A project report on

TELECOM CHURN PREDICTION

Submitted in partial fulfillment of the requirements for the Degree of

B. Tech in **Computer Science and Communication Engineering**

by

ABHINAV KUMAR (1629119)
SAUBHAGYA ASHISH (1629171)
SHUBHANGI GUPTA(1629176)
SRIPRIYA SRIVASTAVA(1629183)
SHEFALI PANDEY(1629191)

under the guidance of

PROF. RAJDEEP CHATTERJEE

School of Computer Engineering
Kalinga Institute of Industrial Technology
Deemed to be University
Bhubaneswar

6th December 2019

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

Deemed to be University U/S 3 of the UGC Act, 1956

CERTIFICATE

This is to certify that the project report entitled " **TELECOM CHURN PREDICTION"** submitted by

ABHINAV KUMAR	**1629119**
SAUBHAGYA ASHISH	**1629171**
SHUBHANGI GUPTA	**1629176**
SRIPRIYA SRIVASTAVA	**1629183**
SHEFALI PANDEY	**1629191**

in partial fulfillment of the requirements for the award of the **Degree of Bachelor of Technology** in **Computer Science and Communication Engineering** is a bonafide record of the work carried out under my guidance and supervision at School of Computer Engineering , Kalinga Institute of Industrial Technology, Deemed to be University.

PROF. RAJDEEP CHATTERJEE

.....

The Project was evaluated by us on

EXAMINER'S NAME:

SIGNATURE:

ACKNOWLEDGEMENTS

We take this opportunity to express our deep gratitude to all those helping hands without whom this project would have not been what it is. We take immense pleasure to express our thankfulness to our mentor Prof. Rajdeep Chatterjee for his constant motivation, timely suggestions which made our sail smooth through the odds we faced in our project. I would like to thank our Dean Prof. Samresh Mishra for giving us this opportunity to enhance our skills by giving this project and for motivating us throughout the time and also thank the complete faculty of Computer Science and Engineering who have provided us the knowledge to successfully complete the project. Last but not the least we would like to thank our friends who not only provided their helping hands in our project as and when required but also for becoming the first user of this portal and providing us with all the feedback which helped us improve our portal.

ABHINAV KUMAR
SAUBHAGYA ASHISH
SHUBHANGI GUPTA
SRIPRIYA SRIVASTAVA
SHEFALI PANDEY

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, a methodical churn prediction model is required.

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 project are **Random Forests, SVM (Support Vector Machines) and XG-Boost**. The use of re-sampling method in order to solve the problem of class imbalance is also discussed. 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, various implications of Synthetic Minority Over-sampling Technique (SMOTE) in order to reduce the imbalance in data in collaboration with the help of few feature reduction techniques such as Co-relation feature extraction method are discussed.

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

TABLE OF CONTENTS

List of Figures

List of Tables

List of graphs

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 it is called 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 in the particular situation. This technique is applied through learning from historical data.

1.1 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 connection 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.

1.2 Motivation

Many approaches were applied to predict churn in telecom companies. Most of these approaches uses 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.

Ionuț Brândușoiu, Gavril Toderean, Horia Beleiu [1] 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.

Adnan Idris, Asifullah Khan [2] proposed an approach based on genetic programming with Ada-Boost to model the churn problem in telecommunications.

J. Burez, D. Van den Poel [3] 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 research is to help companies earn more profit. It can be aimed towards the following objectives-:

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.

1.4 Proposed Model

In this work, the implication of Synthetic Minority Over-Sampling Technique (SMOTE) is used in order to reduce the imbalance in data in collaboration with different feature reduction techniques such as Co-relation feature extraction. Techniques including Random forest [12][13], SVM[16] and XGBoost [10][11][15] were used for predicting the churn in telecom industry. Various performance measures such as accuracy, AUC-ROC [14], precision and recall were used for evaluating the performance of the predictive models.

1.5 Organization of the Study

- Chapter One starts with an introduction about churn management. The problem statement, objectives, significance and scope to the study.
- Chapter Two describes the study of the existing systems and techniques taken into account prior to development of the proposed system.
- Chapter Three provides a detailed walk through of the software engineering methodology adopted to implement the model, the concept of customer, what customer churn is all about, its causes, effects and its managements were reviewed.
- Chapter Four provides with the machine learning techniques used to develop the model. The various modules and their interactions are depicted using relevant descriptive diagrams.
- Chapter Five describes the testing of model to make it error free, results and observations.
- Chapter Six includes the conclusion along with the future scope. Answers to many objectives are presented in this chapter.

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.

Anuj Sharma, Dr. Prabin Kumar Panigrahi [4] 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.

Ammara Ahmed, D. Maheswari Linen [5] summarized that the churn prediction techniques in order to have a deeper understanding of the customer churn and it shows that most accurate churn prediction is given by the hybrid models rather than single algorithms so that telecom industries become aware of the needs of high risk customers and enhance their services to overturn the churn decision.

P.K.D.N.M. Alwis, B.T.G.S. Kumara, H.A.C.S. Hapuarachchi [6] classified the relevant variables with the use of the Pearson chi-square test, cluster analysis, and association rule mining. Using the Weka, the cluster results produced the involvement of customers, interest areas and reasons for the churn decision to enhance marketing and promotional activities. Using the Rapid miner, the association rule mining with the FP-Growth component was expressed rules to identify interestingness patterns and trends in the collected data have a huge influence on the revenues and growth of the telecommunication companies.

