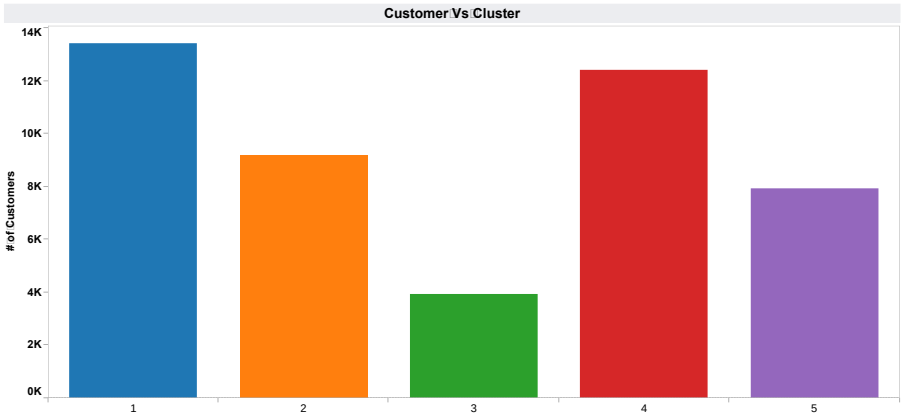


Figure 46: Cluster Analysis vs # of Customers

Interpretation

Customers fall into 5 different clusters and most of the customers are in 1st and 4th clusters.

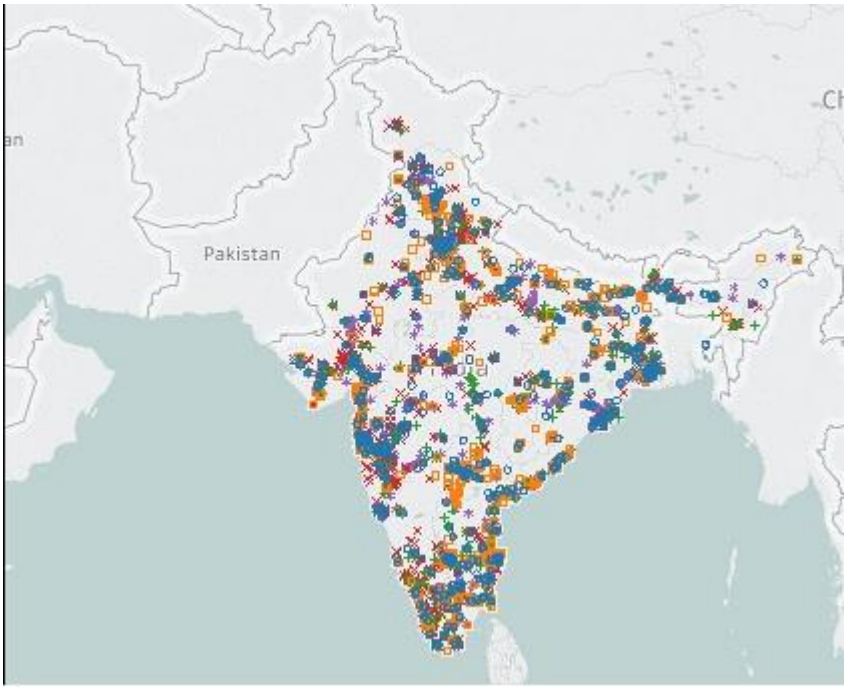


Figure 47: Geographical distribution of Call Drop Category

Customers are distributed across different regions and there is no identification of present of any cluster in only one area. Customers are classified majorly into two different categories

- Cluster 1: Indoor customers
- Cluster 2: Airtel customers
- Cluster 3: Customers facing network issue
- Cluster 4: Air voice, Vodafone and Idea customers
- Cluster 5: Outdoor and Travelling customers

Cluster 1 - Indoor Customers of 4G & Unknown network type

		Airvoice	BSNL	Idea	RJio
4G	Indoor, Poor Network			13	1,555
	Indoor, Satisfactory	765	108	601	10,100
Unknown	Indoor, Poor Network				44
	Indoor, Satisfactory				226

Figure 48: Cluster 1

Cluster 2 - Airtel Customers with 2G/4G

	Airtel			
	2G	3G	4G	Unknown
Indoor, Poor Network	139	546	627	299
Indoor, Satisfactory	173	1,783	1,706	1,486
Outdoor, Poor Network	39	139		104
Outdoor, Satisfactory	62	467		666
Travelling, Poor Network	7	71	53	32
Travelling, Satisfactory	23	270	296	200

Figure 49: Cluster 2

Cluster 3 - Customers facing network problems

		Airtel	Airvoice	BSNL	Idea	RJio
2G	Indoor, Call Dropped	18	17	44	21	
	Outdoor, Call Dropped	9	11	8	4	
	Travelling, Call Dropped	4	4	7	1	
3G	Indoor, Call Dropped	238	191	85	59	
	Outdoor, Call Dropped	63	119	9	10	
	Travelling, Call Dropped	39	33	12	5	
4G	Indoor, Call Dropped	247	113	9	29	1,014
	Indoor, Poor Network		176	12	74	
	Outdoor, Call Dropped	63	48	3	10	279
	Travelling, Call Dropped	36	21	2	3	112
	Travelling, Poor Network		17	1		
Unknown	Indoor, Call Dropped	176	77	68	103	23
	Outdoor, Call Dropped	58	19	28	14	8
	Travelling, Call Dropped	27	8	9	8	4
	Travelling, Poor Network					2

Figure 50: Cluster 3

Cluster 4 - Airvoice, BSNL & Idea customers

		Airvoice	BSNL	Idea
2G	Indoor, Poor Network	45	208	64
	Indoor, Satisfactory	193	724	106
	Outdoor, Poor Network	22	22	16
	Outdoor, Satisfactory	58	111	46
	Travelling, Poor Network	2	7	6
	Travelling, Satisfactory	52	88	20
3G	Indoor, Poor Network	388	219	153
	Indoor, Satisfactory	958	1,268	849
	Outdoor, Poor Network	138	44	38
	Outdoor, Satisfactory	558	221	245
	Travelling, Poor Network	42	12	30
	Travelling, Satisfactory	250	163	69
4G	Indoor, Satisfactory		58	
	Travelling, Satisfactory		7	
Unknown	Indoor, Poor Network	217	226	93
	Indoor, Satisfactory	888	1,268	580
	Outdoor, Poor Network	60	77	1
	Outdoor, Satisfactory	372	230	369
	Travelling, Poor Network	15	10	25
	Travelling, Satisfactory	197	190	106

Figure 51: Cluster 4

Cluster 5 - Outdoor and Travelling customers with 4G & Unknown Network

		Airtel	Airvoice	BSNL	Idea	RJio
4G	Outdoor, Satisfactory	691	312	95	218	3,202
	Travelling, Satisfactory		168	67	82	1,436
	Outdoor, Poor Network	146	69	2	38	999
	Travelling, Poor Network			1	23	209
Unknown	Outdoor, Satisfactory				20	58
	Travelling, Satisfactory					21
	Outdoor, Poor Network				22	46
	Travelling, Poor Network					9

Figure 52: Cluster 5

Accuracy of K-means algorithm is **83.6**%.

