**A Major Project Report**

On

# PATRON DEFECTION DATA ANALYSIS

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

**BACHELOR OF TECHNOLOGY**

**In**

**COMPUTER SCIENCE AND ENGINEERING**

**By**


| | |
|---|---|
| **G. Lakshmi Sravani** | **(20NN1A0575)** |
| **S. Lavanya** | **(20NN1A05A4)** |
| **S. Vijaya Lakshmi** | **(20NN1A05B1)** |
| **S. Sri Lakshmi Prasanna** | **(21NN5A0508)** |

**Under the Esteemed Guidance of**

**Dr. A. Naresh**

**Associate Professor**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**VIGNAN'S NIRULA INSTITUTE OF TECHNOLOGY AND SCIENCE FOR WOMEN PEDAPALAKALURU, GUNTUR-522005**

**(Approved by AICTE, NEW DELHI and Affiliated to JNTUK, Kakinada)**


**2020-2024**

**VIGNAN'S NIRULA INSTITUE OF TECHNOLOGY AND SCIENCE FOR WOMEN PEDAPALAKALURU, GUNTUR-522005**

**(Approved by AICTE, NEW DELHI and Affiliated to JNTUK, Kakinada)**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

# CERTIFICATE

This is to certify that the major project entitled **"Patron Defection Dta Analysis"**, is a bonafide work of **G.Lakshmi Sravani (20NN1A0575), S.Lavanya (20NN1A05A4), S.Vijaya Lakshmi (20NN1A05B1) and S.Sri Lakshmi Prasanna (21NN5A0508)** submitted to the faculty of Computer Science And Engineering, in the partial fulfilment of the requirements for the award of degree of **BACHELOR OF TECHNOLOGY** in **COMPUTER SCIENCE AND ENGINEERING** from **VIGNAN'S NIRULA INSTITUTE OF TECHNOLOGY AND SCIENCE FOR WOMEN, GUNTUR.**

| | |
|---|---|
| **Project Guide** | **Head of the Department** |
| **Dr. A. Naresh** | **Dr. V. Lakshman Narayana** |
| **Associate Professor** | **Associate Professor** |

**EXTERNAL EXAMINER**

## DECLARATION

We hereby declare that the work described in this major project work, entitled **“Patron defection data analysis”** which is submitted by us in partial fulfilment for the award of **Bachelor of Technology** in the Department of **Computer Science and Engineering** to the **Vignan’s Nirula Institute of Technology and Science for women**, affiliated to Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, is the result of work done by us under the guidance of **Dr.A.Naresh,** Associate Professor.

The work is original and has not been submitted for any Degree/ Diploma of this or any other university.

**Place:**

**Date:**

| | |
|---|---|
| **G.Lakshmi Sravani** | **(20NN1A0575)** |
| **S.Lavanya** | **(20NN1A05A4)** |
| **S.Vijaya Lakshmi** | **(20NN1A05B1)** |
| **S.Sri Lakshmi Prasanna** | **(21NN5A0508)** |

## ACKNOWLEDGEMENT

We express our heartfelt gratitude to our beloved principal **Dr. P. Radhika** for giving a chance to study in our esteemed institution and providing us all the required resources.

We would like to thank **Dr. V. Lakshman Narayana, Associate Professor, Head of the Department of Computer science and Engineering**, for her extended and continuous support, valuable guidance and timely advices in the completion of this mini project thesis.

We wish to express our profound sense of sincere gratitude to our Project Guide **Dr.A.Naresh, Associate Professor, Department of Computer Science and Engineering**, without whose help, guidance and motivation this mini project thesis could not have been completed the mini project successfully.

We also thank all the faculty of the Department of Computer Science and Engineering fortheir help and guidance of numerous occasions, which has given us the cogency to build- up adamant aspiration over the completion of our mini project thesis.

Finally, we thank one and all who directly or indirectly helped us to complete our mini project thesis successfully.



| | |
|---|---|
| **G.Lakshmi Sravani** | **(20NN1A0575)** |
| **S.Lavanya** | **(20NN1A05A4)** |
| **S.Vijaya Lakshmi** | **(20NN1A05B1)** |
| **S.Sri Lakshmi Prasanna** | **(21NN5A0508)** |

# Patron Defection Data Analysis

## TABLE OF CONTENTS

# ABSTRACT

The modern business environment presents service providers with increasing competition as well as changing customer preferences, which makes it imperative to lower customer churn—the phenomenon when customers stop doing business with a company. This research investigates the effectiveness of different machine learning techniques for churn prediction in a range of businesses, highlighting the necessity of keeping current customers in the face of growing markets and rising acquisition prices. Among the well-known approaches examined are ensemble methods, K-nearest neighbors, neural networks, decision trees, random forests, gradient boosting machines, support vector machines, and gradient boosted trees. Furthermore, it emphasizes how crucial feature engineering is to improving churn prediction models by obtaining relevant consumer information including usage patterns, transaction histories, and demographics. By implementing these techniques, service providers can proactively.