C. Wei, I. Chiu [7] 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.

R. Jadhav and U. T. Pawar [8] 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.

N. Kamalraj, A.Malathi [9] 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

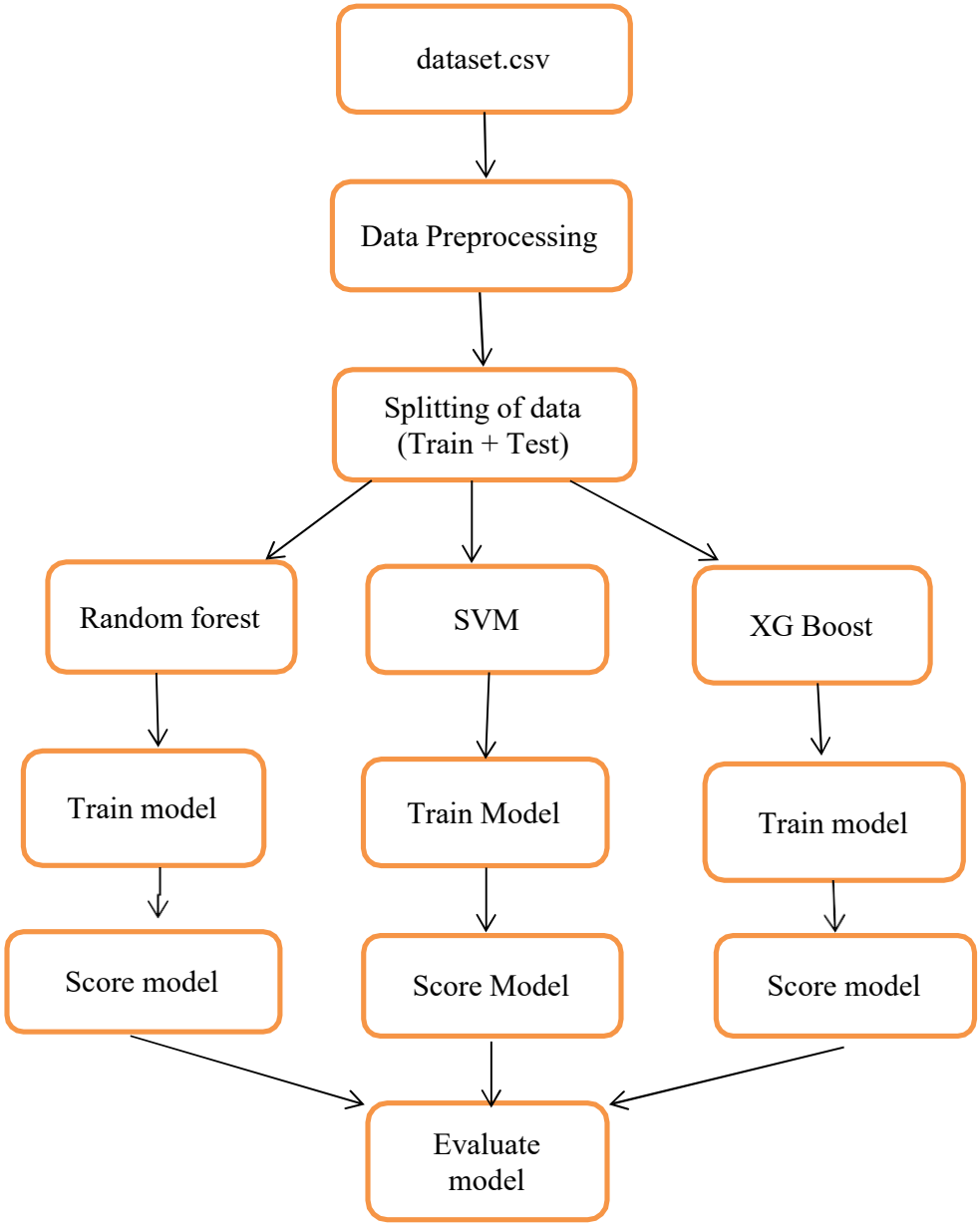


Figure 1: Architecture of the proposed model

In figure 1, we have a telecom churn dataset which is split into training and testing data set. Training methods are applied on this split dataset using various training models. This gives the value of performance measures such as accuracy, precision, recall etc. which is used for evaluation of the model whether there is improvement in the performance or not.

3.1.1 Data Set Description

The used dataset consists of 21 features for 7043 different customers. It consists of various features such as customer ID, gender, partner, dependents, tenure, phone service, multiple lines, internet service, online security, online backup, device protection, tech-support, streaming tv, streaming movies, contract, paperless billing, payment method, monthly charges, total charges and churn. There are three numerical features such as tenure, monthly charges and total charges. Rest of the features are Categorical i.e. paperless billing, streaming movies, payment method etc.

3.2 Performance Check Measures

In this project performance measures used are 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.

-Accuracy

It is the ratio of correct predictions out of all the predictions.

Accuracy=TP+TN/(TP+TN+FP+FN)

-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.

3.3 Steps of Application

1. **Feature Extraction:** This step involves considering the useful features from the dataset which can be used for prediction of the desired feature (churn in this case).

2. **Method:** This includes the use of various methods for the prediction of the desired feature i.e. Churn, this involves techniques such as Random forest, SVM, XGBoost.

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.

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

4.1.1 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.

How it works:

Random Forest is a supervised learning algorithm. 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:

- 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 considered as very handy and easy to use algorithm, because it's default hyper parameters often produce a good prediction result. The number of hyper parameters 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 too 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.

The pseudo code for random forest algorithm can split into two stages.

- Random forest creation pseudo code.
- Pseudo code to perform prediction from the created random forest classifier.

First, let's begin with random forest creation pseudo code

Random Forest pseudo code:

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 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 pseudo code:

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.

4.1.2 SVM (Support Vector Machines)

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, plotting of each data item takes place as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Then classification is performed by finding the hyper-plane that differentiate the two classes very well.

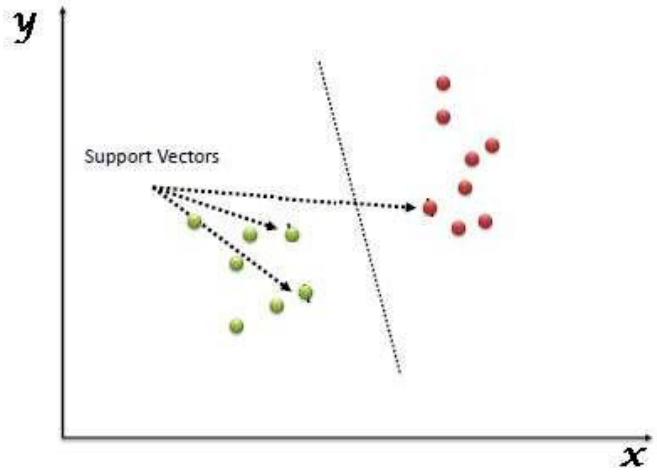


Figure 2: Support Vector Machine

From Figure 2, we infer that 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 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 hyper planes 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.

