Telecommunication Customer Retention

A Dissertation submitted

for the partial fulfillment of the degree of

Bachelor of Engineering in

Information Technology

(Session 2017-2018)

Guided By:

Dr. Bhawna Nigam

Submitted by:

Arpit Nema (14I8171)
Ayush Gupta(14I8074)
Pranav Singh(14I8032)

Department of Information Technology

Institute of Engineering & Technology

Devi Ahilya Vishwavidyalaya, Indore (M.P.)

(www.iet.dauniv.ac.in)

April 2018

Dissertation Approval Sheet

The dissertation **"Telecommunication Customer Retention"** submitted by **Arpit Nema, Ayush Gupta and Pranav Singh** is approved as partial fulfillment for the award of **Bachelor of Engineering in Information Technology Engineering** degree by **Devi Ahilya Vishwavidyalaya, Indore**.

Internal Examiner **External Examiner**

Director
Institute of Engineering & Technology
Devi Ahilya Vishwavidyalaya,
Indore (M.P.)

Recommendation

The dissertation entitled **"Telecommunication Customer Retention"** submitted by **Arpit Nema, Ayush Gupta and Pranav Singh** is a satisfactory account of the bonafide work done under my supervision is recommended towards the partial fulfillment for the award of **Bachelor of Engineering in Information Technology** degree by **Devi Ahilya Vishwavidyalaya, Indore.**

Date:

Dr. Bhawna Nigam
Project Guide

Endorsed By:

Head, Department of Information Technology

Candidate Declaration

We hereby declare that the work which is being presented in this project entitled **"Telecommunication Customer Retention"** in partial fulfillment of degree of Bachelor of Engineering in **Information Technology** is an authentic record of our own work carried out under the supervision and guidance of **Dr. Bhawna Nigam, Assistant Professor** in Department of **Information Technology**, Institute of Engineering and Technology, Devi Ahilya Vishwavidyalaya, Indore .

We are fully responsible for the matter embodied in this project in case of any discrepancy found in the project and the project has not been submitted for the award of any other degree.

Date:

Place:

Arpit Nema (14I8171)

Ayush Gupta(14I8074)

Pranav Singh(14I8032)

ACKNOWLEDGEMENT

Words can never express the extent of indebtness but we still wish to express our deep sense of gratitude to all the people who helped us in the completion of this project.

I want to express my heart-felt gratitude to **Dr. Bhawna Nigam** for her advises and unremitting support over the last one year. She trained me how to do research and how to write, encouraged me not to be intimidated by the difficult problems. She is so generous with her time and ideas. Her intellectual creativity, perseverance and commitment would benefit me for the rest of my life. I could never thank him enough for being an excellent mentor and a wonderful teacher.

My thanks also go to the other members of my college **Dr. Vrinda Tokekar** for the discussion during the course of the project, for the reading of the draft of this thesis and providing valuable feedback.

I want to thank the director of our college **Dr. Sanjeev Tokekar** for assisting us. Therefore, my thanks go to them for making this project possible.

I also thank my team members and friends for supporting me and helping me out with the testing phase as well as rest of the project.

Finally, I want to thank my parents for giving me strength and love.

ABSTRACT

The landscape of the telecommunication industry in India has been changed drastically since the deregulation of telecommunication sector in early 1990s. Number of service providers has been increased from one, i.e. state monopoly, to more than 70 within a short period of time. With the increased competition telecom service providers find it difficult to retain the existing customers.

A telecommunications company offers a wide range of services to millions of customers but as common in the industry, a large number of them terminate their contracts and move to competitors. A company can start a strategic initiative to increase its retention capability by gaining a deep knowledge of the reasons why customers churn and thus can be able to forecast who is going to churn on an individual basis.

The telecom company can follow 3 steps approach to solve the problem: understanding why customers churned in the past(rules of behavior, ranking of drivers), forecasting who is going to churn and making specific actions(marketing and sale) that prevent each individual decision to churn.

In the survey done in the Telecom industry, it is stated that the cost of acquiring a new customer is far more that retaining the existing one. Therefore, by collecting knowledge from the telecom industries can help in predicting the association of the customers as whether or not they will leave the company. Our paper proposes a new framework for the churn prediction model and implements it using the R studio software. The efficiency and the performance of Decision tree, Logistic regression and neural network techniques have been compared.

TABLE OF CONTENTS Page No

List of figures

List of Table

Chapter-1

Introduction

Chapter-1
Introduction

1.1 Overview and issues involved

There are a number of telecommunication networks that are available and we have the luxury to choose the one we want based on our requirements. The increased number of telecoms are a challenge to the telecom companies and many companies are facing huge revenue losses , to keep the customers many companies invest a huge revenue in the beginning and thus it becomes very important for the customers to expand the business and get back the amount that has been invested in the business .The increase in the number of churn customers is become the present day challenge to the telecom industry and such customers create financial burden to the company, identifying such customers is the objective of this research paper.

In the survey done in the Telecom industry, it is stated that the cost of acquiring a new customer is far more that retaining the existing one. Therefore, by collecting knowledge from the telecom industries can help in predicting the association of the customers as whether or not they will leave the company. The required action needs to be undertaken by the telecom industries in order to initiate the acquisition of their associated customers for making their market value stagnant.

1.2 Problem definition

A telecommunications company is concerned about the number of customers leaving their landline business for cable competitors. They need to understand who is leaving. Imagine that you're an analyst at this company and you have to find out who is leaving and why.

We have Telco customer churn data set which provides info to help us predict behavior to retain customers. We can analyze all relevant customer data and develop focused customer retention programs.

The dataset includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents.

1.3 Objectives

- Our prime objective is to generate models which predict whether customer will churn before he actually leaves the company from the given dataset.
- We also focus on generating model which help managers take decision without going deep into technicalities.
- We will train data and then test it. Further, it's performance will be evaluated by applying logistic regression model, decision tree model, etc. which will help predict customers that can churn in near future.

1.4 Proposed solution

To predict whether customer will churn before it actually leave the company, we will use machine learning classification techniques to classify customers as it will churn or not.

We have chosen 3 step approach to solve the problem:

- Understanding why customers churned in the past (rules of the behavior, ranking of drivers)
- forecasting who is going to churn
- making specific actions (marketing & sale) that prevent each individual decision to churn.

Steps:

1.4.1 Churn Management

Since acquiring new customers is challenging it is very important to retain the current customers. Churn can be reduced by analyzing the past history of the potential customers systematically. Large data is maintained about the customers and on performing a proper analysis on the same it is possible to predict the probable customers that might churn. The information that is available can be analyzed in different ways and thereby provide various ways for the operators to envisage the churning and evade the same.

1.4.1.1 Data collection

For analysis the data that is available in the telecom dataset has been used and prediction has been done for the same.

1.4.1.2 Data preparation

Before the data can be analyzed we have to clean the data and keep it ready so that the desired results can be derived from it. Data has be clean so that the redundancy and errors can be removed because having such data will lead to incorrect results as well.

In this paper a Churn Analysis has been applied on Telecom data, here the agenda is to know the possible customers that might churn from the service provider. R programing is used for the same this will help give a statistical computing for the data available. The end result would give us the probability of churn for each customer.

1.4.1.3 Prediction

The business is interested in the final product and it is very important to represent your result in a “graphical representation” such a way that it is understandable and the result helps business make the needed predictions which in turn brings profit. There are many tools that help achieve the same for example, Tableau, Power BI, qlikview etc.

1.4.1.4 Data Visualization Tools

The best way to get your message across is to use visualization tools, by representing data visually it is possible to uncover the surprising patterns and the patterns that would go unnoticed if we took the stats alone.

We have treated this churn prediction problem as a classification problem. Therefore, in this study, we examine three different classification approaches to improve the prediction model (Logistic Regression, decision tree and neural networks). Parameters like accuracy, precision and recall) were used to gauge the accuracy of the models.

Chapter-2

Literature Survey

Chapter-2

Literature Survey

2.1 Churn: Importance and Analysis

- **Churn Rate**

Churn Rate is the percentage of subscribers to a service that discontinue their subscription to that service in a given time period. In order for a company to expand, its growth rate (i.e. its number of new customers) must exceed its churn rate. Churn rate is an important consideration in the telephone and cell phone services industry. In many geographical areas, several companies are competing for customers, making it easy for people to transfer from one provider to another.

- **Importance of Churn Rate**

Churn rates are often used to indicate the strength of a company's customer service division and its overall growth prospects. Lower churn rates suggest a company is, or will be, in a better or stronger competitive state. Customer loss impacts service providers significantly as they often make a significant investment to acquire new customers.

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

- **Churn Rate Analysis**

Clearly, churn rate is a critical metric for any subscription business. But there are also a variety of opinions about how to calculate it[41].

CLTV: Understanding customer lifetime value (LTV) is one of the most complex and important analyses a business can take on. Every part of your organization affects the outcome of the calculation: acquisition costs, revenue, customer service, and returns. It's an accurate approach to customer churn prediction- at the core it has the ability to predict which customers will churn. The approach takes into consideration both micro-segmentation and their behavior pattern. By merging the most exacting micro-segmentation available anywhere with a deep understanding of how customers move from one micro-segment to another over time – including the ability to predict those moves before they occur – an unprecedented degree of accuracy in customer churn prediction is attainable. Figuring out which one will stay for long and will reap how much revenue, helps the service provider to judge whether spending on a customer is worth the effort or not.

CVM: Customer value management (CVM) is a holistic way of evaluating individual subscribers in terms of their overall profitability- now and in the future. CVM has the potential to boost earnings. This measure covers subscribers at every stage of their relationship with the operator. Relying on a combination of tactics, including customer payback period, budget rebalancing, tailored customer rewards, and cross- and up-selling campaigns. CVM technique help companies analyze which customers are the most valuable, and why. Indeed, this approach is a key capability in a world where the potential customer base simply isn't getting any bigger.

Predictive Churn Modeling: Predictive technology is a body of tools capable of discovering and analyzing patterns in data so that past behavior can be used to forecast likely future behavior. Predictive technology is increasingly used for forecasting in most of the Telecom companies' balance sheet. The raw data can be processed to get predictions about consumer behavior for future campaigns.

Postpaid and blended churn rates: This churn rate is based upon the losses of both pre-paid and contract customer. Post-paid subscribers are a telecom company's one of the biggest revenue segments since they have a significant lifetime value for telecom companies. Their discontinuation of services accounts for a major loss in company's revenue.

ARPU: Average Revenue per User or ARPU or average revenue per unit is an expression of the income generated by a typical subscriber or device per unit time in a telecommunications network. ARPU provides an indication of the effectiveness with which revenue-generating potential is exploited. The ARPU can be broken down according to income-producing categories or according to diverse factors such as geographic location, user age, user occupation, user income and the total time per month each user spends on the system.

AMPU: The Average Margin per User is calculated on the basis of net profit rather than total income. In recent years, some telecommunications carriers have increased their reliance on AMPU rather than ARPU to maximize their returns as niche markets become saturated. By breaking down customer sales by margin rather than by revenue, companies that have lower sales volumes but create larger margins can be considered more efficient and arguably more profitable than their high-volume competitors.

Real-Time Data: Real-time data in customer churn makes the best possible solution today, as it is based on up-to-the-moment information about a subscriber. Achieving real-time data enables the company to immediately adjust it's offers and solutions in response to the reason of dissatisfaction/ discontinuation of services. Deploying analytics and systems that trigger the moment your subscriber is shifting to your competitor, helps process the retention effective and faster.

Binary classification method: This method uses a gain/loss matrix, which incorporates the gain of targeting and retaining the most valuable churners and the cost of incentives to the targeted customers. This approach leads to far more profitable retention campaigns than the traditional churn modeling approaches.

General Signs: Customers today are highly conscious of what they need and what is available in the market. The telecom players should always lookout for signs that the customer may be planning to shift. These can easily be picked up from sales support interaction with them- when he bluntly says he is shifting to a competitor, when he is

quoting what other players offer, when he is enquiring about MNP or when he is simply calling competitor's phone line looking for alternatives to his problem.

- **Solutions to Reduce Churn**

Telecom players use a variety of different metrics to determine when customers are about to churn, or leave. It is profitable for companies to explore the reasons why customers are leaving, and then target at-risk customers with enticing offers.

1. There are a number of different tactics companies use to maintain their customer bases. One of the most important is simply providing **efficient customer service**. Providing clients with an easy way to get questions answered and issues handled is the key to maintaining cellular clients.
2. **Value- added services** serve as a subscriber retention tool, especially for established players. While for newer entrants, it will become a part of the marketing strategy to attract customers. If VAS providers leverage the opportunities to tie up with operators, there could be a major increase in the uptake of their services.
3. A commonly used tactic is for a carrier to **offer upgrades** on the client's existing account. Expanding on services offered and giving better rates or discounts to the client often improves customer retention rates.
4. Another tactic is offering free access or **reduced rates** on smartphone applications. The increasing regular use by customers of cellphone applications makes free access to such applications an enticing bonus for many customers.
5. Competing cellular providers aggressively market **special deals** to churn customers away from their current provider. Common practices include offering free phones and buying out any existing service contract. The cellular service business is highly competitive and will likely remain so; therefore, churn rates will continue to be an important focus for cellular providers.

6. **Personalized Tariff plans** and service recommendations to each subset of subscribers because a one-size-fit strategy is no longer suitable for telecom sector, every user has a different purpose and usage pattern.
7. By **leveraging the user traffic**, operators' strategic and technical teams can make clear and decisive decisions to reduce costs by millions of dollars without jeopardizing quality.
8. Fighting wireless churn with trendy smartphones and **fast data network**
9. **One-on-one Marketing** is one of the best tactics to reduce churn rate. Make sure that customers are communicated the new services offering based on their usage analysis and trends and should be given proactive information on the plans which will benefit the customer.
10. Effective communication is one way to reduce churn. **Being proactive** in addressing difficulties and issues faced by your customers not only helps in building trust and reliability but also ensures a strong working relationship.
11. **Cultivate loyalty** with attractive smartphone portfolios and strong mobile data networks to support those devices. Loyal customers are less likely to churn because they are more invested in your business relationship and companies have built up a long history of delivering good results and keeping promises.

2.2 Methods of Churn Prediction

In the past few years, many different researches have been conducted in this domain and researchers have investigated many different approaches and parameters to gauge the churn of a customers for a Telco company. Firms in telecommunication sector have detailed call records. These firms can segment their customers by using call records for developing price and promotion strategies.

Richter et al.[5], tried an unconventional approach for predicting churners. With the analysis of customers' call data, they detected the customers who were friends with each other. They claimed that group of friends tend to churn together; thus, group of customers should be considered together. They found that small social groups often have dominant leader that affects group's decision to churn. The study indicated that possibility to churn

for small groups are 2.7 times higher than for larger groups. In our study, we did not have the call details of the customers that represent social connectivity; therefore we were not able to consider this social group approach.