K-Nearest Neighbor Algorithm

The dataset is split into training and testing for evaluation purpose randomly in the ratio of 70:30.
KNN algorithm tries to predict the customers who are all falling to similar category.
KNN algorithm were able to group customers with highest accuracy of nearly 99.98%.
Confusion matrix of KNN algorithm is as follow

```
[[10743     0     0]
 [    0  1055     0]
 [    0     0  2263]]
```

Figure 53:Confusion Matrix

	precision	recall	f1-score	support
1	1.00	1.00	1.00	10743
2	1.00	1.00	1.00	1055
3	1.00	1.00	1.00	2263
accuracy			1.00	14061
macro avg	1.00	1.00	1.00	14061
weighted avg	1.00	1.00	1.00	14061

Figure 54: KNN Score

The accuracy of the model is 99.98%

Cross validation

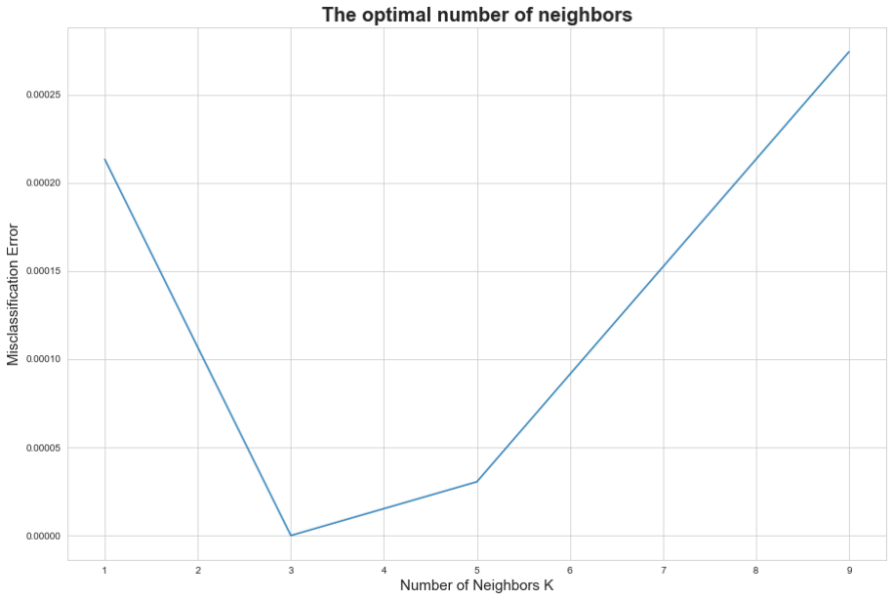


Figure 55: Cross Validation

The optimal value of KNN neighbor is 3. So, when the model ran with K=3 neighbors, the accuracy has just increased by 0.01%.

Dataset: Network Auditing

Classification of No attack and attack connection helps us to audit our network performance and project our future connection from attacks.

Customer give more importance to quality of network and security of the network. Performing an internal audit in network can explain the connection attacks and identify the key variables for attack less connections.

Feature Engineering

In order to obtain an accurate model, it's necessary to refine the data that adds more contribution to the model. Below performing three ways in selecting the best variables for models.

1. Chi Square Test
2. Extra Trees Classifier model
3. Variation Inflation Factor

Data Preparation

Taking up "target class" as y and dropping "class", "class_types" and "target class" from x.

Method 1: Chi Square Test

Chi-square test is used for categorical features in a dataset. We calculate Chi-square between each feature and the target and select the desired number of features with best Chi-square scores. It determines if the association between two categorical variables of the sample would reflect their real association in the population.

Picking up the best feature using Chi square test and giving feature scores for each variable.
Feature scores denotes the importance of each feature of the data towards the output.

Below are the feature scores:

	Specs	Scores			
0	duration	7082810.00	**21**	count	6525742.00
1	protocol_type_icmp	4728.71	**22**	srv_count	14.23
2	protocol_type_tcp	65.72	**23**	serror_rate	37365.42
3	src_bytes	3340325000.00	**24**	srv_serror_rate	37451.90
4	dst_bytes	1746096000.00	**25**	rerror_rate	6923.53
5	land	6.51	**26**	srv_rerror_rate	6997.53
6	wrong_fragment	3282.73	**27**	same_srv_rate	20826.50
7	urgent	1.82	**28**	diff_srv_rate	2694.28
8	hot	487.62	**29**	srv_diff_host_rate	1245.34
9	num_failed_logins	2.97	**30**	dst_host_count	957432.80
10	logged_in	36259.15	**31**	dst_host_srv_count	6968705.00
11	num_compromised	26894.33	**32**	dst_host_same_srv_rate	23447.78
12	root_shell	51.77	**33**	dst_host_diff_srv_rate	3197.89
13	su_attempted	117.30	**34**	dst_host_same_src_port_rate	692.74
14	num_root	32550.48	**35**	dst_host_srv_diff_host_rate	190.57
15	num_file_creations	1053.56	**36**	dst_host_serror_rate	37226.05
16	num_shells	13.47	**37**	dst_host_srv_serror_rate	38544.41
17	num_access_files	409.03	**38**	dst_host_rerror_rate	6354.83
18	num_outbound_cmds	NaN	**39**	dst_host_srv_rerror_rate	6867.10
19	is_host_login	0.87	**40**	diff_level	4889.75
20	is_guest_login	192.53			

Figure 56: Feature Scores

Picking up the top scores

	Specs	Scores
3	src_bytes	3.340325e+09
4	dst_bytes	1.746096e+09
0	duration	7.082810e+06
31	dst_host_srv_count	6.968705e+06
21	count	6.525742e+06
30	dst_host_count	9.574328e+05
37	dst_host_srv_serror_rate	3.854441e+04
24	srv_serror_rate	3.745190e+04
23	serror_rate	3.736542e+04
36	dst_host_serror_rate	3.722605e+04
10	logged_in	3.625915e+04
14	num_root	3.255048e+04
11	num_compromised	2.689433e+04
32	dst_host_same_srv_rate	2.344778e+04
27	same_srv_rate	2.082650e+04
26	srv_rerror_rate	6.997528e+03
25	rerror_rate	6.923532e+03
39	dst_host_srv_rerror_rate	6.867099e+03
38	dst_host_rerror_rate	6.354825e+03
40	diff_level	4.889752e+03
1	protocol_type_icmp	4.728708e+03
6	wrong_fragment	3.282727e+03
33	dst_host_diff_srv_rate	3.197891e+03
28	diff_srv_rate	2.694281e+03
29	srv_diff_host_rate	1.245339e+03

Figure 57: Top Features

Method 2: **Extra Trees Classifier model**

- Bagged decision trees like Random Forest and Extra Trees can be used to estimate the importance of features.
- Extra Trees Classifier is a randomized decision tree classifier which samples a random subset of the feature-space when deciding to make the next split.
- Helps in aggregating multiple de-correlated decision trees collected in “forest”.

