TELCO CUSTOMER CHURN ANALYSIS USING VARIOUS MACHINE LEARNING ALGORITHMS

A PROJECT REPORT

Submitted to Kannur University in partial fulfillment of the requirement for the award of Degree of

Master of Science in Statistics

Offered in the Department of Statistical Sciences

Kannur University

Submitted by

NIVEDITHA V M

Department of Statistical Sciences
Kannur University Campus
Mangattuparamba-670567

KANNUR UNIVERSITY

DEPARTMENT OF STATISTICAL SCIENCES

MANGATTUPARAMBA

CERTIFICATE

This is to certify that **NIVEDITHA V M** have done the project work entitled **TELCO CUSTOMER CHURN ANALYSIS USING VARIOUS MACHINE LEARNING ALGORITHMS** for the partial fulfillment of the requirements for the award of the degree of Master of Science in Statistics from Kannur University under any supervision and guidance.

Dr. Sebastian George

Head of the Department
Department of Statistical Sciences
Kannur University
Mangattuparamba

Name of Examiners:

1.

2.

Place:
Date:

DECLARATION

I hereby declare that the project entitled **TELCO CUSTOMER CHURN ANALYSIS USING VARIOUS MACHINE LEARNING ALGORITHMS** is a record of original work done by me under the supervision and guidance of Dr Tulasi B (Assistantant professor) and Dr Sharon Varghese (Assistantant professor), Department of statistics and Data Science, CHRIST (Deemed to be University), Bangalore , in partial fulfilment of the requirements of the Degree of Master of Science in Statistics of Kannur University, and it has not previously formed the basis for the award of any other Degree or Diploma or Associateship or Fellowship or other similar title to any candidate of this or any other university.

NIVEDITHA V M
Department of Statistical Sciences
Kannur University
Mangattuparamba

Approved by,

ACKNOWLEDGEMENT

The successful completion of this dissertation owes to the inspiration and constant support that I received from various sources. I avail this opportunity to express my sincere gratitude to all those who helped me directly or indirectly for the completion of work.

I wish to express my deep gratitude to my guides **Dr Tulasi B**, Assistant Professor and **Dr Sharon Varghese**, Assistant Professor, Department of statistics and Data Science, CHRIST (Deemed to be University), Bangalore for their valuable guidance that I have received throughout the period of the study.

I place my sincere thanks to **Dr. Deepthi Das**, Head of Department of statistics and Data Science, CHRIST (Deemed to be University), Bangalore and **Dr.Sebastian George**, Head of Department of Statistics,Kannur University.

I sincerely express my gratitude to faculty members of the Department of statistics and Data Science, CHRIST (Deemed to be University), Bangalore, my teachers, my parents, and my friends and all those who have encouraged me in this endeavor.

Above all I am greatly thankful to the grace of Almighty God for the successful completion of my training programme.

NIVEDITHA V M

Contents

1 INTRODUCTION

1.1 Customer churn

The expansion of business,commerce environment and the advent of various communication platforms such as the internet has made it an easy experience for customers to be informed of similar services and products offered by different companies in the shortest time.Accordingly the customers who are dissatisfied can switch to another company which leads to the rise of customer churn of the prior company.Several companies across different sectors including banking sectors ,airline services and telecommunication are directly affected by customer churning.These companies increasingly focus on creating and sustaining long term connections with their current customers.

Customer churn in the telecom industry describes a situation where a customer stops the services of one telecom company during the contract and switches to a competitor to obtain a better cheaper and mores satisfactory service for customer's needs [11].Acquiring new customers was the important strategy for increasing revenue quickly at early stage but when telecom industry become saturated it's focus shifted to the prevention of customer churn.The cost of acquiring new customers is six times higher than maintaining new customers (provided by Castanedo [6] at 2014).The customer churn has become one of the challenging issues due to its significant increase in the telecommunication sector [7]. The telecommunication sector is experiencing significant customer churn due to tough competition on crowded markets, a dynamic environment and the introduction of new and tempting packages. One of the major key purpose of **customer churn prediction(CCP)** is to help in creating new strategies that retain customers and thus will increases business revenue and industrial recognition. Financial loss due to the decline of profits and negative effects on the current customers are the most important problems why the customer churn is affecting for companies.

In recent decades the service providers are increasingly working on long term relations with their customers.Therefore service providers maintain customer relationship management(CRM) in which every event related to each particular cost is recorded [9].The CRM is a tool which is used to learn more about consumer's needs and their behaviour in order to establish a stronger relationship with them. [3].The CRM database are useful to anticipate and respond to customer's needs through the application of business process and machine learning techniques to determine customers behaviour. [10]

Types of customer churn include the following:
Voluntary customer churn. When customer churn is voluntary, it is the purchaser who makes the decision to stop buying the product or service. This may be because the customer no longer has a need for it or has decided to purchase the product or service from another vendor. Voluntary churn is often caused by the customer's perception that the vendor's products to do not align with the customer's needs or values.

Involuntary customer churn. Customer churn can also be involuntary. In this case, it is the seller who decides not to continue a business relationship with the customer. Typically, this type of customer churn occurs because the customer has not met previous financial or logistical responsibilities.

The **Downgrade churn** happens due to a customer choosing a starting plan from a premium plan leading to something called downgrade MRR. (Monthly Recurring Revenue).This mainly happens due to the non-alignment of product features and value add metrics and pricing factors.Companies can deal with downgrade churn by setting the right pricing and packaging products better.Also, looking for ways to make the product features for the customer attractable with improved features will prevent the downgrade.If sales representatives and marketers can understand why customers churn, they can provide other stakeholders within the company with insight into how the organization's products and services can be improved.

In the telecommunications industry, churn rates are calculated by dividing the number of customers who cancel their account in a given month by the total number of customers at the beginning of that same month, then multiplying the result by 100.

For example, if there were 10,000 customers at the beginning of January and 3,000 had cancelled their accounts by the end of January, then the churn rate would be calculated as follows:

Customer Lost / Total Customers * 100 = Churn Rate

3,000 / 10,000 * 100 = 30% Churn Rate

Here are some reasons that explain what churn analysis does for companies and why they are important.

1. **Shows the vulnerabilities of a product:**

Churn analysis will reveal trends that are most likely the intent of customers to leave a brand. These trends can be anything from pricing strategies to new product feature launches that are not in sync with the customers. Also, making a subsequent impact would be this analysis will reveal customer engagement with the brand in their journey. This information is particularly useful to improve the product according to customer's requirements.