Utku Yabas[1] used weka , an open source software for data mining for churn prediction problem. After the preprocessing, they have focused on many ensemble methods. Algorithm that gave them highest accuracy was random forest. Random Forests is built from many decision trees. At each node of single decision trees, number of random features are chosen and best split on these m is used to split that node. Each single tree votes for the prediction. Most popular prediction among many decision trees is the output of the forest. This model achieved 76.51% accuracy on churn prediction.

By using Data Mining techniques, the subscribers who are intended not to make any payments, can be detected from before. And also, financial losses can be prevented. For this type of analysis, Deviation Determination method is applied. According to usage patterns subscribers are divided into specific clusters. The ones showing inconsistent features are determined and will be reviewed. By using Data Mining Techniques, International Roaming Agreements can also be optimized.

In their study Ren, Zheng and Wu [3] presented a clustering method based on genetic algorithm for telecommunication customer subdivision. First, the features of telecommunication customers (such as the calling behavior and consuming behavior) are extracted. Then, the similarities between the multidimensional feature vectors of telecommunication customers are computed and mapped as the distance between samples on a two-dimensional plane. Finally, the distances are adjusted to approximate the similarities gradually by genetic algorithm.

In response to the difficulty of churner prediction, Chang's [4] study applies data mining techniques to build a model for churner prediction. Through an analysis result from a big Taiwan telecom provider, the results indicated that the proposed approach has pretty good prediction accuracy by using customer demography, billing information, call detail records, and service changed log to build churn prediction mode by using Artificial Neural Networks.

Wei and Chiu [8] proposed design, and experimentally evaluate a churn-prediction technique that predicts churning from subscriber contractual information and call pattern changes extracted from call details. This proposed technique is capable of identifying potential churners at the contract level for a specific prediction time-period. The largest telecom company in Taiwan that has 21 million subscribers, were selected for application. They expressed that, the more call records they have, the more accurate results they can have from Churn analysis.

Vodafone which bought Telsim applies data mining for sales, marketing, financial management, future prediction, and for many different needs. Vodafone detects peak hours by using its databases and makes more workforces ready to avoid any disruption in communication. Also, Vodafone determines average of prepaid minutes purchased and finds subscribers who will likely churn.

2.3 Problems in Existing Solutions

Above mentioned methodologies had few problems. Some of them are mentioned below:

- Best way of churn prediction is to directly get hint about customers sentiments by listening to their calls with their groups. But it is equally difficult to get call details of the customers that represent social connectivity.
- **Problem with Weka tool:** Problem with this solution is that it suffers from bias problem on classification. Resampling technique could improve the score by increasing the no. of positive example and decreasing the no. of negative example which would add some cost on positive examples. It is also sometimes difficult to compare the quality of the clusters produced.
- Information collected through data mining intended for the ethical purposes can be misused. This information may be exploited by unethical people or businesses to take benefits of vulnerable people or discriminate against a group of people.

- In addition, data mining technique is not perfectly accurate. Therefore, if inaccurate information is used for decision-making, it will cause serious consequence.

- If we use Support Vector Machine, it suffers from few problems. Choosing a “good” kernel function is not easy. Long training time on large data sets is required which causes delay. It becomes very difficult to interpret and understand final model.Since the final model is not so easy to see, we can not do small calibrations to the model hence its tough to incorporate our business logic SVMs do not perform well on highly skewed/imbalanced data sets. These are training data sets in which the number of samples that fall in one of the classes far outnumber those that are a member of the other class. Customer churn data sets are typically in this group because when you collect the training set, among a million customers during a particular time period, there would be very few who have actually churned. SVMs are not efficient if the number of features are very huge in number compared to the training samples.

Chapter-3

Methodology

Chapter-3

Methodology

3.1 Architectural Analysis

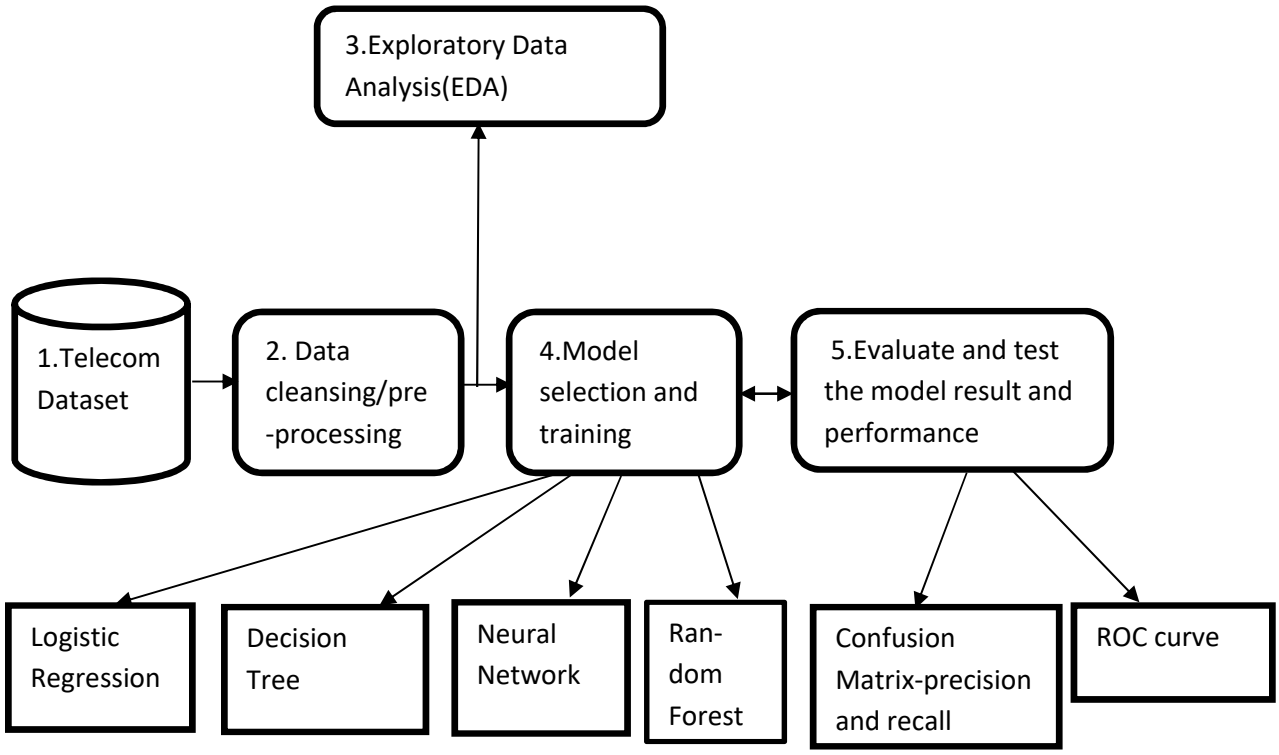


Fig 3.1 Architecture diagram

3.1.1 Telecom Dataset

First, we obtained the database from https://www.ibm.com/communities/analytics/watson-analytics-blog/predictive-insights-in-the-telco-customer-churn-data-set/ which contained parameters like gender, phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies, contract , payment method, tenure ,monthly charges etc. Once, all the entries were complete the data was checked for NA (Not Available) values and all such records were removed. It was available in .csv format.

We load the dataset in R using read.csv command.

3.1.2 Data cleaning/pre-processing

In this step, we treat the missing value by either removing them or replacing them with mean, median or any appropriate value. We convert the variables of dataset into appropriate data type.

3.1.3 Exploratory Data Analysis(EDA)

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. In this step we have done the univariate analysis and bivariate analysis for the variables in the dataset.

3.1.4 Model selection and training

Model selection is a task of selecting a statistical model from a set of candidate models, given data. In this project, we have used logistic regression, decision tree, neural network and random forest models as they work very efficiently for classification problem.

In training step, we divide the data into two parts-one for training and another part for validation. Training set is used to make predictive model whereas testing set is used to evaluate the performance of our model. For this purpose, K-fold cross validation is used which splits entire dataset into k-'folds' randomly. For each k-fold, model is built on k-1 folds of dataset. Then, we test model to check the effectiveness for kth fold. Error on each of the predictions is recorded.This process is repeated until each of the k-folds has served as the test set. Average of k recorded errors is called cross-validation error and will serve as your performance metric for the model.

3.1.4.1 Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.).

The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest:

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic}$$

And

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

Rather than choosing parameters that minimize the sum of squared errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values.

Example: #logistic regression model

```
model <- glm (Recommended ~ .-ID, data = dresstrain, family = binomial)
summary(model)
predict <- predict(model, type = 'response')
 #confusion matrix
table(dresstrain$Recommended, predict > 0.5)
#ROC Curve
library(ROCR)
ROCRpred <- prediction(predict, dresstrain$Recommended)
ROCRperf <- performance(ROCRpred, 'tpr','fpr')
plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7))
#plot glm
```

```
library(ggplot2)
ggplot(dresstrain, aes(x=Rating, y=Recommended)) + geom_point() +
stat_smooth(method="glm", family="binomial", se=FALSE)
```

3.1.4.2 Neural Networks

An **Artificial Neural Network**, often just called a neural network, is a mathematical model inspired by biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. In most cases a neural network is an adaptive system that changes its structure during a learning phase. Neural networks are used to model complex relationships between inputs and outputs or to find patterns in data.

The inspiration for neural networks came from examination of central nervous systems. In an artificial neural network, simple artificial nodes, called "neurons", "neurodes", "processing elements" or "units", are connected together to form a network which mimics a biological neural network.

There is no single formal definition of what an artificial neural network is. Generally, it involves a network of simple processing elements that exhibit complex global behavior determined by the connections between the processing elements and element parameters. Artificial neural networks are used with algorithms designed to alter the strength of the connections in the network to produce a desired signal flow.

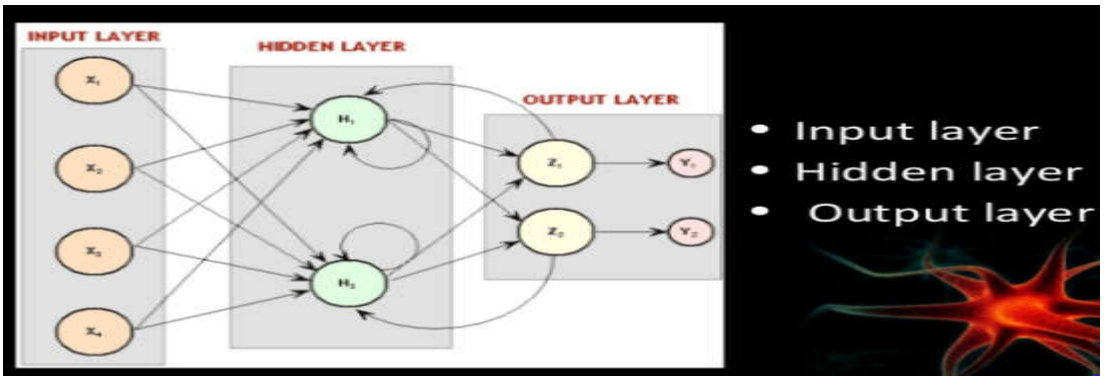


Fig 3.2 Artificial Neuron System

Advantages of neural network:

- A neural network can perform tasks that a linear program cannot.
- When an element of neural network fails ,it can continue without any problem due to its parallel nature.
- A neural networks learns and doesn't need to be reprogrammed. It works even in presence of noise with good quality output.

Example: #Neural Network

```
# Load neuralnet R library
library(neuralnet)
# Build a Neural Network having 1 hidden layer with 2 nodes
nn = neuralnet(Personal.Loan ~. , data=train_Data, hidden=2)
# See covariate and result varaibls of neuralnet model
out <- cbind(nn$covariate, nn$net.result[[1]])
out
 # Plot the neural network
 plot(nn)
```

3.1.4.3 Decision tree

Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a discrete set of values are called **classification trees**; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called **regression trees.**

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision

taken after computing all attributes). The paths from root to leaf represent classification rules.

In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

Decision tree learning is a method commonly used in data mining.[1] The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown in the diagram at right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

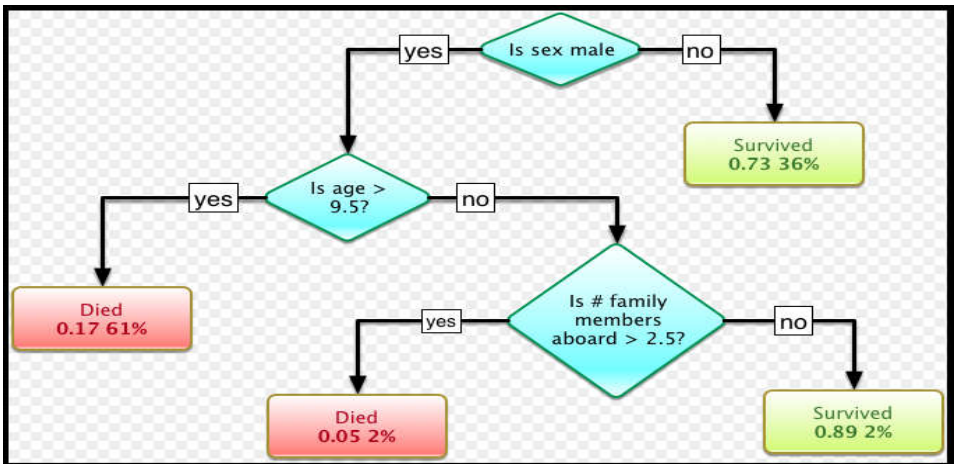


Fig 3.3 Decision Tree

Ex. A tree showing survival of passengers on the Titanic. The figures under the leaves show the probability of survival and the percentage of observations in the leaf.

A decision tree is a simple representation for classifying examples. For this section, assume that all of the input features have finite discrete domains, and there is a single target feature called the "classification". Each element of the domain of the classification is called a class. A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with an input feature are labeled with each of the possible values of the target or output feature or the arc leads

to a subordinate decision node on a different input feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes[why?].

Example: #decision tree

```
library(rpart)
x <- cbind(x_train,y_train)
# grow tree
fit <- rpart(y_train ~ ., data = x,method="class")
summary(fit)
#Predict Output
predicted= predict(fit,x_test)
```

3.1.4.4 Random Forest

Random Forest is a versatile machine learning method capable of performing both regression and classification tasks. It also undertakes dimensional reduction methods, treats missing values, outlier values and other essential steps of data exploration, and does a fairly good job. It is a type of ensemble learning method, where a group of weak models combine to form a powerful model.

In Random Forest, we grow multiple trees as opposed to a single tree in CART model. To classify a new object based on attributes, each tree gives a classification and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest) and in case of regression, it takes the average of outputs by different trees. It works in the following manner. Each tree is planted & grown as follows:

1. Assume number of cases in the training set is N. Then, sample of these N cases is taken at random but *with replacement*. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m<M is specified such that at each node, m variables are selected at random out of the M. The best split on these m is used to split the node. The value of m is held constant while we grow the forest.
3. Each tree is grown to the largest extent possible and there is no pruning.

4. Predict new data by aggregating the predictions of the ntree trees (i.e., majority votes for classification, average for regression).

