

RAMAIAH

Institute of Technology

**“Telecom Churn Prediction”**

**A report submitted in partial fulfillment of the requirements**

**Of**

**Project Based Learning Course**

**In**

**Sixth Semester**

**By**

|  |  |
| --- | --- |
| **1MS16IS012** | **Ankit Anand** |
| **1MS16IS013** | **Ansh Arora** |
| **1MS17IS406** | **Jakku Yamuna** |

**Under the guidance of**

**Lincy Meera Mathews**

Associate Professor

Dept. of ISE, RIT

**DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING**

**RAMAIAH INSTITUTE OF TECHNOLOGY**

(AUTONOMOUS INSTITUTE AFFILIATED TO VTU)

VIDYA SOUDHA

M. S. RAMAIAH NAGAR, M. S. R. I. T. POST, BANGALORE – 560054

**2018-2019**

**ACKNOWLEDGEMENTS**

We would like to express our deepest appreciation to all those who provided us the possibility to complete our task.  We appreciate our project guide, **Lincy Meera Mathews**, Assistant Professor, Dept. of ISE, whose guidance and contribution in stimulating suggestions and encouragement, helped us to coordinate our project in an efficient way.

We would like to express our gratitude to **Dr. Vijaya Kumar**, Head of Dept., ISE for his valuable encouragement.

We would like to express our special gratitude and thanks to faculties, library staff and laboratory staffs of ISE department for their extended long-term support and coordination.

We also appreciate our team members for their exquisite coordination and contribution in the completion of this project.

**ABSTRACT**

The telecommunication industry has a very tough competition with the fellow competitors in order to retain their customers, and has thus become one of the research fields in machine learning. In order to monitor the customers' churn behaviour closely and efficiently, we require a methodical churn prediction model.

For the last two decades, mobile communication has become one of the dominant medium of communication. In many countries, including the developed ones, the market is saturated to the extent that each new customer must be won over from the competitors. At the same time, public policies and standardization of mobile communication now allow customers to easily switch over from one carrier to another, resulting in a fluid market.

Since the cost of winning a new customer is far greater than the cost of retaining an existing one, mobile carriers have now shifted their focus from customer acquisition to customer retention. As a result, churn prediction has emerged as the most crucial Business Intelligence (BI) application that aims at identifying customers who are about to transfer their business to a competitor i.e. to churn. This project aims to present commonly used machine learning techniques for the identification of customers who are about to churn. Based on historical data, these methods try to find patterns which can identify possible churners. Some of the well-known algorithms used during this research are **Random Forests, SVM (Support Vector Machines) and XGBoost**. We also discuss the use of re-sampling method in order to solve the problem of class imbalance.

The main obstructions in achieving the desired results and performances in a classifier are due to the large feature space and imbalanced class distribution. In this project, we explore various implications of Synthetic Minority Over-sampling Technique (SMOTE) to reduce the imbalance in data in collaboration with the help of few feature reduction techniques such as Co-relation feature extraction method.

Prediction the performance of the classifiers is evaluated through measures such as Area Under the Curve (AUC), accuracy, precision and recall. We finally conclude with the help of simulations that method proposed based on SMOTE, co-relation, and ensembling, performs quite well in order to predict churners as compared when we were simply applying learners on the unrefined dataset. Therefore, this methodology can be helpful for the telecommunication industry to predict churn.

**TABLE OF CONTENTS**

**1. Introduction 6-8**

1.1 Problem Statement 6

1.2 Motivation 7

1.3 Scope and Objective 7

1.4 Proposed Model 8

1.5 Organization of report 8

**2. Literature Survey 9-10**

2.1Summary 10

**3. System Analysis and Design 11-14**

3.1 System Architecture 11

3.2 Feature Extraction 12

3.3 Performance check measures 13

3.4 Techniques Applied 13-14

3.5 Summary 14

**4. Modelling and Implementation 15-21**

4.1 Algorithms 15-19

4.2 Use Case Diagram 20

4.3 Sequence Diagram 21

4.4 Collaboration Diagram 21

**5. Testing, Results and Discussion 22-27**

5.1 Testing 22

5.2 Results and Discussion 22-27

**6. Conclusion and Future Works 28-29**

**Reference 30**

**List of Figures**

Figure 1 Architecture of the proposed model

Figure 2 Use Case Diagram Figure 3 Sequence diagram Figure 4 Collaboration diagram

**List of Tables**

Table 1 Performance metrics for algorithms used

**List of graphs**

Graph1: The number of customer churn Vs different parameters.(a) Gender (b) Senior Citizen (c) Partner (d) Dependents

Graph 2: The number of customer churn Vs different parameters. (a) Phone Service (b) Multiple Lines (c) Internet service (d) Online security

Graph 3: The number of customer churn Vs different parameters. (a) Online Backup (b) Device Protection (c) Tech Support (d) Streaming TV

Graph 4: The number of customer churn Vs different parameters. (a) Streaming Movies (b) Contract (c) Paperless Billing (d) Payment Method

Graph 5: Customer churn based on tenure (in months)

Graph 6: Customer churn based on Monthly charges (in rupees)

Graph 7: Graph depicting the correlation between the features

Graph 8: Graph depicting the important features for random forest in descending order

Graph 9: Graph depicting the important features for xgboost in descending order

Graph 10: The accuracy measure for each algorithm (a) XBG (b) SVM (c) Random forest

**Chapter 1**

# Introduction

The telecommunications sector has become one of the main industries in most of the developed countries. Factors such as increasing number of operators and technical progress in the field has resulted in a rise in the level of competition. It has become very difficult to survive in this era of competition as a result of which various companies are working hard to remain in the competition depending on multiple strategies. Three main strategies have been proposed for such purpose:

(1) acquire new customers,

(2) up sell the existing customers, and

(3) increase the retention period of customers.

While comparing these strategies, it has been found that the third strategy is the most profitable strategy, thus proving that retaining an existing customer costs much less than acquiring a new customer. In order to apply the third strategy, companies have to decrease the potential of customer’s churn, known as “the customer movement from one provider to another”

**Customer Churn**

When a customer shifts from one service provider to other competitor providing better services in the market, then we call it as customer churning or the customers are said to have churned the former service provider. It is a key challenge in high competitive markets, which is highly and carefully observed in telecom sector. Customers’ churn is a major concern to be considered in service sectors which have high competitive services. If predicting the customers who are most likely to leave the company is done at an early phase, then it will represent large additional revenue source potentially. Many machine learning techniques is highly efficient for predicting the particular situation. This technique is applied through learning from historical data.

* 1. **Problem Statement**

Churn (loss of customers to competition) is a big problem for telecom companies where the company finds it difficult and much expensive in acquiring a new customer than to keep an existing one from leaving. This report is about enabling churn reduction in the telecom industry using machine learning techniques.

Many of the telecom companies suffer from churn. The churn rate has a very strong impact on the life time value of the customer because it affects many factors such as the length of service and the future revenue of the company. For example if a company has 25% churn rate then the average customer lifetime is 4 years; similarly a company with a churn rate of 50%, has an average customer lifetime of 2 years. It has been estimated that 75 percent of subscribers coming up with a new connections every year are coming from another wireless provider, which means they are churners. Telecom companies spend hundreds of dollars in acquiring a new customer and when that customer leaves, the company loses resources spent to acquire that customer as well as future revenue from that customer. Churn erodes profitability.