Observation

```
ExtraTreesClassifier(bootstrap=False, class_weight=None, criterion='gini',
                     max_depth=None, max_features='auto', max_leaf_nodes=None,
                     min_impurity_decrease=0.0, min_impurity_split=None,
                     min_samples_leaf=1, min_samples_split=2,
                     min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=None,
                     oob_score=False, random_state=None, verbose=0,
                     warm_start=False)
```

Figure 58: Output

Feature importance

```
[1.69815472e-03 5.81425156e-02 8.64038813e-03 1.05692973e-02
 3.84282292e-03 5.18961072e-05 7.39326735e-03 3.25062547e-06
 5.81420774e-03 4.00016581e-04 7.17304783e-02 6.35054322e-03
 1.44663257e-04 2.43720106e-05 2.12840852e-04 9.39219439e-05
 1.49942601e-05 6.67633231e-05 0.00000000e+00 0.00000000e+00
 6.89889723e-04 1.37380132e-02 8.48391186e-03 4.90152216e-02
 6.42693868e-02 2.38811931e-02 1.73196833e-02 1.42639070e-01
 2.04633675e-03 6.01315548e-03 8.80579539e-03 1.23223338e-01
 7.36328497e-02 5.84304099e-03 3.15845488e-02 1.66327897e-02
 1.60084432e-02 1.01280015e-01 4.76687905e-02 1.00978206e-02
 6.19323120e-02]
```

Figure 59: Feature Importance

Plotting the feature importance

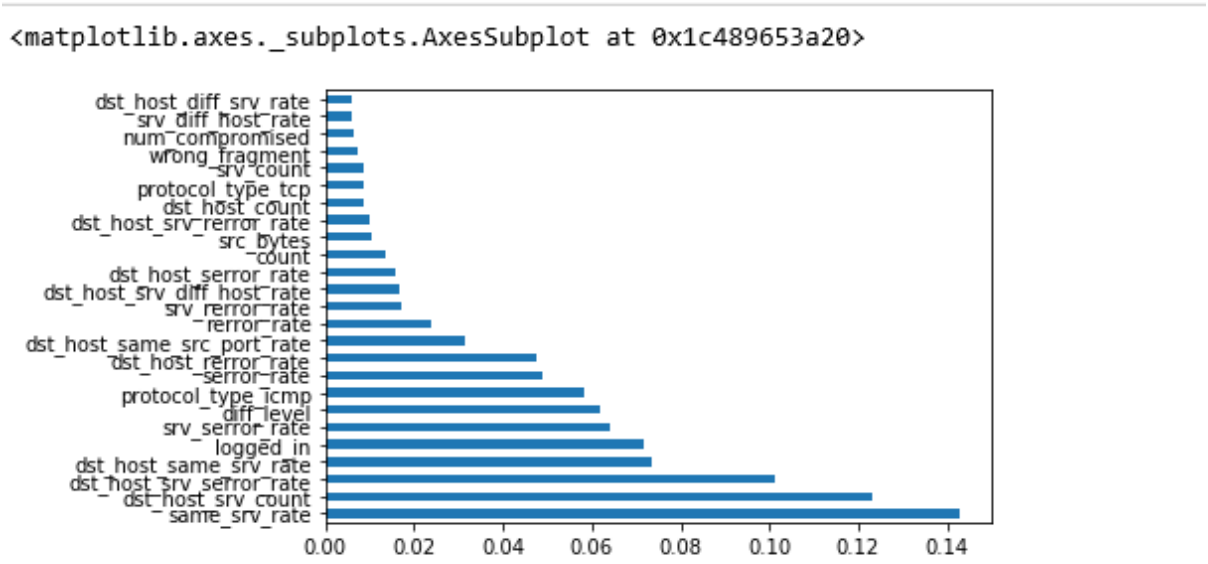


Figure 60: Feature Importance Distribution

- Same_srv_rate (% of connection of same service) is the most important feature in splitting the connection with attacks and no attacks followed by dst_hst_srv_Count.
- Including the above variables in model building can helps us classify better with good accuracy.

Method 3: Variation Inflation Factor

For model to perform better there must not be collinearity within the independent variables. However, multicollinearity check has to be done and removing the highly correlated variables from the model can be robust and classify better. Variation Inflation factor-VIF quantifies the severity of multicollinearity.

Heat Map of correlation

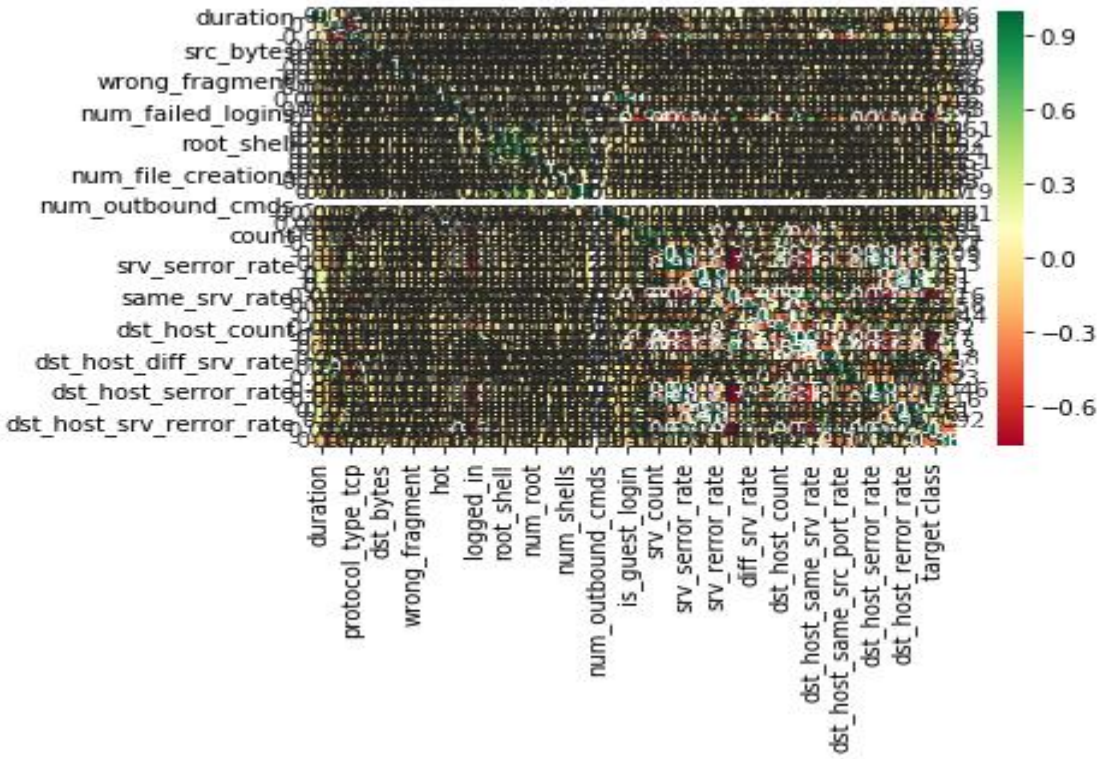


Figure 61: Correlation

Calculate the VIF factor and removing variables with VIF more than 5 in order to control multicollinearity.

VIF Factor calculation

[1.3194366422995418, 2.724010721441517, 47.126179452869515, 1.0073227798276154, 1.0015355007532891, 1.0191269269613754, 1.1537564245722014, 1.0284909000718814, 3.948619464707921, 1.0328957337050568, 18.47561560795527, 894.8444047432919, 1.6407497325603333, 2.6276863320350303, 921.8664655695586, 1.0461234764916223, 1.0251902573477987, 1.8927277425277274, nan, 1.0005740927173696, 3.999958539701492, 7.4125248074624235, 4.654634014210323, 147.6647103593568, 174.3220777420197, 71.78786147159116, 76.14713379626919, 31.364580104165405, 2.2208011729017656, 1.593057794929994, 8.590064152649477, 17.238370429995097, 27.230908829362406, 3.3347659863079016, 2.810761218828737, 2.0319867643512044, 61.120540060524355, 87.7198541847881, 10.974500351772244, 23.243892423585933, 39.72053019569353]

Variables with VIF>5 are removed and remaining variables are selected for modelling.