Advantages of Random Forest:

- This algorithm can solve both type of problems i.e. classification and regression and does a decent estimation at both fronts.
- One of benefits of Random forest which excites me most is, the power of handle large data set with higher dimensionality. It can handle thousands of input variables and identify most significant variables so it is considered as one of the dimensionality reduction methods. Further, the model outputs Importance of variable, which can be a very handy feature (on some random data set).
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.
- It has methods for balancing errors in data sets where classes are imbalanced.
- The capabilities of the above can be extended to unlabeled data, leading to unsupervised clustering, data views and outlier detection.
- Random Forest involves sampling of the input data with replacement called as bootstrap sampling. Here one third of the data is not used for training and can be used to testing. These are called the out of bag samples. Error estimated on these out of bag samples is known as *out of bag error*. Study of error estimates by Out of bag, gives evidence to show that the out-of-bag estimate is as accurate as using a test set of the same size as the training set. Therefore, using the out-of-bag error estimate removes the need for a set aside test set.

Disadvantages of Random Forest:

It surely does a good job at classification but not as good as for regression problem as it does not give precise continuous nature predictions. In case of regression, it doesn't predict beyond the range in the training data, and that they may over-fit data sets that are particularly noisy.

Random Forest can feel like a black box approach for statistical modelers – you have very little control on what the model does. You can at best – try different parameters and random seeds!

Example: #Random forest

```
library(randomForest)
set.seed(71)
rf<-randomForest(Creditability~.,data=mydata,ntree=500)
print(rf)
```

3.1.5 Evaluate and test the model result and performance

Predictive Modeling works on constructive feedback principle. You build a model. Get feedback from metrics, make improvements and continue until you achieve a desirable accuracy. Evaluation metrics explain the performance of a model. An important aspects of evaluation metrics is their capability to discriminate among model results. We have used confusion matrix, ROC curve and AUC curve as evaluation metrices.

Parameters of judgement:

3.1.5.1 Confusion Matrix

A confusion matrix shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data. The matrix is *N*x*N*, where *N* is the number of target values (classes). Performance of such models is commonly evaluated using the data in the matrix. The following table displays a 2x2 confusion matrix for two classes (Positive and Negative).

Confusion Matrix		**Target**			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

Fig 3.4 Confusion Matrix

- **Accuracy**: the proportion of the total number of predictions that were correct.

- **Positive Predictive Value** or **Precision**: The proportion of positive cases that were correctly identified.
- **Negative Predictive Value** : the proportion of negative cases that were correctly identified.
- **Sensitivity** or **Recall** : the proportion of actual positive cases which are correctly identified.
- **Specificity** : the proportion of actual negative cases which are correctly identified.

Example:

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	70	20	*Positive Predictive Value*	0.78
	Negative	30	80	*Negative Predictive Value*	0.73
		Sensitivity	*Specificity*	**Accuracy** = 0.75	
		0.70	0.80		

Fig 3.5 Example of Confusion Matrix

3.1.5.2 ROC Chart

The ROC chart is similar to the gain or lift charts in that they provide a means of comparison between classification models. The ROC chart shows false positive rate (1-specificity) on X-axis, the probability of target=1 when its true value is 0, against true positive rate (sensitivity) on Y-axis, the probability of target=1 when its true value is 1. Ideally, the curve will climb quickly toward the top-left meaning the model correctly predicted the cases.

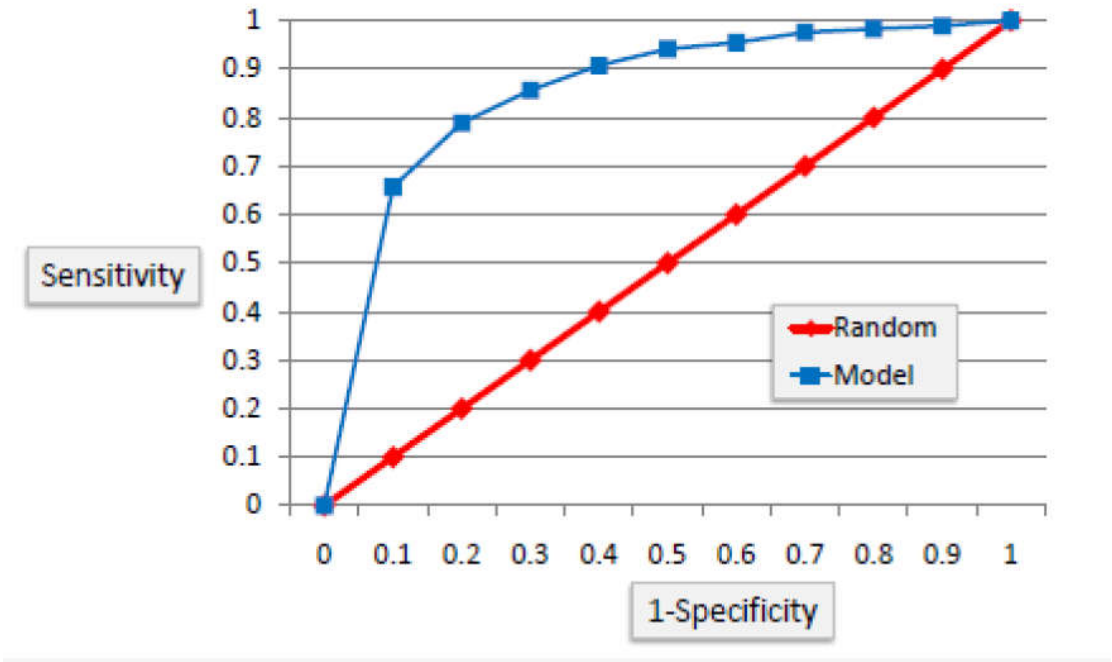


Fig 3.6 ROC curve

3.1.5.3 Area Under the Curve (AUC)

Area under ROC curve is often used as a measure of quality of the classification models. A random classifier has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1. In practice, most of the classification models have an AUC between 0.5 and 1.

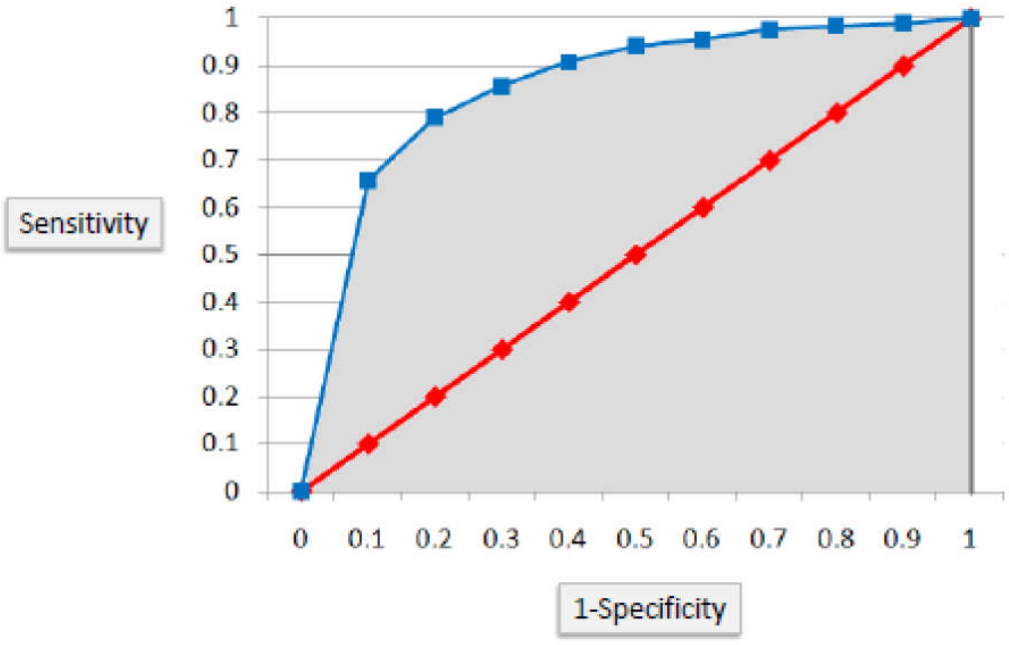


Fig 3.7 ROC curve showing Area Under Curve

An area under the ROC curve of 0.8, for example, means that a randomly selected case from the group with the target equals 1 has a score larger than that for a randomly chosen case from the group with the target equals 0 in 80% of the time. When a classifier cannot distinguish between the two groups, the area will be equal to 0.5 (the ROC curve will coincide with the diagonal). When there is a perfect separation of the two groups, i.e., no overlapping of the distributions, the area under the ROC curve reaches to 1 (the ROC curve will reach the upper left corner of the plot).

Chapter-4

Experiment Results

Chapter-4

Experiment and Results

4.1. About Dataset

The dataset includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents.

customer	gender	SeniorCiti	Partner	Depender	tenure	PhoneSer	MultipleLi	InternetSe	OnlineSec	OnlineBac	DevicePro	TechSupp	Streaming	Streaming	Contract	Paperless	PaymentM	MonthlyC	TotalChar	Churn
7590-VHV	Female	0	Yes	No	1	No	No phone	DSL	No	Yes	No	No	No	No	Month-to-	Yes	Electronic	29.85	29.85	No
5575-GNV	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed ch	56.95	1889.5	No
3668-QPY	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-	Yes	Mailed ch	53.85	108.15	Yes
7795-CFO	Male	0	No	No	45	No	No phone	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank trans	42.3	1840.75	No
9237-HQI	Female	0	No	No	2	Yes	No	Fiber opti	No	No	No	No	No	No	Month-to-	Yes	Electronic	70.7	151.65	Yes
9305-CDS	Female	0	No	No	8	Yes	Yes	Fiber opti	No	No	Yes	No	Yes	Yes	Month-to-	Yes	Electronic	99.65	820.5	Yes
1452-KIO	Male	0	No	Yes	22	Yes	Yes	Fiber opti	No	Yes	No	No	Yes	No	Month-to-	Yes	Credit car	89.1	1949.4	No
6713-OKO	Female	0	No	No	10	No	No phone	DSL	Yes	No	No	No	No	No	Month-to-	No	Mailed ch	29.75	301.9	No
7892-POO	Female	0	Yes	No	28	Yes	Yes	Fiber opti	No	No	Yes	Yes	Yes	Yes	Month-to-	Yes	Electronic	104.8	3046.05	Yes
6388-TAB	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank trans	56.15	3487.95	No
9763-GRS	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No	Month-to-	Yes	Mailed ch	49.95	587.45	No
7469-LKB	Male	0	No	No	16	Yes	No	No	No intern	No intern	No intern	No intern	No intern	No intern	Two year	No	Credit car	18.95	326.8	No
8091-TTV	Male	0	Yes	No	58	Yes	Yes	Fiber opti	No	No	Yes	No	Yes	Yes	One year	No	Credit car	100.35	5681.1	No
0280-XJG	Male	0	No	No	49	Yes	Yes	Fiber opti	No	Yes	Yes	No	Yes	Yes	Month-to-	Yes	Bank trans	103.7	5036.3	Yes
5129-JLP	Male	0	No	No	25	Yes	No	Fiber opti	Yes	No	Yes	Yes	Yes	Yes	Month-to-	Yes	Electronic	105.5	2686.05	No
3655-SNQ	Female	0	Yes	Yes	69	Yes	Yes	Fiber opti	Yes	Yes	Yes	Yes	Yes	Yes	Two year	No	Credit car	113.25	7895.15	No
8191-XWS	Female	0	No	No	52	Yes	No	No	No intern	No intern	No intern	No intern	No intern	No intern	One year	No	Mailed ch	20.65	1022.95	No
9959-WOF	Male	0	No	Yes	71	Yes	Yes	Fiber opti	Yes	No	Yes	No	Yes	Yes	Two year	No	Bank trans	106.7	7382.25	No
4190-MFL	Female	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No	No	Month-to-	No	Credit car	55.2	528.35	Yes
4183-MYF	Female	0	No	No	21	Yes	No	Fiber opti	No	Yes	Yes	No	No	Yes	Month-to-	Yes	Electronic	90.05	1862.9	No
8779-QRD	Male	1	No	No	1	No	No phone	DSL	No	No	Yes	No	No	Yes	Month-to-	Yes	Electronic	39.65	39.65	Yes
1680-VDC	Male	0	Yes	No	12	Yes	No	No	No intern	No intern	No intern	No intern	No intern	No intern	One year	No	Bank trans	19.8	202.25	No
1066-JKS	Male	0	No	No	1	Yes	No	No	No intern	No intern	No intern	No intern	No intern	No intern	Month-to-	No	Mailed ch	20.15	20.15	Yes
3638-WEA	Female	0	Yes	No	58	Yes	Yes	DSL	No	Yes	No	Yes	No	No	Two year	Yes	Credit car	59.9	3505.1	No

Fig 4.1 Fields in Dataset

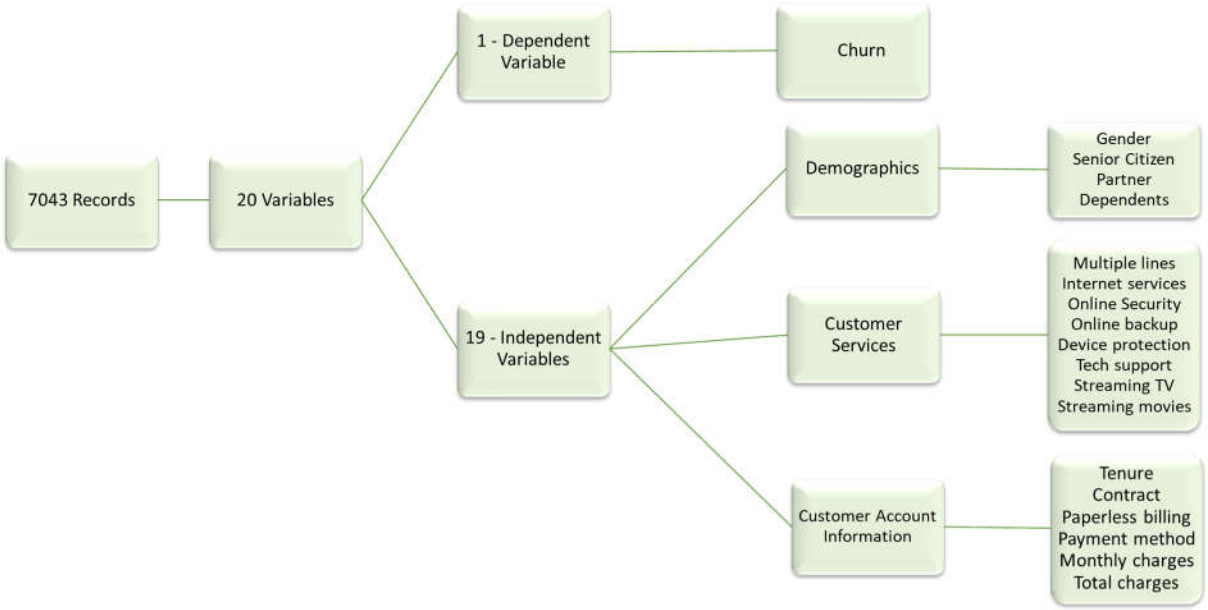


Fig 4.2 Analysis of variables

```
> str(telco)
'data.frame':   7032 obs. of  20 variables:
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1
2 ...
 $ SeniorCitizen   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
 $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
 $ MultipleLines   : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2
 1 3 3 2 3 1 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2
 2 1 2 1 ...
 $ OnlineSecurity  : Factor w/ 3 levels "No","No internet service",..: 1 3
3 3 1 1 1 3 1 3 ...
 $ OnlineBackup    : Factor w/ 3 levels "No","No internet service",..: 3 1
3 1 1 1 3 1 1 3 ...
 $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3
1 3 1 3 1 1 3 1 ...
 $ TechSupport     : Factor w/ 3 levels "No","No internet service",..: 1 1
1 3 1 1 1 1 3 1 ...
 $ StreamingTV     : Factor w/ 3 levels "No","No internet service",..: 1 1
1 1 1 3 3 1 3 1 ...
 $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1
1 1 1 3 1 1 3 1 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1
1 1 2 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4
 4 1 3 3 2 4 3 1 ...
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
```

Fig 4.3 Structure of Dataset

- **Training and Testing data**

  In order to assess the performance of the classifier we split the date into training and testing set .We are dividing the data set into training and testing data in 80:20 respectively.