* 1. **Motivation**

Many approaches were applied to predict churn in telecom companies. Most of these approaches have used machine learning.

Telecom companies have used two approaches to address churn - (a) Untargeted approach and (b) Targeted approach. The untargeted approach relies on superior product and mass advertising to increase brand loyalty and thus retain customers. The targeted approach relies on identifying customers who are likely to churn, and provide suitable intervention to encourage them to stay. Gavril et al. presented an advanced methodology of data mining to predict churn for prepaid customers using dataset for call details of 3333 customers with 21 features, and a dependent churn parameter with two values: Yes/No.

Idris proposed an approach based on genetic programming with Ada-Boost to model the churn problem in telecommunications.

Burez and Van den Poel studied the problem of unbalance datasets in churn prediction models and compared performance of Random Sampling, Advanced Under-Sampling, Gradient Boosting Model, and Weighted Random Forests.

**1.3 Scope and Objectives:**

The importance of this type of research is to help companies earn more profit. It can be applied with the following aim-:

1 To define and explain the related terms in churn prediction model.

2. To propose the novel framework that uses churn prediction in Data Handling.

3. To evaluate the techniques used in the churn prediction.

4. To determine the extent of Customer Churn.

5. To determine the causes of Customer Churn.

6. To examine the effects of Customer Churn on the telecom industry.

7. To find out features that can be utilized in order to build a predictive model for customer

churn in mobile telephony industry.

* 1. **Proposed Model**

In this work, we explore the implication of Synthetic Minority Over-sampling Technique (SMOTE) to reduce the imbalance in data in collaboration with different feature reduction techniques such as Co-relation feature extraction. We have used Random forest, SVM and XGBoost technique for analyzing the performance on the dataset.

Prediction performance of the classifiers is evaluated through measures such as Area Under the Curve (AUC), accuracy, precision and recall. We finally conclude with the help of various simulations that our proposed method which is based on SMOTE, co-relation, and ensembling, performs well for predicting churners as against simply applying learners on the unrefined dataset.

* 1. **Organization of the Study**

This project report starts with an introduction about churn management. The problem statement, objectives, significance and scope to the study are presented in Chapter One. Few Literatures were reviewed in Chapter Two. The concept of customer, what customer churn is all about, causes, effects and its managements were reviewed. The machine learning techniques which were used to develop the model was also reviewed together. The methodology, which describes the procedures, tools and techniques used in the project are all presented in Chapter Three. Chapter Four contains the analyzed data. Answers to many objectives are presented in this chapter with the summary of findings and conclusions in further chapters.

**Chapter 2**

# Literature Review

Writing survey is basically writing down the past work which is important for the theme of the project. This section considers the past research work from which certain facts can be drawn. This section consists of a group involving Theoretical information and Methodologies identified with the help of imbalanced information.

[Anuj Sharma](https://arxiv.org/search/cs?searchtype=author&query=Sharma%2C+A), [Dr. Prabin Kumar Panigrahi](https://arxiv.org/search/cs?searchtype=author&query=Panigrahi%2C+D+P+K), Marketing literature states that it costs more when we engage a new customer at the place of retaining an existing loyal customer. The ability to correctly predict customer churn is necessary as churn management is an important activity for companies to retain loyal customers. As the cellular network services market becoming more competitive, customer churn management has become a crucial task for mobile communication operators. This paper proposes and presents a neural network based approach for predicting customer churn in subscription of cellular wireless services. The results of experiments indicate that neural network based approach can be used to predict customer churn.

Hadden et al. (2007), stated that mobile telecommunication has become one of the dominant communication medium over the last few decades. In many countries, especially developed ones, the market has reached a degree of saturation where each new customer must be won over from the competitors. At the same time, public regulations and the standardization of mobile communication now allow customers to easily move from one carrier to another, resulting in a very fluid market. Since the cost of winning a new customer is far greater than the cost of preserving an existing one, mobile carriers have been shifting considerable attention from customer acquisition to customer retention (Fildes, 2002).

As a result, churn prediction has become a crucial mobile Business Intelligence (BI) application which aims at identifying customers who will transfer their business to another competitor.

Song et al. (2007) found out that a good churn prediction system should provide a sufficient long horizon forecast for predictions and pinpointing the potential churners successfully. Once a churner is identified, the department for retention usually makes a contact and, if the customer is established or found to be a churn risk, then it takes appropriate measures for preserving the business. Thus, a sufficient long forecast horizon is an advantage as the customer is to actually make the churn decision, the easier it is to prevent the particular decision at a significantly lower cost. So, efforts for retention are given limited resources and therefore only a small fraction of the subscriber list or pool can be contacted at any particular time.

Web Chin-Ping Wei and I-Tang Chiu proposed a churn prediction technique for retention of customers. They used decision tree approach C4.5 on customer call details. Yi-Fan wang, Ding-An chlang and Mei-Hua Hsu designed and discussed a Recommender system for customer churn by using a decision tree algorithm. Data that has been used for the analysis has covered over about 60,000 transactions and of more than 4000 members, over a period of three months.

Jadhav and Pawar made a decision support system using machine learning technique. The churn behaviour of customers whether they will churn or not is predicted by this technique. The authors have used Back propagation algorithm on a customer billing data.

Tomas Philip Rúnarsson Olafur Magnússon Birgis Hrafnkelsson designed a churn prediction model that gives the probabilities that customers will churn in the near future as the output. In this particular work, the training data is being used for building classifiers by using machine learning methods.

N. Kamalraj and A. Malathi carried out their research for a better understanding of churn prediction using machine learning techniques. Telecommunication industry can further use the approach for retention of customer activities.

**2.1 Summary:**

The summarized Literature review explains about what is imbalanced data, methods to deal with the data in different area. We also describe the approaches to deal with imbalanced issue and some future directions.

**Chapter 3**

# System Analysis and Design

**3.1 System Architecture**

Telecom.csv

Data Preprocessing

Split data

XG Boost

SVM

Random forest

Train model

Train Model

Train model

Score model

Score Model

Score model

Evaluate model

**Figure 1: Architecture of the proposed model**

In the above figure, we have a telecom dataset named as dataset.csv which is split into training and testing data set. Training methods are applied on this split dataset using various training models which gives value of performance measures such as accuracy, precision, recall etc. which is used for evaluation of the model whether there is improvement in the model or not.

**3.1.1 Original Set:**

In the used dataset we have 21 features for 7043 customers. In the dataset the variable churn has already been defined.

**3.2 Feature Extraction**

The features extracted for this project are explained below:

**Mean:**

The mean is defined as the average of all the values of a particular feature. It can be calculated by the formula given below:

Ā=(1/n)∑Ai

where n is the number of values.

**Standard Deviation:**

This normalizes the facts and tell us how far our data is from the mean value. Larger the value of standard deviation, farther the mean and vice versa. This is calculated using the formula:

SD=√∑(x-xi)2 /n

**Correlation:**

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. The main result of a correlation is called the **correlation coefficient** (or "r"). It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related.

**3.3 Performance Check Measures**

In this project we have used performance measures such as AUC-ROC, accuracy, recall and precision. The following elements were used to calculate the above measures: TP (true positive), TN (true negative), FP (false positive) and FN (false negative).

* TP=positive values correctly recognized as positive.
* TN=negative values correctly recognized as negative.
* FP=negative values recognized as positive.
* FN=positive values recognized as negative.