Variables Selected

1. duration
2. protocol_type_udp
3. src_bytes
4. dst_bytes
5. land
6. wrong_fragment
7. urgent
8. hot
9. num_failed_logins
10. logged_in
11. num_compromised
12. root_shell
13. su_attempted
14. num_file_creations
15. num_shells
16. num_access_files
17. is_host_login
18. is_guest_login
19. srv_count
20. serror_rate
21. diff_srv_rate
22. srv_diff_host_rate
23. dst_host_diff_srv_rate
24. dst_host_same_src_port_rate
25. dst_host_srv_diff_host_rate
26. dst_host_srv_rerror_rate

Correlation matrix: Below Correlation matrix of the above selected 26 variables, shows that the multicollinearity is reduced.

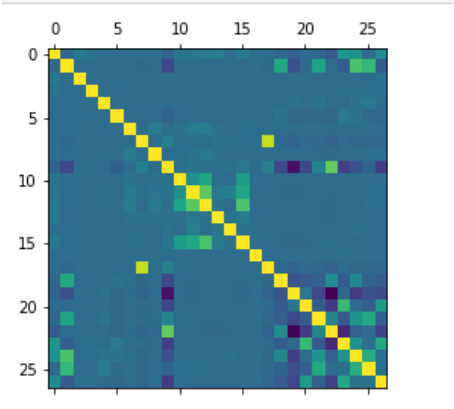
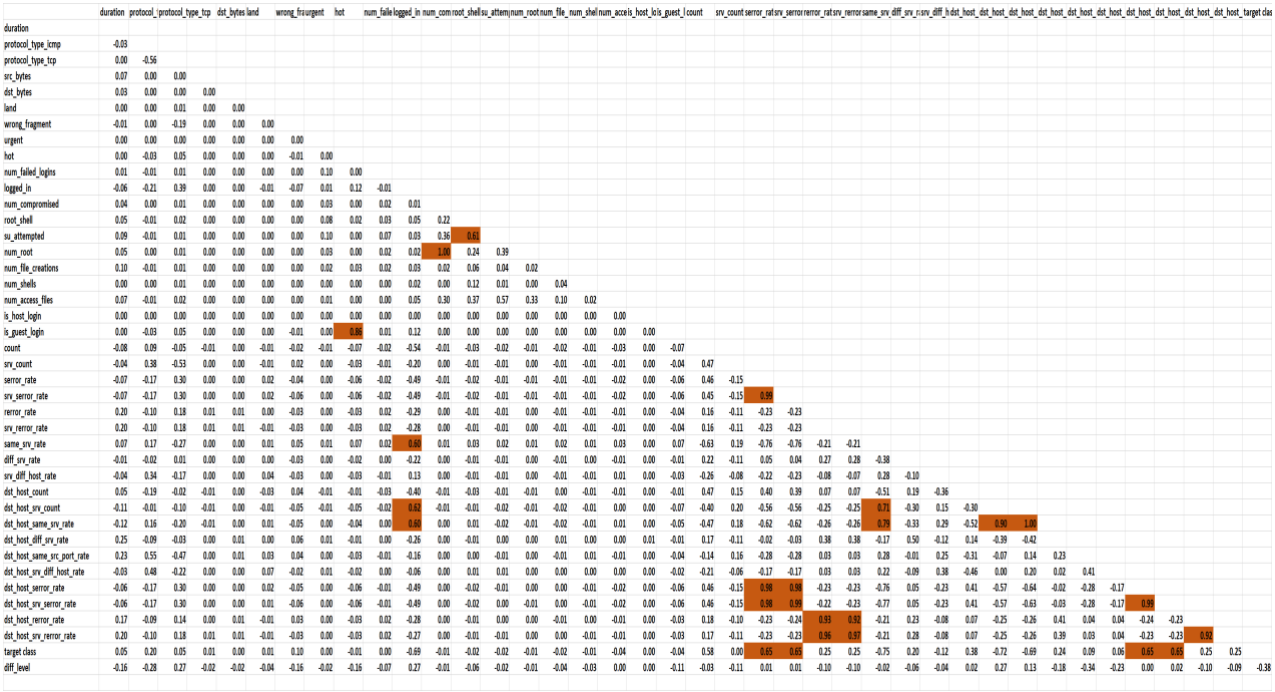