  We achieved this ratio by using **K-fold cross validation** with 5 folds.

4.2 Data Preprocessing

- **Demographic Variable**

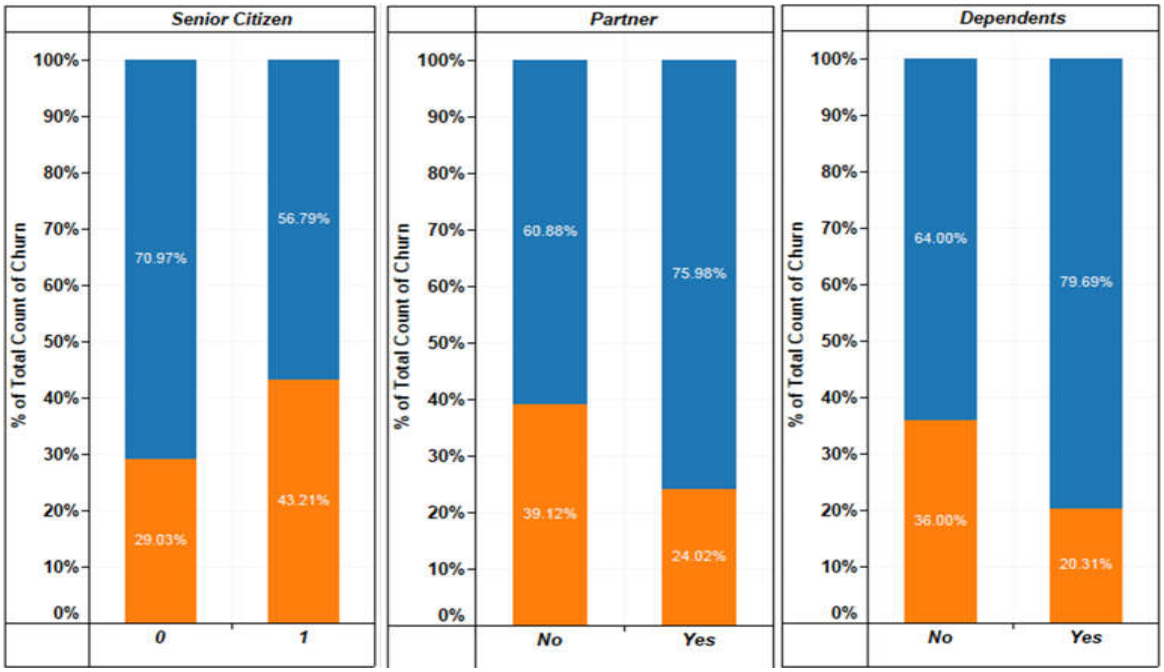


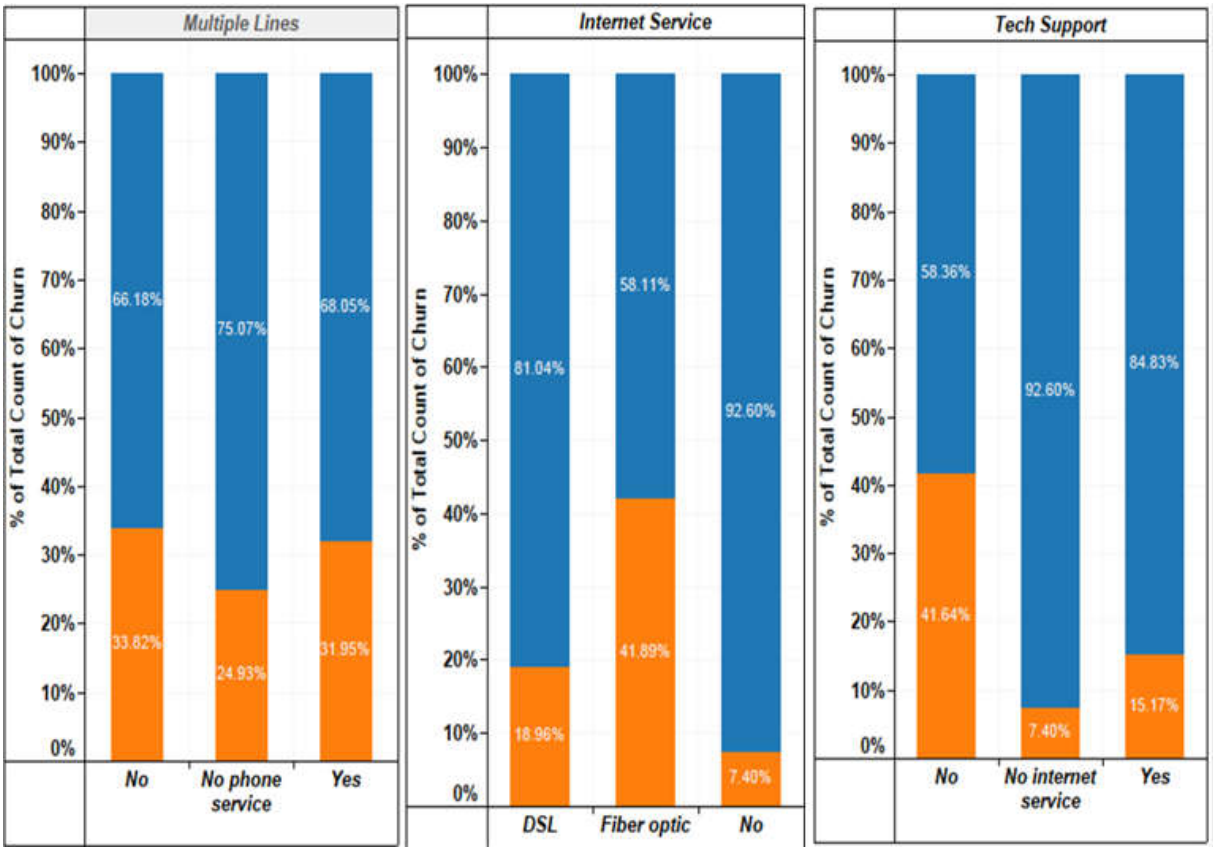
Fig 4.4 Demographic Variable

From Fig 4.4, we conclude that

Here, 0 represents person is not a senior citizen while 1 denotes he/she is a senior citizen.

- Churn rate for senior citizens is 50% more than that of others.
- Customers without partners churn more.
- Customers with dependents churn less.

- **Customer Service**



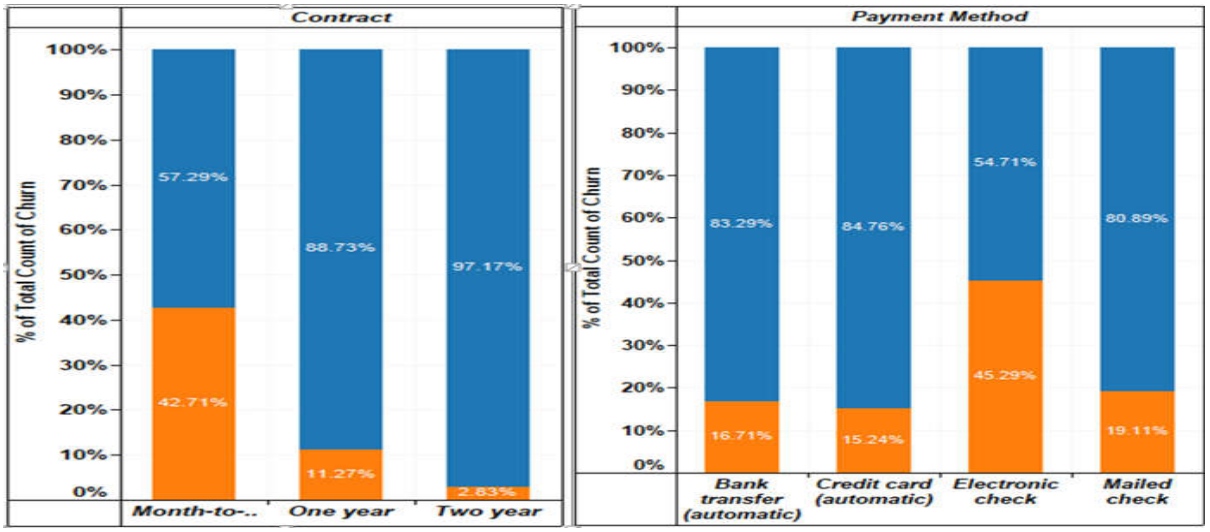


Fig 4.5 Univariate analysis on basis of Customer service

From Fig 4.5, we conclude that

- Customers with Fiber optic as Internet Service have a churn rate that is more than double of the churn rate of customers with DSL as internet service.
- Customers with no internet service have a very low churn of only 7%.
- Customers with tech support churn less.
- As the contact duration increases, the churn rate goes down.
- Based on payment method, Electronic check has the highest churn followed by Mailed check which is less than half on the former.

- **Bivariate Analysis**

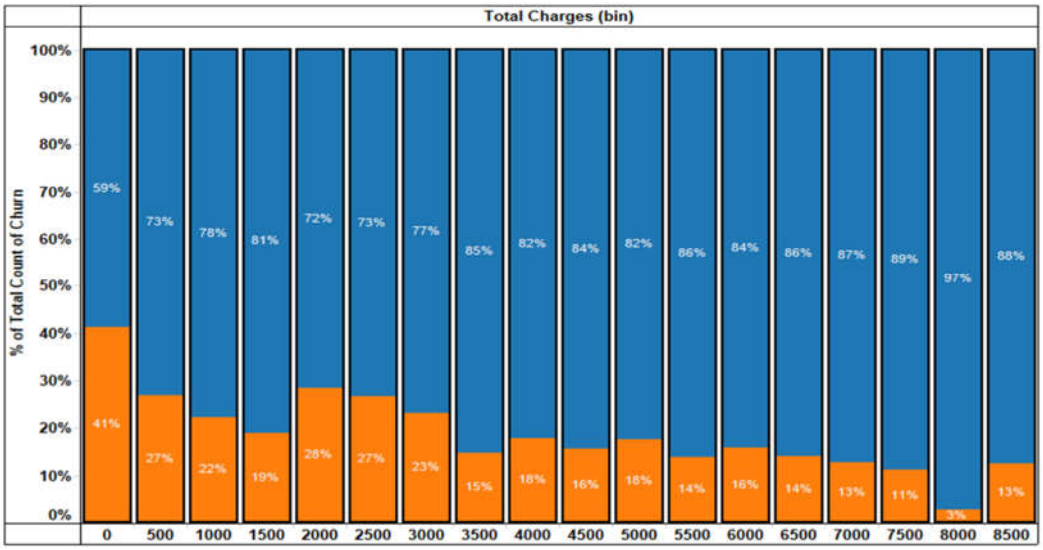


Fig 4.6 Bivariate Analysis of categorical variable

- Churn rate is decreasing with increase in total charges of the services.

4.3 Implementation

4.3.1 Logistic Regression Model

1. Model 1

```
LogReg_Model <- glm(formula = Churn ~.,
                    data = TrainData,family="binomial")
summary(LogReg_Model)
```

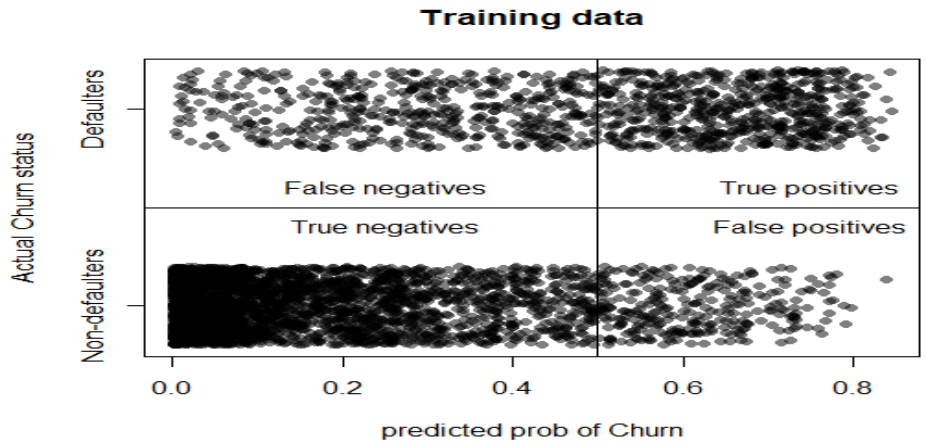


Fig 4.7 Training Data

Fig 4.7 shows precision of the model we implemented by comparing it with actual results.

- True negative- Customers who were not likely to churn and our model predicted them correctly
- True positive- Customers who were likely to churn and our model predicted them correctly.
- False positive- Customers who were likely to churn but our model predicted about them incorrectly.
- False negative- Customers who were not likely to churn but our model predicted about them incorrectly.

Higher density of dots in True positive and True negative indicates that our model had a decent accuracy.