2. **New Customer Opportunities:** Tailoring the customer journey by understanding the necessities of the customer at every buying stage is very important to enhance the customer service experience. A churn analysis lets you analyze the repeated patterns and tracks the evolving buying behavior at every customer interaction point in the stage for improvements. Strategies such as personalizing messages and offering customer-specific experiences make them feel privileged.

3. **Prediction for lower future Churn rate:** Analysis of old customer data is a big part of customer churn rate prediction. Understanding customers at every stage using metrics like customer lifetime value scores gives you a count of loyal and inactive customers. Doing this will give you the time to predict customer churn much earlier and start retention programs or strategies.

1.2 Machine Learning

Artificial intelligence means replicating human intellectual processes through machines, especially computers.Machine learning is a subset of Artificial intelligence that provide system the ability to automatically learn and improve from experience without being explicitly programmed.Machine learning(ML) explores the study and construction of algorithms that can learn from data and make predictions on data.Based on more data, machine learning can change actions and responses which will make it more efficient and adaptable.

There are 3 types of machine learning Supervised learning, unsupervised learning and reinforcement learning.In supervised learning we train the computer using the labelled training set.Filtering spam mails from important mail is an application of supervised learning.In unsupervised learning the training is done using unlabelled training set as the Flipkart and Amazon figure out products suitable for customers is one of it's application.In the case of reinforcement learning the computer learns from the mistakes and experiences.In gaming the strength of the opponent increases as gamer get better.This is an application of the reinforcement learning.The supervised learning predict the output while unsupervised will find the hidden structure in data. The right ML solution is determined based on the problem statement ,size quality and nature of data and complexity of algorithms.

There are different types of machine learning algorithms designed in such these times to help solve real-world complex problems. The ml algorithms are automated and self-modifying to continue improving over time. Classification,regression and clustering are the types of machine learning algorithms. Classification algorithm is one among the supervised machine learning algorithm. It is the process of recognizing, understanding, and grouping ideas and objects into preset categories.Classification is a form of "pattern recognition," with classification algorithms applied to the training data to find the same pattern. Regression is used when the predicted data is numerical.When a value needs to be predicted in stock prices according to it's demand we use the regression algorithm.Linear regression, Ridge regression,LASSO regression are some of the example for it. Clustering is used when data needs to be organised to find patterns in the case of product recommendation.K-means algorithm is one of the example for it.

Classification algorithm is used when the output is categorical like "YES" or "NO".Here we use classification algorithm since we have to predict that the customer had churned or not.Decision tree ,random forest ,logistic regression, naive bayes,K-nearest neighbour and support vector machine are the common classification algorithms.

A **decision tree** algorithm is a machine learning algorithm that uses a decision making tree to make predictions. It follows a tree-like model of decisions and their possible consequences. The algorithm works by recursively splitting the data into subsets based on the most significant feature at each node of the tree.An example of a decision tree is a flowchart that helps a person decide what to wear based on the weather conditions.The purpose of a decision tree is to

make decisions or predictions by learning from past data. It helps to understand the relationships between input variables and their outcomes and identify the most significant features that contribute to the final decision.It starts with a root node and ends with a decision made by leaves.

Random Forest is one of the most popular and commonly used algorithms for data analysis. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification. It performs better for classification and regression tasks.

Logistic Regression is a "Supervised machine learning" algorithm that can be used to model the probability of a certain class or event. It is used when the data is linearly separable and the outcome is binary or dichotomous in nature. That means Logistic regression is usually used for Binary classification problems. There are 2 types of logistic regression Simple Logistic Regression where a single independent variable is used to predict the output and Multiple logistic regression where multiple independent variables are used to predict the output

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K-NN algorithm

Naïve Bayes re is also known as a probabilistic classifier since it is based on Bayes' Theorem. It would be difficult to explain this algorithm without explaining the basics of Bayesian statistics. This theorem, also known as Bayes' Rule, allows us to "invert" conditional probabilities.Bayes' Theorem is distinguished by its use of sequential events, where additional information later acquired impacts the initial probability.

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithm whose goal is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.

The ensemble methods in machine learning combine the insights obtained from multiple learning models to facilitate accurate and improved decisions. In learning models, noise, variance, and bias are the major sources of error. The ensemble methods in machine learning help minimize these error-causing factors, thereby ensuring the accuracy and stability of machine learning (ML) algorithms. When we are taking into account different views and ideas from a wide range of people to fix issues that are limiting the user experience we found out the importance of ensemble learning.To get combined strength of the models free from individual model variances and biases we learn several simple models and combine their output to produce the final decision.And this is what

ensembling do.

There are 2 types of ensemble methods : **Sequential ensemble method** and **Parallel ensemble method**.Sequential ensemble model works based on the dependence between the base learners while parallel ensemble method works based on independence between the base learners.Ensemble model is the application of multiple models to obtain better performances than from a single model.Performance of the ensemble model depends on 2 major factors. That is Robustness and accuracy.

Ensemble can be created by combining all weak learners once weak learners are properly combined we can obtain more accurate and or robust models.**Model averaging** is an approach to ensemble learning where each ensemble member contributes an equal amount to the final prediction.In the case of regression the ensemble prediction is calculated as the average of member predictions.limitation of this prediction is that equal weights are assigned to different models despite some models performing better than others.**Weighted averaging** is a extension of a model averaging where the contribution of each member to final prediction is weighted by the performance of model.model weights are positive and small.

2 REVIEW OF LITERATURE

The objective of a research conducted by Abdelrahim Kasim [1] and peers is to predict the customer churn happens in a telecom company named SyriaTel using various ML algorithms like decision tree,random forest and also the analysis was done by gradient boosted machine tree(GBM) and extreme gradient boosting(XGBOOST).Decision tree, Random forest, GBM and XGBOOST are the algorithms used for this.In order to build the churn predictive system a big data platform called horton works data platform(HDP) was chosen. This customized package of installed systems and tools is called STYL-BD framework.Also i HDFS(Hadoop distributed file system) to store data, Spark execution engine to process data and Yarn to manage resources.Apache Flume is used to move unstructured and semi structured data from STYL-BD to HDFS and Apache SQOOP was used to move structured data. While having a comparison between all the algorithms used,XGBOOST algorithm outperform the rest of the tested algorithm with an AUC(Area Under Curve) value of 93.3% so that it can be chosen to be the classification algorithm in this proposed predictive model. GBM occupied second place with AUC 90.89% and Random forest and Decision tree in the last place with AUC 87.76 and 83 percentage respectively.