**-Recall**

It is the ratio of correctly predicted positive values out of all positive samples. It is also known as True Positive Rate and is given by:

Recall=TP/(TP+FN)

**-Precision**

It is the ratio of correctly identified positive samples as positive out of all positive predictions. It is also called as positive prediction value and is given by:

Precision=TP/(TP+FP)

**-Area under curve(AUC)-Receiver Operating Characteristics(ROC):**

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.

**3.4 Techniques Applied**

**1. Features-** This step involves considering the feature set involved in the selected dataset which can be used for prediction of the desired feature (churn in this case).

**2. Method:** This includes the various methods which are being used for the accomplishment of the project. This project involves techniques such as Random forest, SVM, XGBoost, Smote etc.

**3. Performance Check:** This step involves the checking and validation of the technique applied on the selected dataset for the performance of the model. It is done with the help of various performance check metrics such as accuracy, precision, recall, AUC.

**4.Evaluation and improvement:** This step involves the evaluation of the various metrics described in the above step and using their results for further improvement in the metrics selected by applying other techniques.

**3.5 Summary**

The proposed architecture explains the use case diagram. The system architecture consists of four major modules: Training the model, categorizing, checking the performance and improvement.

**Chapter 4**

# Modelling and Implementation

Implementation is the process of converting the designed system architecture into working modules where it is made sure that all the functional and non-functional requirements are met.

**4.1 Algorithms**

**1. Random forest**

**Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it’s simplicity and the fact that it can be used for both classification and regression tasks. In this post, you are going to learn, how the random forest algorithm works and several other important things about it.**

**How it works:**

Random Forest is a supervised learning algorithm. Like you can already see from its name, it creates a forest and makes it somehow random. The forest it builds, is an ensemble of Decision Trees, most of the time trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

**Advantages and Disadvantages:**

* Like it is already mentioned, an advantage of random forest is that it can be used for both regression and classification tasks and that it’s easy to view the relative importance it assigns to the input features.
* Random Forest is also considered as very handy and easy to use algorithm, because it’s default hyperparameters often produce a good prediction result. The number of hyperparameters is also not that high and they are straightforward to understand.
* One of the big problems in machine learning is overfitting, but most of the time this won’t happen in random forest classifier. That’s because if there are enough trees in the forest, the classifier won’t overfit the model.
* The main limitation of Random Forest is that a large number of trees can make the algorithm to slow and ineffective for real-time predictions. In general, these algorithms are fast to train, but quite slow to create predictions once they are trained. A more accurate prediction requires more trees, which results in a slower model. In most real-world applications the random forest algorithm is fast enough, but there can certainly be situations where run-time performance is important and other approaches would be preferred.
* And of course Random Forest is a predictive modeling tool and not a descriptive tool. That means, if you are looking for a description of the relationships in your data, other approaches would be preferred.

**ALGORITHM:**

The pseudocode for random forest algorithm can split into two stages.

* **Random forest** creation pseudocode.
* **Pseudocode to perform prediction** from the created random forest classifier.

First, let’s begin with random forest creation pseudocode

#### Random Forest pseudocode:

1. Randomly select **“k”** features from total **“m”** features, where **k << m.**
2. Among the**“k”** features, calculate the node **“d”** using the best split point.
3. Split the node into **daughter nodes** using the **best split**.
4. Repeat **1 to 3** steps until “l” number of nodes has been reached.
5. Build forest by repeating steps **1 to 4** for “n” number times to create **“n” number of trees**.

The beginning of random forest algorithm starts with randomly selecting **“k”** features out of total **“m”** features. In the image, you can observe that we are randomly taking features and observations.

In the next stage, we are using the randomly selected **“k”** features to find the root node by using the [best split](https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/) approach.

The next stage, we will be calculating the daughter nodes using the same best split approach until we form the tree with a root node and having the target as the leaf node.

Finally, we repeat 1 to 4 stages to create **“n”** randomly created trees. This randomly created trees forms the **random forest.**

#### Random forest prediction pseudocode:

To perform prediction using the trained random forest algorithm uses the below pseudocode.

1. Takes the **test features** and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
2. Calculate the **votes** for each predicted target.
3. Consider the **high voted** predicted target as the **final prediction** from the random forest algorithm.

**2. SVM (Support Vector Machines)**

“Support Vector Machine” (SVM) is a supervised [machine learning algorithm](https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2?utm_source=blog&utm_medium=understandingsupportvectormachinearticle) which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_1.png)

Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line).

## How does it work?

**1.Identify the right hyper-plane:** Identify the right hyper-plane to classify the groups.

**2.Identify the right hyper-plane:** Here, maximizing the distances between nearest data point (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as **Margin**.

**3.Identify the right hyper-plane:** Use the rules as discussed in previous section to identify the right hyper-plane.

**4.Find the hyper-plane to segregate to classes:** Use quadratic or cubic hyperplanes instead of linear planes wherever necessary.

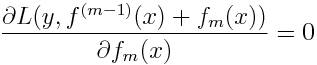
**Pros and Cons associated with SVM**

* **Pros:**
  + It works really well with clear margin of separation
  + It is effective in high dimensional spaces.
  + It is effective in cases where number of dimensions is greater than the number of samples.
  + It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
* **Cons:**
  + It doesn’t perform well, when we have large data set because the required training time is higher
  + It also doesn’t perform very well, when the data set has more noise i.e. target classes are overlapping
  + SVM doesn’t directly provide probability estimates, these are calculated using an expensive five-fold cross-validation. It is related SVC method of Python scikit-learn library.

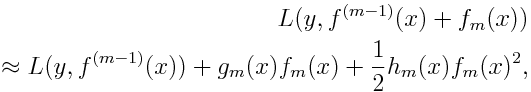
**3. XG Boost**

XGBoost, short for “Extreme Gradient Boosting”, was introduced by Chen in 2014. Since its introduction, XGBoost has become one of the most popular machine learning algorithm. In this post, we will dive deeply into the algorithm itself and try to figure out how XGBoost differs from the traditional boosting algorithms GBM.

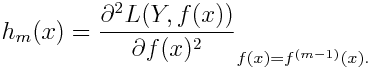
As mentioned in the previous post, GBM divides the optimization problem into two parts by first determining the direction of the step and then optimizing the step length. Different from GBM, XGBoost tries to determine the step directly by solving



for each x in the data set. By doing second-order Taylor expansion of the loss function around the current estimate f(m-1)(x), we get



where g\_m(x) is the gradient, same as the one in GBM, and h\_m(x) is the Hessian (second order derivative) at the current estimate:



**How it works?**

In XGBoost, we fit a model on the gradient of loss generated from the previous step. In XGBoost, we just modified our gradient boosting algorithm so that it works with any differentiable loss function.