2. **Model 2 : using stepAIC**

```
> model.step<- stepAIC(model1)
Start:  AIC=3571.79
Churn ~ gender + SeniorCitizen + Partner + Dependents + tenure +
    PhoneService + MultipleLines + InternetService + OnlineSecurity +
    OnlineBackup + DeviceProtection + TechSupport + StreamingTV +
    StreamingMovies + Contract + PaperlessBilling + PaymentMethod +
    MonthlyCharges + TotalCharges
```

3. **Model 3**

```
Step:  AIC=3571.79
Churn ~ gender + SeniorCitizen + Partner + Dependents + tenure +
    MultipleLines + InternetService + OnlineSecurity + OnlineBackup +
    DeviceProtection + TechSupport + StreamingTV + StreamingMovies +
    Contract + PaperlessBilling + PaymentMethod + MonthlyCharges +
    TotalCharges
```

4. **Model4**

```
Step:  AIC=3569.83
Churn ~ gender + SeniorCitizen + Dependents + tenure + MultipleLines +
    InternetService + OnlineSecurity + OnlineBackup + DeviceProtection
    TechSupport + StreamingTV + StreamingMovies + Contract +
    PaperlessBilling + PaymentMethod + MonthlyCharges + TotalCharges

                  Df Deviance    AIC
- OnlineBackup     1   3523.9 3567.9
- gender           1   3524.3 3568.3
- DeviceProtection 1   3524.5 3568.5
- OnlineSecurity   1   3524.6 3568.6
- MonthlyCharges   1   3524.8 3568.8
- StreamingTV      1   3524.9 3568.9
- TechSupport      1   3525.4 3569.4
- StreamingMovies  1   3525.6 3569.6
<none>                 3523.8 3569.8
- SeniorCitizen    1   3526.8 3570.8
- InternetService  1   3526.9 3570.9
- Dependents       1   3527.2 3571.2
- MultipleLines    2   3540.7 3582.7
- PaperlessBilling 1   3540.2 3584.2
- PaymentMethod    3   3545.0 3585.0
- TotalCharges     1   3546.0 3590.0
- Contract         2   3561.2 3603.2
- tenure           1   3609.1 3653.1
```

5. **Model 5:**

```
Step:  AIC=3566.37
Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines +
    InternetService + OnlineSecurity + DeviceProtection + TechSupport +
    StreamingTV + StreamingMovies + Contract + PaperlessBilling +
    PaymentMethod + MonthlyCharges + TotalCharges
```

6. **Model 6**

```
Step:  AIC=3565.54
Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines +
    InternetService + DeviceProtection + TechSupport + StreamingTV +
    StreamingMovies + Contract + PaperlessBilling + PaymentMethod +
    MonthlyCharges + TotalCharges
```

Model	AIC
Model1	3571.83
Model2	3571.79
Model3	3571.54
Model4	3569.79
Model5	3566.37
Model6	3565.54

Table 4.1 Accuracy of models used in Logistic Regression

Table 4.1 shows 6 different model and among which we get 6 different AIC value. AIC. The Akaike information criterion (**AIC**) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection. The lesser the AIC value, the better will be the model.

Since model 6 has the min AIC that is 468.34 it is selected.Model 6:

```
glm(formula = Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines +
    InternetService + DeviceProtection + StreamingTV + StreamingMovies +
    Contract + PaperlessBilling + PaymentMethod + MonthlyCharges +
    TotalCharges, family = "binomial", data = training_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8986  -0.6960  -0.3012   0.7253   3.3218

Coefficients: (3 not defined because of singularities)
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                             2.733e+00  6.044e-01   4.523 6.11e-06 ***
SeniorCitizen1                          1.955e-01  1.071e-01   1.826  0.06778 .
DependentsYes                          -1.930e-01  1.041e-01  -1.853  0.06382 .
tenure                                 -6.697e-02  8.005e-03  -8.366  < 2e-16 ***
MultipleLinesNo phone service          -5.960e-01  2.882e-01  -2.068  0.03865 *
MultipleLinesYes                        6.365e-01  1.164e-01   5.467 4.58e-08 ***
InternetServiceFiber optic              2.592e+00  2.989e-01   8.671  < 2e-16 ***
InternetServiceNo                      -2.343e+00  3.800e-01  -6.167 6.98e-10 ***
DeviceProtectionNo internet service            NA         NA      NA       NA
DeviceProtectionYes                     3.446e-01  1.172e-01   2.940  0.00329 **
StreamingTVNo internet service                 NA         NA      NA       NA
StreamingTVYes                          7.356e-01  1.530e-01   4.809 1.52e-06 ***
StreamingMoviesNo internet service             NA         NA      NA       NA
StreamingMoviesYes                      8.717e-01  1.585e-01   5.501 3.78e-08 ***
Contract0ne year                       -6.093e-01  1.389e-01  -4.386 1.16e-05 ***
ContractTwo year                       -1.194e+00  2.170e-01  -5.503 3.73e-08 ***
PaperlessBillingYes                     3.908e-01  9.537e-02   4.098 4.17e-05 ***
PaymentMethodCredit card (automatic)   -2.166e-01  1.467e-01  -1.476  0.13983
PaymentMethodElectronic check           2.609e-01  1.195e-01   2.183  0.02907 *
PaymentMethodMailed check              -2.004e-01  1.460e-01  -1.373  0.16976
MonthlyCharges                         -7.098e-02  1.184e-02  -5.994 2.04e-09 ***
TotalCharges                            4.134e-04  9.035e-05   4.575 4.75e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model6 summary suggest that all the factors are important such tenure, contract, internet service. Factor such as internet service are positive means as these factor increases the chances of churn increases. While tenure, contract has negative coefficient means higher tenure or contract decrease the chance of churn.

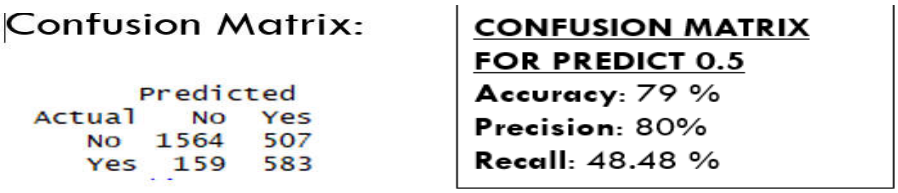


Fig 4.8 Confusion Matrix for Logistic Regression

From Fig 4.8, we get confusion matrix from which we can derive accuracy, precision and recall of our model as shown in figure.

ROC CURVE:

```
library(ROCR)
library(ggplot2)
predicted <- predict(LogReg_updated2,type="response")
prob <- prediction(predicted, TrainData$Churn)
tprfpr <- performance(prob, "tpr", "fpr")
```

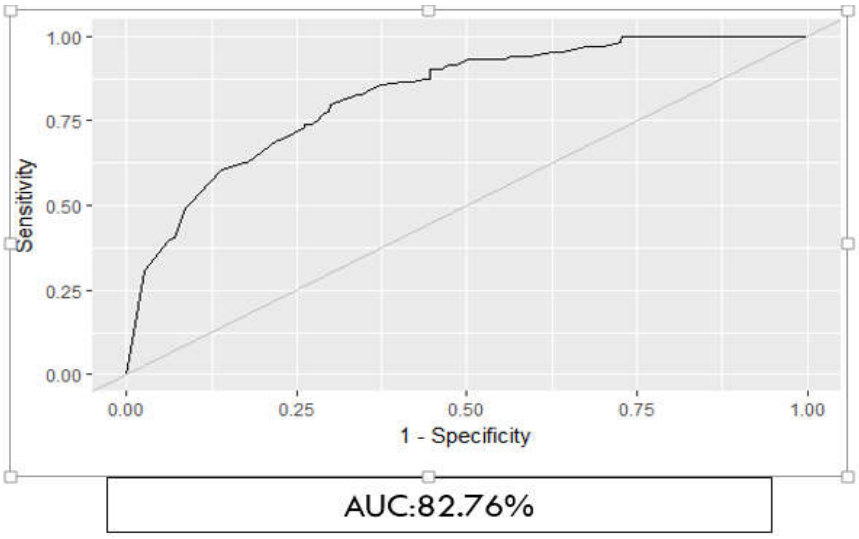


Fig 4.9 ROC Curve for Logistic Regression

Fig 4.9 is a plot of ROC curve which shows that Area Under Curve (AUC) is 82.76% which means prediction probability of our model will be 0.82.

4.3.2 Classification and Regression Trees set

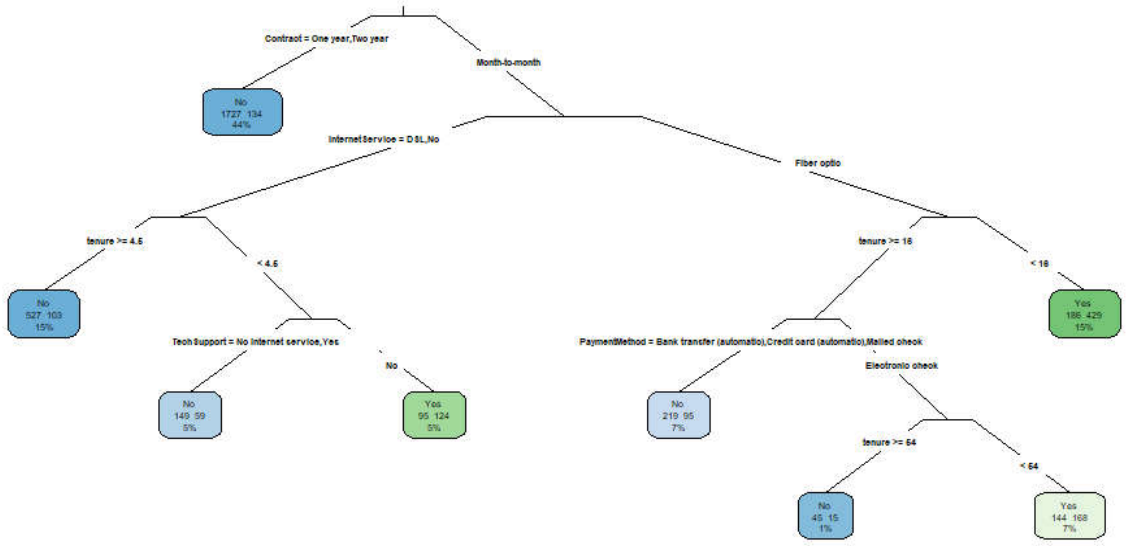
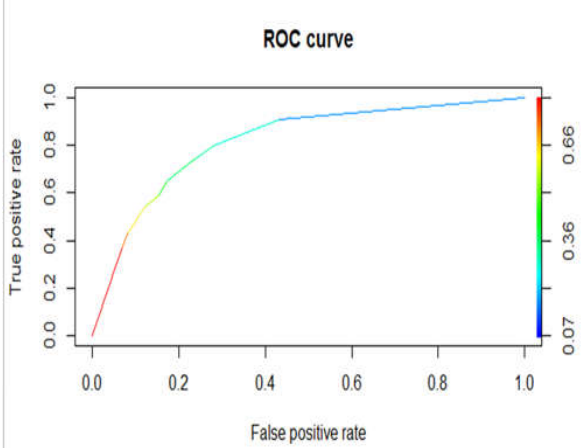


Fig 4.10 Decision tree

Our decision tree defines contract with customers as initial splitting variable, next splitting variable is internet service, then, tech support and so on and the leaf node represents whether customer will churn or not.

```
library(rpart)
library(rpart.plot)
dtree <- rpart(Churn ~.,data = training_data,method = "class",minbucket=30)
dtree
summary(dtree)

rpart.plot(dtree,type=3,extra=101,fallen.leaves = FALSE)
#Predicting on Train
```



Predicted

Actual	No	Yes
No	1371	190
Yes	254	298

	Accuracy (%)	Recall (%)	Precision (%)
Rpart	78.99	53.99	61.07

Fig 4.11 ROC Curve for Classification Tree

In Fig 4.11, ROC curve for decision tree is plotted and we get Area Under Curve equal to 81.48%.

4.3.3 Random Forest

```
> print(rfModel)

Call:
 randomForest(formula = Churn ~ ., data = training_data, nodesize = 25,
ntree = 2000)
               Type of random forest: classification
                     Number of trees: 2000
No. of variables tried at each split: 4

        OOB estimate of  error rate: 19.74%
Confusion matrix:
      No Yes class.error
No  2805 287  0.09282018
Yes  546 581  0.48447205
```

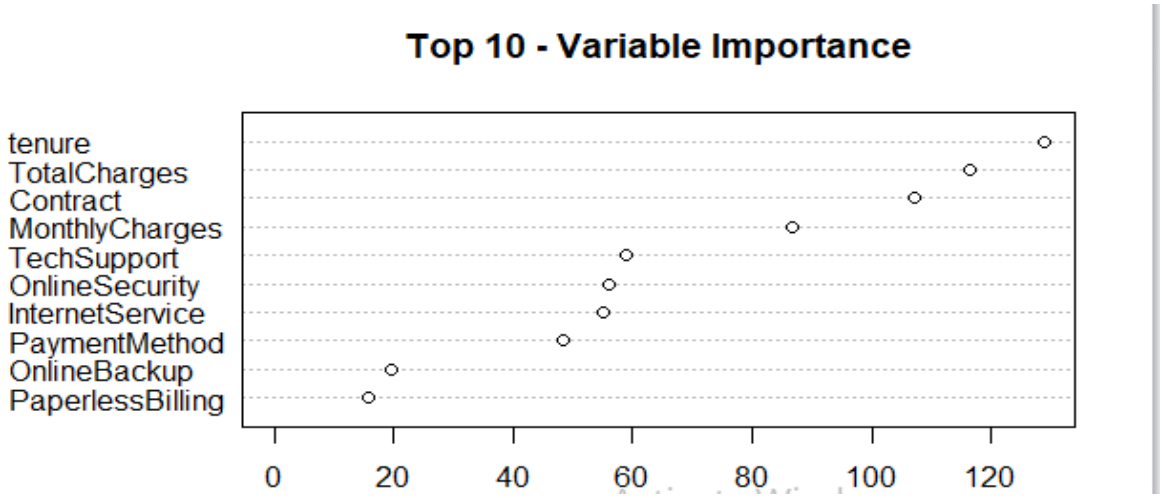
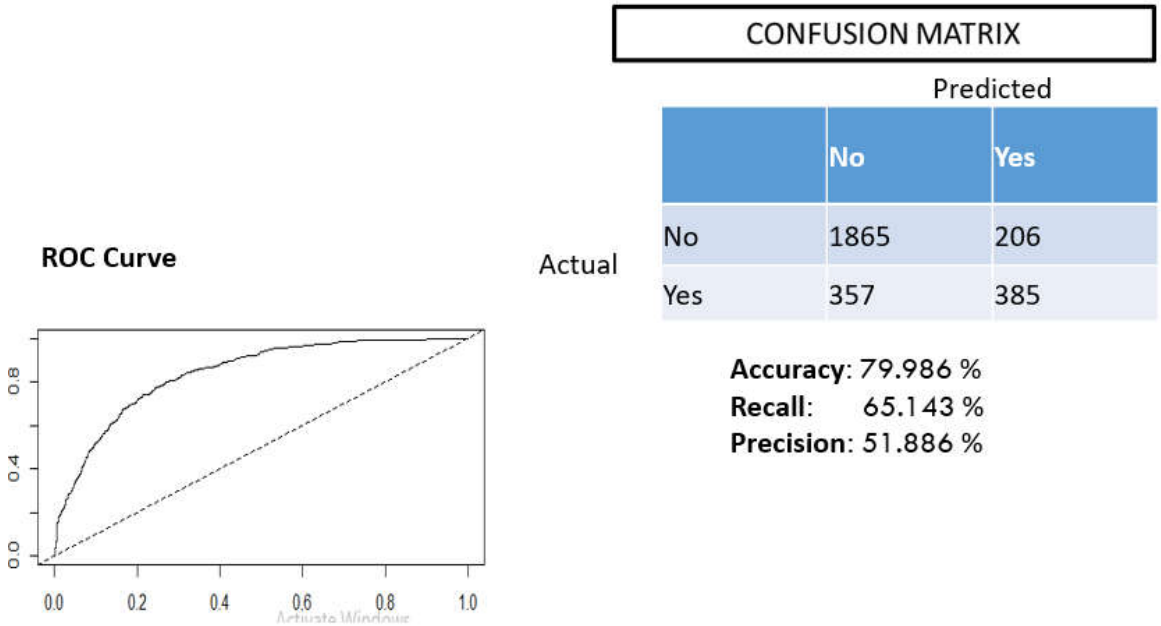


Fig 4.12 Variables in order of importance

Random Forest generates multiple decision tree using different splitting variables and has identified 10 variables of importance as shown in Fig 4.12.



CONFUSION MATRIX

		Predicted	
		No	Yes
Actual	No	1865	206
	Yes	357	385

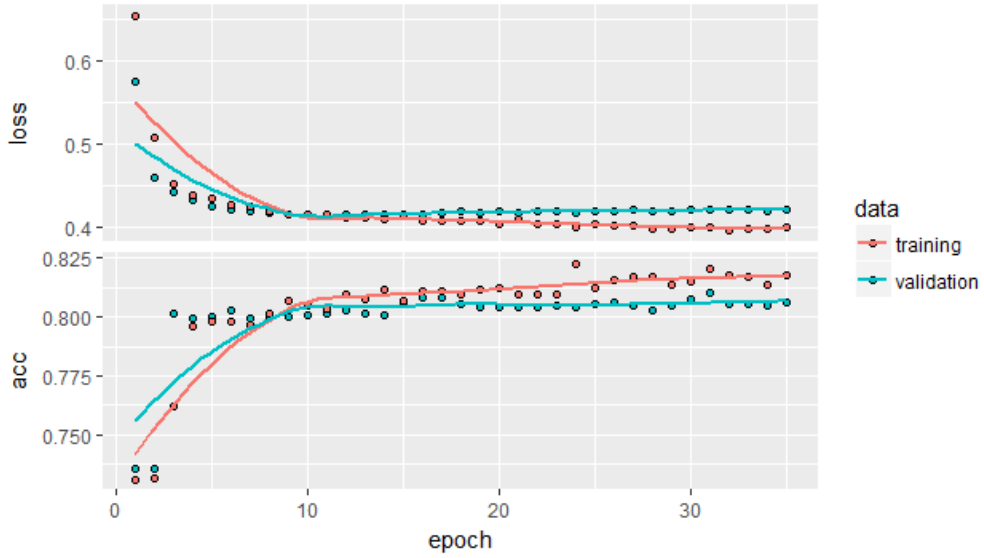
Accuracy: 79.986 %
Recall: 65.143 %
Precision: 51.886 %

Fig 4.13 ROC curve for Random Forest

In Fig 4.13, Area Under Curve for ROC curve is 84.11% which indicates correct prediction probability of model is 0.84.

4.3.4 Artificial Neural Network with keras