The research paper by Yajun Liu [2] and peers proposes an integrated customers churn management framework for the telecom industy combining machine learning algorithms such as gradient boosting decision tree and support vector machine random forest and aiming to achieve and reduce customer churn. Support vector machine, Random forest ,K-means algorithm and ensemble algorithms are used.Dataset used here is a 3 month period of customer data provided by a telecom company.Feature selection and extraction was done using scatter plots ,bar charts and cross tabulation.k-means algorithm is used to cluster users with different cost and analyse factors affecting different consumer groups.SMOTE algorithm was used to balance the data set for different consumer groups.

Objective of the research conducted by Adnan Amin [3] and others is to find out if it is possible to develop a JIT approach using cross company data for CCP in telecommunication sector and also to compare the performance of JIT homogeneous and heterogeneous ensemble models for CCP in telecommunication sector. Support vector machine , Neural net, Naive baye's classifier and K-nearest neighbour are the algorithms used.To classify the customers to churn and non churn category is the main objective of CCP.Numerous ML techniques are used but these conventional approaches require large amount of historical data. By Just in time approach we need to classify that with limited historical data.minimal redundancy maximal relevance(mRMR) is introduced for feature selection. The empirical results indicate that the prediction performance of SVM as a base classifier in the heterogeneous ensemble method is generally better as compared to applying SVM as an individual or homogeneous ensembles for CCP. The results reveal the effectiveness of the heterogeneous method

in JIT-CCP as more useful and practical alternative framework for CCP in telecommunication sector.

There is a study by Muhammad Usman Tariq [4] and others for which the aim is to adopt advanced ML algorithms and HPO techniques to solve customer churn prediction problems in telecom sector and also implementing a controlled-ratio under sampling technique to solve the problem of class imbalance. This research is conducted by 3 ML methods of Random forest, SVM and KNN are optimized by 3 HPO techniques GS,RS and GA respectively. To apply ML models for practical problems of detecting the churn risk customers in the telecom sector, their hyper parameters need to be tuned to fit specific data sets. In this study RF optimized by GS HPO technique has shown it's superior abilities in predicting customer churn in telecom sectors ,regardless of being tested on small or bigger data sets. The purpose of this model is to assist the E-business to predict the churned users using machine learning and also have an objective to monitor customer behaviour and to perform decision making.2-Dimensional convolution neural network , Apache spark parallel and distributed framework are used.The model is accurate having 96.3% accuracy.The training and validation loss is 0.4%

A study conducted by Young jung suh [5] proposed feature engineering by considering the existing domain knowledge.The churn prediction model defined on this paper was verified by considering the actual operational dataset.To ensemble machine learning algorithms like random forest which aim to predict customer churn in the best way, there are many studies ranging from RFM (Recency,Frequency,Monetary)models.This prediction study was conducted based on customer behaviour information of actual water purifier rental company,Where due to the characters of customer's who use the rental business churn occurs frequently.

3 METHEDOLOGY

After reading the research articles it is finalised that the main methods or algorithms used in this project are Logistic regression , Decision tree, Random Forest, Support vector machine and few of the ensembling methods.Here by analysing the details of the customers provided by the telecom company it is to be founded that whether a new customer will churn or not and therefore every algorithms used in this project are of classification type where the target variable is categorical.That is the output have value "YES" or "NO".

3.1 Logistic Regression

Regression is a statistical relationship between 2 or more variables where a change in independent variable is associated with a change in dependent variable.The independent variable is also called explanatory variables and dependent variable is also called response variable.When the response variable is categorical in nature we use logistic regression.Linear regression answers the question " how much" while logistic regression answers the question "will it happen or not?"

Logistic regression is one of the most simple and commonly used machine learning algorithms for 2 class classification.It is a statistical method to predict binary classes.Just like linear regression assumes that the data follows a linear function, in logistic regression models the data follows the sigmoid function. Output of the linear regression is discrete while output of the logistic regression is continuous.The sigmoid function also called logistic function gives an 'S'shaped curve that can take any real valued number and map it into a value between 0 and 1.

If the curve goes to infinity y predicted will be 1. If it goes to -infinity y predicted will be 0.If the output of the sigmoid function is more than 0.5 we can classify the outcome as Yes or 1, and if it is less than 0.5 we can classify it as NO or 0.If we try to use the linear regression cost function to generate $J(\theta)$ in a logistic regression problem , we would end up with a non convex function.A weird shaped graph with no-easy to find minimum global point.

$logit(pi) = 1/(1 + exp(-pi))$

$ln[(\pi)/1-\pi)] = \beta_0+\beta_1 x_1+...+\beta_k x_k$ here logit(pi) is the dependent variable and X is the independent variable.

Hosmerlemeshow test is a popular method to asses model fit.It is used frequently in risk prediction models(statistical model that aim to predict the probability of future events).It calculate the observed event rates match the expected event rates in population subgroups.

Logistic regression is used to make predictions about a categorical variables versus a continuous one. On the basis of the categories, Logistic Regression can be classified into three types:

- Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
- Multinomial:In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
- Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

Logistic regression belongs to the supervised machine learning model.With machine learning to estimate β coefficient of the model, the negative log likelihood is used as the loss function using the process of gradient descent to find global maximum.

Since logistic regression is a probability calculation ,based on the probability we need to decide what outcome should be.So there has to be a threshold value.To understand logistic regression we need to take consider of the odds of success.

odds(θ) = (probability of an event happening)/(Probability of an event not happening)

OR $\theta = \frac{P}{1-P}$

The values of odds range from 0 to ∞.The probability change from 0 to 1. Now the odds of success will be

$$\begin{aligned}
log[p(x)/1-p(x] &= \beta_0 + \beta_1 x \\
Exponetiating \quad both \quad sides & \\
e^{ln[p(x)/1-p(x)]} &= e^{\beta_0+\beta_1 x} \\
p(x)/1-p(x) &= e^{\beta_0+\beta_1 x} \\
let \quad Y &= e^{\beta_0+\beta_1 x} \\
Then \quad p(x)/1-p(x) &= Y \\
p(x) &= Y(1-p(x)) \\
&= Y - YP(x) \\
P(x) + YP(x) &= Y \\
P(x)[1+Y] &= Y \\
p(x) &= Y/1+Y \\
p(x) &= \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 xx}}
\end{aligned}$$

Hence The equation of the Sigmoid function will be

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 xx}}$$

$$p(x) = \frac{1}{1 + e^{-}\beta_0 + \beta_1 x}$$

Logistic regression and linear regression have many differences.The main difference is that linear regression is used to solve regression problem where the response variable is continuous in nature while logistic regression is used to solve classical problems where we predict discrete values as response variables.Linear regression helps to estimate the dependent variable when there is change in independent variable but logistic regression helps to calculate the possibility of a particular event taking place.Linear regression curve is a straight line but logistic regression curve is sigmoid curve.

