A Project Report

on

Customer Churn Prediction Using Machine Learning

Submitted in partial fulfilment of the requirements for the award of the Degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

G. SRIDATTA ANIRUDHA NIVAS (18FE1A0536)

CH. TEJA VENKAT (18FE1A0514)

J. HEMANTH PURNA KUMAR (19FE5A0503)

G. PRUDHVI (18FE1A0538)

Under the guidance of

Dr. V. SUJATHA, Ph.D.

Professor

Department of Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VIGNAN'S LARA INSTITUTE OF TECHNOLOGY & SCIENCE

(Affiliated to Jawaharlal Nehru Technological University Kakinada, Kakinada)

(An ISO 9001:2015 Certified Institution, Approved by AICTE)

Vadlamudi, Guntur Dist., Andhra Pradesh-522213

June - 2022.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
VIGNAN'S LARA INSTITUTE OF TECHNOLOGY & SCIENCE
(Affiliated to Jawaharlal Nehru Technological University Kakinada, Kakinada)
(An ISO 9001:2015 Certified Institution, Approved by AICTE)
Vadlamudi, Guntur Dist., Andhra Pradesh-522213

CERTIFICATE

This is to certify that the project report entitled **"CUSTOMER CHURN PREDICTION USING MACHINE LEARNING"** is a bonafide work done by **G. SRIDATTA ANIRUDHA NIVAS (18FE1A0536), CH. TEJA VENKAT (18FE1A0514), J. HEMANTH PURNA KUMAR (19FE5A0503), G. PRUDHVI (18FE1A0538)** under my guidance and submitted in fulfilment of the requirements for the award of the degree of Bachelor of Technology in **COMPUTER SCIENCE AND ENGINEERING** from **JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY KAKINADA, KAKINADA**. The work embodied in this project report is not submitted to any University or Institution for the award of any degree or diploma.

Project Guide
Dr. V. Sujatha, Ph.D.
Professor

Head of the Department
Dr. K. Venkateswara Rao, Ph.D.
Professor

External Examiner

DECLARATION

We hereby declare that the project report entitled "**Customer Churn Prediction Using Machine Learning**" is a record of an original work done by us under the guidance of **Dr. V. SUJATHA**, Professor of Computer Science and Engineering and this project report is submitted in the fulfilment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering. The results embodied in this project report are not submitted to any other University or Institute for the award of any Degree or Diploma.

Project Members	**Signature**
G. SRIDATTA ANIRUDHA NIVAS (18FE1A0536)	_____
CH. TEJA VENKAT (18FE1A0514)	_____
J. HEMANTH PURNA KUMAR (19FE5A0503)	_____
G. PRUDHVI (18FE1A0538)	_____

Place: Vadlamudi

Date :

ACKNOWLEDGMENT

The satisfaction that accompanies with the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success.

We are grateful to **Dr. V. SUJATHA**, Professor, Department of Computer Science and Engineering for guiding through this project and for encouraging right from the beginning of the project till successful completion of the project. Every interaction with her was an inspiration.

We thank **Dr. K. VENKATESWARA RAO**, Professor & HOD, Department of Computer Science and Engineering for support and Valuable suggestions.

We also express our thanks to **Dr. K. PHANEENDRA KUMAR**, Principal, Vignan's Lara Institute of Technology & Science for providing the resources to carry out the project.

We also express our sincere thanks to our beloved Chairman **Dr. LAVU RATHAIAH** for providing support and stimulating environment for developing the project.

We also place our floral gratitude to all other teaching and lab technicians for their constant support and advice throughout the project.

Project Members

G. SRIDATTA ANIRUDHA NIVAS (18FE1A0536)
CH. TEJA VENKAT (18FE1A0514)
J. HEMANTH PURNA KUMAR (19FE5A0503)
G. PRUDHVI (18FE1A0538)

TABLE OF CONTENTS

ABSTRACT

In the telecom sector, a huge volume of data is being generated on a daily basis due to a vast client base. Decision makers and business analysts emphasized that attaining new customers is costlier than retaining the existing ones. Business analysts and customer relationship management (CRM) analysers need to know the reasons for churn customers, as well as, behaviour patterns from the existing churn customers' data. This paper proposes a churn prediction model that uses classification, as well as, clustering techniques to identify the churn customers and provides the factors behind the churning of customers in the telecom sector. Feature selection is performed by using information gain and correlation attribute ranking filter. The proposed model first classifies churn customer's data using classification algorithms, in which the Gradient Boosting algorithm performed well. Creating effective retention policies is an essential task of the CRM to prevent churners. After classification, the proposed model segments the churning customer's data by categorizing the churn customers in groups using cosine similarity to provide group-based retention offers. This paper also identified churn factors that are essential in determining the root causes of churn. By knowing the significant churn factors from customers' data, CRM can improve productivity, recommend relevant promotions to the group of likely churn customers based on similar behaviour patterns, and excessively improve marketing campaigns of the company. The proposed churn prediction model is evaluated using metrics, such as accuracy, precision, recall, f-measure, and receiving operating characteristics (ROC) area. The results reveal that the proposed churn prediction model produced better churn classification using the Gradient Boosting algorithm.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

RF	Random Forest
GBM	Gradient Boosting Machine
QOE	Quality Of Experience
CRM	Customer Relationship Management
ML	Machine learning
ROC	Receiving Operating Characteristics

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

In the present world, a huge volume of data is being generated by telecom companies at an exceedingly fast rate. There is a ultimate goal of telecom companies is to maximize their profit and stay alive in a competitive market place [1]. A customer churn happens when a vast percentage of clients are not satisfied with the services of any telecom company. It results in service migration of customers who start switching to other service providers.

There are many reasons for churning. Unlike post-paid customers, prepaid customers are not bound to a service provider and may churn at any time. Churning also impacts the overall reputation of a company which results in its brand loss. A loyal customer, who generates high revenue for the company, gets rarely affected by the competitor companies. Such customers maximize the profit of a company by referring it to their friends, family members and colleagues. Telecom companies consider policy shift when the number of customers drops below a certain level which may result in a huge loss of revenue [3].

Churn prediction is vital in the telecom sector as telecom operators have to retain their valuable customers and enhance their Customer Relationship Management (CRM) administration [5], [6]. The most challenging job for CRM is to retain existing customers [7]. Due to the saturated and competitive market, customers have the option to switch to other service providers. Telecom companies have developed procedures to identify and retain their customers as it is less expensive than attracting the new ones [5]. This is due to the cost involved in advertisements, workforce, and concessions which can scale up to almost five to six times than retaining existing customers [3]. Small attention is needed for identifying the existing churn customers, which can help in overturning the situation. The requirement of retaining customers' needs to develop an accurate and high performance model for identifying churn customers. The proposed model should have the capability to

identify churn customers and then find the reasons behind churn to avoid loss of customers and provide measures to retain them. In addition, it should employ techniques to predict when such a situation is going to arise in the future.

1.2 PROBLEM STATEMENT

In the competitive Telecom industry, public policies and standardization of mobile communication allow customers to easily switch over from one carrier to another, resulting in a strained fluidic market. Churn prediction, or the task of identifying customers who are likely to discontinue use of a service, is an important and lucrative concern of the Telecom industry. The aim of this thesis is to study and analyse customer churn prediction based on mobile data usage volumes with respect to QoE and users' perspective with the help of Churn Prediction model using machine learning techniques.

1.3 OBJECTIVE

A huge volume of data is being generated on a daily basis due to vast client base. Business Analysts and Customer Relationship Management (CRM) need to know the main reason for Churn customers , as well as behavior patterns from existing churn customers data. Here we propose a churn prediction model that uses classification and clustering techniques to identify churn customers and provides factors behind the churning of customers in the telecom sector.

1.4 MACHINE LEARNING

Machine learning (**ML**) is the study of computer algorithms that can improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. Some implementations of machine learning use data and neural networks in a way that mimics the working of a biological brain. In its application across business problems, machine learning is also referred to as predictive analytics.

Advantages of Machine learning

It is automatic: In machine learning, the whole process of data interpretation and analysis is done by computer. No men intervention is required for the prediction or interpretation of data. The whole process of machine learning is machine starts learning and predicting the algorithm or program to give the best result. One of the examples in the Google home that detect the voice and them accordingly finds out the result that the user wants, and antivirus software detects the virus of the computer and fixes it.

It is used in various fields: Machine learning is used in various fields of life like education, medicine, engineering, etc. From a very small application to very big and complicated structured machines that help in the prediction and analysis of data. It not only becomes the healthcare provider but also provides more personal services to the potential customer.

It can handle varieties of data: Even in an uncertain and dynamic environment, it can handle a variety of data. It is multidimensional as well as a multitasker.

Scope of advancement: As humans after gaining experience improve themselves in the same way machine learning improve themselves and become more accurate and efficient in work. This led to better decisions. For example, in the weather forecast, the more data. And experience the machine gets the more advanced forecast it will provide.

Can identify trends and patterns: A machine can learn more when it gets more data and since it gets more data it also learns the pattern and trend for example for a social networking site like Facebook people surf and browses several data and their interest is recorded and understand the pattern and shows the same or similar trend to them to keep their interest within the same app. In this way machine learning help in identifying trends and patterns.

Considered best for Education: Machine learning is considered best for education as education is dynamic and nowadays smart classes, distance learning, and e-learning for students have increased a lot. Smart machine learning will act as a teacher and keep students updated with the current scenario of the world. The same thing happens in shopping or e-business people need to remain updated therefore they are shown the current trends of the world.

MACHINE LEARNING CHALLENGES

1. Poor Quality of Data

Data plays a significant role in the machine learning process. One of the significant issues that machine learning professionals face is the absence of good quality data. Unclean and noisy data can make the whole process extremely exhausting. We don't want our algorithm to make inaccurate or faulty predictions. Hence the quality of data is essential to enhance the output. Therefore, we need to ensure that the process of data preprocessing which includes removing outliers, filtering missing values, and removing unwanted features, is done with the utmost level of perfection.

2. Underfitting of Training Data

This process occurs when data is unable to establish an accurate relationship between input and output variables. It simply means trying to fit in undersized jeans. It signifies the data is too simple to establish a precise relationship. To overcome this issue:

- Maximize the training time
- Enhance the complexity of the model
- Add more features to the data
- Reduce regular parameters
- Increasing the training time of model

3. Overfitting of Training Data

Overfitting refers to a machine learning model trained with a massive amount of data that negatively affect its performance. It is like trying to fit in Oversized jeans. Unfortunately, this is one of the significant issues faced by machine learning professionals. This means that the algorithm is trained with noisy and biased data, which will affect its overall performance. Let's understand this with the help of an example. Let's consider a model trained to differentiate between a cat, a rabbit, a dog, and a tiger. The training data contains 1000 cats, 1000 dogs, 1000 tigers, and 4000 Rabbits. Then there is a considerable probability that it will identify the cat as a rabbit. In this example, we had a vast amount of data, but it was biased; hence the prediction was negatively affected.

We can tackle this issue by:

- Analysing the data with the utmost level of perfection
- Use data augmentation technique
- Remove outliers in the training set
- Select a model with lesser features

4. Machine Learning is a Complex Process

The machine learning industry is young and is continuously changing. Rapid hit and trial experiments are being carried on. The process is transforming, and hence there are high chances of error which makes the learning complex. It includes analyzing the data, removing data bias, training data, applying complex mathematical calculations, and a lot more. Hence it is a really complicated process which is another big challenge for Machine learning professionals.

5. Lack of Training Data

The most important task you need to do in the machine learning process is to train the data to achieve an accurate output. Less amount training data will produce inaccurate or too biased predictions. Let us understand this with the help of an example. Consider a machine learning algorithm similar to training a child. One day you decided to explain to a child how to distinguish between an apple and a watermelon. You will take an apple and a watermelon and show him the difference between both based on their color, shape, and taste. In this way, soon, he will attain perfection in differentiating between the two. But on the other hand, a machine-learning algorithm needs a lot of data to distinguish. For complex problems, it may even require millions of data to be trained. Therefore we need to ensure that Machine learning algorithms are trained with sufficient amounts of data.

6. Slow Implementation

This is one of the common issues faced by machine learning professionals. The machine learning models are highly efficient in providing accurate results, but it takes a tremendous amount of time. Slow programs, data overload, and excessive requirements usually take a lot of time to provide accurate results. Further, it requires constant monitoring and maintenance to deliver the best output.