4.1.3 XG Boost

XGBoost, short for "Extreme Gradient Boosting", was introduced by Chen in 2014.

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

$$\frac{\partial L(y, f^{(m-1)}(x) + f_m(x))}{\partial f_m(x)} = 0$$

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

$$L(y, f^{(m-1)}(x) + f_m(x))$$
$$\approx L(y, f^{(m-1)}(x)) + g_m(x) f_m(x) + \frac{1}{2} h_m(x) f_m(x)^2,$$

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:

$$h_m(x) = \frac{\partial^2 L(Y, f(x))}{\partial f(x)^2}_{f(x)=f^{(m-1)}(x).}$$

How it works?

In XG Boost, model is fit on the gradient of loss generated from the previous step. In XG Boost, the gradient boosting algorithm is modified so that it works with any differentiable loss function.

4.2 Use Case Diagram

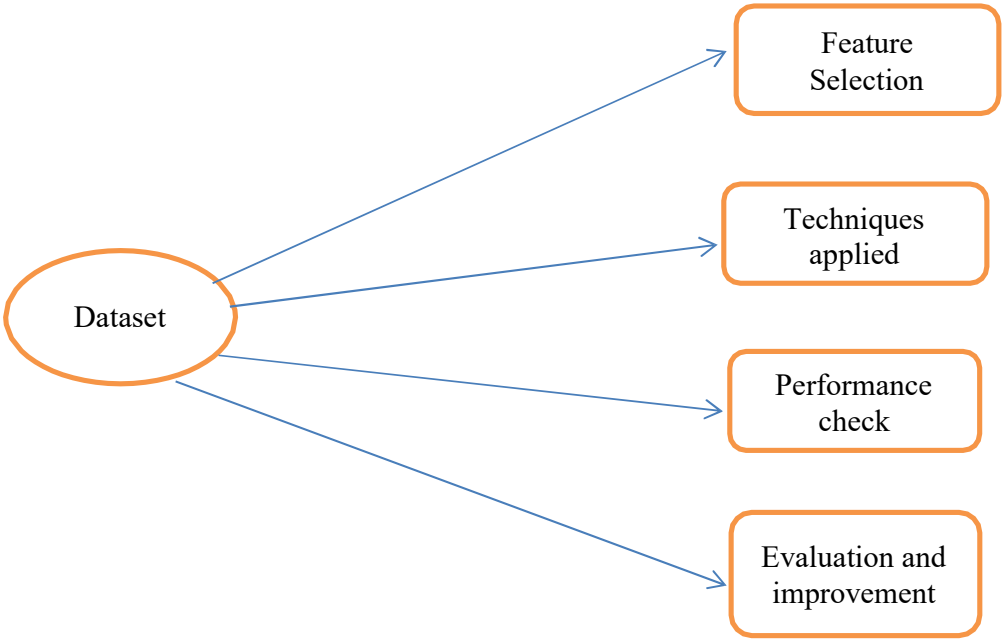


Figure 3: Use case diagram for model selection

Figure 3, explains the use case of the project, i.e. the various stages involved in the process such as feature selection, techniques used, checking the performance and the final evaluation with improvements.

4.3 Sequence Diagram

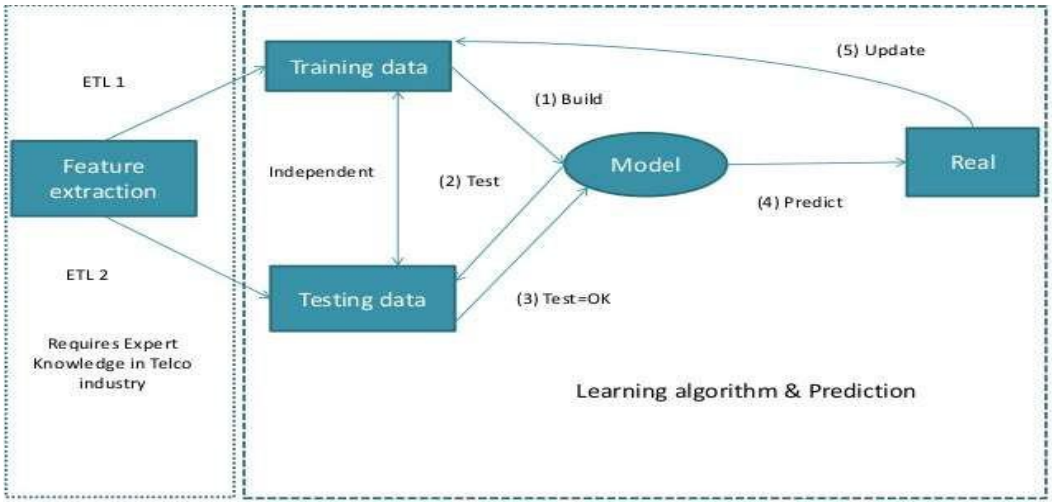


Figure 4: Sequence diagram

Figure 4, shows the order of interaction of steps for the proposed model in order to carry out the functionality.