```
model_keras %>%

  # First hidden layer
  layer_dense(
    units              = 16,
    kernel_initializer = "uniform",
    activation         = "relu",
    input_shape        = ncol(x_train_tbl)) %>%

  # Dropout to prevent overfitting
  layer_dropout(rate = 0.1) %>%

  # Second hidden layer
  layer_dense(
    units              = 16,
    kernel_initializer = "uniform",
    activation         = "relu") %>%

  # Dropout to prevent overfitting
  layer_dropout(rate = 0.1) %>%

  # Output layer
  layer_dense(
    units              = 1,
    kernel_initializer = "uniform",
    activation         = "sigmoid") %>%

  # Compile ANN
  compile(
    optimizer = 'adam',
    loss      = 'binary_crossentropy',
    metrics   = c('accuracy')
  )
```



Confusion Matrix

Prediction	Truth no	Truth yes
no	1866	357
yes	205	385

Accuracy: 80.00%

Precision: 65.30%

Recall : 51.9%

AUC:84.921%

Fig 4.14 Confusion Matrix for Neural Network

Fig 4.14 shows that as neural network model is deployed, initially there is data loss and accuracy is also poor. But, at later stages accuracy increases and become constant while loss decreases and become constant. At last accuracy is 80% as shown in confusion matrix.

4.4 Result

Four models used for analysis are Logistic regression, Decision tree, Random Forest, Artificial Neural Network.

For evaluation of performance of all models, we have applied k-fold cross validation method with fold equals to 5.

	Accuracy	Precision	Recall	AUC
Fold 1	81.4	64.6	58.6	0.86
Fold 2	81.4	68	57.9	0.856
Fold 3	77.7	61.4	49.2	0.809
Fold 4	79.6	68.7	51.9	0.835
Fold 5	80.5	60.9	62.5	0.861

Table 4.2 Evaluation metrics for different samples in Logistic Regression

	Accuracy	Precision	Recall	AUC
Fold 1	77.9	66.6	36	0.787
Fold 2	79.2	69	47.3	0.832
Fold 3	78.6	63.4	53.9	0.806
Fold 4	79.	61.7	44.8	0.785
Fold 5	79.4	61.9	47.3	0.807

Table 4.3 Evaluation metrics for different samples in Decision Tree

	Accuracy	Precision	Recall	AUC
Fold 1	79.2	67.3	45.6	0.838
Fold 2	80	68.8	50.8	0.837
Fold 3	80.5	64.5	52.7	0.845
Fold 4	80.7	69.3	49.4	0.842
Fold 5	80.9	68.1	49.3	0.842

Table 4.4 Evaluation metrics for different samples in Random Forest

	Accuracy	Precision	Recall	AUC
Fold 1	80.5	67.3	51.6	0.838
Fold 2	80	68.8	50.8	0.837
Fold 3	81.25	63.5	52.7	0.855
Fold 4	82	62.3	49.4	0.852
Fold 5	83	64.1	49.3	0.862

Table 4.5 Evaluation metrics for different samples in Neural Network

Four models used for analysis are Logistic regression, Decision tree, Random Forest, Artificial Neural Network. The summary is as follows:

	Accuracy	Precision	Recall	AUC
Logistic Regression	79	80	48.48	0.8276
Decision Tree	78.99	53.99	61.07	0.8148
Random Forest	79.986	65.14	51.88	0.8411
ANN	80	65.3	51.9	0.85061

Table 4.6 Overall Performance of all models

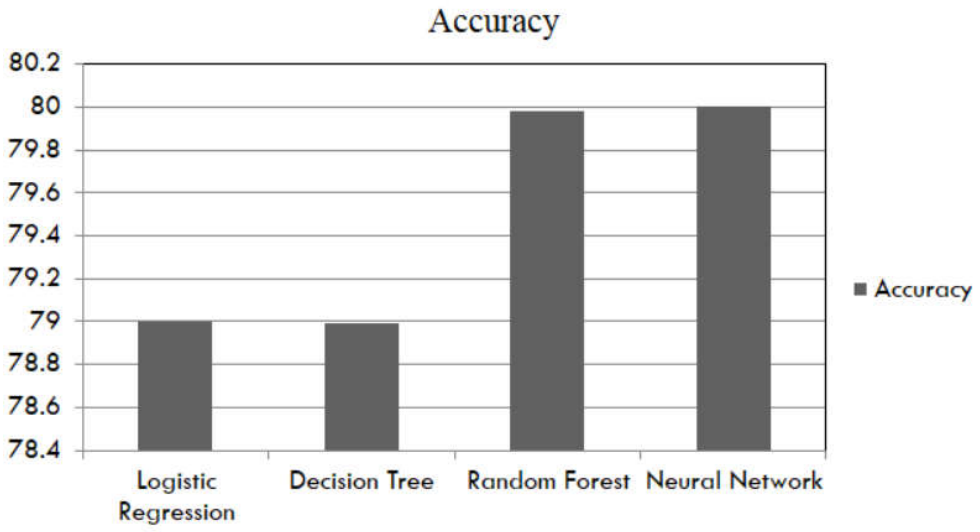


Fig 4.15 Plot for accuracy of all models

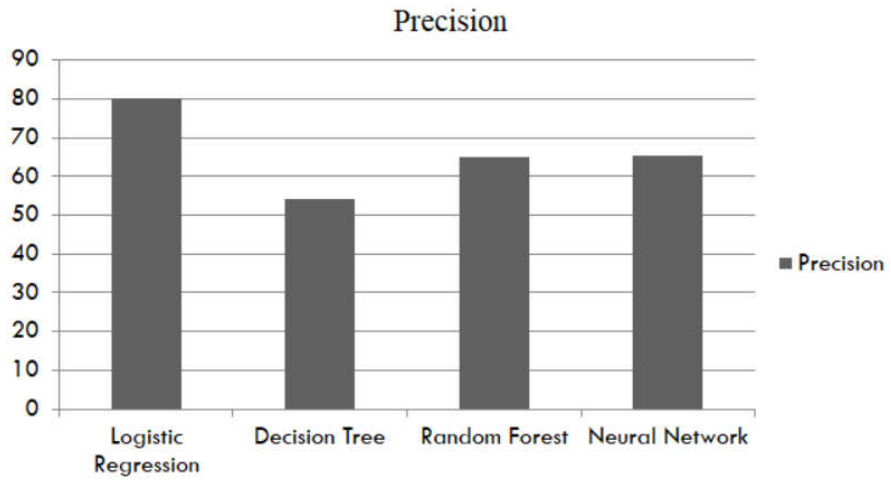


Fig 4.16 Plot for precision of all models

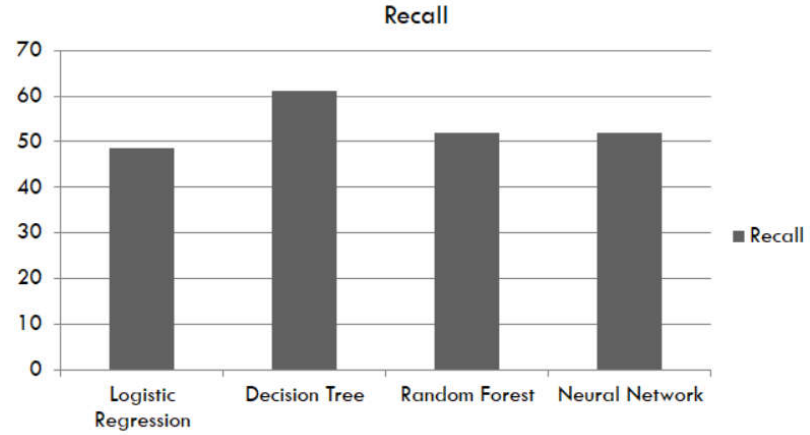


Fig 4.17 Plot for recall of all models

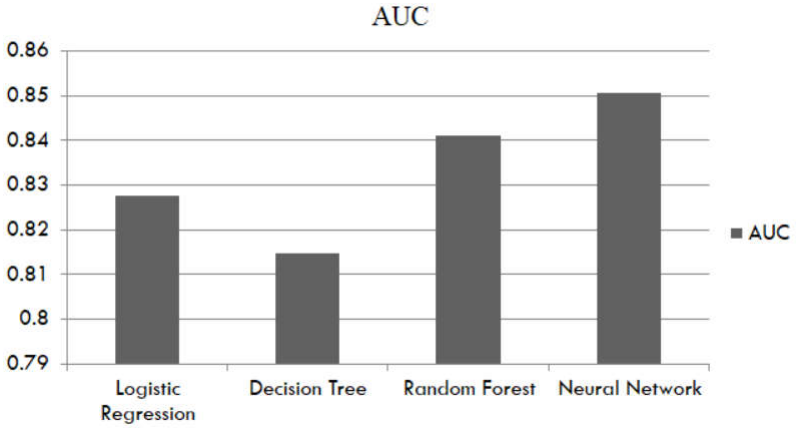


Fig 4.18 Plot for AUC of all models

Chapter-5

Conclusion and Future Enhancements

Chapter-5

Conclusion and Future Enhancements

Telecommunication industry always suffers from a very high churn rates when one industry offers a better plan than the previous there is a high possibility of the customer churning from the present due to a better plan in such a scenario it is very difficult to avoid losses but through prediction we can keep it to a minimal level.

In this project the method we have used is Logistic Regression, Decision trees, Random Forest , Artificial Neural network and this helps to identify the probable churn customers and then make the necessary business decisions.

- **Learning and achievement from the project**

Steps which can help in reducing churn rate:

- Target customers with less than 12 month tenure
- Offer promotion to switch to long term contracts.
- Customers may be dissatisfied with fiber optic service.
- Offer customers a promotion to switch to automatic payments.
- Target users in the lower age demographic.
- Promote online security and other packages that increase retention rates.

- **Limitations**

1. Data mining models have a relatively short expiry life .The mobile market faces new technologies on a daily basis. As a result, historical data become less valuable for prediction.
2. Churn relates to complex interactions within population. Examining all the factors affecting customer churn simultaneously and jointly by building a model is not applicable.
3. The level of the analysis in the data mining models decreases the ability to capture the heterogeneity of the customers. New techniques are thus required to support customer churn analysis to bypass the cons of the data mining techniques.