Here are some of the applications of logistic regression.It helps to determine the kind of weather that can be expected.If we want to predict discrete variable for example whether it rains or not we use logistic regression.If we want to predict what is the temperature tomorrow we use linear regression.Another application of logistic is to identify the different components that are present in the image and helps categorize them.

Miroculus is a company that develops express blood test kits. Its goal is to identify diseases that are affected by genes, such as oncology diseases. The company entered into an agreement with Microsoft to develop an algorithm to identify the relationship between certain micro-RNA and genes.Algorithms such as logistic regression, support vector machine, and random forest were considered as models for finding the link between micro RNA and certain gene. Logistic regression was selected because it demonstrated the best results in speed and accuracy.
Toxic speech detection, topic classification for questions to support, and email sorting are examples where logistic regression shows good results.

The assumptions for Logistic regression are as follows:

- Independent observations Each observation is independent of the other. meaning there is no correlation between any input variables.
- Binary dependent variables:It takes the assumption that the dependent variable must be binary or dichotomous, meaning it can take only two values. For more than two categories softmax functions are used.
- Linearity relationship between independent variables and log odds: The relationship between the independent variables and the log odds of the dependent variable should be linear.
- No outliers: There should be no outliers in the data set.
- Large sample size: The sample size is sufficiently large.

3.2 Decision Tree

Decision Tree is a tree shaped diagram used to determine a course of action.Each branch of a tree represents a possible decision, occurrence or reaction. As it's name suggest, it is an algorithm that has shaped like tree which helps to make certain decision by monitoring the different attributes provided.Decision tree can be used both in classification and regression problem.In classification a classified tree will determine a set of logical "if-then" conditions to classify problems.For example discriminating between 3 types of flowers based on certain features.A regression tree is used when the target variable is numerical or continuous in nature.We fit a regression model to the target variable using each of the independent variables. each split is made based on the sum of squared error.Since the decision tree does not make any assumptions about data it is also known as CART(classification and regression tree).

Advantages of decision tree are

- It is simple to understand, interpret and visualize.
- Little effort required for data preparation.
- We can handle both numerical and categorical data.
- Non linear parameters don't affect it's performance.Even though a curve does not fit the data we can use decision tree for effective decision or prediction.

Now some of the disadvantages of the decision tree are

- Over fitting occurs when the tree is too complex or when algorithm captures noise or irrelevant data in the data set.This can lead to poor generalised performance on new data.That is the algorithm might not take right predictions if it is trained over the noisy data.
- The model can get unstable due to small variation in data.Small variation in data can lead to different tree structures which in turn can make things difficult.
- A highly complicated decision tree tends to have a low bias which makes it difficult for model to work with new data. decision tree can be biased towards features that appear earlier in the tree or have more splits, leading to sub-optimal performance.
- The algorithm is not suited for huge data sets with a multitude of important features. It has very limited expressiveness.Not well suited for

complex problems with continuous data.They work best with categorical data or discrete data and are less effective for problems that require more sophisticated modelling approach.

Here are some of the important terms that are in the decision tree

- Nodes: Leaf nodes carries the classification or the decision so it is the final end at the bottom The decision node has 2 or more branches this is where the breaking group into different parts are occurred and finally the most decision node is known as the root node.
- Entropy: Entropy is the measure of randomness or unpredictability in the data set.The more the classes in the data set the more the entropy is.On each step of decision tree entropy will decreases.At the leaf nodes we have got the distinct classes and there is no randomness at all so the entropy for these leaf nodes will remain zero.The formula for calculating entropy is

$$H_i = \sum_{k=1}^{n} P_i K log_2 P_i K$$

  P_iK is the probability of positive and negative class i at the particular node.

- Information gain : It is a measure of decrease in entropy after the data set is split.
  Entropy(T)- Entropy at node before split (Parent node)
  Entropy(T_v) - Entropy's after split(Child node), T-total no of split T_V -total no of instances after split

$$IG(T, A) \quad = \quad Entropy(T) - \sum \frac{T_v}{T} Entropy(T_v)$$

- Gini-Impurity : Calculates the purity of the split at nodes of the decision tree. Gini-impurity at the ith node is

$$G_i \quad = \quad 1 - \sum_{k=1}^{n} P_i k^2$$

  Unlike entropy gini impurity varies from 0 to 0.5. Node is pure when gini impurity is equal to 0.That is all instances are of same class.
  P_iK is te ratio of class K instances among the training instances in the ith node.

Working of a decision tree
Partitioning: It refers to the process of splitting the data set into subsets. The decision of making strategic splits greatly affects the accuracy of the tree. Many algorithms are used by the tree to split a node into sub-nodes which results in an overall increase in the clarity of the node with respect to the target variable. Various Algorithms like the chi-square and Gini index are used for this purpose and the algorithm with the best efficiency is chosen.

Pruning: This refers to the process wherein the branch nodes are turned into leaf nodes which results in the shortening of the branches of the tree. The essence behind this idea is that over fitting is avoided by simpler trees as most complex classification trees may fit the training data well but do an underwhelming job in classifying new values.

Selection of the tree: The main goal of this process is to select the smallest tree that fits the data due to the reasons discussed in the pruning section.

3.3 Random Forest

Random forest is an ensemble machine learning algorithm.It operates by building multiple decision tree's.The decision of the majority of the tree is chosen by the random forest as the final decision.Decision trees are highly sensitive to the training data which could result in high variance.Random forest or Random decision forest operates by considering multiple decision trees during the training phase.Random forest work for both classification and regression problems.Regression problems have continuous or numerical valued output variable.More the no:of trees in random forest, more will be the accuracy of the prediction.

Some of the advantages of the random forest are

- Use of multiple tree's reduce the risk of over fitting
- Runs efficiently on large database.for large data it produces highly accurate predictions.
- It estimates missing data.Random forest maintain high accuracy when a large proportion of data is missing.

Some of it's disadvantages are

- Since it builds numerous trees and we need to combine their output it require much computational power and also resources.

- Need more time for training since it combine a lot of decision trees to determine the class.
- It suffer while interpreting and fails to determine the importance of each attribute since it ensemble decision trees.

Random forest is called random since it uses 2 random process that is bootstrapping and random feature selection.Bootstrapping means instead of training on all the observations, each tree of random forest is trained on a subset of the observations.The chosen subset is called the bag, and the remaining are called out of bag samples.Bootstrapping ensures that we are not using the same data for every tree so in a way it helps our model to be less sensitive to the original training data.