Figure 62: Correlation Matrix

Feature Selection



	duration	protocol_type_icmp	protocol_type_tcp	dst_bytes	land	wrong_fragment	urgent	hot	num_failed_logins	logged_in	num_compromised	root_shell	su_attempted	num_root	num_file_creations	num_shells	num_access_files	is_host_login	is_guest_login	count	srv_count	serror_rate	srv_serror_rate	rerror_rate	srv_rerror_rate	same_srv_rate	diff_srv_rate	srv_diff_host_rate	dst_host_count	dst_host_srv_count	dst_host_same_srv_rate	dst_host_diff_srv_rate	dst_host_same_src_port_rate	dst_host_srv_diff_host_rate	dst_host_serror_rate	dst_host_srv_serror_rate	dst_host_rerror_rate	dst_host_srv_rerror_rate	target class	diff_level
duration																																								
protocol_type_icmp	-0.03																																							
protocol_type_tcp	0.00	-0.56																																						
src_bytes	0.07	0.00	0.00																																					
dst_bytes	0.03	0.00	0.00	0.00																																				
land	0.00	0.00	0.01	0.00	0.00																																			
wrong_fragment	-0.01	0.00	-0.19	0.00	0.00	0.00																																		
urgent	0.00	0.00	0.00	0.00	0.00	0.00	0.00																																	
hot	0.00	-0.03	0.05	0.00	0.00	0.00	-0.01	0.00																																
num_failed_logins	0.01	-0.01	0.01	0.00	0.00	0.00	0.00	0.10	0.00																															
logged_in	-0.06	-0.21	0.39	0.00	0.00	-0.01	-0.07	0.01	0.12	-0.01																														
num_compromised	0.04	0.00	0.01	0.00	0.00	0.00	0.00	0.03	0.00	0.02	0.01																													
root_shell	0.05	-0.01	0.02	0.00	0.00	0.00	0.00	0.08	0.02	0.03	0.05	0.22																												
su_attempted	0.09	-0.01	0.01	0.00	0.00	0.00	0.00	0.10	0.00	0.07	0.03	0.36	0.65																											
num_root	0.05	0.00	0.01	0.00	0.00	0.00	0.00	0.03	0.00	0.02	0.02	1.00	0.24	0.39																										
num_file_creations	0.10	-0.01	0.01	0.00	0.00	0.00	0.00	0.02	0.03	0.02	0.03	0.02	0.06	0.04	0.02																									
num_shells	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.12	0.01	0.00	0.04																								
num_access_files	0.07	-0.01	0.02	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.05	0.30	0.37	0.57	0.33	0.10	0.02																							
is_host_login	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00																						
is_guest_login	0.00	-0.03	0.05	0.00	0.00	0.00	-0.01	0.00	0.66	0.01	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00																					
count	-0.08	0.09	-0.05	-0.01	0.00	-0.01	-0.02	-0.01	-0.07	-0.02	-0.54	-0.01	-0.03	-0.02	-0.01	-0.02	-0.01	-0.03	0.00	-0.07																				
srv_count	-0.04	0.38	-0.53	0.00	0.00	-0.01	0.02	0.00	-0.03	-0.01	-0.20	0.00	-0.01	-0.01	0.00	-0.01	-0.01	-0.01	0.00	-0.04	0.47																			
serror_rate	-0.07	-0.17	0.30	0.00	0.00	0.02	-0.04	0.00	-0.06	-0.02	-0.49	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.02	0.00	-0.06	0.46	-0.15																		
srv_serror_rate	-0.07	-0.17	0.30	0.00	0.00	0.02	-0.06	0.00	-0.06	-0.02	-0.49	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.02	0.00	-0.06	0.45	-0.15	0.99																	
rerror_rate	0.20	-0.10	0.18	0.01	0.01	0.00	-0.03	0.00	-0.03	0.02	-0.29	0.00	-0.01	-0.01	0.00	-0.01	-0.01	-0.01	0.00	-0.04	0.16	-0.11	-0.23	-0.23																
srv_rerror_rate	0.20	-0.10	0.18	0.01	0.01	-0.01	-0.03	0.00	-0.03	0.02	-0.28	0.00	-0.01	-0.01	0.00	-0.01	-0.01	-0.01	0.00	-0.04	0.16	-0.11	-0.23	-0.23																
same_srv_rate	0.07	0.17	-0.27	0.00	0.00	0.01	0.05	0.01	0.07	0.02	0.60	0.01	0.03	0.02	0.01	0.02	0.01	0.03	0.00	0.07	-0.63	0.19	-0.76	-0.76	-0.21	-0.21														
diff_srv_rate	-0.01	-0.02	0.01	0.00	0.00	0.00	-0.03	0.00	-0.02	0.00	-0.22	0.00	-0.01	-0.01	0.00	-0.01	0.00	-0.01	0.00	-0.01	0.22	-0.11	0.05	0.04	0.27	0.28	-0.38													
srv_diff_host_rate	-0.04	0.34	-0.17	0.00	0.00	0.04	-0.03	0.00	-0.03	-0.01	0.13	0.00	-0.01	-0.01	0.00	-0.01	-0.01	0.01	0.00	-0.03	-0.26	-0.08	-0.22	-0.23	-0.08	-0.07	0.28	-0.10												
dst_host_count	0.05	-0.19	-0.02	-0.01	0.00	-0.03	0.04	-0.01	-0.01	-0.03	-0.40	-0.01	-0.03	-0.01	-0.01	0.00	-0.01	-0.01	0.00	-0.01	0.47	0.15	0.40	0.39	0.07	0.07	-0.51	0.19	-0.36											
dst_host_srv_count	-0.11	-0.01	-0.10	-0.01	0.00	-0.01	-0.05	-0.01	-0.05	-0.02	0.62	-0.01	-0.01	-0.02	-0.01	-0.02	-0.01	0.00	0.00	-0.07	-0.40	0.20	-0.56	-0.56	-0.25	-0.25	0.71	-0.30	0.15	-0.30										
dst_host_same_srv_rate	-0.12	0.16	-0.20	-0.01	0.00	0.01	-0.05	0.00	-0.04	0.00	0.62	0.00	0.01	-0.02	-0.01	-0.02	-0.01	0.01	0.00	-0.05	-0.47	0.18	-0.62	-0.62	-0.26	-0.26	0.79	-0.33	0.29	-0.52	0.90	1.00								
dst_host_diff_srv_rate	0.25	-0.09	-0.03	0.00	0.01	0.00	0.06	0.01	-0.01	0.00	-0.26	0.00	-0.01	0.00	0.00	0.01	0.00	0.00	0.01	-0.01	0.17	-0.11	-0.02	-0.03	0.38	0.38	-0.17	0.50	-0.12	0.14	-0.39	-0.42								
dst_host_same_src_port_rate	0.23	0.55	-0.47	0.00	0.01	0.03	0.04	0.00	-0.03	-0.01	-0.16	0.00	0.00	-0.01	0.00	-0.01	0.01	-0.01	0.00	-0.04	-0.14	0.16	-0.28	-0.28	0.03	0.03	0.28	-0.01	0.25	-0.31	-0.07	0.14	0.23							
dst_host_srv_diff_host_rate	-0.03	0.48	-0.22	0.00	0.00	0.07	-0.02	0.01	-0.02	0.00	-0.06	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	-0.02	-0.21	-0.06	-0.17	-0.17	0.03	0.03	0.22	-0.09	0.38	-0.46	0.00	0.20	0.02	0.41						
dst_host_serror_rate	-0.06	-0.17	0.30	0.00	0.00	0.02	-0.05	0.00	-0.06	-0.01	-0.49	0.00	-0.02	-0.01	0.00	0.00	-0.01	-0.02	0.00	-0.06	0.46	-0.15	0.98	0.98	-0.23	-0.23	-0.76	0.05	-0.23	0.41	-0.57	-0.64	-0.02	-0.28	-0.17					
dst_host_srv_serror_rate	-0.06	-0.17	0.30	0.00	0.00	0.01	-0.06	0.00	-0.06	-0.01	-0.49	0.00	-0.02	0.00	-0.01	0.00	-0.01	-0.02	0.00	-0.06	0.46	-0.15	0.98	0.99	-0.22	-0.23	-0.77	0.05	-0.23	0.41	-0.57	-0.63	-0.03	-0.28	-0.17	0.99				
dst_host_rerror_rate	0.17	-0.09	0.14	0.00	0.01	-0.01	0.03	0.00	-0.03	0.02	-0.28	0.00	-0.01	0.00	0.00	-0.01	0.00	-0.01	0.00	-0.03	0.18	-0.10	-0.23	-0.24	0.93	0.92	-0.21	0.23	-0.08	0.07	-0.25	-0.26	0.41	0.04	0.04	-0.24	-0.23			
dst_host_srv_rerror_rate	0.20	-0.10	0.18	0.01	0.01	-0.01	-0.03	0.00	-0.03	0.02	-0.27	0.00	-0.01	-0.01	0.00	0.00	-0.01	-0.01	0.00	-0.03	0.17	-0.11	-0.23	-0.23	0.96	0.97	-0.21	0.28	-0.08	0.07	-0.25	-0.26	0.39	0.03	0.04	-0.23	-0.23	0.92		
target class	0.05	0.20	0.05	0.01	0.00	0.01	0.10	0.00	-0.01	0.00	-0.69	-0.01	-0.02	-0.02	-0.01	-0.02	-0.01	-0.04	0.00	-0.04	0.58	0.00	0.65	0.65	0.25	0.25	-0.75	0.20	-0.12	0.38	-0.72	-0.69	0.24	0.09	0.06	0.65	0.65	0.25	0.25	
diff_level	-0.16	-0.28	0.27	-0.02	-0.02	-0.04	-0.16	-0.02	-0.16	-0.07	0.27	-0.01	-0.06	-0.02	-0.01	-0.04	-0.03	0.00	0.00	-0.11	-0.03	-0.11	0.01	0.01	-0.10	-0.10	-0.02	-0.06	-0.04	0.02	0.27	0.13	-0.18	-0.34	-0.23	0.00	0.02	-0.10	-0.09	-0.38

Interpretation

- The variable num_compromised and num_root are highly correlated. Hence consider any one variable.
- Logged_in is highly correlated with same_srv_rate , dst_host_srv_count, dst_host_same_srv_rate
- Is guest_login is correlated with is_hot connections.
- serror_rate is highly correlated to srv_serror_rate, dst_host_serror_rate, dst_host_srv_serror_rate,
- same_srv_rate is highly correated to dst_host_srv_count and dst_host_same_srv_rate.
- dst_host_srv_serror_rate is highly corelated to dst_host_serror_rate.
- dst_host_rerror_rate is highly correlated to dst_host_srv_rerror_rate.