7. Imperfections in the Algorithm When Data Grows

So you have found quality data, trained it amazingly, and the predictions are really concise and accurate. Yay, you have learned how to create a machine learning algorithm!! But wait, there is a twist; the model may become useless in the future as data grows. The best model of the present may become inaccurate in the coming Future and require further rearrangement. So you need regular monitoring and maintenance to keep the algorithm working. This is one of the most exhausting issues faced by machine learning professionals.

1.5 APPLICATIONS OF MACHINE LEARNING

1. Image Recognition

Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, Automatic friend tagging suggestion

Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's face detection and recognition algorithm.

It is based on the Facebook project named "**Deep Face**," which is responsible for face recognition and person identification in the picture.

2. Speech Recognition:

While using Google, we get an option of "Search by voice," it comes under speech recognition, and it's a popular application of machine learning.

Speech recognition is a process of converting voice instructions into text, and it is also known as "Speech to text", or "Computer speech recognition." At present, machine learning algorithms are widely used by various applications of speech recognition. Google assistant, Siri, Cortana, and Alexa are using speech recognition technology to follow the voice instructions.

3. Traffic prediction:

If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions.

It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways:

- Real Time location of the vehicle form Google Map app and sensors
- Average time has taken on past days at the same time.

Everyone who is using Google Map is helping this app to make it better. It takes information from the user and sends back to its database to improve the performance.

4. Product recommendations:

Machine learning is widely used by various e-commerce and entertainment companies such as Amazon, Netflix, etc., for product recommendation to the user. Whenever we search for some product on Amazon, then we started getting an advertisement for the same product while internet surfing on the same browser and this is because of machine learning.

Google understands the user interest using various machine learning algorithms and suggests the product as per customer interest.

As similar, when we use Netflix, we find some recommendations for entertainment series, movies, etc., and this is also done with the help of machine learning.

5. Self-driving cars:

One of the most exciting applications of machine learning is self-driving cars. Machine learning plays a significant role in self-driving cars. Tesla, the most popular car manufacturing company is working on self-driving car. It is using unsupervised learning method to train the car models to detect people and objects while driving.

6. Email Spam and Malware Filtering:

Whenever we receive a new email, it is filtered automatically as important, normal, and spam. We always receive an important mail in our inbox with the important symbol and spam emails in our spam box, and the technology behind this is Machine learning. Below are some spam filters used by Gmail:

- Content Filter

- Header filter
- General blacklists filter
- Rules-based filters
- Permission filters

Some machine learning algorithms such as Multi-Layer Perceptron, Decision tree, and Naïve Bayes classifier are used for email spam filtering and malware detection.

7. Virtual Personal Assistant:

We have various virtual personal assistants such as Google assistant, Alexa, Cortana, Siri. As the name suggests, they help us in finding the information using our voice instruction. These assistants can help us in various ways just by our voice instructions such as Play music, call someone, Open an email, Scheduling an appointment, etc.

These virtual assistants use machine learning algorithms as an important part.

These assistant record our voice instructions, send it over the server on a cloud, and decode it using ML algorithms and act accordingly.

8. Online Fraud Detection:

Machine learning is making our online transaction safe and secure by detecting fraud transaction. Whenever we perform some online transaction, there may be various ways that a fraudulent transaction can take place such as fake accounts, fake ids, and steal money in the middle of a transaction. So to detect this, Feed Forward Neural network helps us by checking whether it is a genuine transaction or a fraud transaction.

For each genuine transaction, the output is converted into some hash values, and these values become the input for the next round. For each genuine transaction, there is a specific pattern which gets change for the fraud transaction hence, it detects it and makes our online transactions more secure.

9. Stock Market trading:

Machine learning is widely used in stock market trading. In the stock market, there is always a risk of up and downs in shares, so for this machine learning's long short term memory neural network is used for the prediction of stock market trends.

10. Medical Diagnosis:

In medical science, machine learning is used for diseases diagnoses. With this, medical technology is growing very fast and able to build 3D models that can predict the exact position of lesions in the brain.

It helps in finding brain tumors and other brain-related diseases easily.

11. Automatic Language Translation:

Nowadays, if we visit a new place and we are not aware of the language then it is not a problem at all, as for this also machine learning helps us by converting the text into our known languages. Google's GNMT (Google Neural Machine Translation) provide this feature, which is a Neural Machine Learning that translates the text into our familiar language, and it called as automatic translation.

The technology behind the automatic translation is a sequence to sequence learning algorithm, which is used with image recognition and translates the text from one language to another language.

CHAPTER 2

LITERATURE SURVEY

2.1 LITERATURE STUDY

1. **A. Amin et al.,** "Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods ", Int. J. Inf. Manage., vol. 46, pp. 304–319, Jun 2019. Cross-Company Churn Prediction (CCCP) is a domain of research where one company (target) is lacking enough data and can use data from another company (source) to predict customer churn successfully. To support CCCP, the cross-company data is usually transformed to a set of similar normal distribution of target company data prior to building a CCCP model. However, it is still unclear which data transformation method is most effective in CCCP. Also, the impact of data transformation methods on CCCP model performance using different classifiers have not been comprehensively explored in the telecommunication sector. In this study, we devised a model for CCCP using data transformation methods (i.e., log, z-score, rank and box-cox) and presented not only an extensive comparison to validate the impact of these transformation methods in CCCP, but also evaluated the performance of underlying baseline classifiers (i.e., Naive Bayes (NB), K- Nearest Neighbor (KNN), Gradient Boosted Tree (GBT), Single Rule Induction (SRI) and Deep learner Neural net (DP)) for customer churn prediction in telecommunication sector using the above mentioned data transformation methods.

2. **S, Babu, B.N. Ananthanarayan and V. Ramesh,** "A Survey on factors impacting churn in telecommunication using datamining techniques", Int. J. Eng. Res. Technol., vol. 3, no. 3, pp.17451748, Mar.2014. Customer churn is the central concern of most companies which are active in industries with low switching cost. Telecommunication industry can be considered to be top, among the industries which suffer from this issue. Mobile Service Providers have implemented CRM (Customer Relationship Management) with intention to reduce the number of Customer Churn. However, Still the Telecom Industry facing with high churn rate. The objectives of

this research study are to identify the high impact factors that cause customer churn in Mobile Service Provider Industry. Questionnaire survey from the sample of 750 was conducted under the category of Students, Professionals (Salaried) and Business Persons. The collected data was analyzed using Decision Tree (ID3 Algorithm) in WEKA in order to predict the high impact factors that cause customer churn. Based on the study Network Quality, Call Facilities , Internet Facilitie and Booster Facilities were the high impact factors that cause customer churn in telecommunication.

3. **A. Sharma and P. K. Kuma**r. (Sep. 2013). ‘‘A neural network based approach for predicting customer churn in cellular network servicesMarketing literature states that it is more costly to engage a new customer than to retain an existing loyal customer. Churn prediction models are developed by academics and practitioners to effectively manage and control customer churn in order to retain existing customers. As churn management is an important activity for companies to retain loyal customers, the ability to correctly predict customer churn is necessary. As the cellular network services market becoming more competitive, customer churn management has become a crucial task for mobile communication operators. This paper proposes a neural network (NN) based approach to predict customer churn in subscription of cellular wireless services. The results of experiments indicate that neural network based approach can predict customer churn with accuracy more than 92%. Further, it was observed that medium sized NNs perform best for the customer churn prediction when different neural network’s topologies were experimented.

4. **C. Geppert,** ‘‘Customer churn management: Retaining high-margin customers with customer relationship management techniques,’’ KPMG & Associates Yarhands Dissou Arthur/Kwaku Ahenkrah/David Asamoah, 2002. The Mobile Telecommunications industry in Ghana has over the years seen telecommunication firms such as Ghana Telecom (GT) now Vodafone Ghanahas faced a lot of competition from the its competitors. Even though Vodafone the then Ghana telecom was the incumbent but was over taken by Scancom Ghana (now trading under the brand name MTN).Due to the increase in competition among the telecommunication

industries has made it important for the companies to focus on the factors customer retention. The paper outlines the causes of customer churn behavior as outlines in her study of service industry identification of additional factors to the telecommunication industry in Ghana, Susan Keaveney (1995). The main aim of the paper is to identify the most important factors that causes customer to switch and also access the relationship that exist between the factors and the switching likelihood. Using descriptive statistics, relative importance index of the various factors were also computed to obtain the most important factors. The paper then present chi-square test of independence between then customer switching likelihood against the factors that causes switching. The paper reported that the high tariff was the most important factor most Ghanaians consider when considering to switch from one network to another. hidden charges, service disruptions, unreliable help lines, inadequate or incomplete information provided by service providers and finally attractive features offered by other networks were the first five factors ranked to be the most important factors. the confirmatory test of independence using chi-square test of independent confirm the truth that the factors truly contribute to customer churn. The paper finally made recommendation the telecommunication industry in Ghana on their customer retention strategies.

5. **Y. Huang and T. Kechadi**, ‘‘An effective hybrid learning system for telecommunication churnprediction,’’ Expert Syst. Appl., vol. 40, no. 14, pp. 5635–5647, Oct. 2013. Customer churn has emerged as a critical issue for Customer Relationship Management and customer retention in the telecommunications industry, thus churn prediction is necessary and valuable to retain the customers and reduce the losses. Moreover, high predictive accuracy and good interpretability of the results are two key measures of a classification model. More studies have shown that single model-based classification methods may not be good enough to achieve a satisfactory result. To obtain more accurate predictive results, we present a novel hybrid model-based learning system, which integrates the supervised and unsupervised techniques for predicting customer behaviour. The system combines a

modified k-means clustering algorithm and a classic rule inductive technique (FOIL).Three sets of experiments were carried out on telecom datasets. One set of the experiments is for verifying that the weighted k-means clustering can lead to a better data partitioning results; the second set of experiments is for evaluating the classification results, and comparing it to other well-known modelling techniques; the last set of experiment compares the proposed hybrid-model system with several other recently proposed hybrid classification approaches. We also performed a comparative study on a set of benchmarks obtained from the UCI repository. All the results show that the hybrid model-based learning system is very promising and outperform the existing models.

6. **M. Kaur, K. Singh, and N. Sharma**, ‘‘Data mining as a tool to predict the churn behaviour among Indian bank customers,’’ Int. J. Recent Innov. Trends Comput.Commun., vol. 1, no. 9, pp. 720– 725, Sep. 2013.The socio economic growth of the country is mainly dependent on the services sector. The financial sector is one of these services sector. Data mining is evolving into a strategically important dimension for many business organizations including banking sector. The churn problem in banking sector can be resolved using data mining techniques. The customer churn is a common measure of lost customers. By minimizing

customer churn a company can maximize its profits. Companies have recognized that existing customers are most valuable assets. Customer relationship management (CRM) can be defined as the process of acquiring, retaining and growing profitable customer which requires a clear focus on service attributes that represent value to the customer and creates loyalty. Customer retention is critical for a good marketing and a customer relationship management strategy. The prevention of customer churn through customer retention is a core issue of Customer relationship management. Predictive data mining techniques are useful to convert the meaningful data into knowledge. In this analysis the data has been analyzed the decision trees algorithm (J48) and the support vector machines(SMO).

2.2 EXISTING METHODS

The following are the various existing methods. The most commonly used classification algorithms along with the python code: Logistic Regression, Naïve Bayes, Convolution Neural Network, K-Nearest Neighbours, Decision Tree, Random Forest, and Support Vector Machine.

LOGISTIC REGRESSION

Logistic regression is a machine learning algorithm for classification. In this algorithm, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function.

NAÏVE BAYES

Naive Bayes algorithm based on Bayes' theorem with the assumption of independence between every pair of features. Naive Bayes classifiers work well in many real-world situations such as document classification and spam filtering.

K-NEAREST NEIGHBOURS

Neighbours based classification is a type of lazy learning as it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the k nearest neighbours of each point.

DECISION TREE

Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data.

RANDOM FOREST

Random forest classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the

original input sample size but the samples are drawn with replacement.

SUPPORT VECTOR MACHINE

Support vector machine is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

2.3 LIMITATIONS OF EXISTING METHODS

1. In the existing methods, the authors used the algorithms like naïve bayes, decision trees etc.by using the above algorithms the results weren't upto mark due to various factors.
2. In some cases, the performance of the model was low due to imbalanced dataset and no methods weren't applied to overcome this problem.
3. There is no user interaction with the existing models and hence we don't know how the model works for the new values.
4. The the feature selection process can be improved using optimization algorithms which increases the classification accuracy. within the performance evaluation, gain measure and ROC curve was used.