**4.2 Use Case Diagram**

Feature extraction

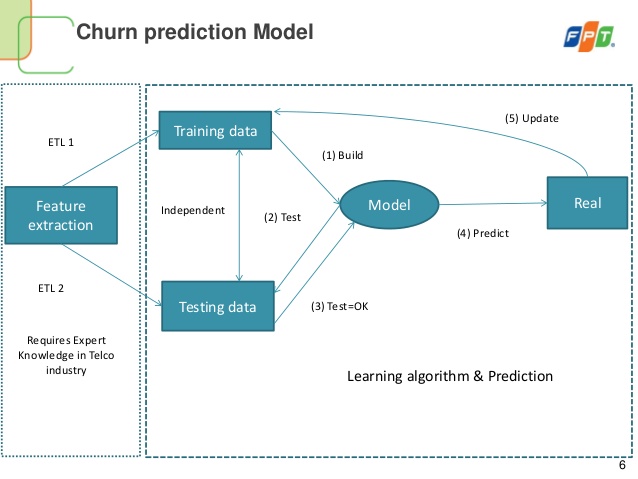
Techniques applied

Performance check

Evaluation and improvement

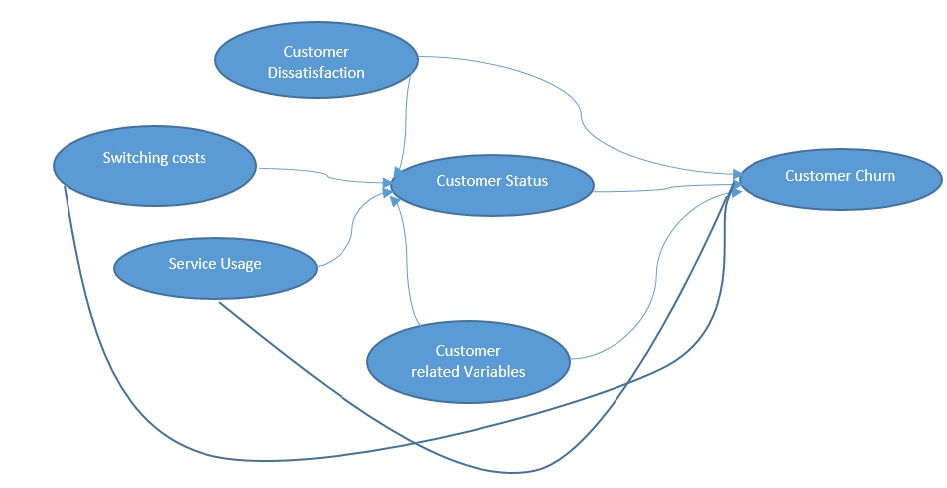
**Figure 2: Use case diagram for model selection**

**4.3 Sequence Diagram**



**Figure 3: Sequence diagram**

## 4.4 Collaboration Diagram



**Figure 4**: **Collaboration diagram**

**Chapter 5**

# Testing, Results and Discussion

## 5.1 Testing

Testing is one of the way of assessing the system which helps to detect the quality of the software, the methods we follow and evaluate the expected output and the actual input.

Verification and validation are the process in software testing where we verify various things and validate them. Some of the conditions are stated at the development phase which must be satisfied by the product is called verification. The requirement must be specified at the end of the development phase which assures the validation.

## 5.2 Results and Discussion

The results of the various applied algorithms are presented here with the help of various performance metrics such as accuracy, precision, recall and AUC-ROC in the form of a table.

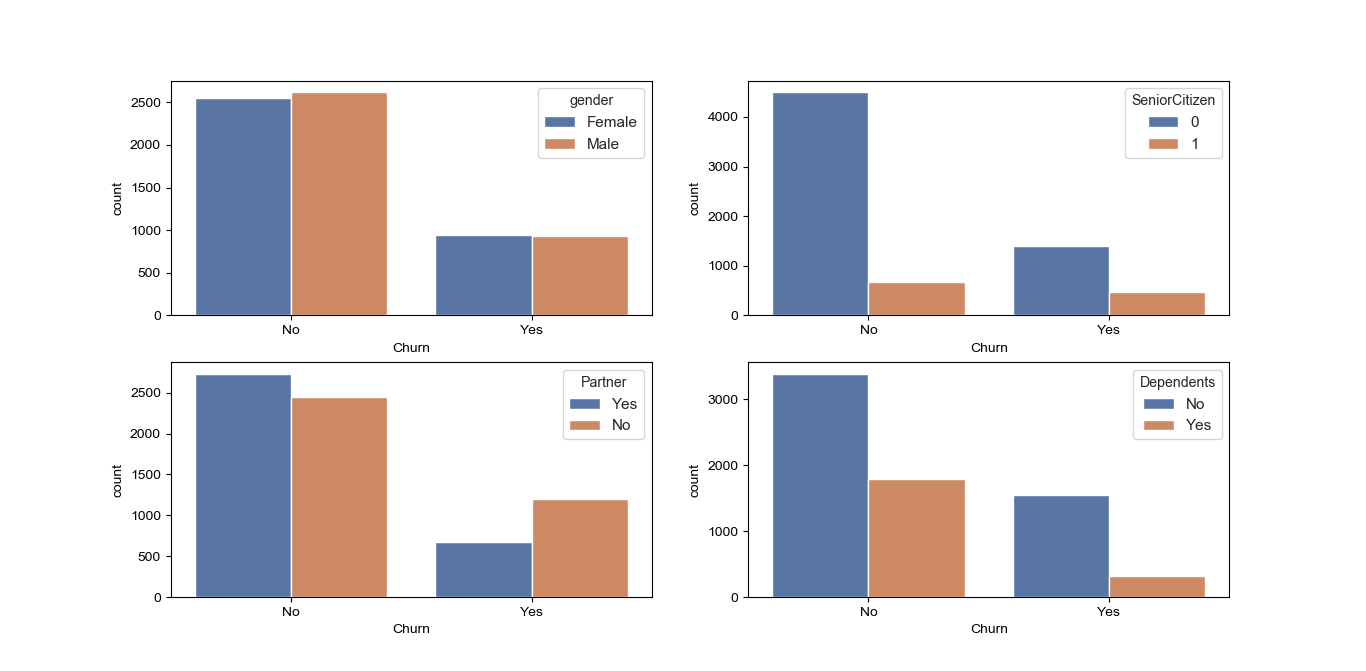
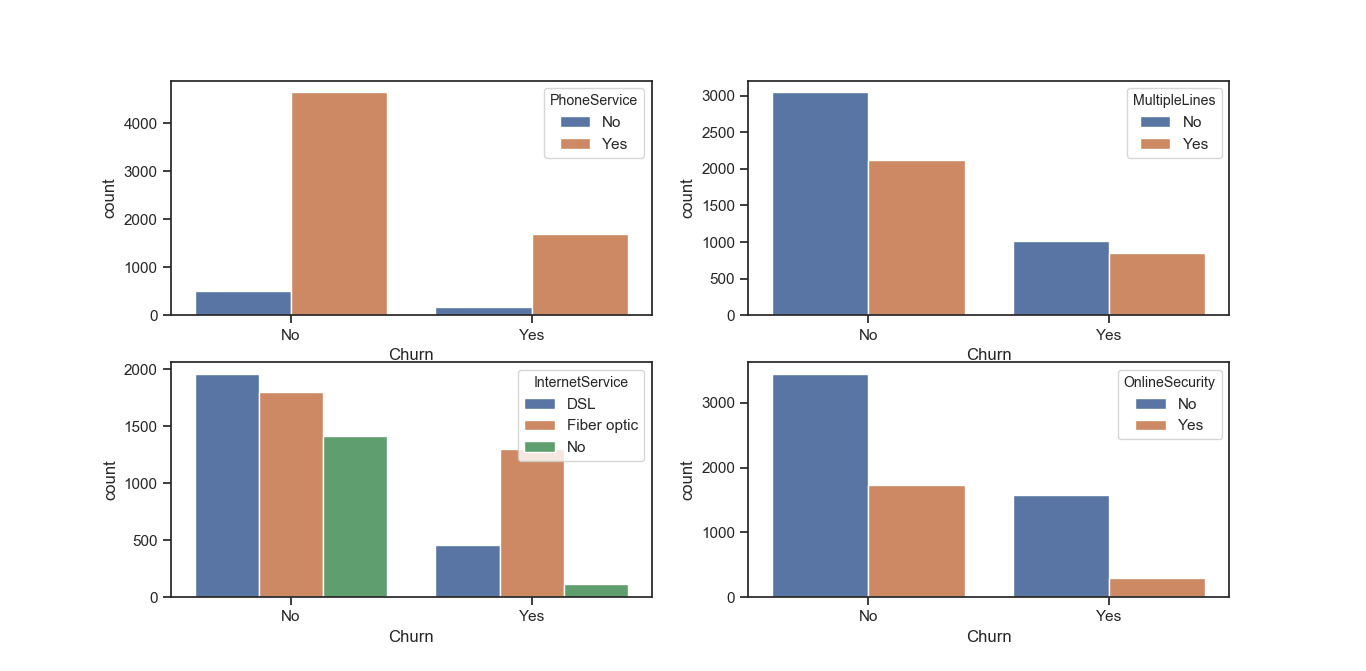
The table consists of the performance metrics as the columns and the various algorithms as the rows.