4.4 Collaboration Diagram

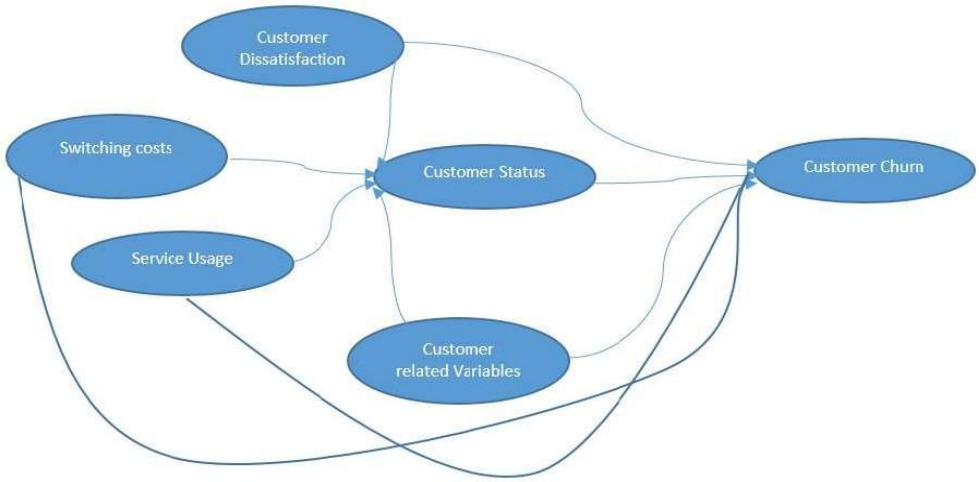


Figure 5: **Collaboration diagram**

Figure 5, shows how various factors and software objects interact with each other and how does it benefit the users of the project.

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	60.0	58.2	72.5
2.SVM	77.217	53.7	74.0	76.1
3.XG-Boost	82.079	68.6	54.3	72.9

Table 1: Performance metrics for used algorithms

Observation from Table 1:

In terms of accuracy, XG-Boost performs the best and returns the best accuracy value of 82.079 % among the three used models. Random forests gave an accuracy of nearly 80 % whereas SVM gave an accuracy of 77.217 %.

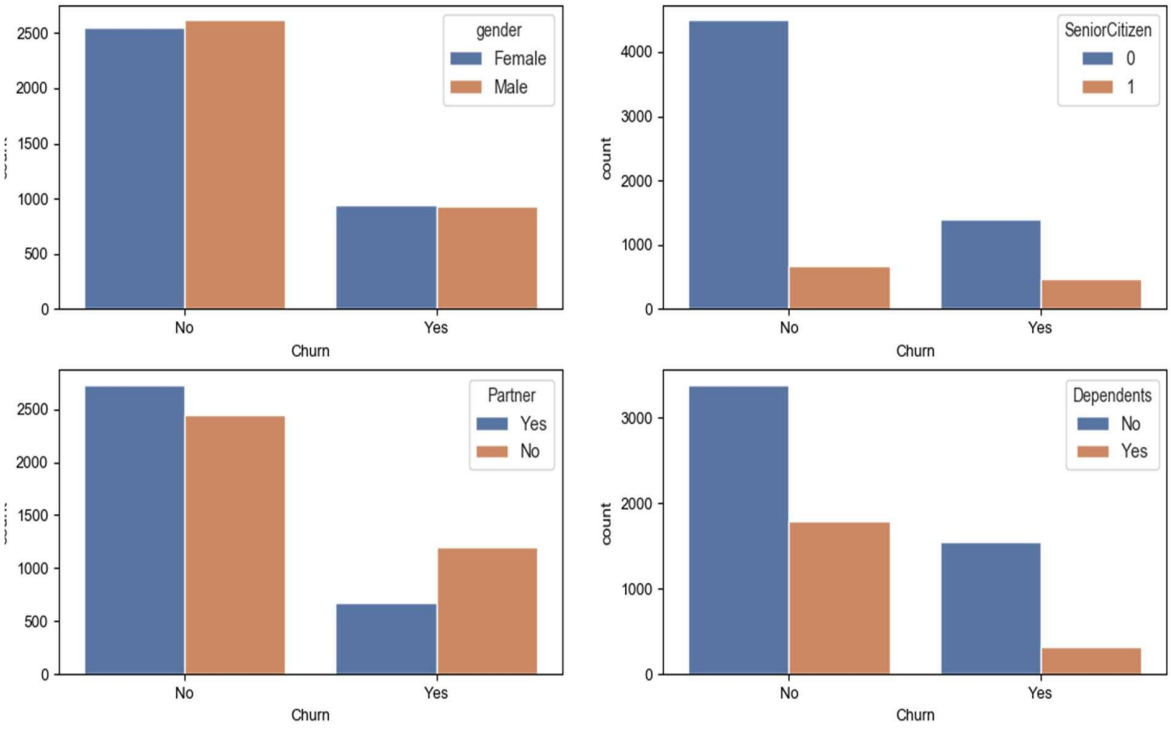
Similarly, XG-boost gave the best value for precision as well followed by random forests and SVM i.e. 68.6 %, 60 % and 53.7 % respectively.

SVM gave the best value of recall followed by random forests and XG boost.

In terms of area under curve, SVM gave the best value followed by XG boost and random forests.

Graphs by taking into account the various features in the dataset Vs the Churn. Also, a bar graph is plotted for showing the important features considered by random forest and XG boost. A bar graph is plotted to show the accuracy obtained for different algorithms.