CHAPTER 3

PROPOSED METHODOLOGY

3.1 INTRODUCTION

Gradient boosting is a method standing out for its prediction speed and accuracy, particularly with large and complex datasets. From Kaggle competitions to machine learning solutions for business, this algorithm has produced the best results. We already know that errors play a major role in any machine learning algorithm. There are mainly two types of error, bias error and variance error. Gradient boost algorithm helps us minimize bias error of the model

Before getting into the details of this algorithm we must have some knowledge about AdaBoost Algorithm which is again a boosting method. This algorithm starts by building a decis Gradient boosting is a method standing out for its prediction speed and accuracy, particularly with large and complex datasets. From Kaggle competitions to machine learning solutions for business, this algorithm has produced the best results. We already know that errors play a major role in any machine learning algorithm. There are mainly two types of error, bias error and variance error. Gradient boost algorithm helps us minimize bias error of the model

Before getting into the details of this algorithm we must have some knowledge about AdaBoost Algorithm which is again a boosting method. This algorithm starts by building a decision stump and then assigning equal weights to all the data points. Then it increases the weights for all the points which are misclassified and lowers the weight for those that are

easy to classify or are correctly classified. A new decision stump is made for these weighted data points. The idea behind this is to improve the predictions made by the first stump.

ion stump and then assigning equal weights to all the data points. Then it increases the weights for all the points which are misclassified and lowers the weight for those that are easy to classify or are correctly classified. A new decision stump is made for these weighted data points. The idea behind this is to improve the predictions made by the first stump.

3.2 BLOCK DIAGRAM

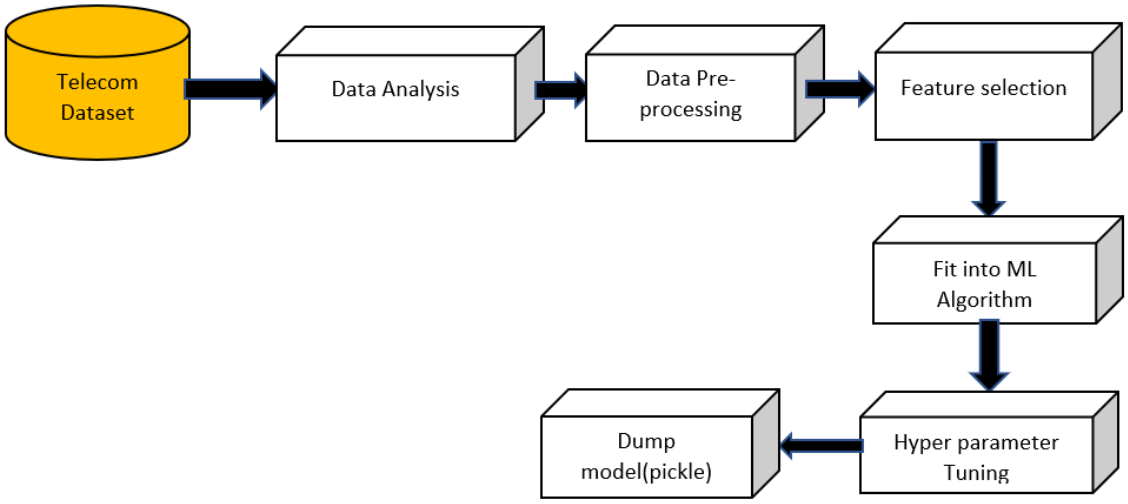


Fig 3.1 Block Diagram

As shown in the above figure, if you send an attributes as input it will undergo many steps and then send an output. Between input and output layers there are some other layers which are referred as hidden layers. The telcom dataset undergoes various steps like data analysis, data pre-processing,Feauture selection ,Fit into ML algorithm,Hyperparameter Tuning and finally dumping the model.

3.3 PROPOSED SYSTEM

The proposed system consists of 8 phases

About Dataset

The data set contains 7043 rows (customers) and 21 columns (features) and is taken from https://www.kaggle.com/datasets/blastchar/telco-customer-churn

The data set includes information about:

● Customers who left within the last month – the column is named Churn

● Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies

● Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges

● Demographic info about customers – gender, age range, and if they need partners and dependents

● Churn - dependent feature ('Yes' denotes customers left, 'No' denotes customer stay here)

EDA (Exploratory Data Analysis)

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

The main aim of this procedure is to draw useful insights from the dataset. Below are the few insights drawn from the dataset using EDA.

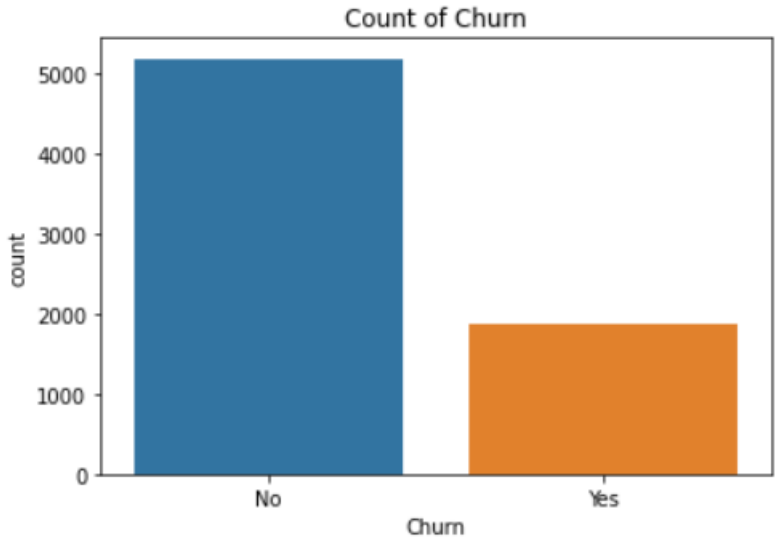
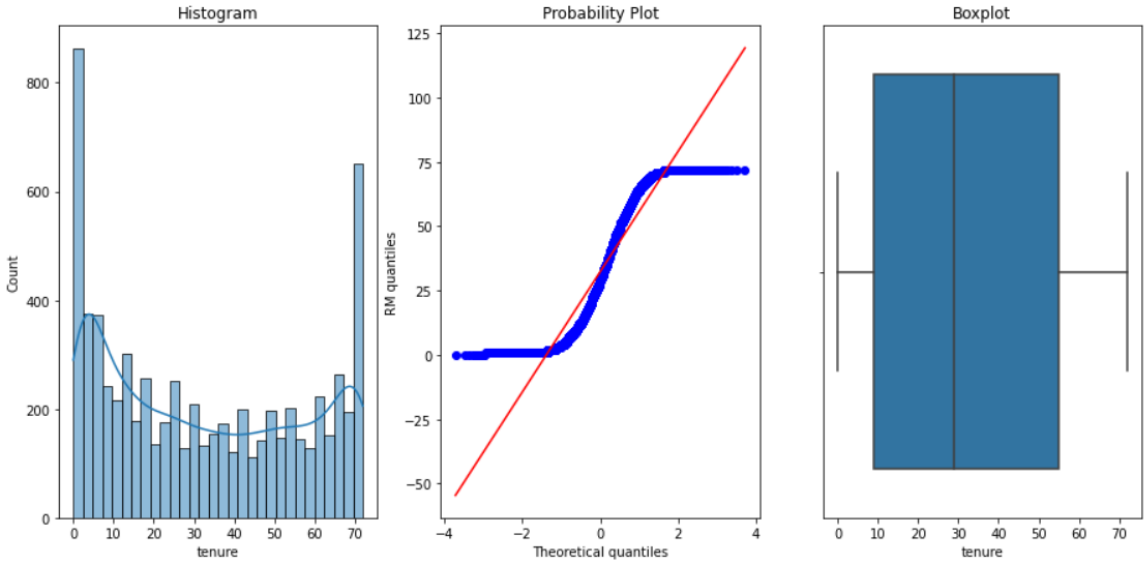


Fig.3.2 Bar graph of distribution of classes in dataset

From the above graph ,1869 of customer are left about 26.5 percentage from overall, this looks like an imbalance dataset.



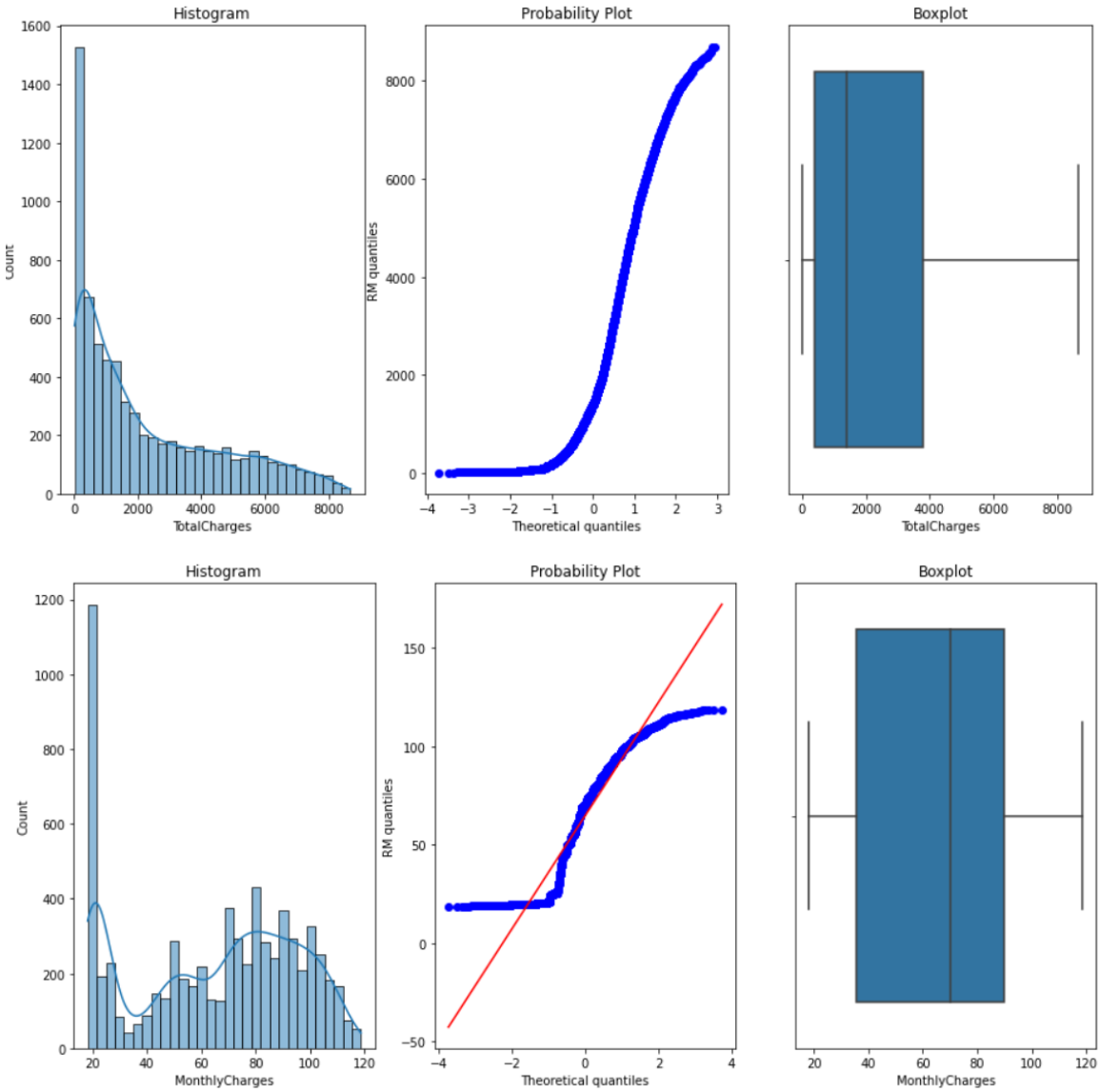


Fig 3.3 Boxplot, Histogram, Probability plot among various attributes

After plotting histogram probability distribution and box plot to find numerical value are in normally distribution and our dataset has no outlier dataset. So, we don't want to remove the outlier in our dataset.