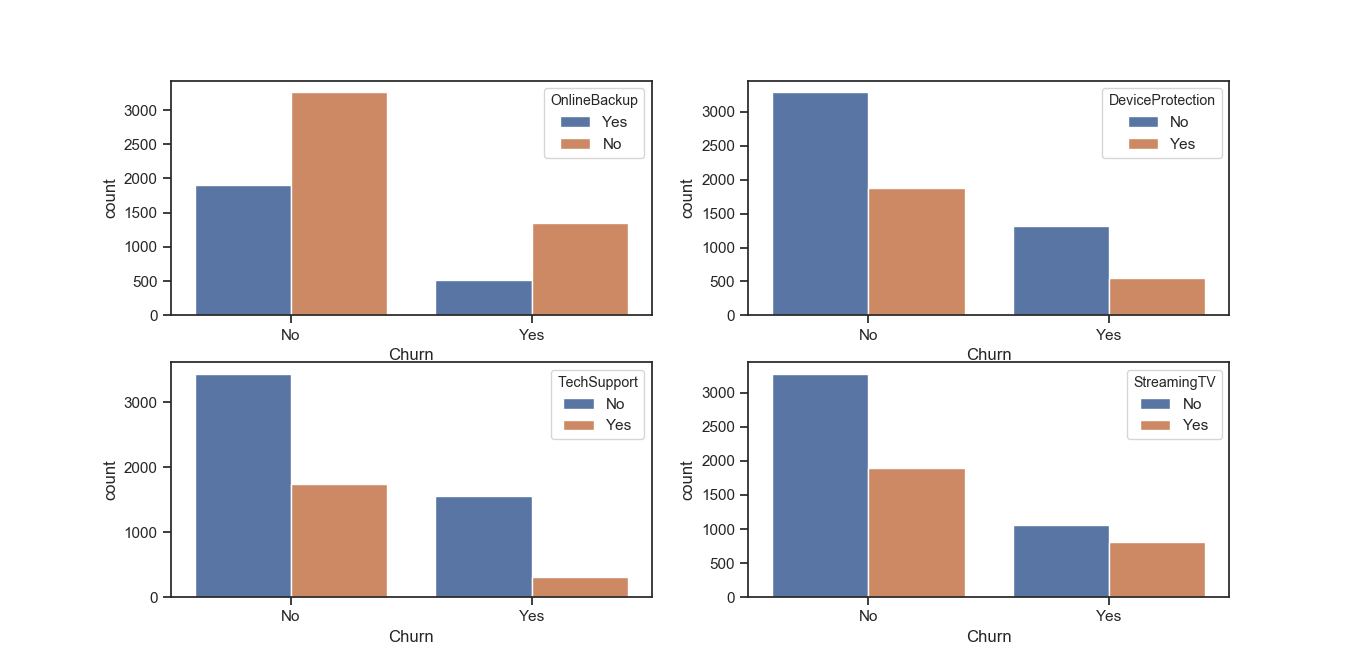
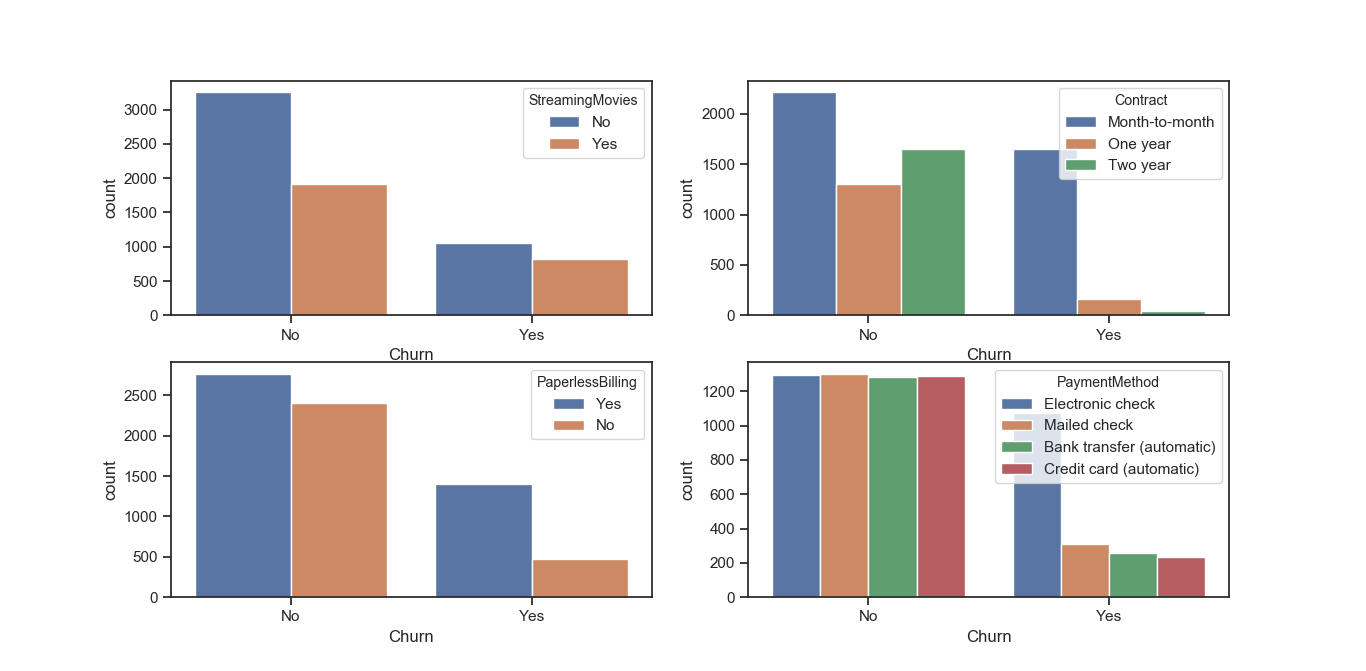
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithms** | **Accuracy**  **Value (%)** | **Precision value** | **Recall Value** | **AUC-ROC value** |
| **1.Random forests** | 79.524 | 0.600 | 0.582 | 0.725 |
| **2.SVM** | 77.217 | 0.537 | 0.740 | 0.761 |
| **3.XG-Boost** | 82.079 | 0.686 | 0.543 | 0.729 |