Below are the important variables in the dataset

1. num_compromised
2. Logged_in
3. same_srv_rate
4. is_host_login
5. is_guest_login
6. serror_rate
7. diff_srv_rate
8. dst_host_srv_serror_rate
9. dst_host_rerror_rate

Model Building

Below are the types of models built for assessing the data and to the finest accuracy.

1. Logistic Regression in R
2. Random Forest
3. GBM
4. SVM

Test Train Split:

- Partition the data set by allocating 70% -for training data and 30% -for validating the results.

Logistic Regression

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	18150	2233
1	2053	15356

Table 51: Logistic Regression - Confusion Matrix

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.886589754	0.882072491	0.873045654	0.877535859

Table 52: Scores - Logistic Regression

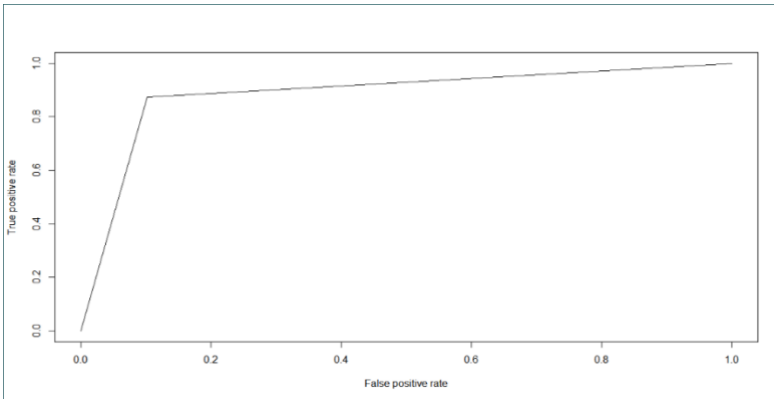


Figure 63: ROC Curve

AUC	0.845

Individual Coefficient's of the logit model:

```
                              Estimate Std. Error    z value Pr(>|z|)
(Intercept)                 -1.969e+15  1.070e+06 -1.841e+09   <2e-16 ***
duration                    -5.414e+10  9.567e+01 -5.660e+08   <2e-16 ***
protocol_type_udp1           5.107e+14  1.113e+06  4.587e+08   <2e-16 ***
src_bytes                    6.931e+06  3.386e-02  2.047e+08   <2e-16 ***
dst_bytes                    2.640e+06  4.707e-02  5.610e+07   <2e-16 ***
land1                       -3.840e+15  1.636e+07 -2.347e+08   <2e-16 ***
wrong_fragment               1.605e+15  9.287e+05  1.728e+09   <2e-16 ***
urgent                      -3.572e+14  1.567e+07 -2.280e+07   <2e-16 ***
hot                          2.839e+14  2.070e+05  1.371e+09   <2e-16 ***
num_failed_logins            3.823e+14  4.849e+06  7.885e+07   <2e-16 ***
logged_in1                   2.554e+14  1.025e+06  2.490e+08   <2e-16 ***
num_compromised              2.178e+11  3.359e+04  6.483e+06   <2e-16 ***
root_shell1                  6.651e+14  8.277e+06  8.035e+07   <2e-16 ***
su_attempted1               -1.756e+15  1.647e+07 -1.066e+08   <2e-16 ***
su_attempted2               -4.628e+15  1.910e+07 -2.424e+08   <2e-16 ***
num_file_creations          -2.987e+14  5.022e+05 -5.947e+08   <2e-16 ***
num_shells                   5.376e+14  1.018e+07  5.283e+07   <2e-16 ***
num_access_files            -5.752e+13  3.120e+06 -1.844e+07   <2e-16 ***
is_host_login1              -3.971e+15  6.712e+07 -5.917e+07   <2e-16 ***
is_guest_login1             -6.150e+15  4.611e+06 -1.334e+09   <2e-16 ***
srv_count                    2.309e+12  3.723e+03  6.202e+08   <2e-16 ***
serror_rate                  4.599e+15  1.118e+06  4.113e+09   <2e-16 ***
diff_srv_rate                3.729e+14  1.519e+06  2.456e+08   <2e-16 ***
srv_diff_host_rate          -2.212e+14  9.888e+05 -2.237e+08   <2e-16 ***
dst_host_diff_srv_rate       1.252e+15  1.595e+06  7.849e+08   <2e-16 ***
dst_host_same_src_port_rate  1.651e+15  9.845e+05  1.677e+09   <2e-16 ***
dst_host_srv_diff_host_rate -1.926e+14  2.450e+06 -7.860e+07   <2e-16 ***
dst_host_srv_rerror_rate     2.826e+15  1.208e+06  2.340e+09   <2e-16 ***
```

Insights:

- The summary of the Logit model provides a precise value of the coefficients beneficial for the model.

- Based on the p value the most significant variables are selected and processed with CART and decision tree model.

Below are the variables used in the upcoming models

duration
protocol_type_udp
src_bytes
land
hot
logged_in
root_shell
num_file_creations
num_shells
is_guest_login
srv_count
serror_rate
diff_srv_rate
dst_host_diff_srv_rate
dst_host_same_src_port_rate
dst_host_srv_diff_host_rate

Refining Logistic Model with Significant Variables

- Based on the significant values obtained from logistic regression is used in CART Model and evaluated.
- Below are the confusion matrix and the stats:

Logistic Regression

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	17694	2509
1	1298	16291

Table 53: Logistic Regression - Confusion Matrix

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.899264395	0.926203877	0.866542553	0.895380472

Table 54: Scores - Logistic Regression

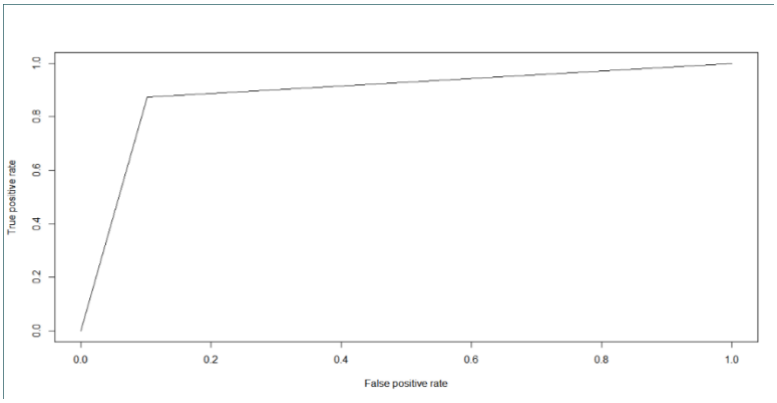


Figure 64: ROC Curve

AUC	0.895

Insights:

- Refining the model with significant variables increased the true Positive and help us identify clearly the attack connections.
- Accuracy of the model is 89.9.