Data cleaning

Data cleaning is the process of preparing data for analysis by weeding out information that is irrelevant or incorrect. This is generally data that can have a negative

impact on the model or algorithm it is fed into by reinforcing a wrong notion.Data cleaning not only refers to removing chunks of unnecessary data, but it's also often associated with fixing incorrect information within the dataset and reducing duplicates.

In this procedure we found the correlation between the independent and dependent features by using heatmap as shown below

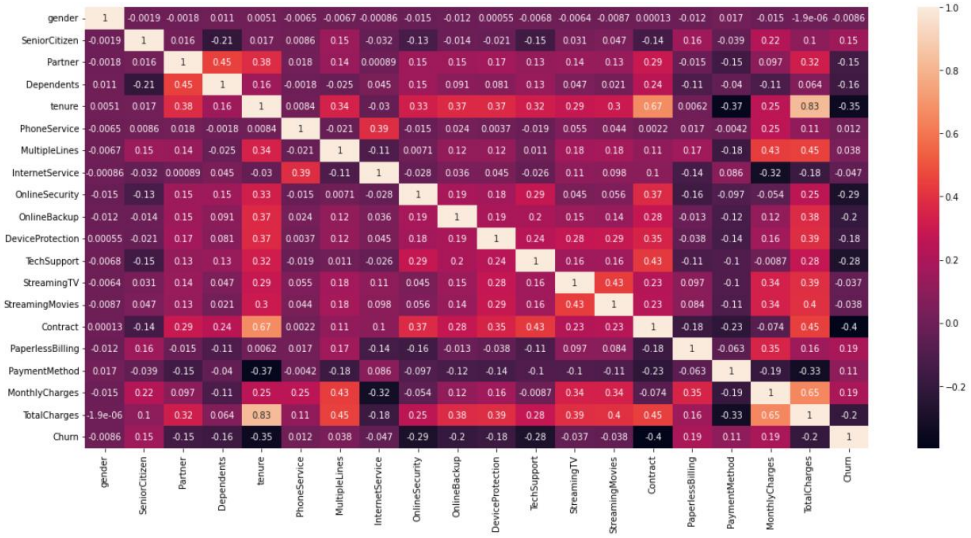


Fig 3.4 Heatmap

Feature Selection

In our proposed system, we mainly select 10 features which have higher correlation by using feature selection methods. From sklearn using feature selection modules importing the SelectKBest to select the important features.

According to the feature selection, we select 10 out of 21 features. These are the 10 features selected [Dependents, tenure, Online Security, Online Backup, Device Protection, Tech Support, Contract, Paperless Billing, Monthly Charges, Total charges] and we split the data into training and testing in 80-20 ratio as follows

X_train, X_test, y_train, y_test = train_test_split (X, y, test_size=0.2)

Fit into Algorithm(ML Algorithm)

We implemented machine learning algorithms such as logistic regression ,Decision tree classifier,Random forest and Gradient tree boosting methods.

Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing t**he logistic function:**

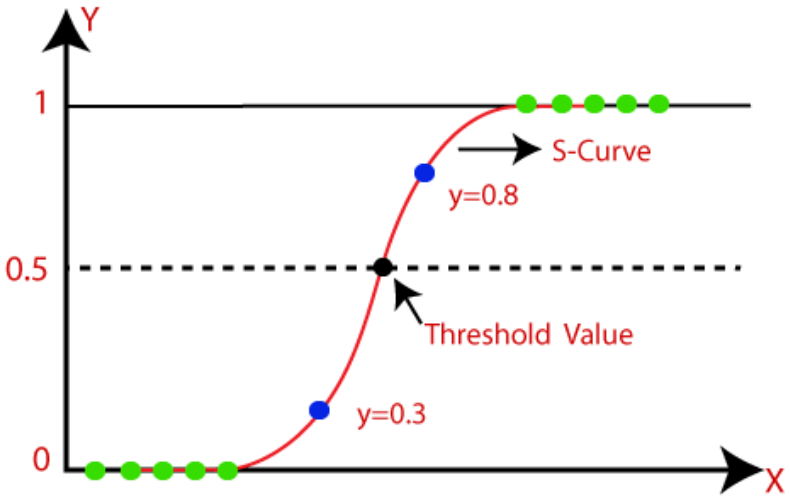


Fig 3.5 logistic function

Decision tree classifier

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset.It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.Below diagram explains the general structure of a decision tree:

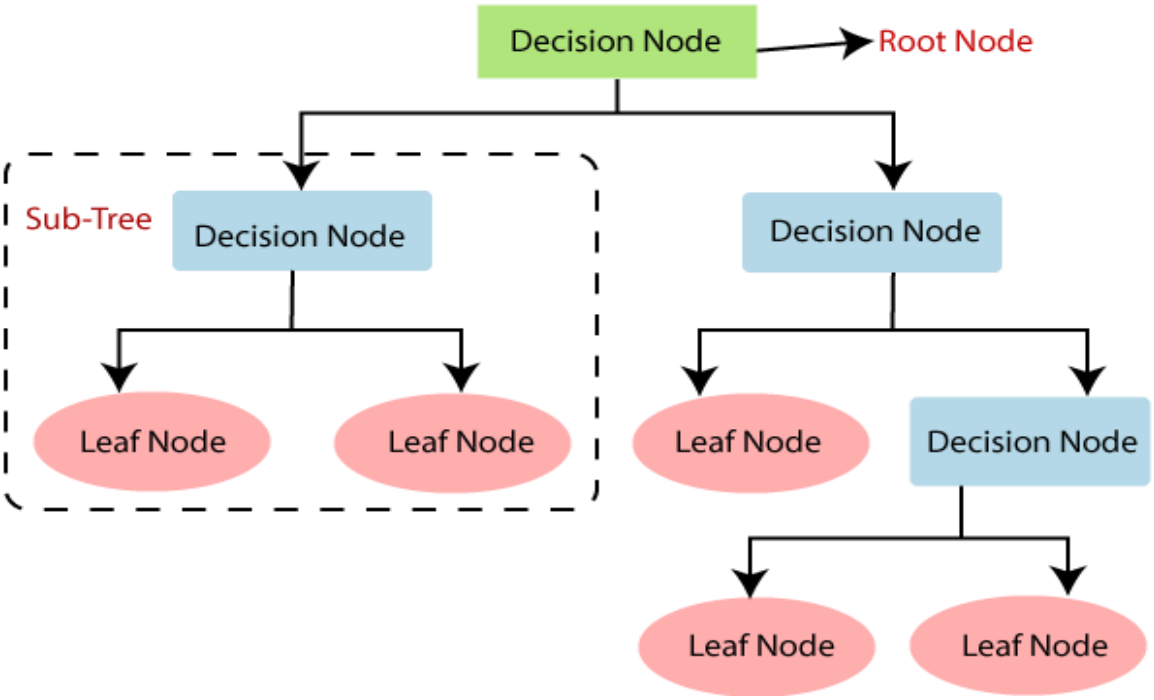


fig 3.6 Decision tree

Random Forest Classifier

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The below diagram explains the working of the Random Forest algorithm:

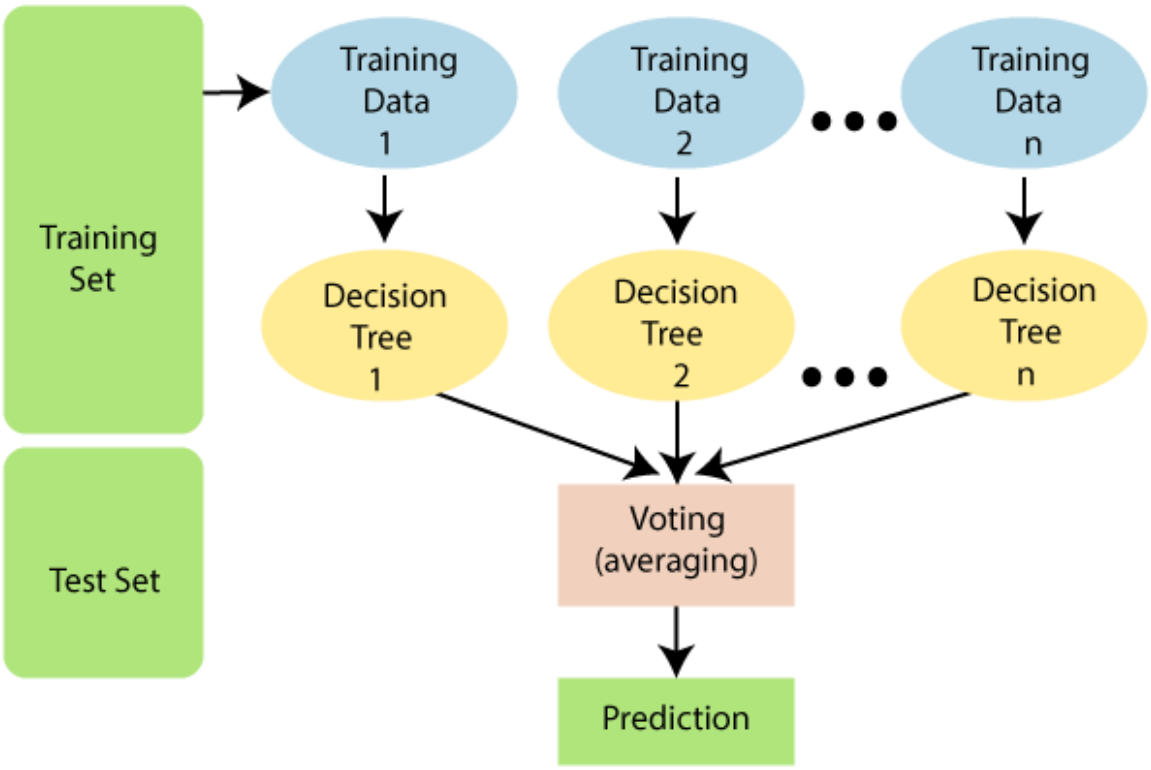


Fig 3.7 Random Forest classifier

Gradient Tree Boosting Algorithm

Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor's error. In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels.

There is a technique called the **Gradient Boosted Trees** whose base learner is CART (Classification and Regression Trees).

The below diagram explains how gradient boosted trees are trained for regression problems. The ensemble consists of N trees. Tree1 is trained using the feature matrix X and the labels y. The predictions labelled y1(hat) are used to determine the training set residual errors r1. Tree2 is then trained using the feature matrix X and the residual errors r1 of Tree1 as labels. The predicted results r1(hat) are then used to determine the residual r2. The process is repeated until all the *N* trees forming the ensemble are trained.

There is an important parameter used in this technique known as **Shrinkage**.

Shrinkage refers to the fact that the prediction of each tree in the ensemble is shrunk after it is multiplied by the learning rate (eta) which ranges between 0 to 1. There is a trade-off between eta and number of estimators, decreasing learning rate needs to be compensated with increasing estimators in order to reach certain model performance. Since all trees are trained now, predictions can be made

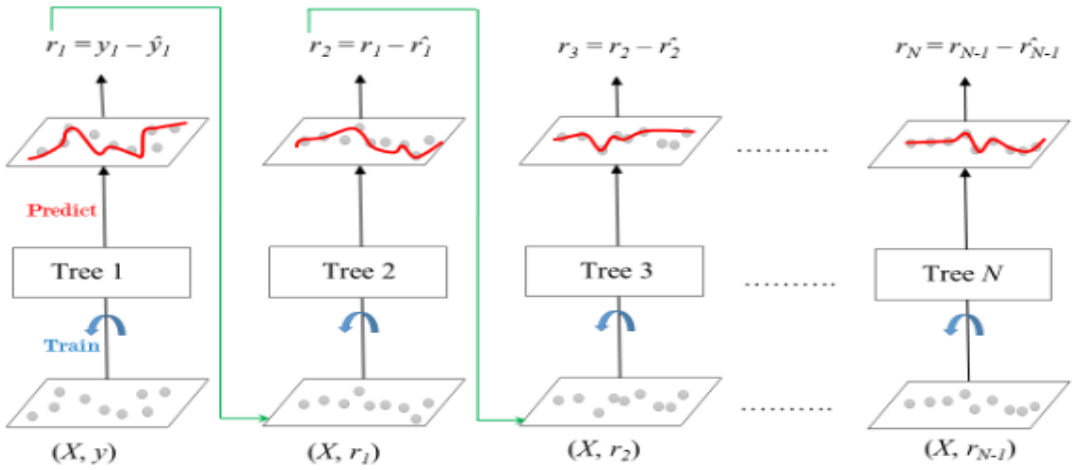


Fig 3.8 Gradient Boosted Trees for Regression

Dump Model (Pickle)

We saved the model using pickle module and this model is able to predict the output based on the user inputs.