## LIST OF FIGURES

# LIST OF ABBREVIATION

| | | |
|---|---|---|
| **ROI** | - | Return On Investment |
| **KNN** | - | K-Nearest Neighbors |
| **CRM** | - | Customer Relationship Management |
| **CCP** | - | Customer Churn Prediction |
| **SOM** | - | Self-Organized Maps |
| **ANN** | - | Artificial Neural Network |
| **UNL** | - | Unified Modeling Language |
| **NBC** | - | Naïve Bayes Classifier |
| **DT** | - | Decision Tree |
| **LR** | - | Logistic Regression |

# Chapter – 1
# INTRODUCTION

# CHAPTER - 1
# INTRODUCTION

## 1. Introduction

In nations with advanced economies, the telecommunications industry has grown to be one of the key sectors. The volume of competition has increased due to advancements in technology and an increase in the number of operators. Businesses are using a variety of tactics and putting in a lot of effort to thrive in this cutthroat environment. To enhance revenue, three primary tactics have been suggested: (1) acquiring new customers; (2) upselling current customers; and (3) lengthening the client retention term. But when these strategies are compared and their respective come back on investment (RoI) is taken into consideration, it becomes clear that the third method is the most effective. It also demonstrates that keeping an existing customer is far less expensive than gathering a new one, and it is also thought to be much simpler compared to the upselling strategy. Using the third technique businesses must lessen the possibility of "the movement of customers from a single provider to another," often known as customer churn.

A major worry in service industries with fierce competition is customer attrition. However, if done early on, identifying the clients who are most likely to abandon the organization might result in a sizable increase in income. Numerous studies have demonstrated how successful machine learning technique is in foreseeing this kind of scenario. This method is used by using lessons learned from historical data.

A common definition of "churning" is the proportion of customers that terminate their contracts because of rivalry. Churners are clients who have left a company because they were dissatisfied with the level of service they received. An analysis of the probability that a client would stop using a good or service is known as a consumer churn analysis.

These days, the market is extremely competitive and dynamic. The availability of a vast array of service providers is the reason behind this. Since they are a company's main source of income, customers are its most important asset. Companies today understand that maintaining the satisfaction of their current customer base is just as important as acquiring new ones.

A churner is someone who is constantly on the go for a multitude of reasons. When an organization can predict customer mentality with accuracy and create connections between loss of customers and controllable factors, customer churn is reduced .

One binary classification task that distinguishes churners from non-churners is predicting churn ratesAny business that wants to beat out new clients must walk through the sales funnel and use its advertising and sales resources to close the loop.In contrast, customer retention is typically more cost-effective because it has already gained the trust and loyalty of existing clients. Thus, it is critical for a firm to anticipate client attrition early.

To mitigate the aforementioned discomfort, enterprises need to possess the ability to precisely predict the purchasing patterns of their clientele. Two strategies exist for reducing client attrition: Above all, be (1) Reactive and (2) Proactive. A proactive agency offers the customer enticing options as soon as they request a cancellation in order to keep their business. Since the proactive method foresees that some consumers may go, it offers a plan for them to adhere to. The binary task classifies the churners and non-churners. We applied a couple of machine learning strategies to address this issue: Utilizing boosting methods such as XGBoost, Logistic Regression (LR), the KNN, the decision tree classifier, random-forest classification algorithm. Additional tree classification algorithm, Ridge classifier, and Bagging Ridge classifier,

# Chapter – 2

# LITERATURE SERVEY

# CHAPTER – 2
# LITERATURE SERVEY

## 2. LITERATURE SURVEY

Comparing oversampling techniques to handle the class imbalance problem a customer churn prediction case study" by Amin A, Anwar S, Adnan A, Nawaz M, Howard N, Qadir J, Hawalah A, and Hussain A aimed to tackle the challenge of class imbalance in customer churn prediction models. By employing oversampling techniques, the researchers sought to rebalance the dataset for more accurate predictions.