**Table 1: Performance metrics for algorithms used**

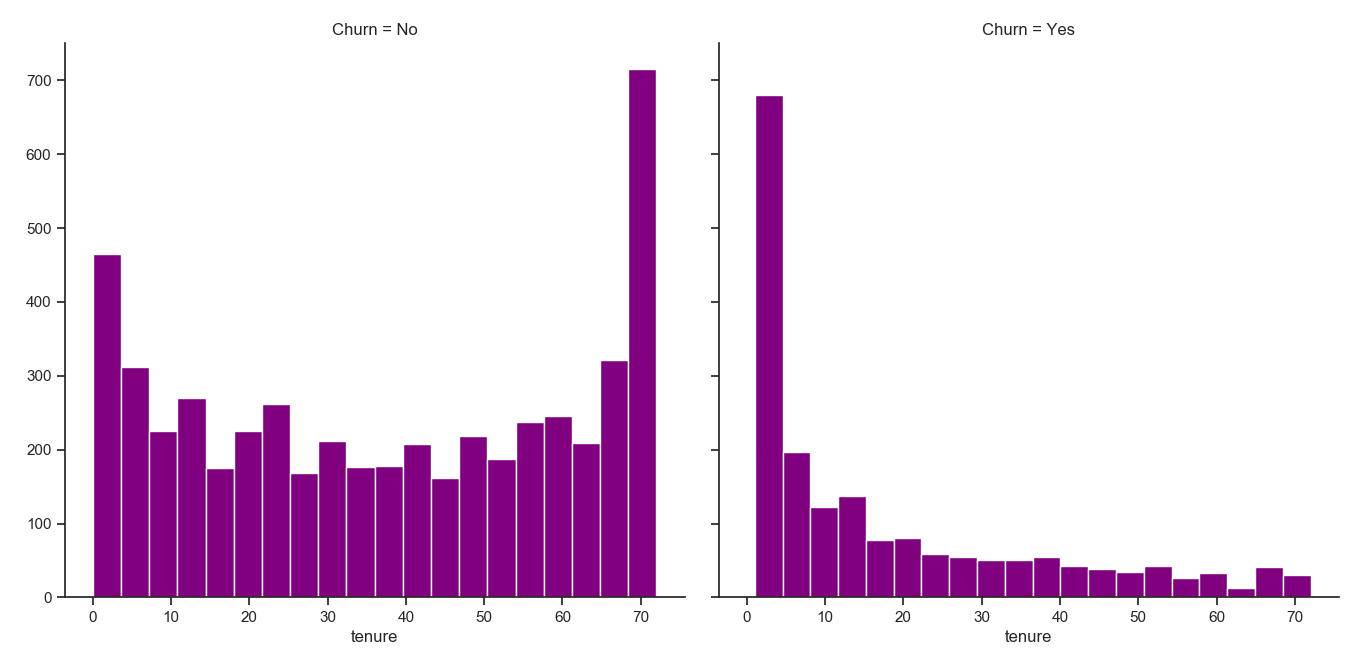
We also plotted various graphs by taking into account the various features in the selected dataset vs Churn. Also, the important features for random forest and xgboost is presented in the form of bar graphs. We plotted the bar graph to show the accuracy obtained from different algorithms.

* 1. Categorical Features Vs Churn
  2. Numerical Features Vs Churn
  3. Correlation Matrix
  4. Important Features
  5. Accuracy Comparison

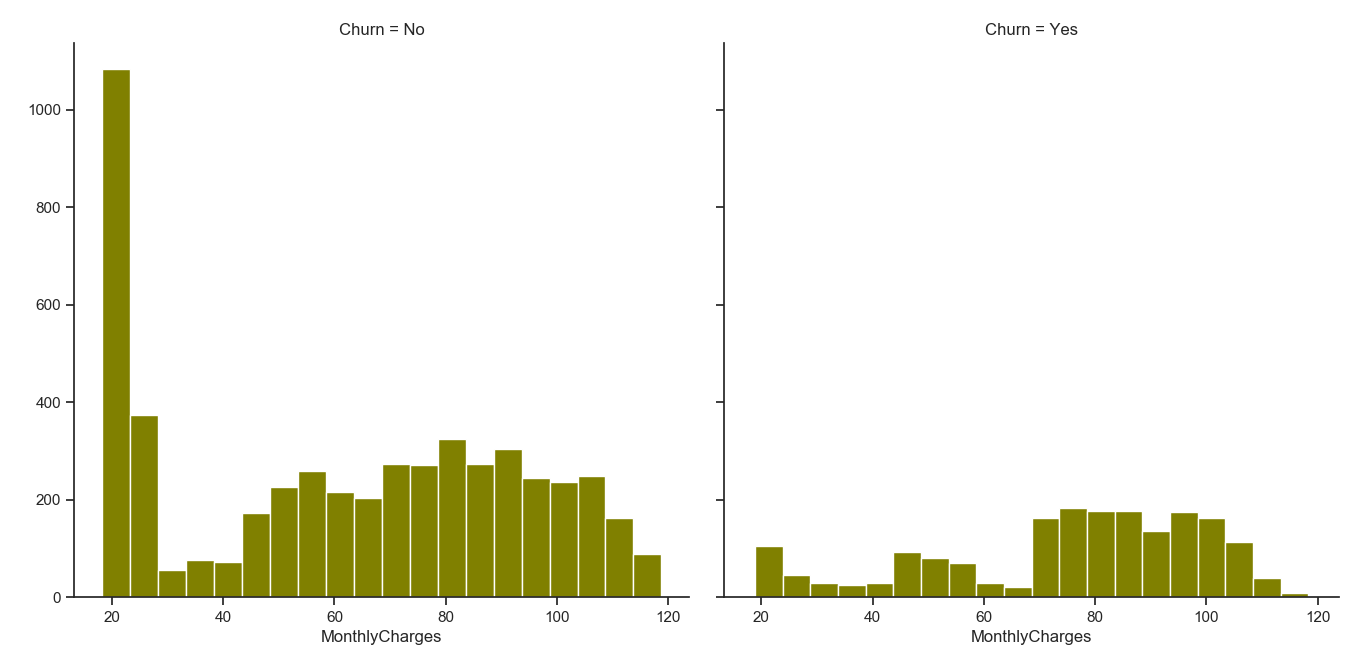
**Graph1:** The number of customer churn Vs different parameters. (a) Gender (b) Senior Citizen (c) Partner (d) Dependents  **Graph 2:** The number of customer churn Vs different parameters. (a) Phone Service (b) Multiple Lines (c) Internet service (d) Online security

**Graph 3:** The number of customer churn Vs different parameters. (a) Online Backup (b) Device Protection (c) Tech Support (d) Streaming TV 