- **Future Scope**

To improve churn analysis:

a. Include customer interaction in customer churn analysis.
b. Include tools that take account of the heterogeneity of customers.
c. Use new and more advanced algorithms using deep learning and hybrid classification techniques.

The result and accuracy can be bettered if we use more variables in the data.

References

[1] Utku Yabas, Hakki C kanyaka, “Churn Prediction in Subscriber Management for Mobile and Wireless Communications Services”, IEEE Globecom Workshops (GC Wkshps), 2013.

[2] Utku Yabas, Hakki C kanyaka, Turker Ince, “Customer Churn Prediction for Telecom Services”, IEEE 36th Annual Computer Software and Applications Conference, 2012.

[3] Preeti K. Dalvi, Siddhi K. Khandge, “Analysis of Customer Churn Prediction in Telecom industry using Decision Tree and Logistic Regression”, 2016 Symposium on Colossal Data Analysis and Networking (CDAN).

[4] Ammar A Ahmed, Dr. D. Maheshwari, “A review and analysis of Churn Prediction Methods for Customer Retention in Telecom Industries”, 4th International Conference on Advanced Computing and Communication Systems (ICACCS), 2017.

[5] Anujkumar Tiwari, Reuben Sam, Shakila Shaikh, “Analysis and Prediction of Churn Customers for Telecommunication Industry”, International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2017.

[6] Chen Yu-bao, Li Bao-sheng, Ge Xin-quan, “Study on Predictive Model of Customer Churn of Mobile Telecommunication Company”, Fourth International Conference on Business Intelligence and Financial Engineering, 2011.

[7] Essam Shaaban, Yehia Helmy, Ayman Khedr, “A Proposed Churn Prediction Model”, International Journal of Engineering Research and Applications(IJERA), Vol. 2, Issue 4, June-July 2012

[8] Gauri D Limaye, Jyoti P Chaudhary, Prof. Sunil K Punjabi, “Churn Prediction using MapReduce and HBase.”, Mar 15 Volume 3 Issue 3 , International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), ISSN: 2321-8169, PP: 1699 – 1703.

[9] Aishwarya Churi, Mayuri Divekar, Sonal Dashpute, Prajakta Kamble, “Analysis of Customer Churn in Mobile Industry using Data Mining”, International Journal of Emerging Technology and Advanced Engineering, Volume 5, Issue 3, March 2015.

[10] N Kamalraj, Dr. A Malathi, “Appyling Data Mining Techniques in Telecom Churn Prediction”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10, 2013.

[11] Rahul J. Jadhav and Usharani T. Pawar, ”Churn Prediction in Telecommunication Using Data Mining Technology”, International Journal of Advanced Computer Science and Applications, Vol. 2, February 2011.

[12] Clement Kirui1, Li Hong, Wilson Cheruiyot and Hillary Kirui, “Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining”, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, March 2013.

[13] Saad Ahmed Qureshi, Ammar Saleem Rehman, Ali Mustafa Qamar, Aatif Kamal, Ahsan Rehman, “Telecommunication Subscribers' Churn Prediction Model Using Machine Learning”, IEEE International Conference on Digital Information Management (ICDIM), Eighth International Conference on, 2013

[14] Kiran Dahiya and Kanika Talwar, “Customer Churn Prediction in Telecommunication Industries using Data Mining Techniques- A Review”,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, 2015.

[15] Nisha Saini and Monika, ”Churn Prediction in Telecommunication Using Classification Techniques Based on Data Mining: A Survey”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 3, March 2015.

[16] Dr. M. Balasubramaniam, M.Selvarani, “Churn Prediction In Mobile Telecom System Using Data Mining Techniques”, International Journal of Scientific and Research Publications, Volume 4, Issue 4, April 2014.

[17] Georges D. Olle Olle and Shuqin Cai , ”A Hybrid Churn Prediction Model in Mobile Telecommunication Industry”, International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 4, No. 1,February 2014.

[18] Dataset https://www.ibm.com/communities/analytics/watson-analytics-blog/predictive-insights-in-the-telco-customer-churn-data-set

[19] Chih-Fong Tsai and Yu-Hsin Lu. "Customer churn prediction by hybrid neural networks." Expert Systems with Applications, vol. 36, no. 10, pp. 12547-12553, 2009.

[20] Amjad Hudaib, Reham Dannoun, Osama Harfoushi,Ruba Obiedat, and HossamFaris. "Hybrid Data Mining Models for Predicting Customer Churn." International Journal of Communications, Network and System Sciences, vol. 8, no. 05, pp. 91, 2015.

[21] Hsiu-Yu Liao, Kuan-Yu Chen, Duen-Ren Liu, and Yi-Ling Chiu. "Customer Churn Prediction in Virtual Worlds." In 2015 IIAI 4th International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 115-120. IEEE, 2015.

[22] Aimée Backiel, Bart Baesens, and GerdaClaeskens. "Predicting time-to-churn of prepaid mobile telephone customers using social network analysis." Journal of the Operational Research Society, pp.1-11, 2016.