Evaluation of churn rate using modified random forest technique in telecom industry. The modified random forest model achieved 92.7% accuracy in predicting customer churn, outperforming the standard random forest model. Feature engineering and normalization were key factors in improving the accuracy of the churn prediction model.

A proposed churn prediction model. The clustering step produces 3 clusters for retention strategy evaluation. The churn prediction model aims to detect customers likely to leave a service provider using Data Mining techniques for classification and clustering. The SVM technique yielded the best accuracy rate of 83.7%, with error rates ranging from 16.3% to 22.1% for different techniques.

Comparative Analysis of Different Machine Learning Algorithms to Predict Online Shoppers Behavior. The study compares various machine learning algorithms to predict online shoppers' behavior. Algorithms like Gradient Boosting (GB), Random Forest (RF), Support Vector Machine (SVM), and others were evaluated. The best performing algorithm, GB classifier, achieved an accuracy of around 91% and a precision of 76%.

Developing a Prediction Model for Customer Churn from Electronic Banking Services. The paper presents a predictive model for customer churn in electronic banking services using logistic regression and decision tree algorithms. It achieved high accuracy with an AUC of 0.929 and precision, recall, and F-measure values of 91.81%, 91.00%, and 90.96% respectively.

Customer Churn Prediction Using Four Machine Learning Algorithms. It utilized feature selection, normalization, and feature engineering techniques to enhance model performance. The experiments showed that Gradient Boosting with feature selection outperformed other algorithms, achieving a 99% F1-score and 99% AUC, indicating the effectiveness of machine learning in predicting customer churn in the telecom sector.

Customer churn prediction in telecom using machine learning in big data platform By leveraging a big data platform, the research aims to improve the efficiency and effectiveness of churn prediction models in the telecommunications sector. However, challenges related to data availability and feature engineering remain as potential limitations in this domain.

Customer Churn Prediction in B2B Contexts. The study proposes a two-stage process involving data mapping and prediction modeling to predict customer churn effectively in B2B contexts. It emphasizes the significance of single customers in B2B businesses and the impact of losing them.

A Churn Model for Swiss Mandatory Health Insurance. The model focuses on actuarial pricing, predicts portfolio structure, and considers premium sensitivity. The accuracy of the model is evaluated using metrics like Pricing Loss, Binomial Deviance, and AUC.

A Survey on Customer Churn Prediction using Machine Learning Techniques. The paper reviews popular machine learning algorithms for churn prediction, emphasizing the importance of retaining customers in a competitive market. It discusses the use of SVM with boosting algorithms to address churn issues and suggests constant model development for improved accuracy.

| **Title** | **Author** | **Year Published** | **Algorithms Used** | **Advantages** | **Accuracy** | **Limitations** |
|---|---|---|---|---|---|---|
| Comparing oversampling techniques to handle the | Jutla DN, Sivakumar SC, Zhan J, Guidibande V, Parsa SPK | 2016 | Gradient Boosting Model, Weighted Random Forests, XGBOOS | Enhanced prediction accuracy in customer churn models. | The study was 93.3% using the XGBOOST algorithm | Limited availability of data due to restrictions from telecom companies |

| | | | | | | |
|---|---|---|---|---|---|---|
| class imbalance problem: a customer churn prediction case study | | | T | | | |
| Evaluation of churn rate using modified random forest technique in telecom industry | P.Swetha, S.Usha, and S.Vijay anand | 2018 | Random Forest | Improved evaluation of churn rate in the telecom industry | An accuracy of 92.7% with the random forest technique | It focuses solely on performance of the random forest model without comparing it to other machine learning algorithms. |
| A proposed churn prediction model | Shaaban E, Helmy Y, Khedr A, Nasr M | 2012 | Churn prediction model- DT, SVM, and NN for classification and K Means for clustering | Predicts customer churn effectively | SVM technique yields an accuracy rate of 83.7% and an error rate of 16.3% | Lacks in-depth interpretation of churn reasons and detailed discussion on classification technique |
| Comparative Analysis of Different Machine Learning Algorithms to Predict Online Shoppers Behaviour | Veena Parihar, Surendra Yadav | 2022 | Gradient Boosting (GB) classifier, Random Forest (RF), Support Vector Machine (SVM), and others | GB classifier showed the best performance with 91% accuracy and 76% precision | The accuracy of the best performing algorithm, Gradient Boosting (GB) classifier, in predicting online shoppers' behavior was around 91% | he study focused on online shopping behavior prediction, may not cover all aspects of consumer behavior |
| Developing a Prediction Model for | Keramati A, Ghaneei H, Mirmohammadi SM | 2016 | Logistic regression, decision tree | Prediction model development for | Achieved an AUC of 0.929 with precision, | A time-consuming process of data extraction. |