The random feature selection helps to decreases the correlation between the trees. if we use every feature most of our trees will have the same decision nodes and they will act very similarly.Multiple trees are trained on different bags and later the results from all the trees are aggregated.

Here are some of the application of Random forest

- In banking Random forest is used to predict fraudulent customers.
- It is used to analyze symptoms of patient and detecting disease.
- In E-commerce the recommended based on customer's activity.

Since every attributes are not taken into consideration Random forest have diversity.Each tree does not consider all the features.The feature space is reduced for random forest.In random forest stability occurs since the result is based on majority voting or averaging.
Since the random forest combines more than one trees to predict the class of the data, there is a chance for some decision trees to predict the correct output, while others will predict the wrong one. But together, all the trees predict the correct output. Therefore, there must be some assumptions satisfied by a better random classifier.

- There should be some actual values in the feature variable of the data set so that the classifier can predict accurate results rather than a guessed result.
- The correlation between the predictions of each tree must be very low.

3.4 Support Vector Machine

Support vector machine or SVM is one of the supervised machine learning algorithm which is used for both classification and regression.The aim of SVM is to fix a decision boundary which make an n-dimensional data into classes so that a new data point can be correctly classified.This best decision boundary is known as a hyperplane.To create a hyperplane we need to choose extreme points or vectors.For which SVM is used. SVM algorithm can be used for face detection image classification text categorization etc.

There are 2 types of SVM

- Linear svm: If we can classify a data into 2 classes using a straight line, then such data set is termed as linearly separable data and linear SVM is used for such linearly separable data.The classifier is called linear SVM classifier.
- Non linear svm: This is used for non-linearly separated data.That is for the data where a straight line cannot be used to classify it. The classifier for non-linearly separated data is called non-linear SVM classifier.

Hyper plane: From the multiple decision boundaries to segregate the classes in n-dimensional space we have to find the best decision boundary that helps to classify the data points which in termed as hyperplane of SVM.Dimension of hyperplane depend on the features in the data set.That is the hyperplane will be a straight line if there are 2 features and it will be 2 dimensional plane if there are 3 features.Hyperplane having maximum margin that is having maximum distance between the data points is said to be optimal hyperplane.The data points or vectors which are close to hyperplane that is the vectors which affect the position of hyperplane are termed as support vector.They are called support vector because these vectors support the hyperplane.

Some of the advantages of support vector machine is

- SVM works good when we can find an optimal hyperplane as decision boundary that is if there is a clear margin which separate data set into classes.
- SVM can be very effective in high dimensional spaces and it is also memory efficient comparing to other models.
- When the dimension is greater than the number of samples, SVM can be more efficient.

Some of it's disadvantages are:

- SVM algorithm does not perform well when there are more noise in the data.That means if there are overlapping target classes.If the no:of features for each data points are more than the training data samples, svm will under perform.
- This SVM algorithm is not suitable for large data sets.
- There is no probabilistic explanation to the classifier since it works by putting data points above and below classifying hyperplane.

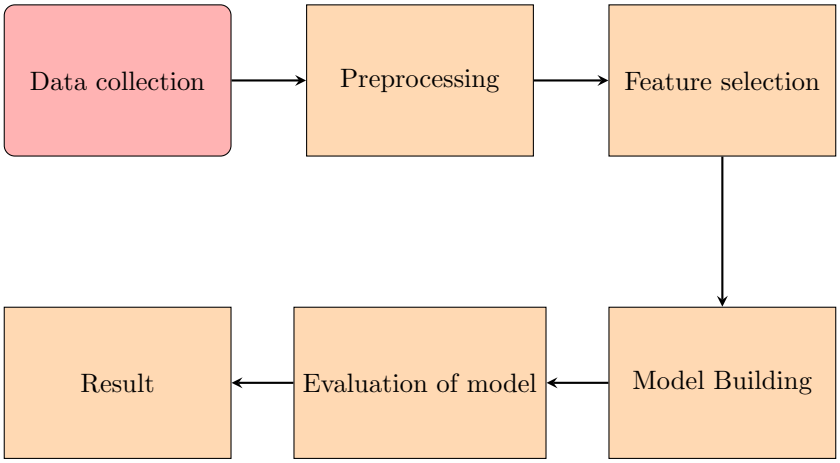
In the analysis of the data it is noted that all of the variables are not needed.For feature selection we uses 2 methods known ad Principal component analysis and Boruta algorithm.
There are 3 types of feature selection.

- Filter method: In this method features are filtered based on general characteristics of the data set.
- Wrapper method: In wrapper method the feature selection algorithm exist as a wrapper around the predictive model algorithm and uses the same model to select features.
- Embedded method: This combine the qualities of filter and wrapper method.

In feature selection, one of the method used is principal component analysis.Principal component analysis (PCA) is a popular technique for analyzing data set which is large and having high dimension.By saving maximum information this PCA increases the interpretability of data.It also make possible to visualize the multidimensional data.This PCA linearly transform the data into a new coordinate system where we can describe the variation in the data using less dimension than the initial data. Many studies use the first two principal components and hence can plot the data in two dimensions.

Another feature selection method used in this analysis is Boruta algorithm.Boruta is a wrapper feature selection algorithm around random forest. This algorithm is important in the situation where we want to build a model for data set which is having several variables.Boruta follows an all-relevant feature selection method where it captures all features which are in some circumstances relevant to the outcome variable. In contrast, most of the traditional feature selection algorithms follow a minimal optimal method where they make the data having less variables. having minimum error.

4 ANALYSIS



4.1 Data collection

For this analysis data of a telecom company had been from kaggle website.The raw data has 7043 rows and 21 columns.Each row in the data represent a customer and each column contains customer attributes.Here the last column named "Churn" is the target variable or the dependent variable for our model.

The first column gives the customer id.Second column gives the gender of the customer that is whether the customer is male or female.60% of the customer in the data are female and remaining 40% are male.third attribute is whether the customer is a senior citizen or not and fourth and fifth attribute tell us the whether the customer has a partner and Dependents or not respectively.No of months the customer has stayed with the company known as tenure is given on the next column and sixth column tells whether the customer has a phone service or not.seventh attribute says whether the customer has a multiple line or not.And it is categorized into YES ,NO and NO PHONE SERVICE.42% of the customer has multiple lines,48% of them does not have it.remaining 10% does not have the phone service.