Creating GUI using Python

This user Interface will serve as a frontend to our model and it takes inputs from the user and predicts our desired output.

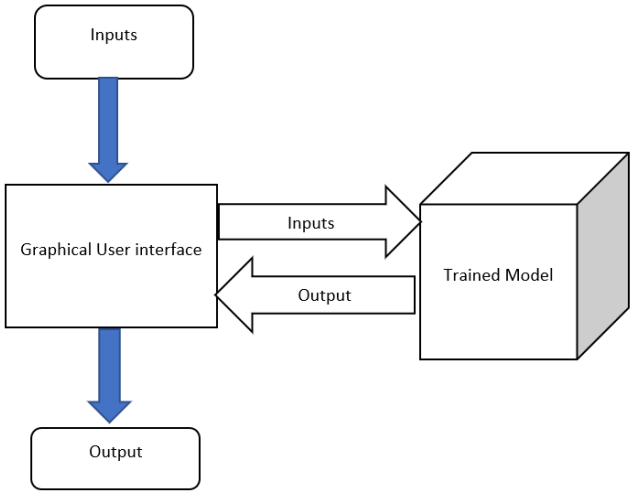


Fig 3.9 UI block diagram

ADVANTAGES OF PROPOSED TECHNIQUE OVER THE EXISTING

The merits of the proposed algorithm have listed as follows:

1. We have applied SelectKBest to perform feature selection and to reduce the dimensions of the data-set, in contrast to existing approaches where prediction accuracy is low due to improper feature selection [5,6].
2. The obtained accuracy was low in existing methods due to an imbalanced dataset, to overcome this problem we implemented SMOTEENN methods to balance the dataset.
3. We have implemented User interface using python, so that the model is well able to work with new values apart from the dataset.
4. Then we have evaluated the algorithms on the test set using confusion matrix and AUC curve, which have been mentioned in the form of graphs and tables in order to compare which algorithm performs best for this particular data-set, in contrast to the existing techniques where obtained results are not properly evaluated.

PUTTING IT ALL TOGETHER

To create our GUI, we utilized tkinter and Python.

To accomplish this task, we:

1. First give the inputs in the dataset and see the results.
2. Try giving new values which are not present in the dataset..

This network is purposely shallow, ensuring that:

1. We reduce the chances of overfitting on our small dataset.
2. The model itself is capable of running in real-time (including on the Raspberry Pi).

Overall, our customer churn model was able to obtain 90% accuracy on our validation set.

To demonstrate the full liveness detection pipeline in action we created a Python + GUI script that loaded our model and applied it to real-time values.

As our demo showed, our model was capable of distinguishing between customers who churn and who don’t churn.

3.4 HARDWARE REQUIREMENTS

- 8GB RAM
- INTEL I5 PROCESSOR
- GPU OF 2 GB RECOMMENDED

3.5 SOFTWARE REQUIREMENTS

PYTHON

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It has a wide range of applications from Web development (like: Django and Bottle), Scientific and mathematical computing (Orange, SciPy, NumPy) to desktop graphical user Interfaces (Pygame, Panda3D).

NUMPY

NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed. This tutorial explains the basics of NumPy such as its architecture and environment. It also discusses the various array functions, types of indexing, etc. An introduction to Matplotlib is also provided. All this is explained with the help of examples for better understanding.

PANDAS

Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users. Pandas were initially developed by Wes McKinney in 2008 while he was working at AQR Capital Management. He convinced the AQR to allow him to open source the Pandas. Another AQR employee, Chang She, joined as the second major contributor to the library in 2012. Over time many versions of pandas have been released. The latest version of the pandas is 1.4.1.

SKLEARN

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

IMBLEARN

Imbalanced-Learn is a Python module that helps in balancing the datasets which are highly skewed or biased towards some classes. Thus, it helps in resampling the classes which are otherwise oversampled or undesampled. If there is a greater imbalance ratio, the output is biased to the class which has a higher number of examples.

CHAPTER 4

SYSTEM DESIGN

4.1 DATA FLOW DIAGRAM

A data-flow diagram is a way of representing a flow of a data of a process or a system. The DFD also provides information about the outputs and inputs of each entity and the process itself. A data-flow diagram has no control flow; there are no decision rules and no loops.

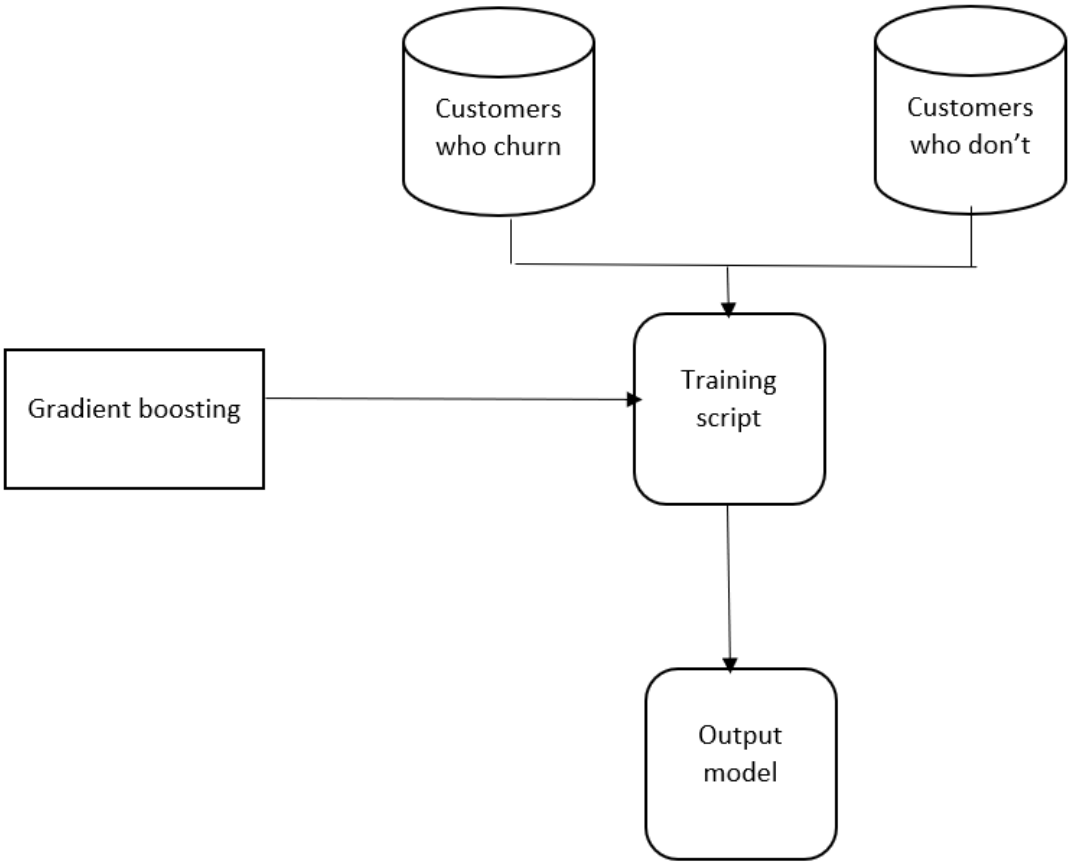


Fig 4.1 The process of training

4.2 UML DIAGRAMS

UML is an acronym that stands for Unified Modeling Language. Simply put, UML is a modern approach to modeling and documenting software. In fact, it's one of the most popular business process modeling techniques. It is based on diagrammatic representations of software components. As the old proverb says: "a picture is worth a thousand words". By using visual representa- tions, we are able to better understand possible flaws or errors in software or business processes. The elements are like components which can be associated in different ways to make a complete UML picture, which is known as diagram. Thus, it is very important to understand the different diagrams to implement the knowledge in real- life systems. Any complex system is best understood by making some kind of diagrams or pictures. These diagrams have a better impact on our understanding. If we look around, we will realize that the diagrams are not a new concept but it is used widely in different forms in different industries. Mainly, UML has been used as a general-purpose modeling language in the field of software engineering. However, it has now found its way into the documenta- tion of several business processes or workflows. For example, activity diagrams, a type of UML diagram, can be used as a replacement for flowcharts. They provide both a more standardized way of modeling workflows as well as a wider range of features to improve readability and efficiency. Use cases are best discovered by examining the actors and defining what the actor will be able to do with the system. Since all the needs of a system typically cannot be covered in one use case, it is usual 7 to have a collection of use cases. Together this use case collection specifies all the ways the system. An association provides a pathway for communication. The communication can be between use cases, actors, classes or interfaces. Associations are the most general of all relationships and consequentially the most semantically weak. If two objects are usually considered independently, the relationship is an association. They provide both a more standardized way of modeling workflows as well as a wider range of features to improve readability and efficiency. Use cases are best discovered by examining the actors and defining what the actor will be able to do with the system. Since all the needs of a system typically cannotbe covered in one use case, it is usual to have a collection of use cases. By default, the

association tool on the toolbox is unidirectional and drawn on a diagram with a single arrow at one end of the association. The end with the arrow indicates who or whatis receiving the communication. A dependency is a relationship between two model elements in which a change to one model element will affect the other model element. Typically, on class diagrams, a dependency relationship indicates that the operations of the client invoke operationsof the supplier. The work flow in this case begins from importing the dataset by the developer and then replacing missing values with mean value of corresponding column, model building, validating that model by generating a confusion matrix and finally predicting the test sample class label. Transitions are used to show the passing of the flow of controlfrom activity to activity. The various UML diagrams are:

1. Usecase diagram

2. Activity diagram

3. Sequence diagram

4. Colloboration diagram

5. Object diagram

6. State chart diagram

7. Class diagram

8. Component diagram

9. Deployment diagram

4.2.1 CLASS DIAGRAM

Class diagram in the Unified Modelling Language is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations, and the relationships among objects.

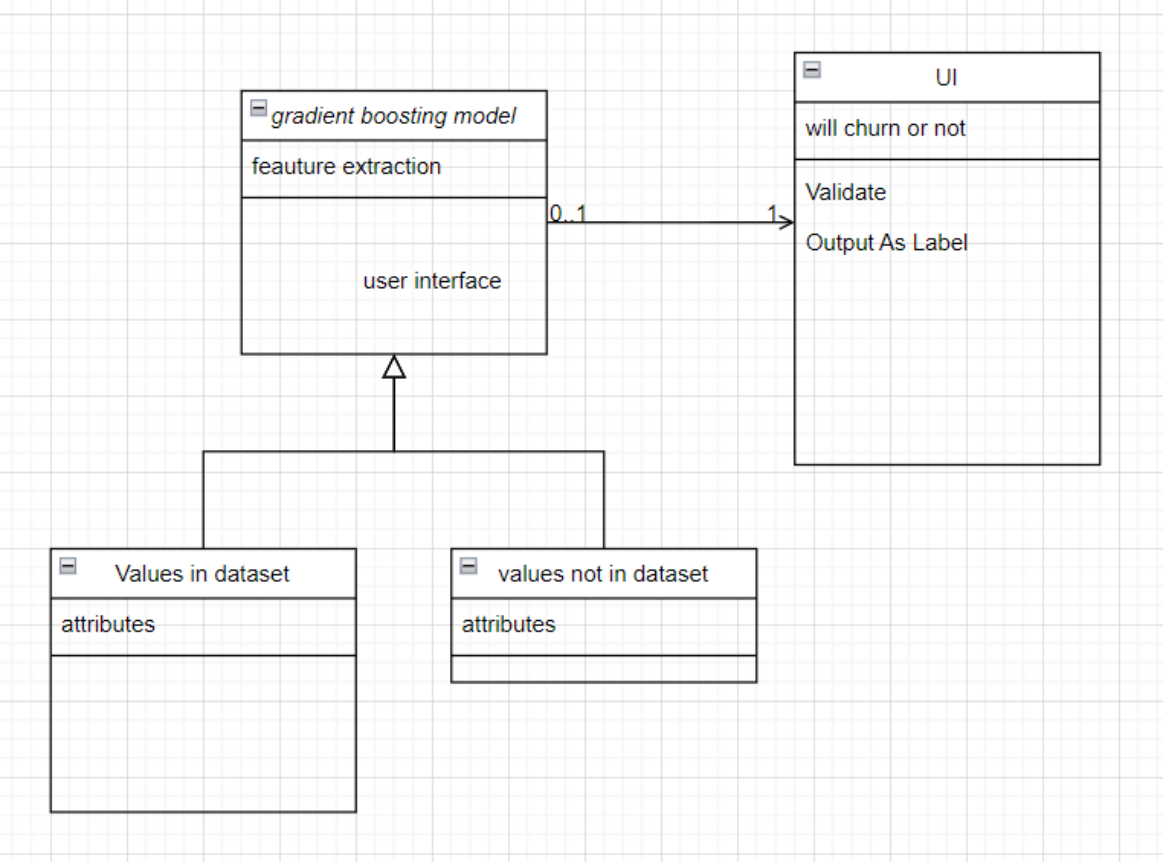


Fig 4.2.1 Class Diagram

4.2.2 USECASE DIAGRAM

A use case diagram is a graph of actors, a set of use cases enclosed by a system boundary, communication (participation) associations between the actors and users and generalization among use cases. The use case model defines the outside (actors) and inside (use case) of the system's behaviour. Actors are not part of the system. Actors represent anyone or anything that interacts with (input to or receive output from) the system. Use-case diagrams can be used during analysis to capture the system requirements and to understand how the system should work. During the design phase, you can use use case diagrams to specify the behaviour of the system as implemented. Use case is a sequence of transactions performed by a system that yields a measurable result of values for a particular actor. The use cases are all the ways the system may be used.