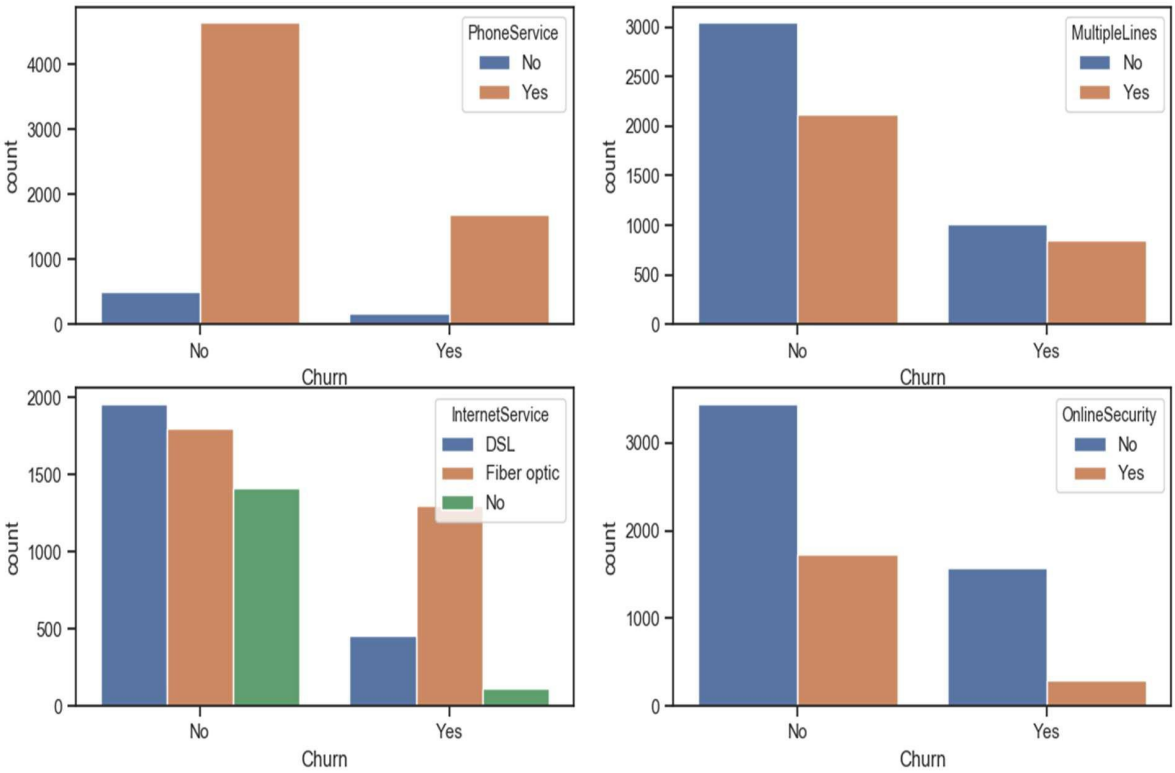
1. Categorical Features VS Churn
2. Numerical Features VS Churn
3. Correlation Matrix
4. Important Features
5. Accuracy Comparison



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

Observation from Graph 1:

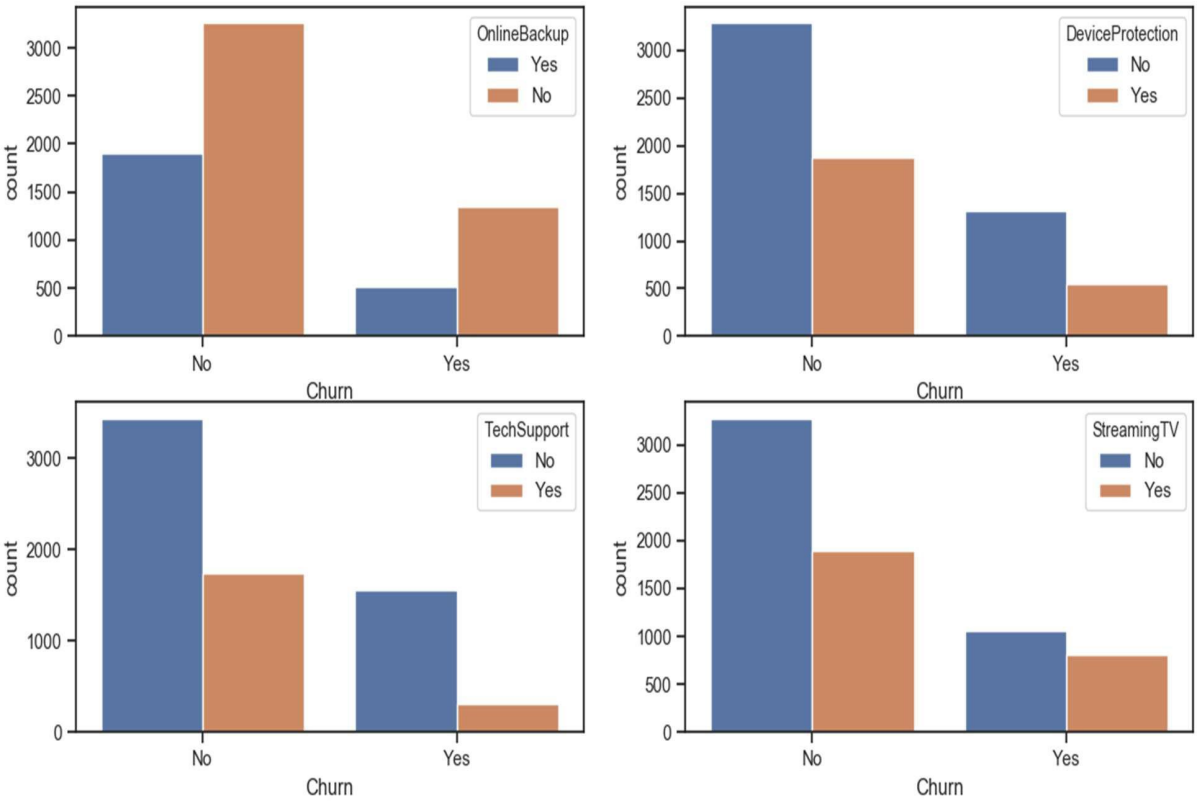
- Gender has no effect on the Churning of customers.
- Senior Citizen has less churn as compared to youngsters.
- Customer without a partner are more probable to churn.
- Customer who don't have dependents are more probable to churn.