Random Forest

Prediction	Reference	
	0	1
0	20153	50
1	87	17502

Table 55: Random Forest - Confusion Matrix

Model	Accuracy	Precision	Recall	F1
Random Forest	0.996374894	0.995053727	0.997151322	0.99610142

Table 56: Scores - Random Forest

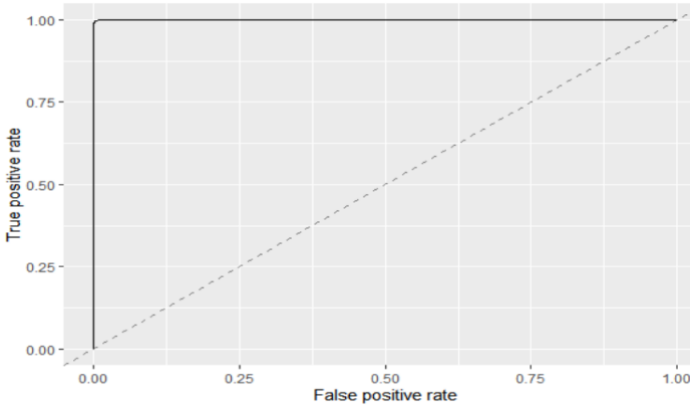


Figure 65: ROC Curve

AUC	0. .998

Insights:

- Random forest with the selected 16 variables increased the true Negatives.
- Accuracy of the model is 99.6.

GBM

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	19549	654
1	705	16884

Table 57: GBM - Confusion Matrix

Model	Accuracy	Precision	Recall	F1
GBM	0.964040008	0.959918131	0.962709545	0.961311811

Table 58: Scores - GBM

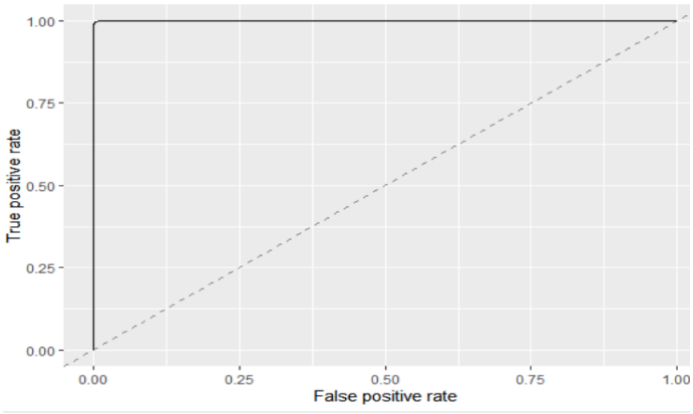


Figure 66: ROC Curve

AUC	0.966

Insights:

- Accuracy of this model is 96.4

SVM

Confusion Matrix and Statistics

Prediction	Reference	
	0	1
0	20074	129
1	311	17278

Table 59: GBM - Confusion Matrix

Model	Accuracy	Precision	Recall	F1
SVM	0.988357324	0.982318495	0.992589188	0.987427135

Table 60: Scores - GBM

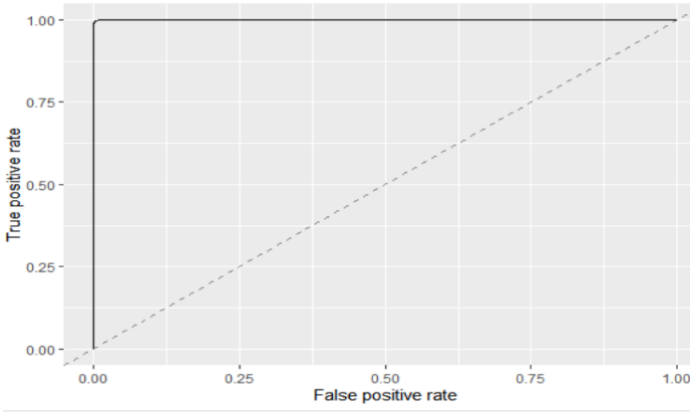


Figure 67: ROC Curve

AUC	0.988

Insights:

Model Accuracy is 98.4.

True Negative rate is high and helps classify the no attack connections better.

Comparison of models:

Table 61: Comparison of models for Network intrusion

	Accuracy	Precision	F1
Logistic Regression	89.9%	92.62%	89.53%
Random Forest	99.63%	99.5%	99.6%
SVM	98.83%	98.23%	98.74%
GBM	96.4%	95.99%	96.13%

Random Forest has performed well with 99.6% accuracy.

Chapter 5: Recommendation and Conclusions

The project outcome is an analytical solution which can be used by the Telecom Operator:

- To understand and gain insights related to its customers
- Understand the reasons behind customer attrition and strategize preventive/corrective measures accordingly.
- Better understand its network and reasons for call drops. These details can be used to build an action plan to improve network performance and enhance network coverage.
- Secure the network to avoid any intrusions causing outage or other service impacts

Dataset 1: Customer Churn

- It is recommended to revisit the service plans and offer better incentives to turn down the attrition.
- It is also strongly advised to discourage offering monthly contracts and focus on long term contracts, so the customers are more likely to see the benefits of the service and are more likely to commit to product.
- Churn in the short term could be also due to improper fit between the service plans and the usage pattern. So, it is extremely important to keep the customers informed of their behaviour pattern with spend analysis and recommend change of options in plan to fit their needs.
- It is highly necessary that the customers are kept engaged always and their complaints are attended to satisfaction at the earliest.
- The additional data and voice charges can be revisited for the customer segments based on usage. To keep customers engaged over the long-term, it is needed to reinforce the core value of the product consistently.
- Look for up-sell and cross-sell opportunities to keep them invested and interested in the product.
- It is recommended to perform customer Journey analytics to identify at risk consumers and thereby take actions to reduce churn. This would help to identify problem and pain areas in customer experience. Service provider should focus on the entire customer journey and not just before the churn.
- Based on the prediction model, drive targeted marketing campaigns and frequent touch points with the customer addressing the concerns which will help to change the propensity to churn.
- Personalised tariff plans to suit the usage and customer behaviour along with better choice of devices
- Premium Tech support can be extended as a privileged free service for long tenured customers. Similarly, streaming and other online services like backup and security can be offered as a clubbed service or a free service for some duration to better understand the product.
- Streaming application content to be reviewed and target based on the age groups
- Service provider can exploit the customer advocates and offer free services or discounts based on the referral count
- Reduce Long distance call charges and Extra data consumption charges based on the data usage rates

Dataset 2 : Network Performance

- Leverage user data traffic monitoring or KPI metrics from user device to track network quality and performance and take actions to fix. Based on the call drop % captured and analysed, the network configuration parameters can be optimized.

Dataset 3 : Network Auditing

- Duration of attack connections are 40 % higher than no attack connections. Hence it is recommended to check the length of connection.