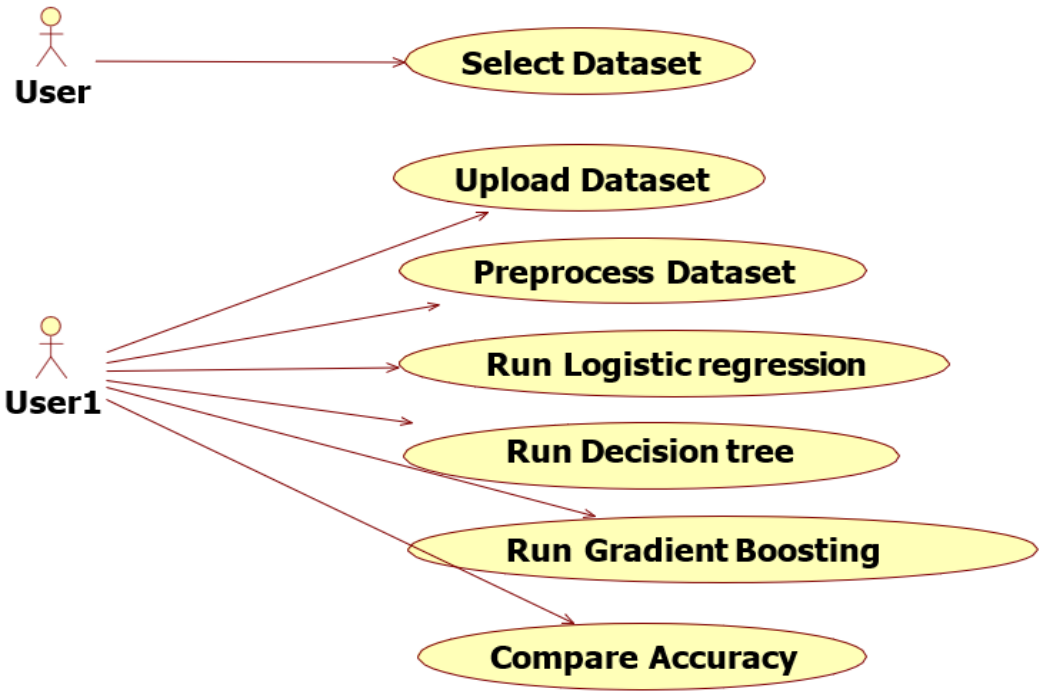


Fig 4.2.2 Use Case Diagram

4.2.3 SEQUENCE DIAGRAM

Sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario.

A sequence diagram is an interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case realizations in the Logical View of the system under development. Sequence diagrams are sometimes called as event diagrams.

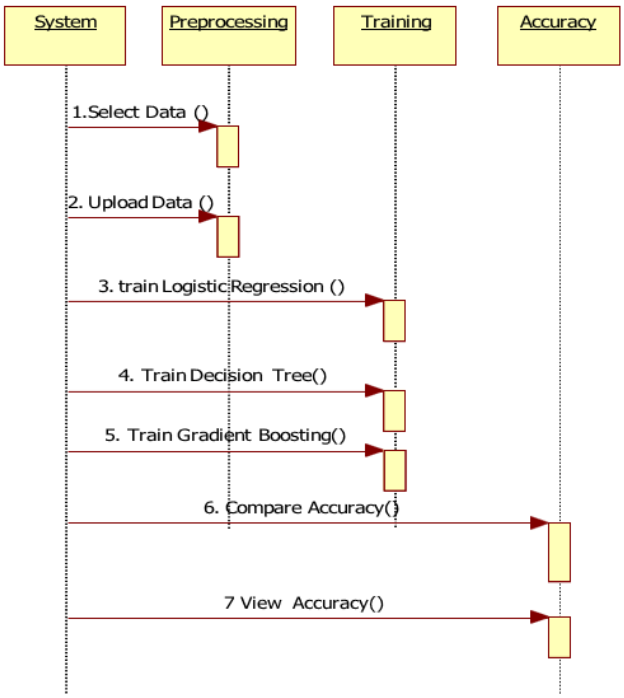


Fig 4.2.3 Sequence Diagram

4.2.4 STATE CHART DIAGRAM

State diagram is a type of diagram used in computer science and related fields to describe the behaviour of systems. State diagrams require that the system described is composed of a finite number of states; sometimes, this is indeed the case, while at other times this is a reasonable abstraction.

A state chart diagram shows the states of a single objects, the events or messages that cause a transition from one state to another and the actions that result from a state change. As I activity diagram, state chart diagram also contains special symbols for start state and stop state.

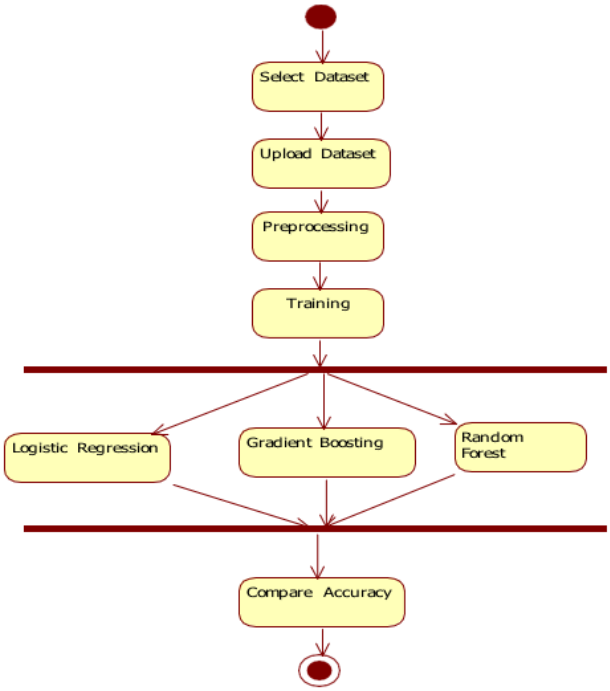


Fig 4.2.4 State Chart Diagram

4.2.5 ACTIVITY DIAGRAM

Activity diagram is another important diagram in UML to describe the dynamic aspects of the system. Activity diagram is basically a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system.

An Activity diagram is a variation of a special case of a state machine, in which the states are activities representing the performance of operations and the transitions are triggered by the completion of the operations. The purpose of Activity diagram is to provide a view of flows and what is going on inside a use case or among several classes. Activity diagrams contain activities, transitions between the activities, decision points, and synchronization bars. An activity represents the performance of some behavior in the workflow. In the UML, activities are represented as rectangles with rounded edges, transitions are drawn as directed arrows, decision points are shown as diamonds, and synchronization bars are drawn as thick horizontal or vertical bars as shown in the following. The activity icon appears as a rectangle with rounded ends with a name and a component for actions.

Swim lanes may be used to partition an activity diagram. This typically is done to show what person or organization is responsible for the activities contained in the swim lane. Swim lanes are helpful when modeling a business workflow because they can represent organizational units or roles within a business model. Swim lanes are very similar to an object because they provide a way to tell who is performing a certain role.

Swim lanes only appear on activity diagrams. When a swim lane is dragged onto an activity diagram, it becomes a swim lane view. Swim lanes appear as small icons in the browse while a swim lane views appear between the thin, vertical lines with a header that can be renamed and relocated. In the UML, activities are represented as rectangles with rounded edges, transitions are drawn as directed arrows, decision points are shown as diamonds, and synchronization bars are drawn as thick horizontal or vertical bars as shown in the following. The activity icon appears as a rectangle with rounded ends with a name and a component for actions.

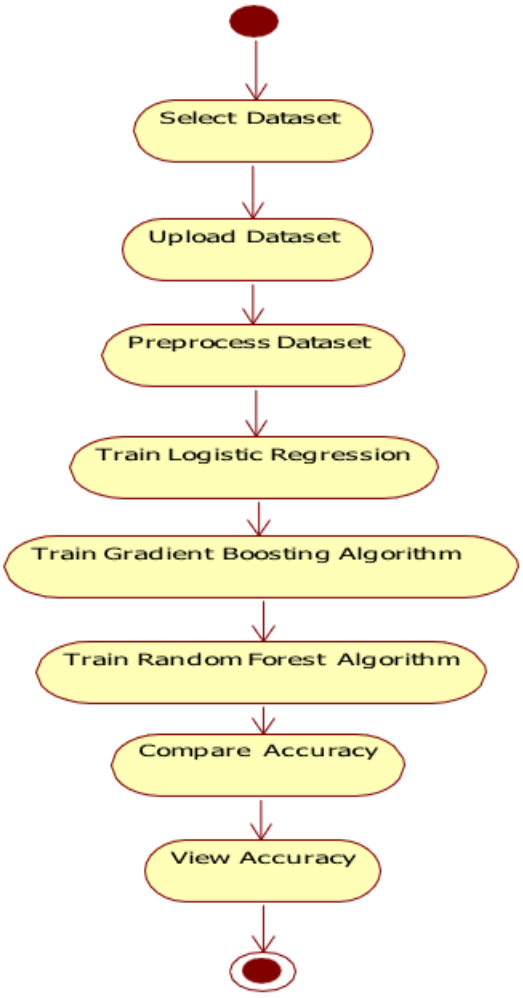


Fig 4.2.5 Activity Diagram

4.2.6 OBJECT DIAGRAM

Object diagrams are derived from class diagrams so object diagrams are dependent upon class diagrams.

Object diagrams represent an instance of a class diagram. The basic concepts are similar for class diagrams and object diagrams. Object diagrams also represent the static view of a system but this static view is a snapshot of the system at a particular moment. Object diagrams are used to render a set of objects and their relationships as an instance.

The purposes of object diagrams are similar to class diagrams. The difference is that a class diagram represents an abstract model consisting of classes and their relationships. However, an object diagram represents an instance at a particular moment, which is concrete in nature.