[23] Pretam Jayaswal, BakshiRohit Prasad, DivyaTomar, and SonaliAgarwal. "An Ensemble Approach for Efficient Churn Prediction in Telecom Industry." International Journal of Database Theory and Application, vol. 9, no. 8, pp. 211-232, 2016.

[24] Umayaparvathi, V., and K. Iyakutti. "A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics." International Research Journal of Engineering and Technology (IRJET), vol 04, no.4, pp. 1065-1070, 2016.

[25] Chih-Ping Wei and I-Tang Chiu. "Turning telecommunications call details to churn prediction: a data mining approach." Expert systems with applications, vol. 23, no. 2, pp. 103-112, 2002.

[26] Shin-Yuan Hung, David C. Yen, and Hsiu-Yu Wang. "Applying data mining to telecom churn management." Expert Systems with Applications, vol. 31, no. 3, pp. 515-524, 2006.

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

[28] Ali Tamaddoni Jahromi, Stanislav Stakhovych, and Michael Ewing. "Managing B2B customer churn, retention and profitability." Industrial Marketing Management, vol. 43, no. 7, pp. 1258-1268, 2014.

[29] Dirk Van den Poel, and Bart Lariviere. "Customer attrition analysis for financial services using proportional hazard models." European journal of operational research, vol. 157, no. 1, pp. 196-217, 2004.

[30] Chih-Ping Wei and I-Tang Chiu. "Turning telecommunications call details to churn prediction: a data mining approach." Expert systems with applications, vol. 23, no. 2, pp. 103-112, 2002.

[31] Shin-Yuan Hung, David C. Yen, and Hsiu-Yu Wang. "Applying data mining to telecom churn management." Expert Systems with Applications, vol. 31, no. 3, pp. 515-524, 2006.

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

[33] Daihwan Min and Lili Wan. "Switching factors of mobile customers in Korea." Journal of Service Science, vol. 1, no. 1, 105-120, 2009.

[34] Wouter Verbeke, David Martens, Christophe Mues, and Bart Baesens. "Building comprehensible customer churn prediction models with advanced rule induction techniques." Expert Systems with Applications, vol. 38, no. 3, pp. 2354-2364, 2011.

[35] T. Sumathi. “Churn Prediction on Huge Sparse Telecom Data Using Meta-heuristic.” International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), vol.5, no.7, pp.574-577, 2016

[36] DongBack Seo, C. Ranganathan, and YairBabad. "Twolevel model of customer retention in the US mobile telecommunications service market." Telecommunications Policy, vol. 32, no. 3, pp. 182-196, 2008.

[37] Chih-Fong Tsai and Yu-Hsin Lu. "Customer churn prediction by hybrid neural networks." Expert Systems with Applications, vol. 36, no. 10, pp. 12547-12553, 2009.

[38] Amjad Hudaib, Reham Dannoun, Osama Harfoushi, Ruba Obiedat, and HossamFaris. "Hybrid Data Mining Models for Predicting Customer Churn." International Journal of Communications, Network and System Sciences, vol. 8, no. 05, pp. 91, 2015.

[39] Hsiu-Yu Liao, Kuan-Yu Chen, Duen-Ren Liu, and Yi- Ling Chiu. "Customer Churn Prediction in Virtual Worlds." In 2015 IIAI 4th International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 115-120. IEEE, 2015

[40] https://en.m.wikipedia.org/wiki/system_architecture

[41] http://blog.mahindracomviva.com/efficient-ways-for-customer-churn-analysis-in-telecom-sector/

[42] https://www.analyticsvidhya.com/blog/tag/naive-bayes

[43] https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/

[44] https://www.r-bloggers.com/random-forests-in-r/

[45] https://www.analyticsvidhya.com/blog/2017/06/getting-started-with-deep-learning-using-keras-in-r/

[46] https://cran.r-project.org/web/packages/kerasR/vignettes/introduction.html

Appendix A:

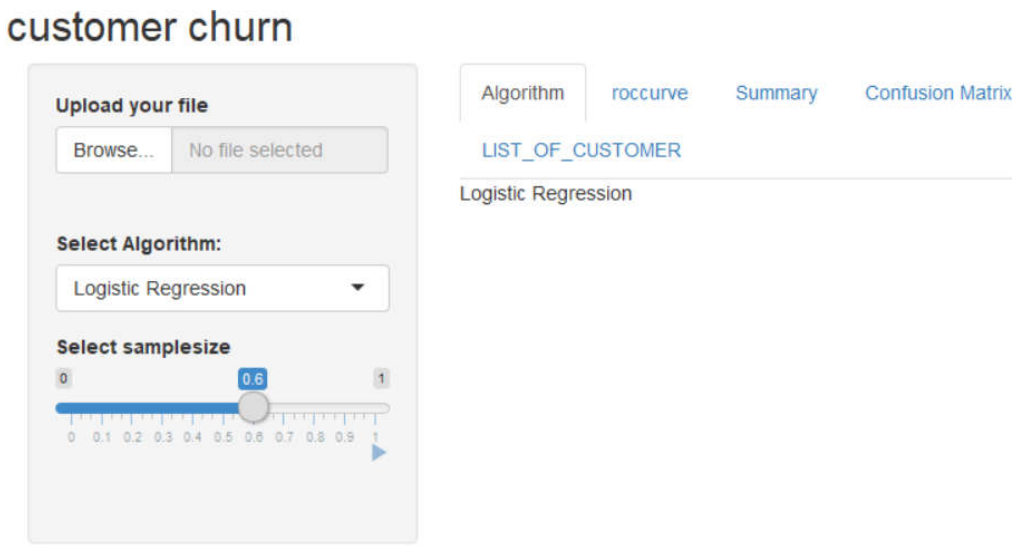


Fig A.1 Homepage

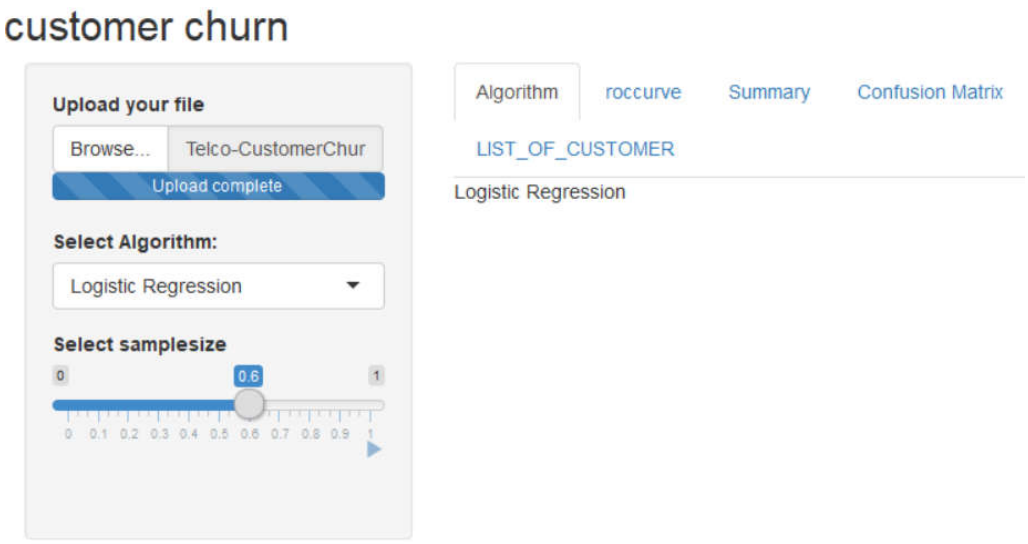


Fig A.2 Loading file

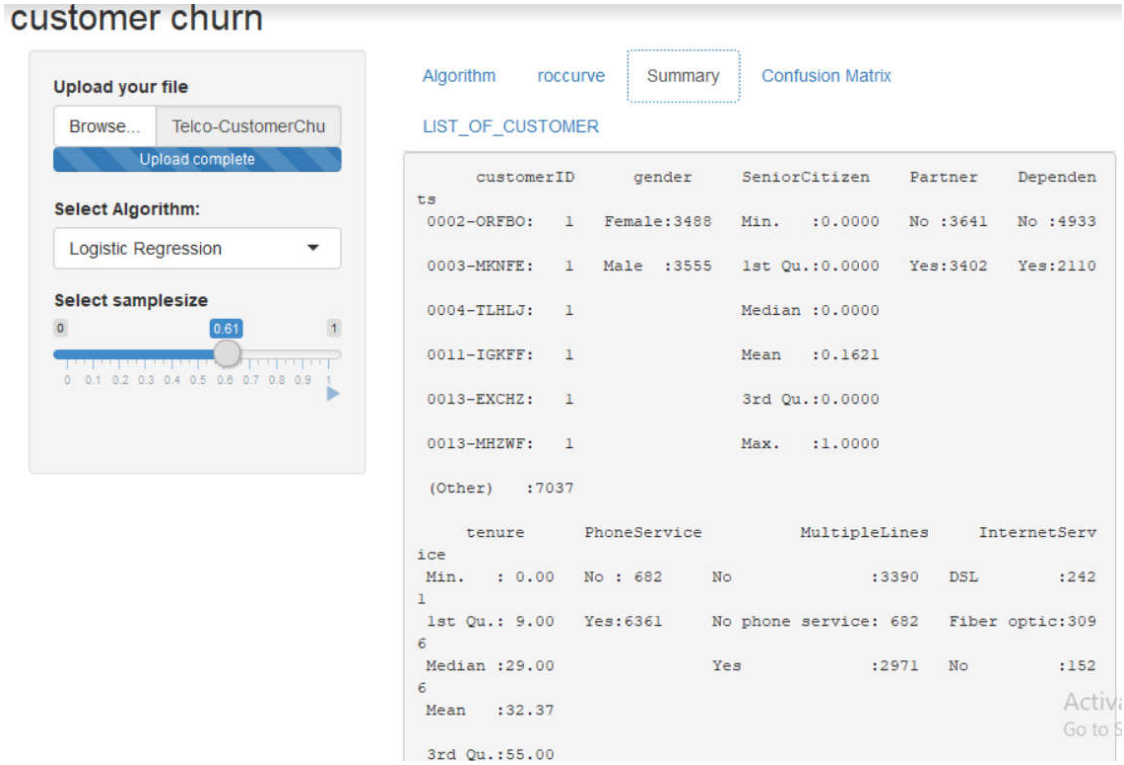


Fig A.3 Data Summary

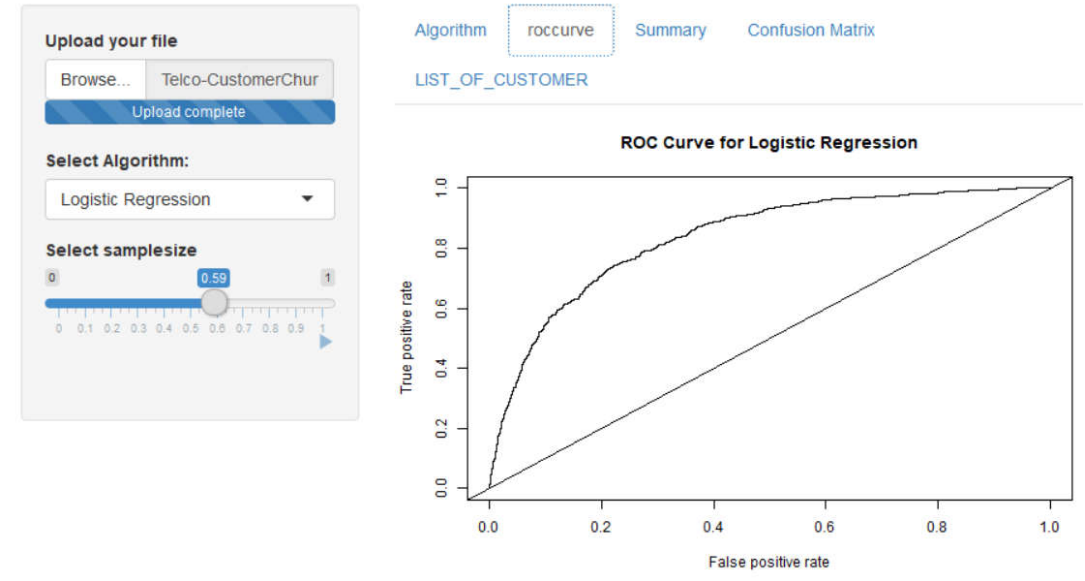


Fig A.4 ROC curve for Logistic Regression

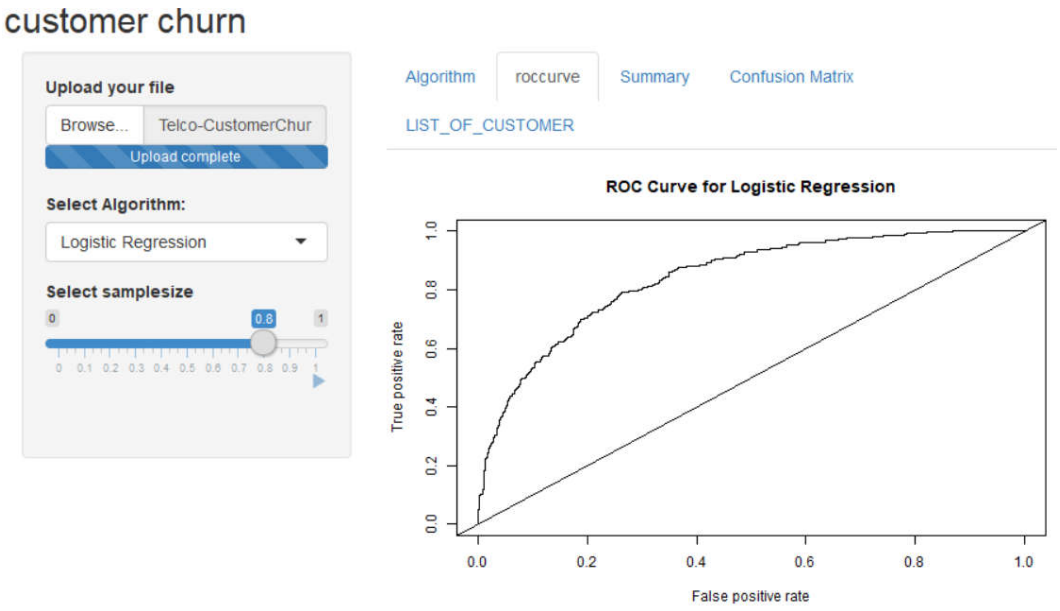
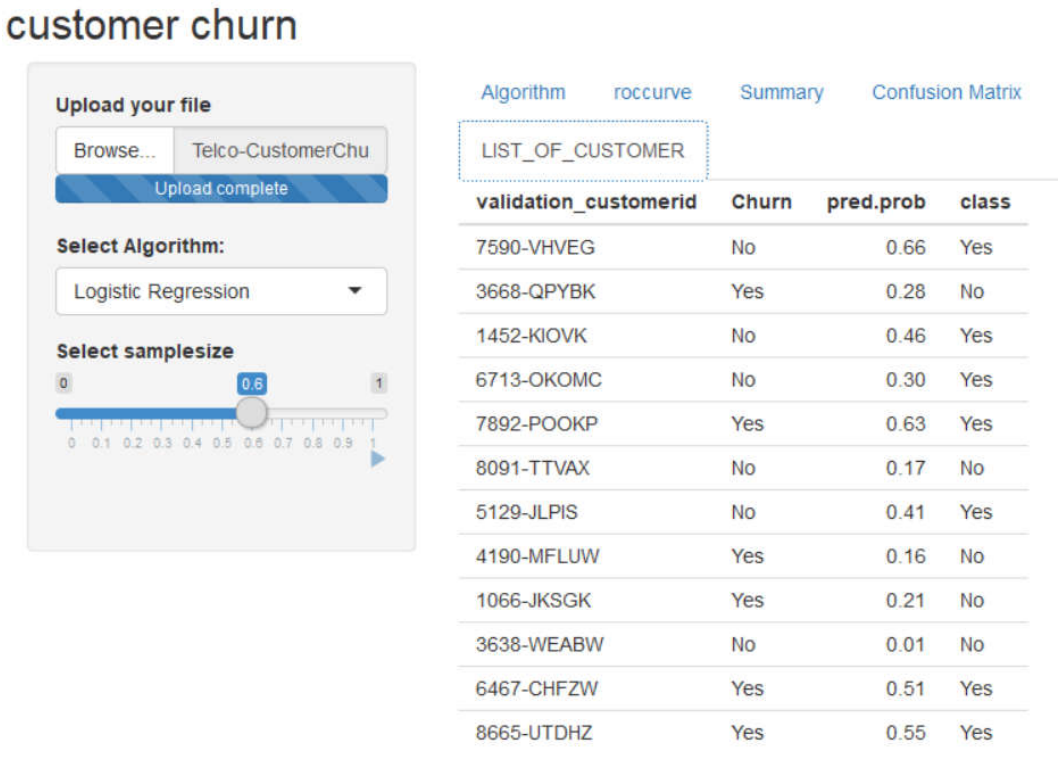


Fig A.5 ROC curve for Logistic Regression with different splitting criteria



validation_customerid	Churn	pred.prob	class
7590-VHVEG	No	0.66	Yes
3668-QPYBK	Yes	0.28	No
1452-KIOVK	No	0.46	Yes
6713-OKOMC	No	0.30	Yes
7892-POOKP	Yes	0.63	Yes
8091-TTVAX	No	0.17	No
5129-JLPIS	No	0.41	Yes
4190-MFLUW	Yes	0.16	No
1066-JKSGK	Yes	0.21	No
3638-WEABW	No	0.01	No
6467-CHFZW	Yes	0.51	Yes
8665-UTDHZ	Yes	0.55	Yes

Fig A.6 Result of Logistic Regression

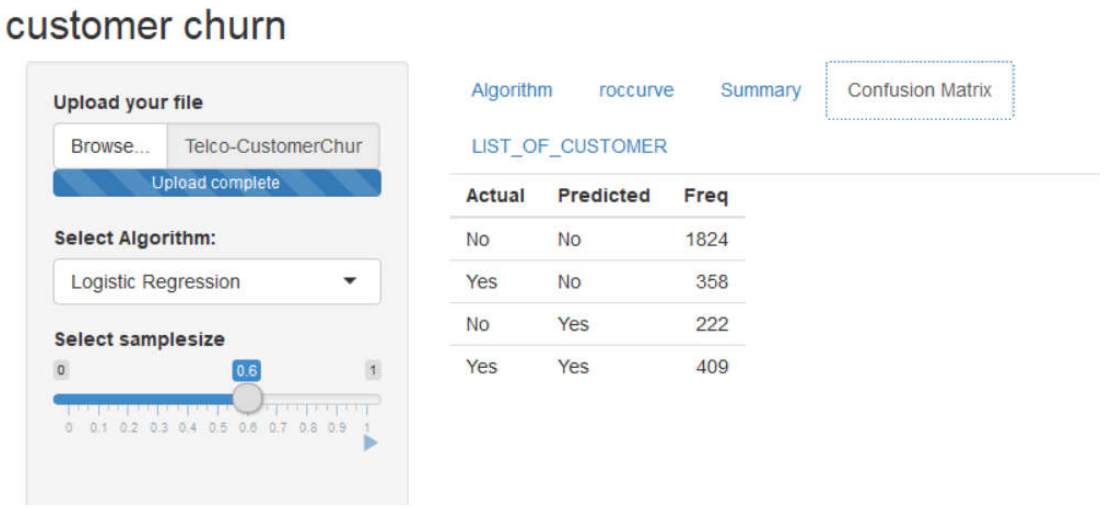


Fig A.7 Confusion Matrix for Logistic Regression

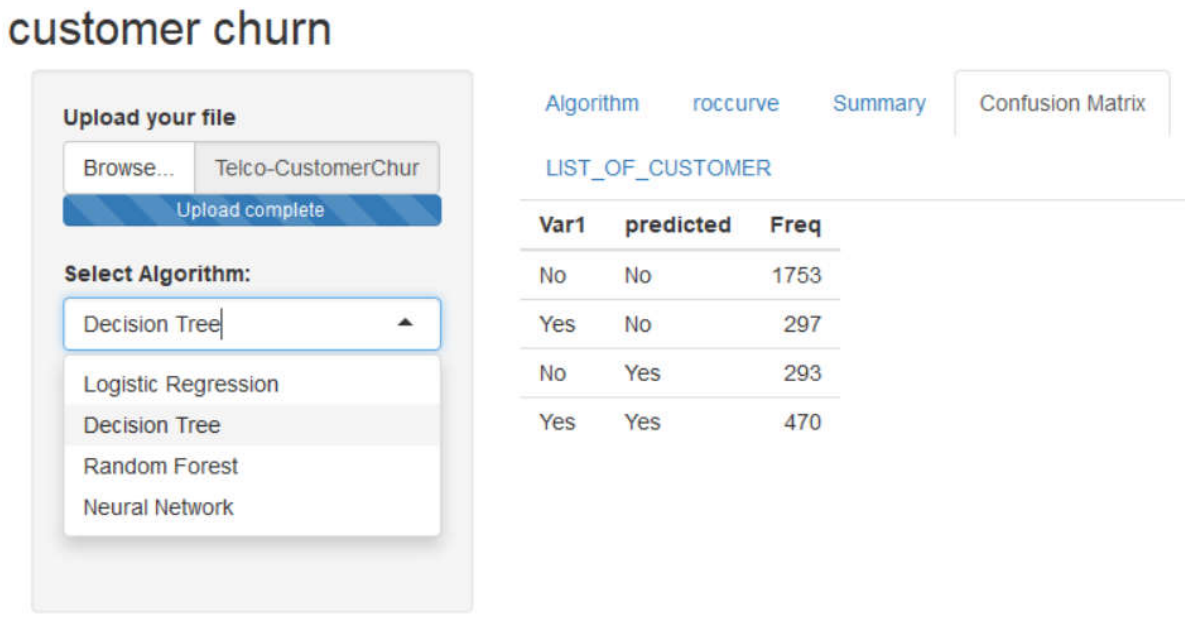


Fig A.8 Confusion Matrix for Decision Tree

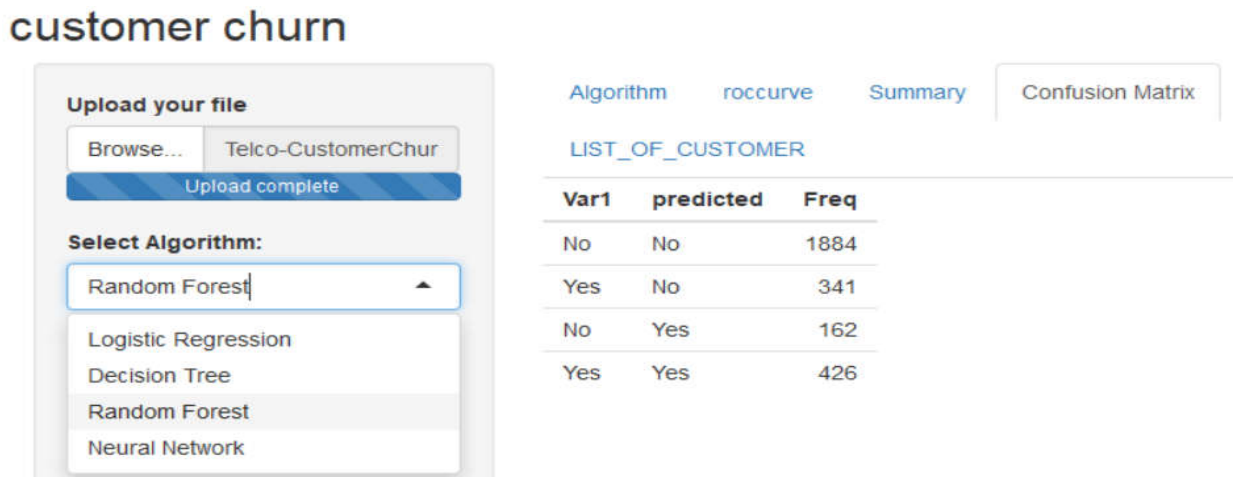


Fig A.9 Confusion Matrix for Random Forest

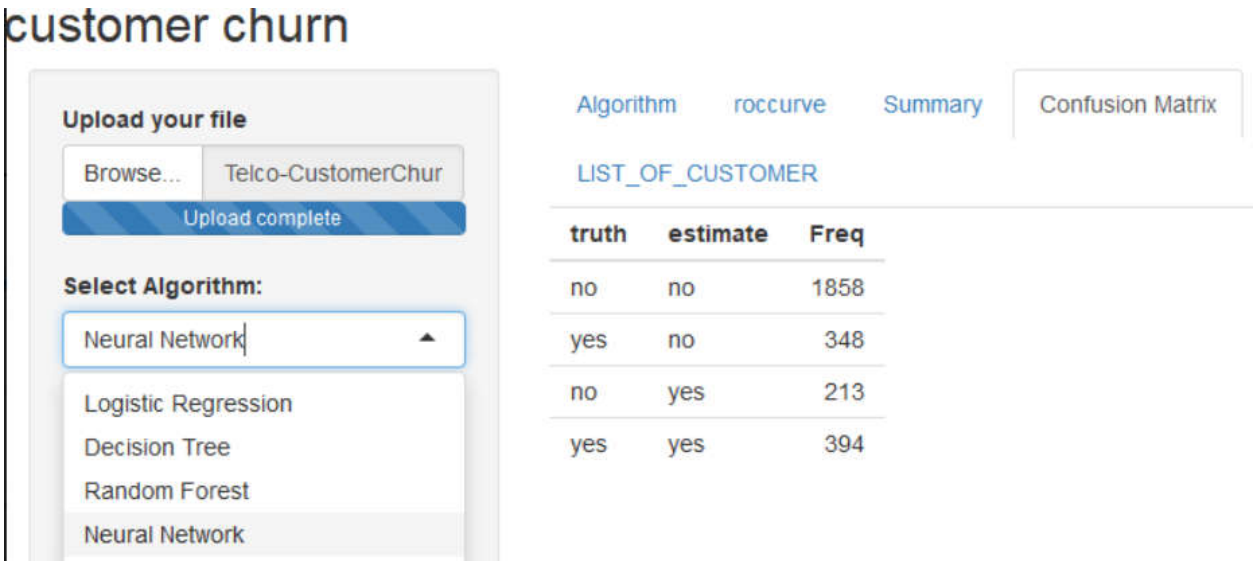


Fig A.10 Confusion Matrix for Neural Network