- 90% of the attack connection via TCP protocol, so TCP connections are more exposed to attack and an alternative protocol can be recommended for such instances to avoid attacks and the network will be more secured for customers to use.
- Src_bytes - Number of data bytes from source to destination is an important feature in classifying the attack and no attack connections. Connection with high volume of src bytes are likely to be attack.
- Diff_srv_rate -% of connections to different services-Connection to different service providers are likely to get attacked. If a customer is establishing a connection to another service provider then there are high chances to get attacked and hence these connections has to e monitored and need to have invoke more security across the line during the connection.

Bibliography

- https://pdfs.semanticscholar.org/1b34/80021c4ab0f632efa99e01a9b073903c5554.pdf?_ga=2.265703823 .103867871.1500221431-557851010.1499886538
- Aleksandar Lazarevic, Levent Ertoz, Vipin Kumar, Aysel Ozgur, Jaideep Srivastava, “A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection”
- http://ijarcet.org/wp-content/uploads/IJARCET-VOL-5-ISSUE-6-1776-1777.pdf
- https://www.researchgate.net/publication/300004210_Feature_Selection_for_Intrusion_Detection_Using_Random_Forest
- https://file.scirp.org/pdf/JIS_2016040716244084.pdf

Annexure

Customer ID: A unique ID that identifies each customer.

Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Gender: The customer's gender: Male, Female

Age: The customer's current age, in years, at the time the fiscal quarter ended.

Senior Citizen: Indicates if the customer is 65 or older: Yes, No

Married: Indicates if the customer is married: Yes, No

Dependents: Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc.

Number of Dependents: Indicates the number of dependents that live with the customer.

CustomerID: A unique ID that identifies each customer.

Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Country: The country of the customer's primary residence.

State: The state of the customer's primary residence.

City: The city of the customer's primary residence.

Zip Code: The zip code of the customer's primary residence.

Lat Long: The combined latitude and longitude of the customer's primary residence.

Latitude: The latitude of the customer's primary residence.

Longitude: The longitude of the customer's primary residence.

ID: A unique ID that identifies each row.

Zip Code: The zip code of the customer's primary residence.

Population: A current population estimate for the entire Zip Code area.

Quarter: The fiscal quarter that the data has been derived from (e.g. Q3).

Referred a Friend: Indicates if the customer has ever referred a friend or family member to this company: Yes, No

Number of Referrals: Indicates the number of referrals to date that the customer has made.

Tenure in Months: Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above.

Offer: Identifies the last marketing offer that the customer accepted, if applicable. Values include None, Offer A, Offer B, Offer C, Offer D, and Offer E.

Phone Service: Indicates if the customer subscribes to home phone service with the company: Yes, No

Avg Monthly Long Distance Charges: Indicates the customer's average long distance charges, calculated to the end of the quarter specified above.

Multiple Lines: Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No

Internet Service: Indicates if the customer subscribes to Internet service with the company: No, DSL, Fiber Optic, Cable.

Avg Monthly GB Download: Indicates the customer's average download volume in gigabytes, calculated to the end of the quarter specified above.

Online Security: Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No

Online Backup: Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No

Device Protection Plan: Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No

Premium Tech Support: Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No

Streaming TV: Indicates if the customer uses their Internet service to stream television programing from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Streaming Movies: Indicates if the customer uses their Internet service to stream movies from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Streaming Music: Indicates if the customer uses their Internet service to stream music from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Unlimited Data: Indicates if the customer has paid an additional monthly fee to have unlimited data downloads/uploads: Yes, No

Contract: Indicates the customer's current contract type: Month-to-Month, One Year, Two Year.

Paperless Billing: Indicates if the customer has chosen paperless billing: Yes, No

Payment Method: Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check

Monthly Charge: Indicates the customer's current total monthly charge for all their services from the company.

Total Charges: Indicates the customer's total charges, calculated to the end of the quarter specified above.

Total Refunds: Indicates the customer's total refunds, calculated to the end of the quarter specified above.

Total Extra Data Charges: Indicates the customer's total charges for extra data downloads above those specified in their plan, by the end of the quarter specified above.

Total Long Distance Charges: Indicates the customer's total charges for long distance above those specified in their plan, by the end of the quarter specified above.

Quarter: The fiscal quarter that the data has been derived from (e.g. Q3).

Satisfaction Score: A customer's overall satisfaction rating of the company from 1 (Very Unsatisfied) to 5 (Very Satisfied).

Satisfaction Score Label: Indicates the text version of the score (1-5) as a text string.

Customer Status: Indicates the status of the customer at the end of the quarter: Churned, Stayed, or Joined

Churn Label: Yes = the customer left the company this quarter. No = the customer remained with the company. Directly related to Churn Value.

Churn Value: 1 = the customer left the company this quarter. 0 = the customer remained with the company. Directly related to Churn Label.

Churn Score: A value from 0-100 that is calculated using the predictive tool IBM SPSS Modeler. The model incorporates multiple factors known to cause churn. The higher the score, the more likely the customer will churn.

Churn Score Category: A calculation that assigns a Churn Score to one of the following categories: 0-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81-90, and 91-100

CLTV: Customer Lifetime Value. A predicted CLTV is calculated using corporate formulas and existing data. The higher the value, the more valuable the customer. High value customers should be monitored for churn.

CLTV Category: A calculation that assigns a CLTV value to one of the following categories: 2000-2500, 2501-3000, 3001-3500, 3501-4000, 4001-4500, 4501-5000, 5001-5500, 5501-6000, 6001-6500, and 6501-7000.

Churn Category: A high-level category for the customer's reason for churning: Attitude, Competitor, Dissatisfaction, Other, Price. When they leave the company, all customers are asked about their reasons for leaving. Directly related to Churn Reason.

Churn Reason: A customer's specific reason for leaving the company. Directly related to Churn Category.

duration : length (number of seconds) of the connection
protocol_type: type of the protocol, e.g. tcp, udp, etc.
Service: network service on the destination, e.g., http, telnet, etc.
src_byte: number of data bytes from source to destination
dst_**bytes**: number of data bytes from destination to source
flag: normal or error status of the connection
land: 1 if connection is from/to the same host/port; 0 otherwise
wrong_fragment: number of ``wrong'' fragments
urgent: number of urgent packets
hot: number of ``hot'' indicators
num_failed_logins: number of failed login attempts
logged_in: 1 if successfully logged in; 0 otherwise
num_compromised: number of ``compromised'' conditions
root_shell:1 if root shell is obtained; 0 otherwise
su_attempted: 1 if ``su root'' command attempted; 0 otherwise
num_root: number of ``root'' accesses
num_file_creations: number of file creation operations
num_shells: number of shell prompts
num_access_files: number of operations on access control files
num_outbound_cmds: number of outbound commands in an ftp session
is_hot_login: 1 if the login belongs to the ``hot'' list; 0 otherwise
is_guest_login: 1 if the login is a ``guest''login; 0 otherwise
count: number of connections to the same host as the current connection in the past two seconds
serror_rate: % of connections that have ``SYN'' errors
rerror_rate: % of connections that have ``REJ'' errors
same_srv_rate: % of connections to the same service
diff_srv_rate: % of connections to different services
srv_count: number of connections to the same service as the current connection in the past two seconds
srv_serror_rate: % of connections that have ``SYN'' errors
srv_rerror_rate: % of connections that have ``REJ'' errors
srv_diff_host_rate: % of connections to different hosts