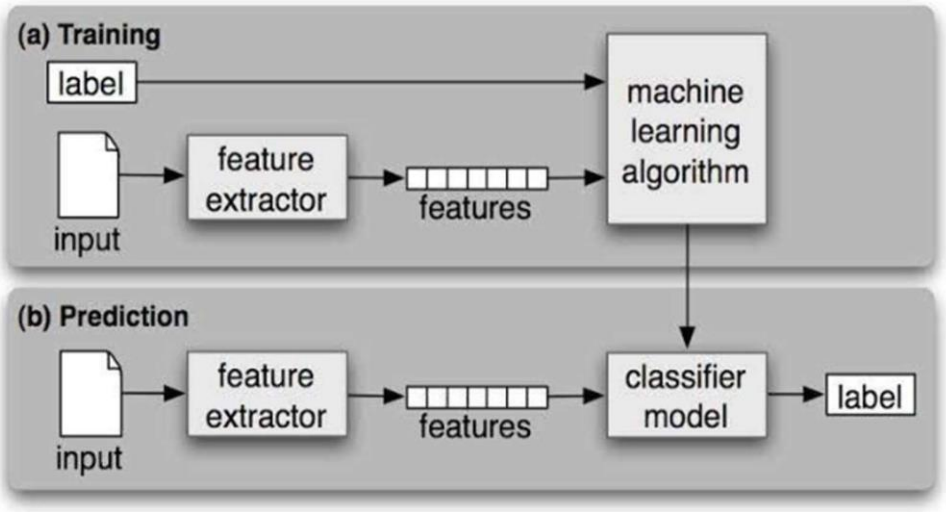


Fig 4.2.6 Object Diagram

CHAPTER 5

TESTING

Software Testing is a process of executing the application with intent to find any software bugs. It is used to check whether the application met its expectations and all the functionalities of the application are working. The final goal of testing is to check whether the application is behaving in the way it is supposed to under specified conditions. All aspects of the code are examined to check the quality of the application. The primary purpose of testing is to detect software failures so that defects may be uncovered and corrected. The test cases are designed in such way that scope of finding the bugs is maximum.

5.1 TESTING LEVELS

Testing Methodologies

The following are the Testing Methodologies:

- Unit Testing.
- Integration Testing.
- User Acceptance Testing.
- Output Testing.
- Validation Testing.

Unit Testing

Unit testing focuses verification effort on the smallest unit of Software design that is the module. Unit testing exercises specific paths in a modules control structure to ensure complete coverage and maximum error detection. This test focuses on each module individually, ensuring that it functions properly as a unit. Hence, the naming is Unit Testing. During this testing, each module is tested individually and the module in-

terfaces are verified for the consistency with design specification. All important processing path are tested for the expected results. All error handling paths are also tested.

Integration Testing

Integration testing addresses the issues associated with the dual problems of verification and program construction. After the software has been integrated a set of high order tests are conducted. The main objective in this testing process is to take unit tested modules and builds a program structure that has been dictated by design. The following are the types of Integration

Testing:

1. Top down Integration:

This method is an incremental approach to the construction of program structure. Modules are integrated by moving downward through the control hierarchy, beginning with the main program module. The module subordinates to the main program module are incorporated into the structure in either a depth first or breadth first manner. In this method, the software is tested from main module and individual stubs are replaced when the test proceeds downwards.

2. Bottom up Integration:

This method begins the construction and testing with the modules at the lowest level in the program structure. Since the modules are integrated from the bottom up, processing required for modules subordinate to a given level is always available and the need for stubs is eliminated. The bottom up integration strategy may be implemented with the following steps:

(a) The low-level modules are combined into clusters into clusters that perform a specific Software sub-function.

(b) A driver (i.e.) the control program for testing is written to coordinate test case input and output.

(c) The cluster is tested.

(d) Drivers are removed and clusters are combined moving upward in the program structure.

The bottom up approaches tests each module individually and then each module is module is integrated with a main module and tested for functionality.

User Acceptance Testing

User Acceptance of a system is the key factor for the success of any system. The system under consideration is tested for user acceptance by constantly keeping in touch with the prospective system users at the time of developing and making changes wherever required. The system developed provides a friendly user interface that can easily be understood even by a person who is new to the system.

Output Testing

After performing the validation testing, the next step is output testing of the proposed system, since no system could be useful if it does not produce the required output in the specified format. Asking the users about the format required by them tests the outputs generated or displayed by the system under consideration. Hence the output format is considered in 2 ways one is on screen and another in printed format.

5.2 SYSTEM TEST CASES

A test case is a set of test data, preconditions, expected results and post conditions, developed for a test scenario to verify compliance with a specific requirement. I have designed and executed a few test cases to check if the project meets the functional requirements.

TEST CONDITION	INPUT SPECIFICATION	OUTPUT SPECIFICATION	PASS/FAIL
Testing with the values in project dataset	Features of customers who churn and don't churn (values in dataset)	Predication label of input values and accuracy	PASS

Table 5.2.1 Test case for values present in dataset

TEST CONDITION	INPUT SPECIFICATION	OUTPUT SPECIFICATION	PASS/FAIL
Testing with the values in project dataset	Features of customers who churn and don't churn (values not in dataset)	Predication label of input values and accuracy	PASS

Table 5.2.2 Test case for values not present in dataset

The testers need to focus on for the following: -

- Test with new data, rather than the original training data. If necessary, split your training set into two groups: one that does training, and one that does test. Better, obtain and use fresh data if you are able.
- Understand the architecture of the network as a part of the testing process. Testers won't necessarily understand how the model was constructed, but need to understand whether it meets requirements. And based on the measurements that they are testing, they may have to recommend a radically different approach, or admit the software is just not capable of doing what it was asked to do with confidence.

CHAPTER 6

RESULTS

The Gradient Boosting model is trained with the dataset. The dataset we considered for the training of Gradient Boosting model consists of two classes. The class labels are Yes or No i.e. Customer will churn or not.

The following is a glance of how our dataset looks like:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupp
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	

5 rows × 21 columns

Fig 6.1 Telcom dataset

We performed data pre-processing on this dataset and made sure it doesn't have any missing values.

GUI with Tkinter

We saved the model using pickle and we use the model as a backend for the graphical user interface. The user interface created is as shown below.

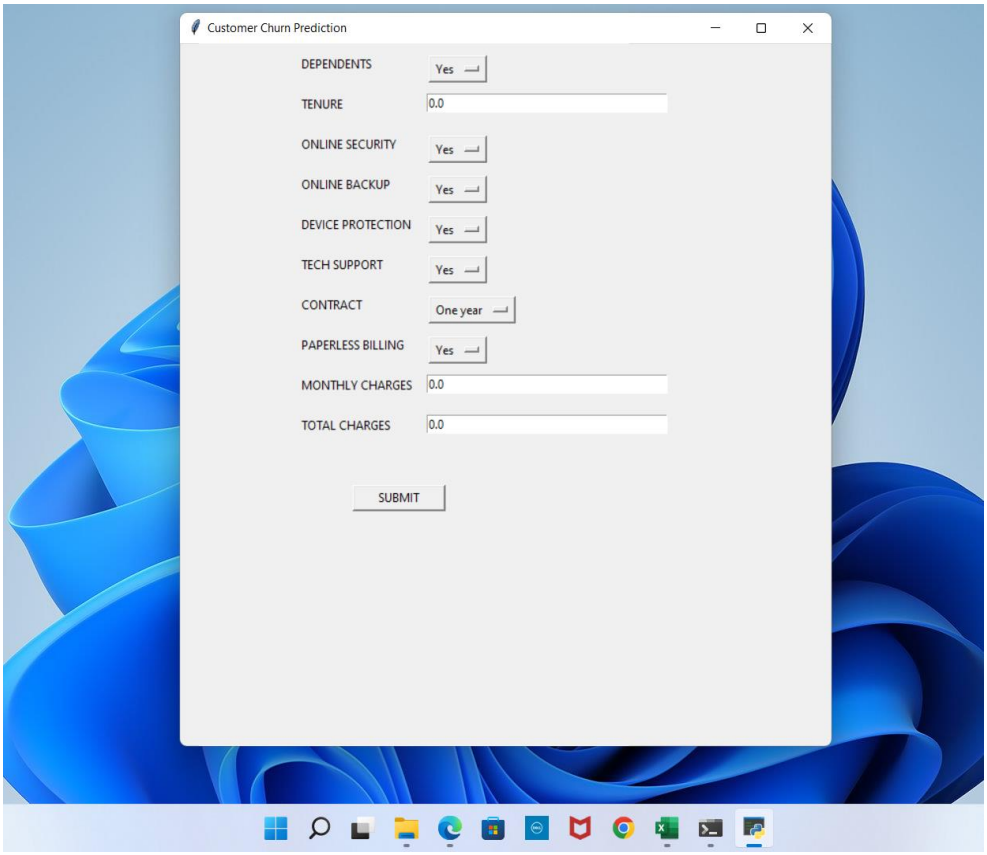


Fig 6.2 Initial GUI

The above screen shows the initial GUI with some default values and it is created using the tkinter module. This UI serves as a front end for our model at the back end. In the next step, we enter some random values as an input to the model.

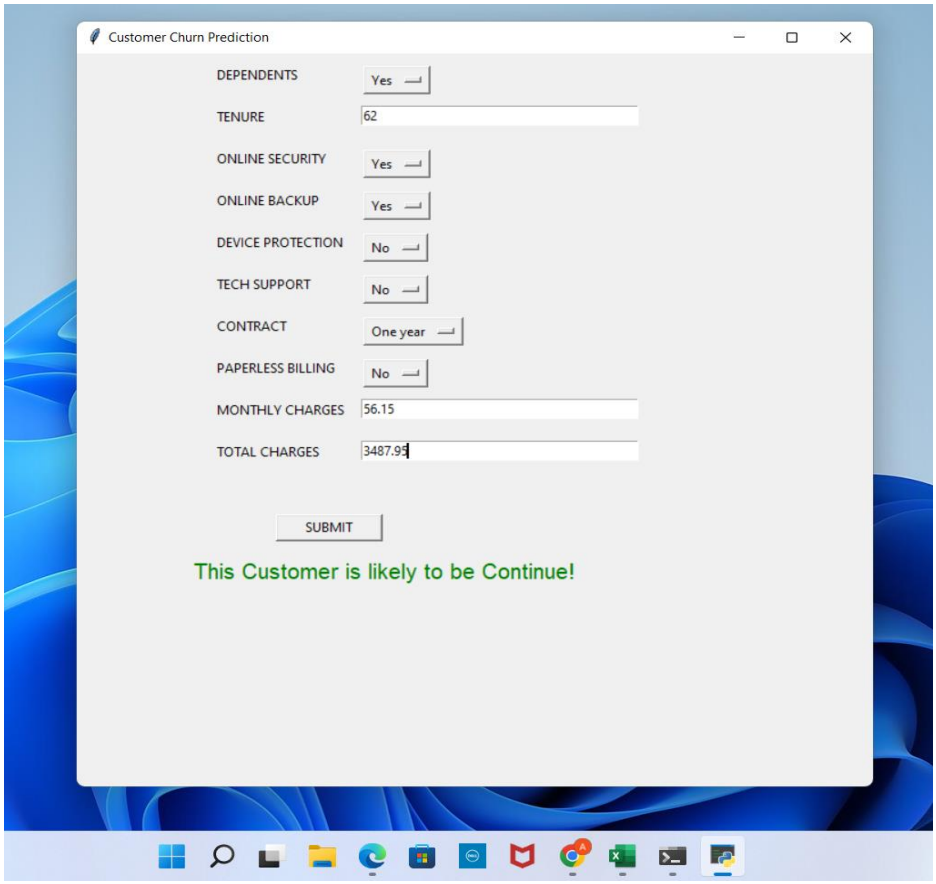


Fig 6.3 GUI showing output

In the above screen for the given inputs the output is “This customer is likely to be continue!” which tells us the customer will not churn. The model was tested with both known and unknown inputs and an accuracy of 96% was achieved for this model.

```
Accuracy score : 0.9638297872340426
Confusion matrix :
 [[422  18]
 [ 16 484]]
Classification report :
              precision    recall  f1-score   support

           0       0.96      0.96      0.96       440
           1       0.96      0.97      0.97       500

    accuracy                           0.96       940
   macro avg       0.96      0.96      0.96       940
weighted avg       0.96      0.96      0.96       940
```

CHAPTER 7

CONCLUSION AND FUTURE SCOPE

In the present competitive market of telecom domain, churn prediction is a significant issue of the CRM to retain valuable customers by identifying a similar group of customers and providing competitive offers/services to the respective groups. Therefore, in this domain, the researchers have been looking at the key factors of churn to retain customers and solve the problems of CRM and decision maker of a company. In this study, a customer churn model is provided for data analytics and validated through standard evaluation metrics. The obtained results show that the proposed churn model performed better by using machine learning techniques, Gradient Boosting produced better F-measure result that is 96%. Finally, guidelines on customer retention are provided for decision-makers of the telecom companies.

FUTURE SCOPE

The future scope of this paper will use hybrid classification techniques to point out the existing association between churn prediction and customer lifetime value. The retention policies need to be considered by selecting appropriate variables from the dataset. The passive and the dynamic nature of the industry ensure that data mining has become an increasingly significant aspect in the telecommunication industry prospect.