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

Observation from Graph 2:

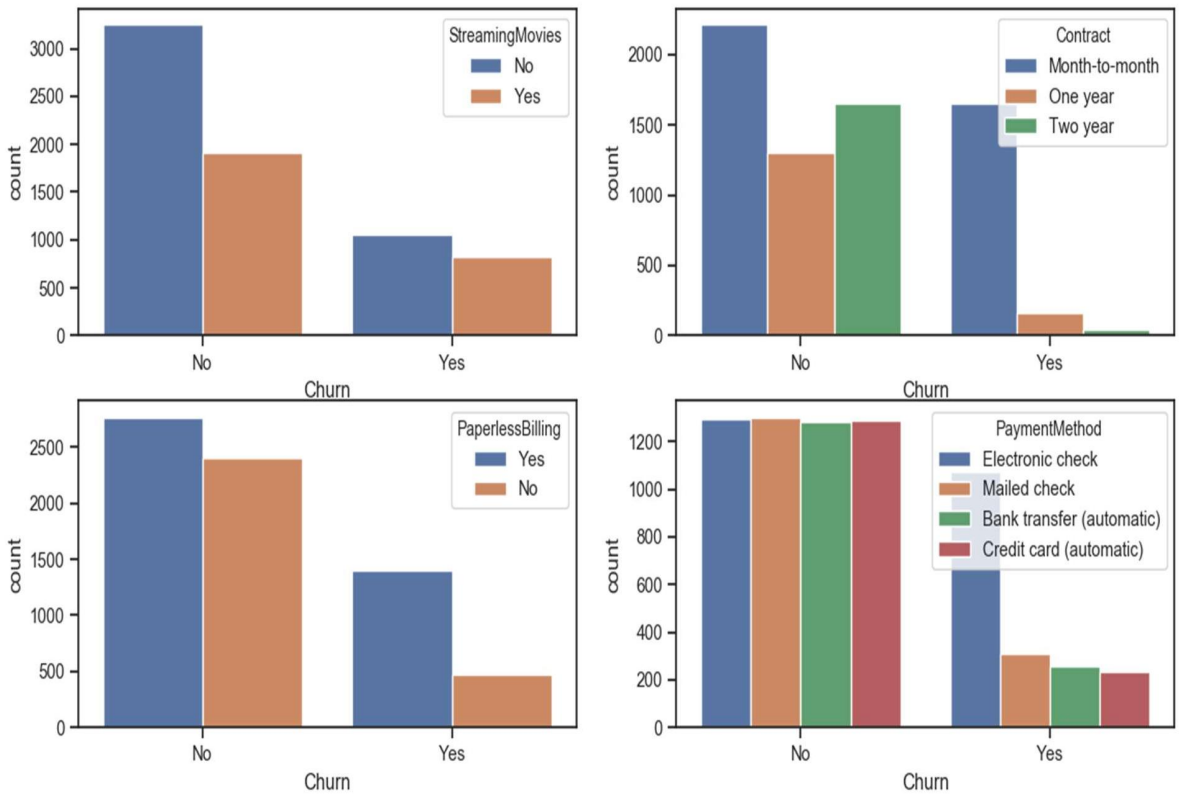
- Customer who have availed Phone Service are more probable to churn.
- Customer who have single line are more probable to churn.
- Customer who use Fiber Optic for Internet Service are more probable to churn.
- Customer who don't have online security service are more probable to churn.



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

Observation from Graph 3:

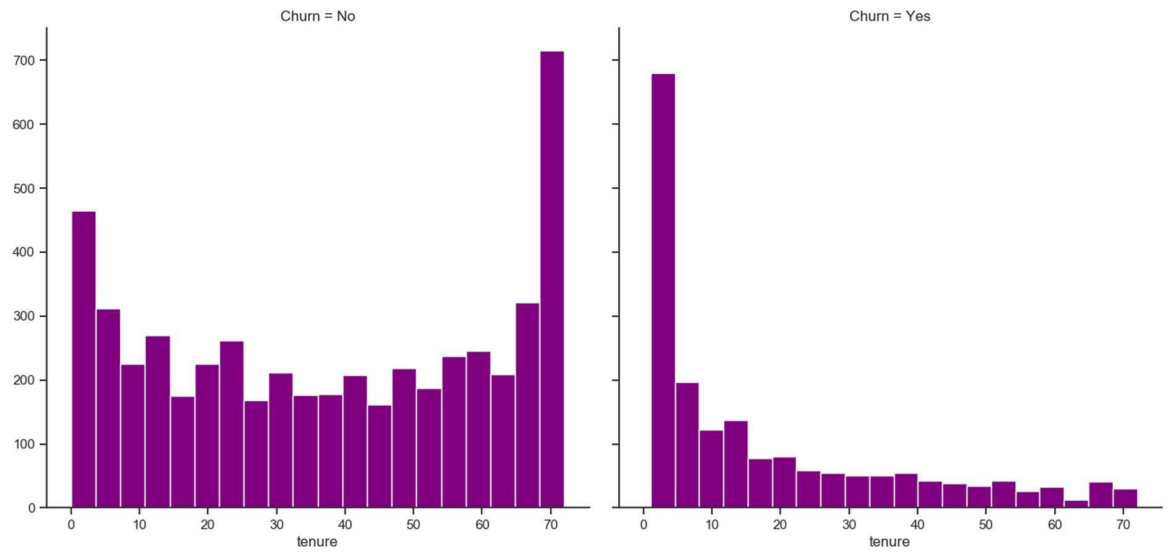
- Customer who don't avail Online Backup are more probable to churn.
- Customer who don't avail Device Protection are more probable to churn.
- Customer who don't avail Tech Support are more probable to churn.
- Customer who don't Stream TV are more probable to churn.