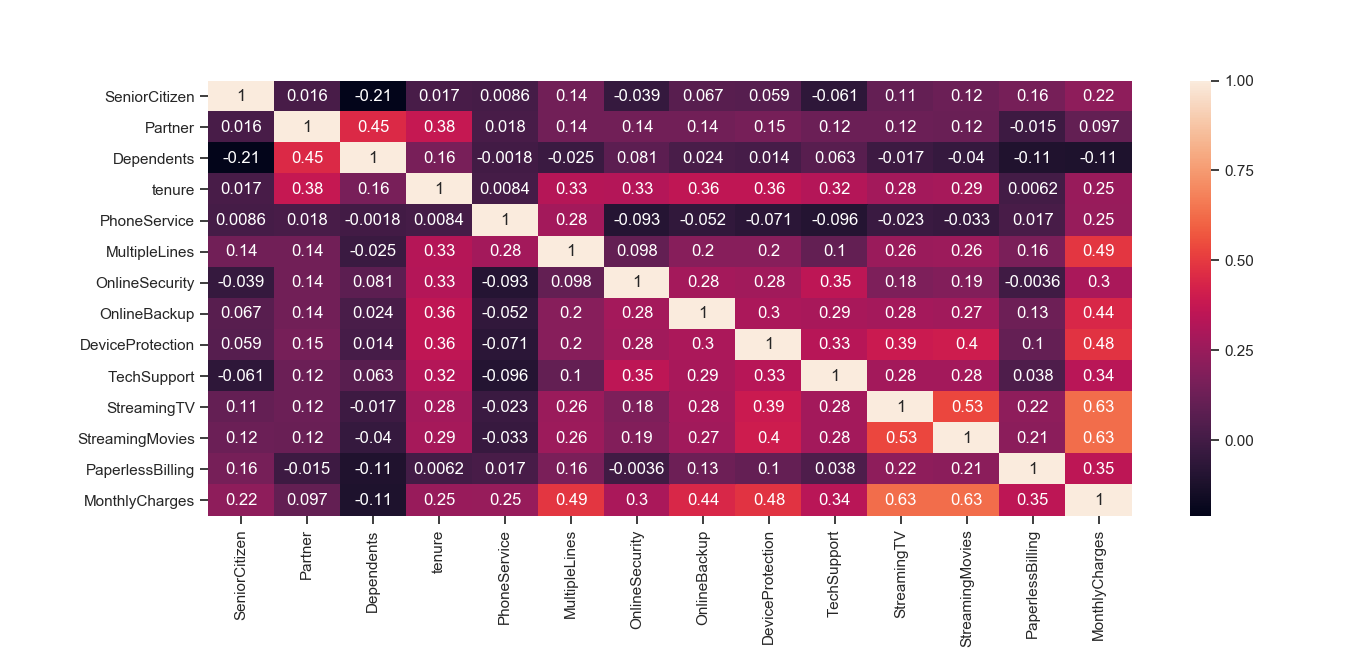
**Graph 4**: The number of customer churn Vs different parameters. (a) Streaming Movies (b) Contract (c) Paperless Billing (d) Payment Method

****

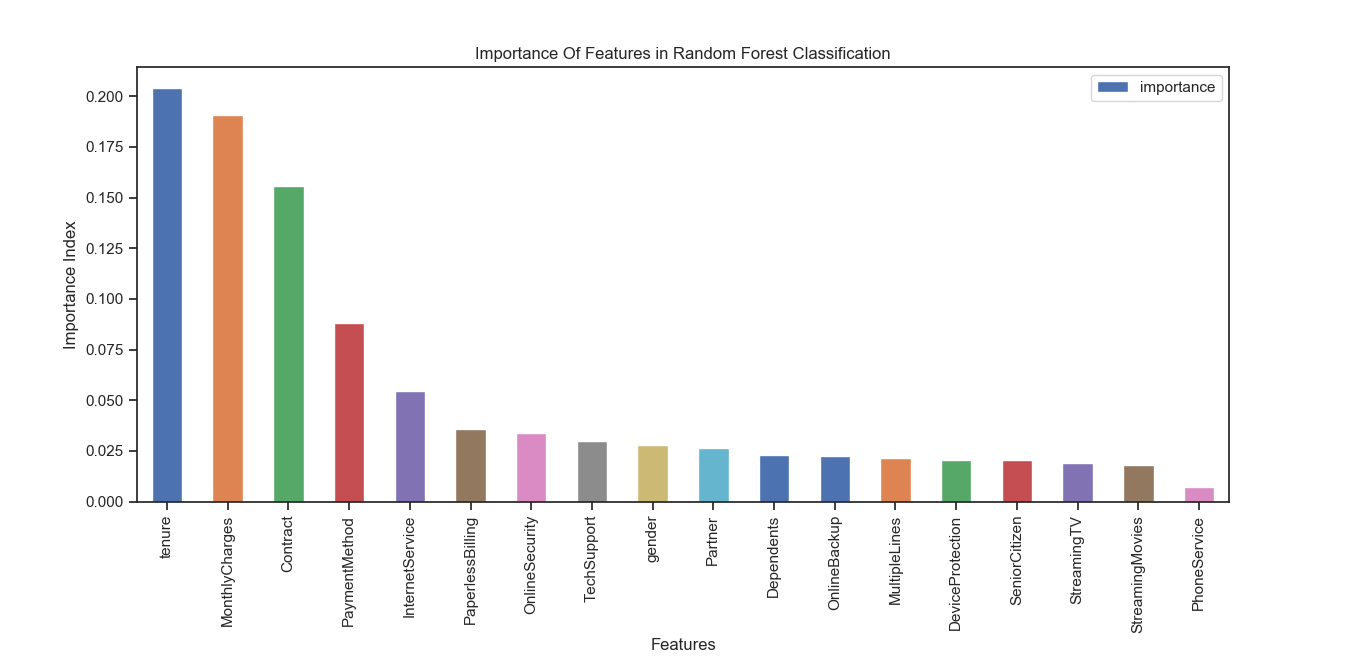
**Graph 5:** Customer churn based on tenure (in months)