Customer's internet service provider is categorized into DSL(digital subscriber line),Fiber optic and No internet service in the eighth column.Among the customer's 34% of them follow DSL internet service.44% of them follow Fiber optic and 22% of them does not have internet services.Whether the customer have online service is divided into 3 categories.29% of them have online services 50% of them does not have that and 21% of them dose not even have the internet services.another column describing whether he or sh have a online backup plan or not. Among the 7043 customers 34% of them have a backup plan and 44% of them does not have and 22% of them does not have internet

services.Device protection of the customer are also given.Next 3 columns describe about if the customer have tech support ,streaming TV and streaming movies or not.29% of them have tech tech support 49% of them does not have and 22% of them does not have internet services.38% of them have streaming TV 40 % of them does not have and 22% of them does not have the internet services.39% of them have streaming movies ,40% of them does not have and 22% of them does not have the internet services.

The contract term of the customer is categorized as month to moth contract,one year and two year contract.55% of them have a month to month contract 24% of them have a one year contract and 21% of them have a 2 year contract.This data also gives the customer have paper less billing or not. Next attribute is the type of payment method of the specific customer.34% of them do their payment through electronic check 23% of them by the mailed check 44% of them by bank transfer and credit card.The amount charged to the customer monthly and total amount charged to the customer is also given as attributes.And finally the target variable gives whether the customer has churned or not.

4.2 Prepossessing

Before going directly to the analysis some pre-processing techniques has been done in R.Summary of this data gives the Maximum,Minimum,Mean,Median,First quartile and Third quartile of the numeric type data.In this data Tenure, Monthly charges and Total charges are the only 3 attribute which is of numeric type.And their summary is given below.

Summary			
Description	Tenure	Monthly charges	Total charges
Minimum	0	18.25	18.8
First quartile	9	35.50	401.4
Median	29	70.35	1397.5
Mean	32.37	64.76	2283.3
Third quartile	55	89.85	3794.7
Maximum	72	118.75	8684.8

Then we eliminate the missing values in the data.There are total 11 missing values and all of them are on the column of total charges. After eliminating the missing values when we continue to clean the data, in the column of multiple lines by changing "no phone service" to "no" we divide the customer into 2 groups one having multiple lines and other not having multiple lines.Similarly

in the columns of online security,online backup,device protection,tech support streaming TV and streaming movies by changing the "no internet service" to "no" we divide the data into 2 groups.Now we group the customer's to 5 groups based on the tenure.We categorize customers into group having 0-12 month,12-24 month,24-48 month,48-60 month,and more than 60 month.And we create a new attribute named tenure group and eliminate the attribute tenure.We also eliminate one unimportant attribute that is customer id.

4.3 Feature selection

After completing the cleaning of the data we go on to the feature selection part.For selecting the features 2 algorithms Principal component analysis and boruta algorithm is used. PCA is used as a feature selection method .But since the correlation between the attributes are very less we reach in a conclusion that we need to take 17 attributes among the 20 to make the principal component a good representative of the attributes.So PCA is not a suitable feature selection method for this data.

```
                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
SS loadings       1.00   1.00   1.00   1.00   1.00   1.00   1.00
Proportion Var    0.05   0.05   0.05   0.05   0.05   0.05   0.05
Cumulative Var    0.05   0.10   0.15   0.20   0.25   0.30   0.35
                Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13
SS loadings       1.00   1.00    1.00    1.00    1.00    1.00
Proportion Var    0.05   0.05    0.05    0.05    0.05    0.05
Cumulative Var    0.40   0.45    0.50    0.55    0.60    0.65
                Comp.14 Comp.15 Comp.16 Comp.17 Comp.18 Comp.19
SS loadings        1.00    1.00    1.00    1.00    1.00    1.00
Proportion Var     0.05    0.05    0.05    0.05    0.05    0.05
Cumulative Var     0.70    0.75    0.80    0.85    0.90    0.95
                Comp.20
SS loadings        1.00
Proportion Var     0.05
Cumulative Var     1.00
```

Figure 1: PCA

By Boruta algorithm it is found that the attribute gender is an unimportant attribute to the analysis and thus we eliminate the "gender" from the data.

```
> boruta<-Boruta(Churn~.,data=telco.churn,doTrace=2,maxRuns=10(
 1. run of importance source...
 2. run of importance source...
 3. run of importance source...
 4. run of importance source...
 5. run of importance source...
 6. run of importance source...
 7. run of importance source...
 8. run of importance source...
 9. run of importance source...
 10. run of importance source...
 11. run of importance source...
After 11 iterations, +1.8 mins:
 confirmed 18 attributes: Contract, Dependents, DeviceProtectic
lyCharges and 13 more;
 rejected 1 attribute: gender;
 no more attributes left.
```

Figure 2: Boruta

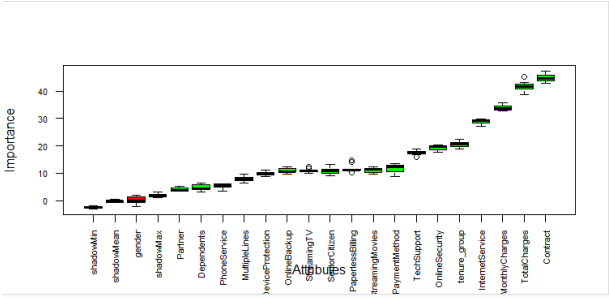


Figure 3: Feature importance using boruta algorithm

4.4 Model Building

First we apply logistic regression algorithm.While applying the algorithm we initially divide the data to testing and training set.Training data is used to train the model and testing data is used to predict the customer churn.We divide the data into 8:2 ratio.That is 80% of the original data is training data and 20% of the original data is testing data.Here the data is having 7043 rows or customers.Among that 4923 customers are in training data and 2109 customers are in testing data.GLM(Generalised linear model) function which is a is a class of regression models that supports non-normal distributions is used to apply logistic regression. For logistic regression the model is

$$log(p/1-p) \quad = \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...$$

where P is the probability that the class 1 ocurs.β_0 is the intercept and β_1, β_2...are the beta coefficients which are the estimates of beta.P-value is used to check whether the attribute is significant for the prediction of the target variable or not.Here the target variable is churn and hence p-value here is used to predict