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

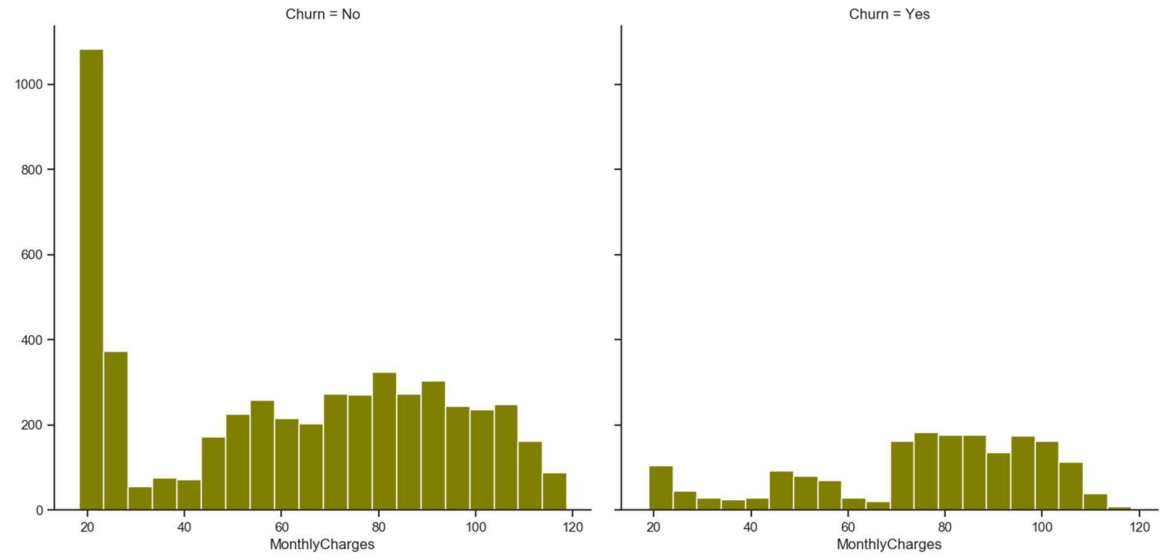
Observation from Graph 4:

- Customer who don't Stream Movies are more probable to churn.
- Customer who have Month-to-Month contract are more probable to churn.
- Customer who have Paperless Billing Method are more probable to churn.
- Customer who use Electronic Check as Payment Method are more probable to churn.



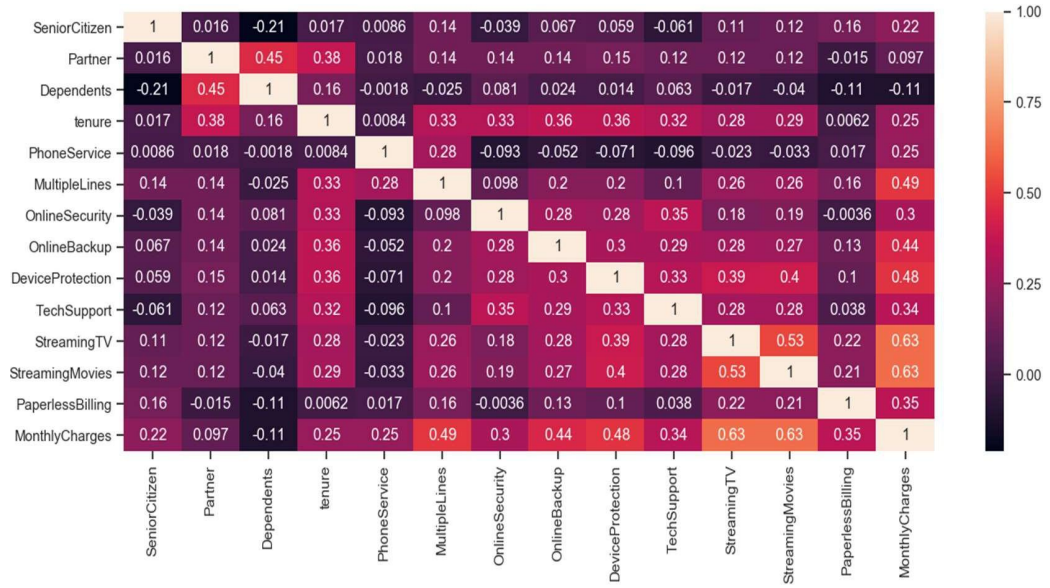
Graph 5: Customer churn based on tenure (in months)

Observation from Graph 5: Customer with tenure period of less than 10 months are more probable to churn.

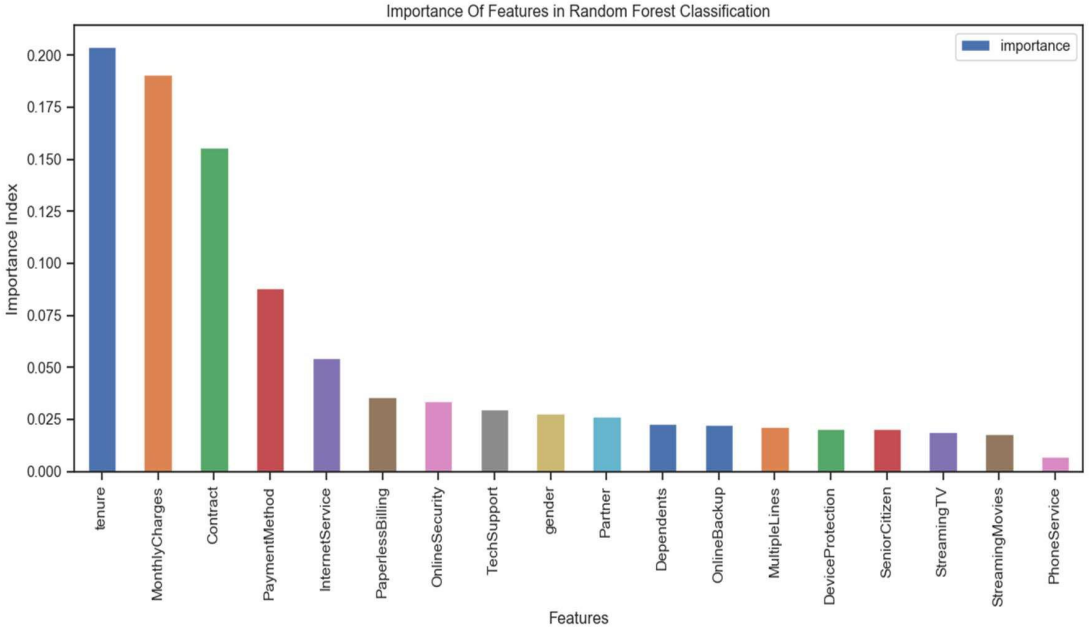


Graph 6: Customer churn based on Monthly charges (in rupees)

Observation from Graph 6: Customer with monthly charges greater than Rs. 70 are more probable to churn.