REFERENCES

[1] S. Babu, D. N. Ananthanarayanan, and V. Ramesh, ‘‘A survey on factors impacting churn in telecommunication using datamining techniques,’’ Int. J. Eng. Res. Technol., vol. 3, no. 3, pp. 1745–1748, Mar. 2014.

[2] C. Geppert, ‘‘Customer churn management: Retaining high-margin customers with customer relationship management techniques,’’ KPMG & Associates Yarhands Dissou Arthur/Kwaku Ahenkrah/David Asamoah, 2002.

[3] W. Verbeke, D. Martens, C. Mues, and B. Baesens, ‘‘Building comprehensible customer churn prediction models with advanced rule induction techniques,’’ Expert Syst. Appl., vol. 38, no. 3, pp. 2354–2364, Mar. 2011.

[4] Y. Huang, B. Huang, and M.-T. Kechadi, ‘‘A rule-based method for customer churn predictionin telecommunication services,’’ in Proc. Pacific– Asia Conf. Knowl. Discovery Data Mining. Berlin, Germany: Springer, 2011, pp. 411–422.

[5] A. Idris and A. Khan, ‘‘Customer churn prediction for telecommunication: Employing various features selection techniques and tree-based ensemble classifiers,’’ in Proc. 15th Int. Multitopic Conf., Dec. 2012, pp. 23–27.

[6] M. Kaur, K. Singh, and N. Sharma, ‘‘Data mining as a tool to predict the churn behaviour among Indian bank customers,’’ Int. J. Recent Innov. Trends Comput. Commun., vol. 1, no. 9, pp. 720– 725, Sep. 2013.

[7] V. L. Miguéis, D. van den Poel, A. S. Camanho, and J. F. e Cunha, ‘‘Modelling partial customer churn: On the value of first product-category purchase sequences,’’ Expert Syst. Appl., vol. 12, no. 12, pp. 11250–11256, Sep. 2012.

[8] D. Manzano-Machob, ‘‘The architecture of a churn prediction system based on stream mining,’’ in Proc. Artif. Intell. Res. Develop., 16th Int. Conf. Catalan Assoc. Artif. Intell., vol. 256, Oct. 2013, p. 157.

[9] P. T. Kotler, Marketing Management: Analysis, Planning, Implementation and Control. London, U.K.: Prentice-Hall, 1994.

[10] F. F. Reichheld and W. E. Sasser, Jr., ‘‘Zero defections: Quality comes to services,’’ Harvard Bus. Rev., vol. 68, no. 5, pp. 105–111, 1990.

[11] J. Hadden, A. Tiwari, R. Roy, and D. Ruta, ‘‘Computer assisted customer churn management: State-of-the-art and future trends,’’ Comput. Oper. Res., vol. 34, no. 10, pp. 2902–2917, Oct. 2007.

[12] H.-S. Kim and C.-H. Yoon, ‘‘Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market,’’ Telecommun. Policy, vol. 28, nos. 9–10, pp. 751–765, Nov. 2004.

[13] Y. Huang and T. Kechadi, ‘‘An effective hybrid learning system for telecommunication churn prediction,’’ Expert Syst. Appl., vol. 40, no. 14, pp. 5635–5647, Oct. 2013.

[14] A. Sharma and P. K. Kumar. (Sep. 2013). ‘‘A neural network based approach for predicting customer churn in cellular network services.’’ [Online]. Available: https://arxiv.org/abs/1309.3945

[15] Ö.G. AliandU.Aritürk, ‘‘Dynamicchurnpredictionframeworkwithmore effective use of rare event data: The case of private banking,’’ Expert Syst. Appl., vol. 41, no. 17, pp. 7889–7903, Dec. 2014. 44

[16] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, ‘‘Customer churn prediction in telecommunication industry using data certainty,’’ J. Bus. Res., vol. 94, pp. 290–301, Jan. 2019.

[17] S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal, and A. Rehman, ‘‘Telecommunication subscribers’ churn prediction model using machine learning,’’ in Proc. 8th Int. Conf. Digit. Inf. Manage., Sep. 2013, pp. 131–136.

[18] V. Lazarov and M. Capota, ‘‘Churn prediction,’’ Bus. Anal. Course, TUM Comput. Sci, Technische Univ. München, Tech. Rep., 2007. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download? doi=10.1.1.462.7201&rep=rep1&type=pdf

[19] R. Vadakattu, B. Panda, S. Narayan, and H. Godhia, ‘‘Enterprise subscription churn prediction,’’ in Proc. IEEE Int. Conf. Big Data, Nov. 2015, pp. 1317–1321

BIBLIOGRAPHY

- MACHINE LEARNING FOR ABSOLUTE BEGINNERS: A Plain English Introduction - (Oliver Theobald)

- MACHINE LEARNING IN ACTION - (Peter Harrington)

- PATTERN RECOGNITION AND MACHINE LEARNING - (Christopher M. Bishop)

URL'S

- https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/
- https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm
- https://www.javatpoint.com/machine-learning-random-forest-algorithm
- https://www.analyticsindiamag.com/7-types-classification-algorithms/
- https://towardsdatascience.com/imbalanced-classification-in-python-smote-enn-method-db5db06b8d50
- https://www.geeksforgeeks.org/hyperparameter-tuning/
- https://www.tutorialspoint.com/python/python_gui_programming.htm#:~:text=Tkinter%20Programming,to%20the%20Tk%20GUI%20toolkit.&text=Import%20the%20Tkinter%20module
- https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/

APPENDIX

MAIN CODE:

```
# Importing neccessary packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import scipy.stats as stats
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split, RandomizedSearchCV
from sklearn.metrics import recall_score, accuracy_score, classification_report,
confusion_matrix
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.tree import DecisionTreeClassifier
from imblearn.combine import SMOTEENN
from sklearn.feature_selection import SelectKBest
from collections import Counter

# ignore warning
import warnings
warnings.filterwarnings('ignore')
import matplotlib.ticker as mtick # for showing percentage in it

### Plotting numerical feature with probability distribution and checking outlier
for feature in numerical_feature:
    if feature != 'SeniorCitizen':
```

```
        plt.figure(figsize=(15,7))

        plt.subplot(1, 3, 1)
        sns.histplot(data=data, x=feature, bins=30, kde=True)
        plt.title('Histogram')

        plt.subplot(1, 3, 2)
        stats.probplot(data[feature], dist="norm", plot=plt)
        plt.ylabel('RM quantiles')

        plt.subplot(1, 3, 3)
        sns.boxplot(x=data[feature])
        plt.title('Boxplot')S

plt.show()

# selects the feature which has more correlation
selection = SelectKBest()  # k=10 default
X = selection.fit_transform(X,y)

# splitting for train and test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

st=SMOTEENN()
X_train_st,y_train_st = st.fit_resample(X_train, y_train)
print("The number of classes before fit {}".format(Counter(y_train)))
print("The number of classes after fit {}".format(Counter(y_train_st)))

X_train_st,y_train_st = st.fit_resample(X_train, y_train)
```

```
# GradientBoostingClassifier
gbc = GradientBoostingClassifier()
gbc.fit(X_train_sap, y_train_sap)
pred = gbc.predict(X_test_sap)

print(f'Accuracy score : {accuracy_score(pred, y_test_sap)}')
print(f'Confusion matrix :\n {confusion_matrix(pred, y_test_sap)}')
print(f'Classification report :\n {classification_report(pred, y_test_sap)}')

param_grid = {'n_estimators':[100, 150, 200, 250, 300],
            'criterion': ['friedman_mse', 'squared_error', 'mse', 'mae'],
            'min_samples_split': [2,3,4,5,6,7,8,9,10],
            'min_samples_leaf': [1,3,5,7,9,11,13,15],'max_leaf_nodes': [3,6,8,9,12,15,18,24],
             'max_depth': [3,5,7,9,11,13,15,17,19],
             'learning_rate': [0.05, 0.1, 0.2, 0.3, 0.4, 0.5],
             'loss': ['deviance', 'exponential']
             }

gbc_optm = RandomizedSearchCV(estimator=gbc,
param_distributions=param_grid,n_iter=100, verbose=3)
gbc_optm.fit(X_train_sap, y_train_sap)

import pickle
filename = 'Model.sav'
pickle.dump(gbc_tunning, open(filename,'wb'))
```

App.py

```
from tkinter import *
import tkinter as tk
import pandas as pd
import numpy as np
import warnings
from sklearn.preprocessing import LabelEncoder
import pickle
warnings.filterwarnings("ignore", category=FutureWarning)
def save():
    a=depvar.get()
    b=tenvar.get()
    c=onsecvar.get()
    d=onbvar.get()
    e=devpvar.get()
    f=techvar.get()
    g=contvar.get()
    h=papvar.get()
    i=monvar.get()
    j=totvar.get()
    model = pickle.load(open('Model.sav', 'rb'))
    data = [[a,b,c,d,e,f,g,h,i,j]]
    df = pd.DataFrame(data, columns=['Dependents', 'tenure', 'OnlineSecurity',
        'OnlineBackup', 'DeviceProtection', 'TechSupport', 'Contract',
        'PaperlessBilling', 'MonthlyCharges', 'TotalCharges'])

    categorical_feature = {feature for feature in df.columns if df[feature].dtypes == 'O'}

    encoder = LabelEncoder()
    for feature in categorical_feature:
```

```
        df[feature] = encoder.fit_transform(df[feature])

    single = model.predict(df)
    probability = model.predict_proba(df)[:, 1]
    probability = probability*100
    if single == 1:
        op1 ="This Customer is likely to be Churned!"
        op2 = f"Confidence level is {np.round(probability[0], 2)}"
    else:
        op1 = "This Customer is likely to be Continue!"
        op2 = f"Confidence level is {np.round(probability[0], 2)}"
    label=Label(top,text = op1,fg='green',font=('Arial',15)).place(x=100,y=480)
    #label1=Label(top,text = op2,fg='green',font=('Arial',15)).place(x=100,y=520)
top = tk.Tk()
top.title('Customer Churn Prediction')
top.geometry("700x700")
Dependents=Label(top,text="DEPENDENTS").place(x=120,y=10)
depvar=tk.StringVar()
depvar.set("Yes")
drop = OptionMenu(top,depvar,"Yes","No",).place(x=250,y=10)
tenvar=tk.DoubleVar()
Tenute=Label(top,text="TENURE").place(x=120,y=50)
entry1=Entry(top,width=40,textvariable=tenvar).place(x=250,y=50)
onsecvar=tk.StringVar()
onlinesecurity=Label(top,text="ONLINE SECURITY").place(x=120,y=90)
onsecvar.set("Yes")
entry1=OptionMenu(top,onsecvar,"Yes","No").place(x=250,y=90)
onbvar=tk.StringVar()
onlineBackup=Label(top,text="ONLINE BACKUP").place(x=120,y=130)
```

```
onbvar.set("Yes")
entry1=OptionMenu(top,onbvar,"Yes","No").place(x=250,y=130)
devpvar=tk.StringVar()
deviceprotection=Label(top,text="DEVICE PROTECTION").place(x=120,y=170)
devpvar.set("Yes")
entry1=OptionMenu(top,devpvar,"Yes","No").place(x=250,y=170)
techvar=tk.StringVar()
techvar.set("Yes")
techsupport=Label(top,text="TECH SUPPORT").place(x=120,y=210)
entry1=OptionMenu(top,techvar,"Yes","No").place(x=250,y=210)
contvar=tk.StringVar()
contvar.set("One year")
contract=Label(top,text="CONTRACT").place(x=120,y=250)
entry1=OptionMenu(top,contvar,"Month-to-month","One Year","Two
years").place(x=250,y=250)
papvar=tk.StringVar()
papvar.set("Yes")
paperlessbilling=Label(top,text="PAPERLESS BILLING").place(x=120,y=290)
entry1=OptionMenu(top,papvar,"Yes","No").place(x=250,y=290)
monvar=tk.DoubleVar()
monthlycharges=Label(top,text="MONTHLY CHARGES").place(x=120,y=330)
entry1=Entry(top,width=40,textvariable=monvar).place(x=250,y=330)
totvar=tk.DoubleVar()
totalcharges=Label(top,text="TOTAL CHARGES").place(x=120,y=370)
entry1=Entry(top,width=40,textvariable=totvar).place(x=250,y=370)
submit=Button(top,text="SUBMIT",width="12",height="1",command=save).place(x=17,
y=440)
top.mainloop()
```