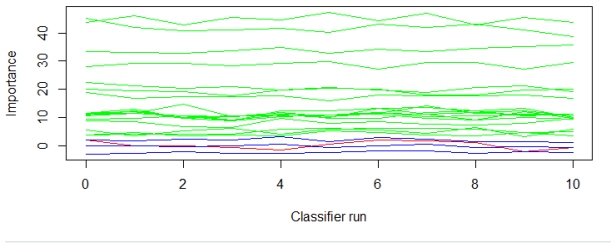


Figure 4: Feature importance using boruta algorithm

whether the particular attribute is significant or not in predicting the customer churn.The null hypothesis $H_0 : \beta_j = 0$ and $H_1 : \beta_j \neq 0$.If the p-value is less than 0.05 then we reject the null hypothesis and the alternate hypothesis $H_1 : \beta_j \neq 0$ is being accepted and the β - coefficients corresponding to the jth attribute is non zero.That means the attribute have a significant contribution for building the model and predicting the output.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.012e+00  5.144e-01   3.911 9.17e-05 ***
gender            2.543e-02  7.786e-02   0.327   0.7440
SeniorCitizen     2.188e-01  1.018e-01   2.150   0.0316 *
Partner          -8.826e-02  9.284e-02  -0.951   0.3418
Dependents       -1.715e-01  1.079e-01  -1.589   0.1120
PhoneService     -1.175e+00  1.711e-01  -6.864 6.70e-12 ***
MultipleLines     1.025e-01  9.833e-02   1.043   0.2971
InternetService   5.303e-02  7.595e-02   0.698   0.4850
OnlineSecurity   -5.464e-01  1.035e-01  -5.280 1.29e-07 ***
OnlineBackup     -3.088e-01  9.535e-02  -3.238   0.0012 **
DeviceProtection -1.830e-01  9.715e-02  -1.883   0.0596 .
TechSupport      -5.379e-01  1.047e-01  -5.139 2.76e-07 ***
StreamingTV      -9.062e-02  1.040e-01  -0.871   0.3837
StreamingMovies   4.334e-02  1.038e-01   0.418   0.6763
Contract         -8.706e-01  9.020e-02  -9.652  < 2e-16 ***
PaperlessBilling  3.699e-01  8.822e-02   4.193 2.76e-05 ***
PaymentMethod     8.578e-02  4.191e-02   2.047   0.0407 *
MonthlyCharges    4.145e-02  2.973e-03  13.941  < 2e-16 ***
TotalCharges     -3.273e-04  3.359e-05  -9.744  < 2e-16 ***
tenure_group     -1.777e-01  3.730e-02  -4.764 1.90e-06 ***
---
```

Figure 5: Logistic regression

From the result we get that the attributes Senior citizen,Phone service, Online security,Online backup,Tech support,Contract,Paperless Billing,Payment method,Monthly charges Total charges and Tenure group are the attributes which are significant.And the remaining Partner,Dependents,Multiple lines,Internet service,Device protection,Streaming TV and Streaming movies are not significant in the model.The accuracy of this logistic model is 0.10648 and that is a very less accuracy to fit a model. So we go on to other algorithm models.

To apply decision tree algorithm we divide the data to testing and training

in the similar way.70% of data is given to training data and 30% of them to testing data.To draw the decision tree we have to change all the variables having character type to factor type.

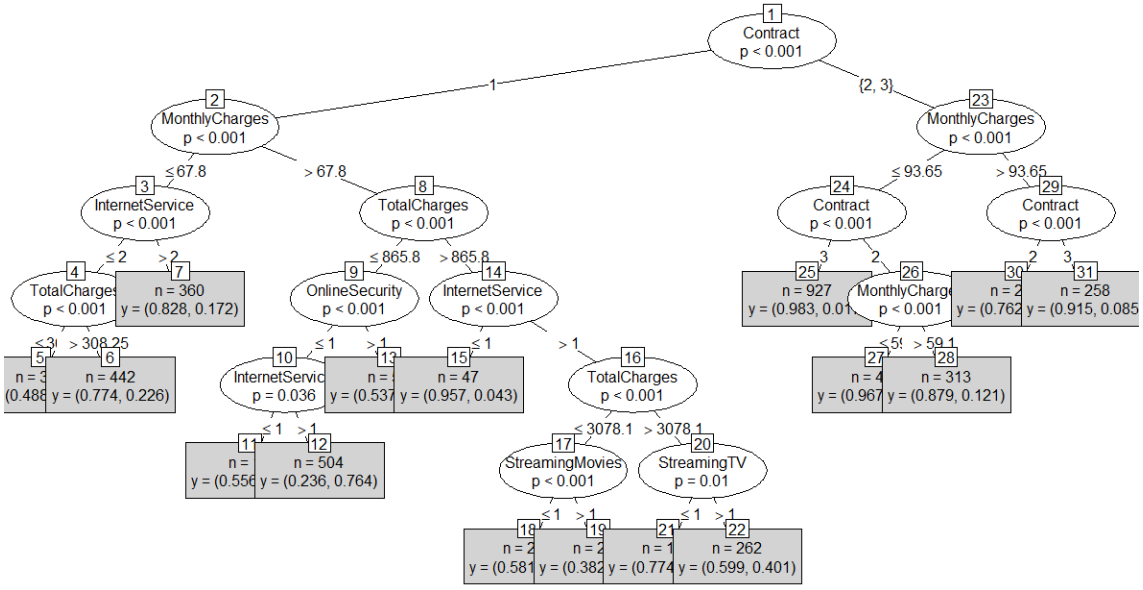


Figure 6: Decision tree

Random forest is applied by giving the maximum number of trees as 500.So it predict the class of a given customer based on the majority results of this 500 trees.Here the out of bag estimate of error rate is 21.31%.The out of bag error(OOB-error) is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample.OOB error estimate of 21.31% is not much significant and the confusion matrix of this model is given by

$$\begin{array}{cc} 3198 & 415 \\ 634 & 675 \end{array}$$

The confusion matrix of every model be in the form

$$\begin{array}{cc} True\ positive(TP) & False\ positive(FP) \\ False\ negative(FN) & True\ postive(TN) \end{array}$$

The accuracy is found from confusion matrix by the equation

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

Here we get the accuracy as 0.790 which is a good one compared to the logic model and decision tree.

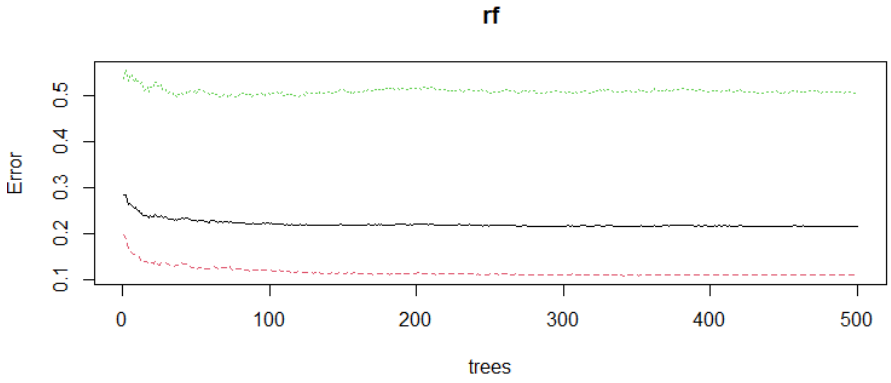


Figure 7: Random Forest

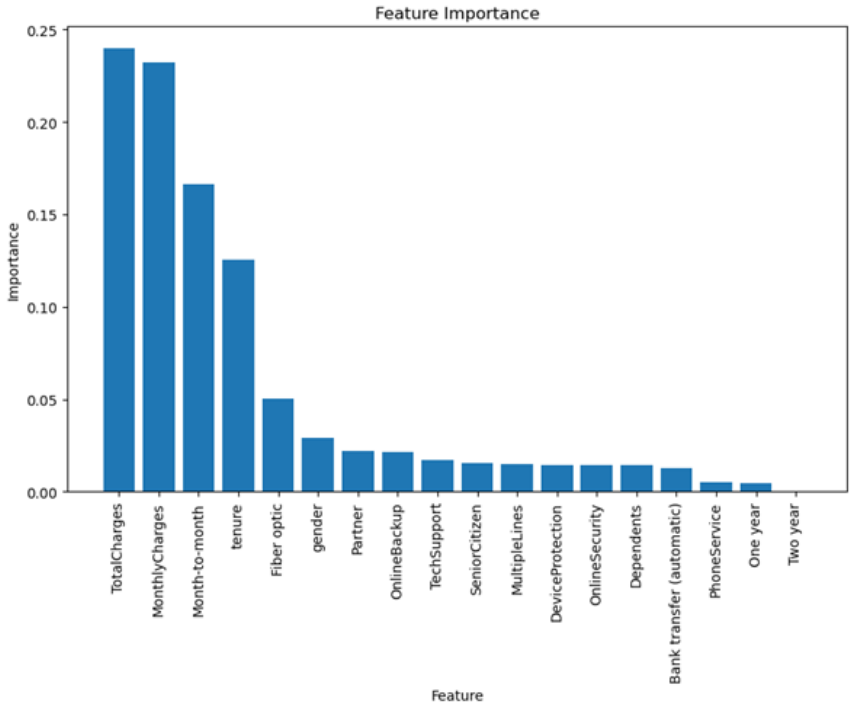
Finally support vector machine is applied to the data.Since the data is linearly separable data we uses linear kernel here.The model gives 2277 support vectors that means there are 2277 data points which are very close to the hyperplane and which determine the position of hyperplane.Large no of support vectors is often a sign of over fitting.The confusion matrix of the predicted model is