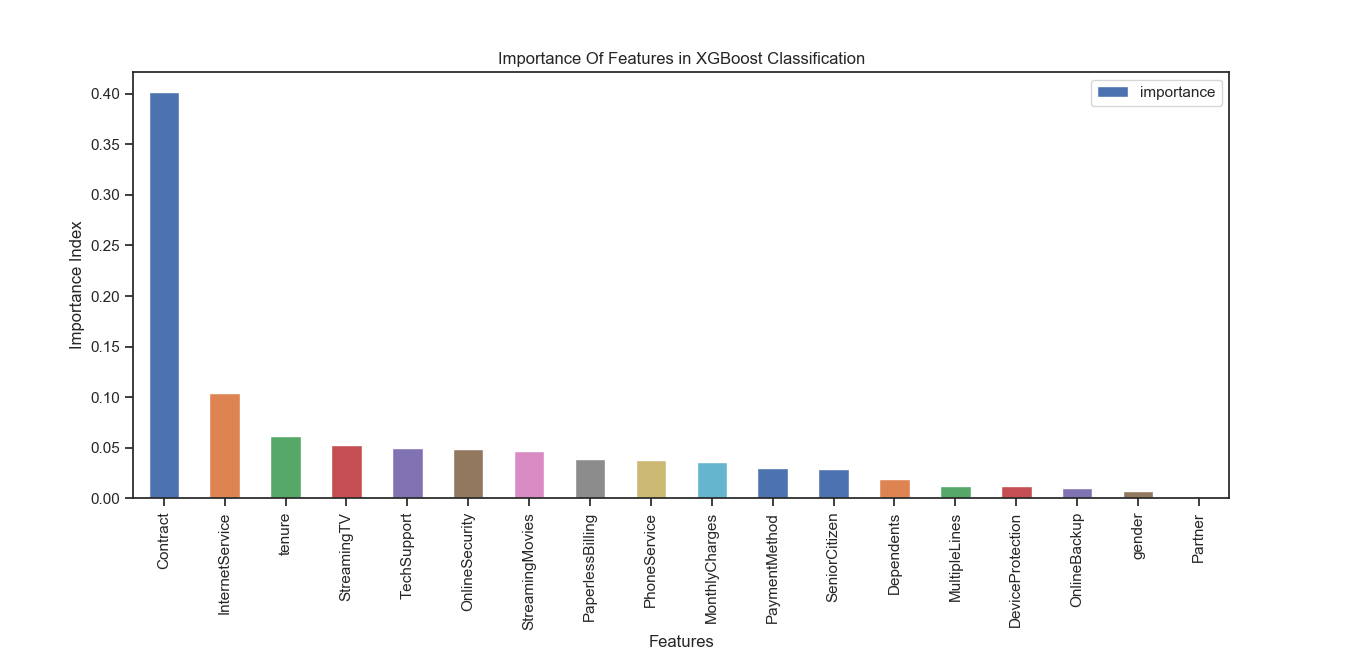
****

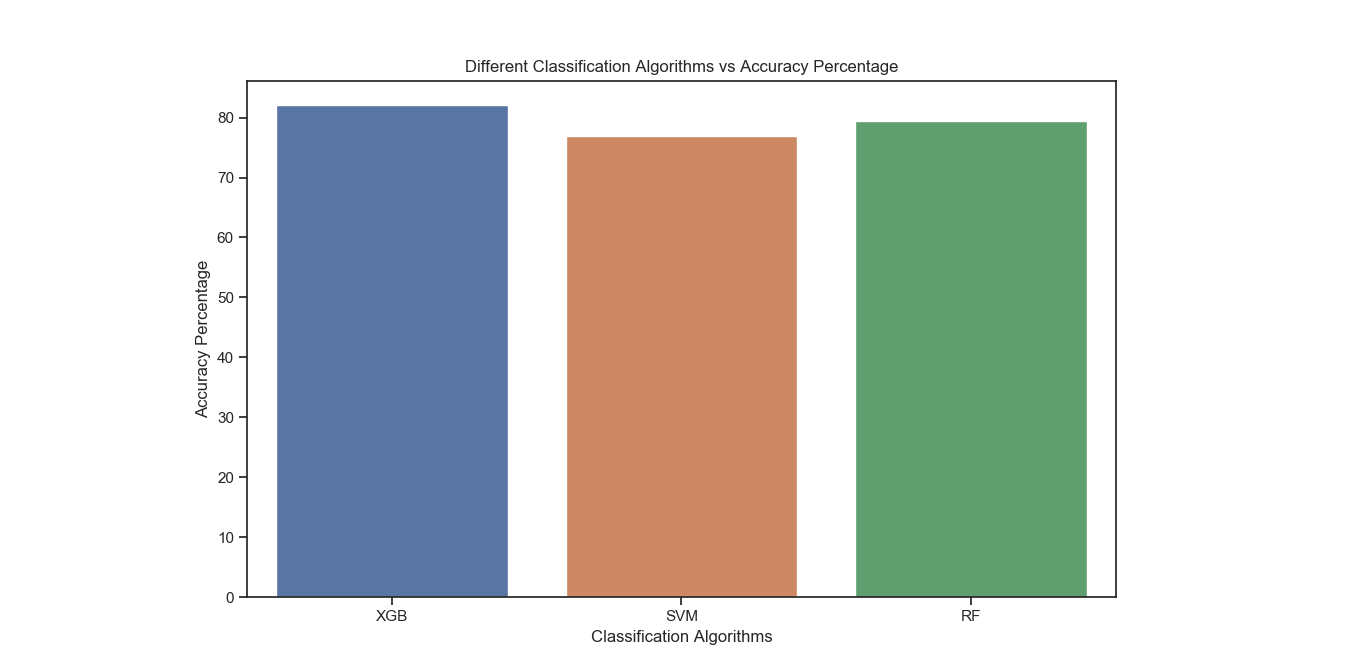
**Graph 6:** Customer churn based on Monthly charges (in rupees)

****

**Graph 7:** Graph depicting the correlation between the features

**Graph 8:** Graph depicting the important features for random forest in descending order

**Graph 9:** Graph depicting the important features for xgboost in descending order

****

**Graph 10**: The accuracy measure for each algorithm (a) XBG (b) SVM (c) Random forest

**Chapter 6**

# Conclusion and Future Work

Churn prediction is a function that involves systematic analysis of customer data for identifying and analyzing patterns and trends of customer loyalty and blend. The detected patterns and trends can be used by telecommunication industries to improve customer relationship and at the same time improve net profit. Identification of churners and non-churners is a time consuming and critical task, that has to be performed carefully, as the future growth of the company relies on the result of such an analysis. This task is considered challenging because of two reasons, (i) customer information volume has increased and (ii) the data available is inconsistent and are incomplete thus making the task of formal analysis a difficult task.

As technology progress, sophisticated data mining and artificial intelligence tools are increasingly accessible to the telecommunication sector. These techniques combined with state-of-the-art computers can process thousands of instructions in seconds, saving precious time. In addition, installing and running software often costs less than hiring and training personnel. Computers are also less prone to errors than human investigators, especially those who work long hours.

The current needs of telecom companies is a tool that can be used to help them to understand customer patterns and locate churners and possible actions that can be taken to convert the churners to non-churners. This tool is called as ‘Customer Loyalty Assessment Model and Actionable Knowledge Discovery System’ and the main goal is to provide timely and pertinent customer information to decision-makers in a company. The present research work focus on developing such a system that can be used by telecom industry easily discover customer patterns and trends, make forecasts, find relationships and possible explanations and identify possible churners.

To obtain an extra edge over competitive business, telecommunication industries are relying more and more on CRM combined with data mining techniques. In this study, customer’s churning behaviour is predicted along with actionable knowledge discovery. The proposed system consists of three main steps, namely, data preprocessing, customer loyal assessment and actionable knowledge discovery. Each of the three steps is treated as a separate research phase and the phases are interconnected to each other, where the output of one phase is taken as input by the next phase.

The various experiments conducted proved that the proposed algorithms and the proposed CRM system for customer loyalty assessment and actionable knowledge discovery are efficient. Experimental results showed that the system is effective in terms of analysis accuracy and speed in identifying common customer behaviour patterns and future churn prediction. The developed system has promising value in the current constantly changing telecommunication industry and can be used as effectively by companies to improve customer relationship and improve business opportunities.

**FUTURE RESEARCH DIRECTIONS**

The following can be considered to improve the proposed customer loyalty assessment model and actionable knowledge discovery system. The proposed models can be further enhanced, if the processes can be parallelized. This is feasible, by identifying operations that are independent to each other and propose a parallel architecture to improve the performance. Amount of memory used for loyalty assessment and action discovery is another area which can be analyzed in future Classification process can be improved by using advanced techniques like ensemble clustering or ensemble classification.

**References**

<https://www.datacamp.com/community/tutorials/xgboost-in-python#hyperparameters>

<https://www.datacamp.com/community/tutorials/xgboost-in-python#visualize>

<https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/>

<https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/>

<https://medium.com/greyatom/lets-learn-about-auc-roc-curve-4a94b4d88152>

<https://machinelearningmastery.com/develop-first-xgboost-model-python-scikit-learn/>