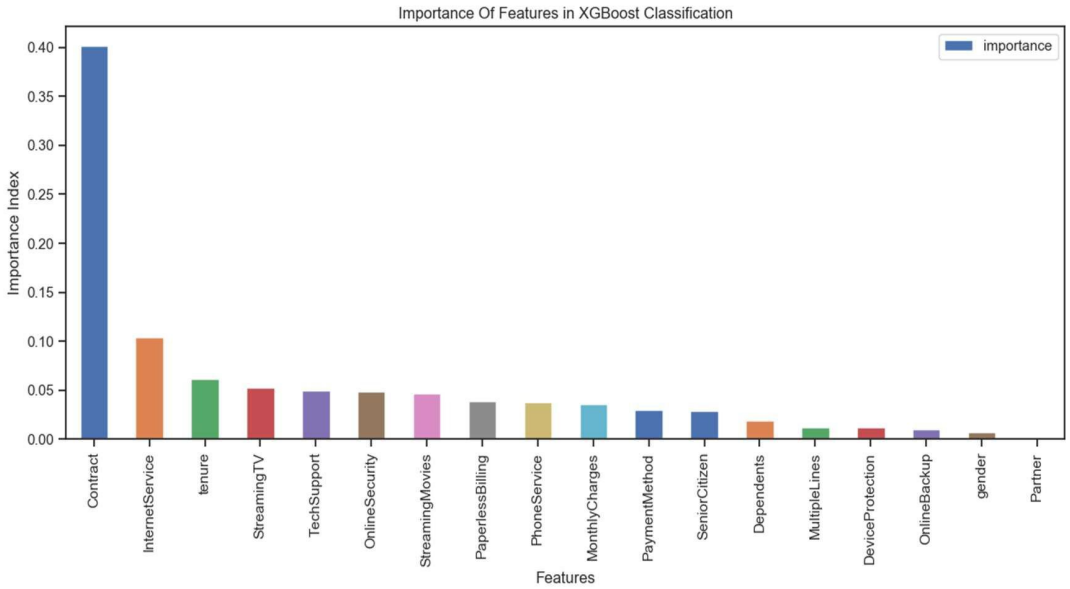


Graph 7: Graph depicting the correlation between the features



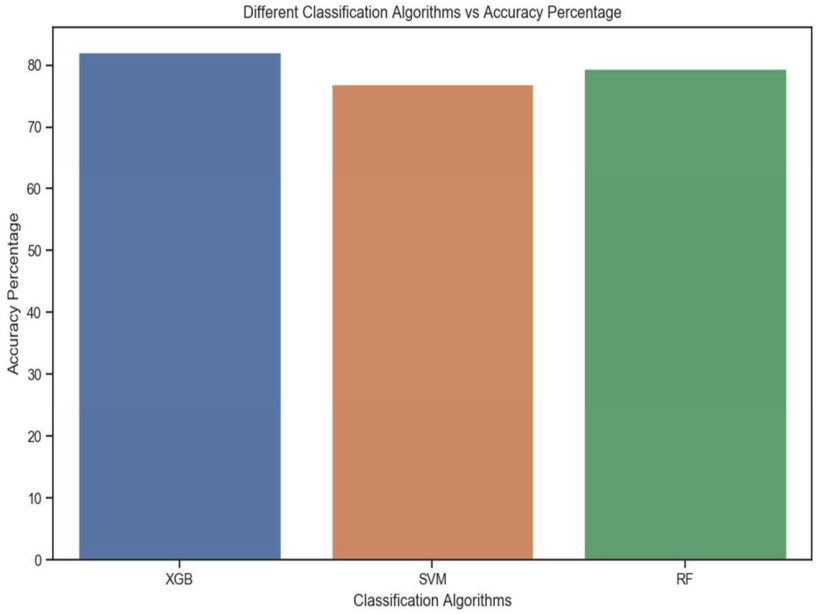
Graph 8: Graph depicting the important features for random forest in descending order

Observation from Graph 8: Most Important Parameter is Tenure according to Random Forest.



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

Observation from Graph 9: Most Important Parameter is Contract according to XGBoost.



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

Observation from Graph 10: XGBoost has better accuracy than SVM & Random Forest.

Chapter 6

Conclusion and Future Work

Churn prediction is a technique 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 are 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

[1] Ionuț Brândușoiu, Gavril Toderean, Horia Beleiu, Methods for churn prediction in the pre paid mobile telecommunication industry

[2] Adnan Idris, Asifullah Khan, Genetic Programming and Adaboosting based churn prediction for Telecom

[3] J. Burez, D. Van den Poel, Handling class imbalance in customer churn prediction

[4] Anuj Sharma, Dr. Prabin Kumar Panigrahi, A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services

[5] Ammara Ahmed, D. Maheswari Linen, A review and analysis of churn prediction methods for customer retention in telecom industries

[6] P.K.D.N.M. Alwis, B.T.G.S. Kumara, H.A.C.S. Hapuarachchi, Customer Churn Analysis and Prediction in Telecommunication for Decision Making

[7] C. Wei, I. Chiu, Turning telecommunication call details to churn prediction: a data mining Approach expert System with applications

[8] R. Jadhav and U. T. Pawar, Churn Prediction in Telecommunication Using Data Mining Technology

[9] N. Kamalraj, A.Malathi, Applying Data Mining Techniques in Telecom Churn Prediction

[10] XG Boost - https://www.datacamp.com/community/tutorials/xgboost-in python#hyperparameters

[11] XG Boost - https://www.datacamp.com/community/tutorials/xgboost-in-python#visualize

[12] Machine Learning - https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/

[13] Random Forest - https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/

[14]ROC Curve - https://medium.com/greyatom/lets-learn-about-auc-roc-curve-4a94b4d88152

[15] XG Boost -https://machinelearningmastery.com/develop-first-xgboost-model-python-scikit-learn/

[16] SVM - https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python