1393	155
264	296

And the accuracy of this model will be around 0.8008 that is 80.08% which is a good accuracy.

5 CONCLUSION

It is crucial to observe the historical behavioural patterns of customers for the churn prediction in telecom sector.This study evaluated the capability of the SVM as a base classifier for some particular data.Initially every analysis should be conducted after feature selection.It is not necessary to select every feature for the analysis.Every feature have an importance in building the model and predicting the future.The feature importance of each attribute in the data is given below.



First logistic regression is applied which ends having the model of accuracy 10.64%.Then the algorithm decision tree is applied which results in a better accuracy of 73.79%.Then group of such decision tree that is the random forest algorithm is used which result in 79.0%.And at last the Support vector machine is applied which gives the better accuracy than all algorithms used here.The attributes used in this data have very less correlation between the remaining one.That is the dependency of the variables each other is small.Hence this project conclude on the inference that for this data SVM having an accuracy of 80.08% is the best model.

Model	Accuracy	Percentage
Logistic regression	0.10648	10.64%
Decision tree	0.7379	73.79%
Random Forest	0.790	79.0%
Support Vector Machine	0.8008	80.08%

By saying the support vector machine shows the high accuracy compared to other models we cannot say that support vector machine is the best model to analyse the data.We can also use some ensembling methods which perform by combineing more than one algorithm and thus give maximum accuracy.

References

[1] Ahmad, A.K., Jafar, A. & Aljoumaa, K. **Customer churn prediction in telecom using machine learning in big data platform***J Big Data 6, 28 (2019)*

[2] Liu, Y., Fan, J., Zhang, J. et al. **Research on telecom customer churn prediction based on ensemble learning**.*J Intell Inf Syst (2022)*

[3] Amin, A., Al-Obeidat, F., Shah, B. et al.**Just-in-time customer churn prediction in the telecommunication sector**. *J Supercomput 76, 3924–3948 (2020)*

[4] Tariq, M.U., Babar, M., Poulin, M. and Khattak, A.S. (2022), **Distributed model for customer churn prediction using convolutional neural network**, *Journal of Modelling in Management, Vol. 17 No. 3, pp. 853-863.*

[5] Suh, Y. **Machine learning based customer churn prediction in home appliance rental business**. *J Big Data 10, 41 (2023).*

[6] Castanedo, F. (2014).**Using Deep Learning to Predict Customer Churn in a Mobile Telecommunication Network**

[6] Chih-Ping Wei, I-Tang Chiu, **Turning telecommunications call details to churn prediction: a data mining approach**, *Expert Systems with Applications*, Volume 23, Issue 2,

[7] Muneer, Amgad & Ali, Rao & Alghamdi, Amal & Mohd Taib, Shakirah & Almaghthwi, Ahmed & Ghaleb, Ebrahim. (2022). Predicting customers churning in banking industry: A machine learning approach. *Indonesian Journal of Electrical Engineering and Computer Science*. 26. 539-549. 10.11591/ijeecs.v26.i1.pp539-549.

[8] Ascarza, E., Iyengar, R., & Schleicher, M. (2016).**The Perils of Proactive Churn Prevention Using Plan Recommendations: Evidence from a Field Experiment**.*Journal of Marketing Research*, 53(1), 46–60

[9] J. David Nuñez-Gonzalez, Manuel Graña, Bruno Apolloni,Reputation features for trust prediction in social networks,*Neurocomputing*,Volume 166,2015,

[10] M.A.H. Farquad, Vadlamani Ravi, S. Bapi Raju,Churn prediction using comprehensible support vector machine: An analytical CRM application,*Applied Soft Computing*,Volume 19,2014,

[11] Chitra, K. and B. Subashini. "Customer Retention in Banking Sector using Predictive Data Mining Technique." (2011).