| | | | | | | |
|---|---|---|---|---|---|---|
| Customer Churn from Electronic Banking Services | | | | electronic banking services | recall, and F-measure values of 91.81%, 91.00%, and 90.96% | |
| Customer Churn Prediction Using Four Machine Learning Algorithms | Alanoud Moraya Aldalan, Abdulaziz Almaleh | 2023 | Random Forest, Decision Tree | Integrating feature selection and normalization in the telecom sector | Achieved high accuracy rates, with the Gradient Boosting model outperforming the other experiments by achieving a 99% F1-score and 99% AUC. | The lack of detailed comparison with a broader range of algorithms beyond the four selected. |
| Customer churn prediction in telecom using machine learning in big data platform | Abdelrahim Kasem Ahmad, Assef Jafar, Kadan Aljoumaa | 2019 | Random Forest, Neural Networks, Support Vector Machine, Gradient Boosting | Enhanced prediction accuracy using machine learning techniques | The accuracy achieved in the telecom sector using machine learning was 93.3% with the XGBOOST algorithm. | Limited availability of data for research purposes |
| Customer Churn Prediction in B2B Contexts | Iris Figalist, Christoph Elsner, Jan Bosch, and Helena Holmström Olsson | 2020 | Enables prediction based on customer and end-user data, considers influencing factors, and helps in retaining | Data mining techniques, Random Forest | Accuracy rates above 80% are considered good for customer churn prediction in B2B contexts | Lack of research on achieving customer churn prediction in B2B contexts, challenges in applying B2C approaches to B2B environments, |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | customers in B2B environments | | | and differences in perceived value between B2B and B2C customers |
| A Churn Model for Swiss Mandatory Health Insurance | Lena Schütte | 2022 | Incorporates actuarial pricing context, Predicts portfolio structure. Develops pricing loss function | Logistic regression, Gradient boosting machine | Model performance evaluated using Pricing Loss, Binomial Deviance, and AUC | Evaluation of model performance with respect to pricing risk not extensively researched |
| A Survey on Customer Churn Prediction using Machine Learning Techniques | Saran Kumar.A and Dr. Chandrakala D. | 2016 | Hybrid Neural Networks, Random Forests, SVM Model with Boosting Algorithms | Providing insights for targeted marketing strategies | SVM with boosting algorithms for higher accuracy and performance as future work | Some models may lack interpretability .Data imbalance can affect prediction accuracy. Constant development needed for improved models |

# Chapter - 3

# SYSTEM ANALYSIS

## 3.1 EXISTING SYSTEM:

The existing system for a Patron Defection Data Analysis project involves a combination of tools and technologies designed to collect, store, analyze, and report on customer data. Data collection tools, such as CRM systems and online surveys, gather information from various sources like transactions and interactions. This data is then stored securely using databases or data warehouses. Before analysis, the data undergoes preprocessing to clean and transform it. Statistical analysis and predictive modeling techniques are then applied to identify patterns and forecast patron defection. Visualization tools help communicate the findings through charts and dashboards, while integration and automation streamline processes. Security measures ensure data privacy and compliance with regulations, and ongoing monitoring and maintenance ensure the system's reliability and performance**.**

### Drawback:

One drawback of the existing system is the potential for data fragmentation and security concerns due to disparate sources and complex integration processes.

## 3.2 PROPOSED SYSTEM:

This paper is a relative study on the implementation of models using Logistic Regression and XGBoost algorithms, applied on a dataset that is taken from Kaggle repository. With respect to the results of accuracy, precision, recall, specificity and False Positive Rate the efficiencyof each algorithm is calculated and compared. All algorithms are written in python and executed in Google Colab, the Scientific Python Development Environment. Our experiments have shown that both Logistic Regression and XGBoost are the best algorithms for prediction with an accuracy of 80.38% and 80.103% each individually. As a result of our study that XGBoost is the extremely fast, stable, faster to tune and robust to randomness, which is well suited for tabular data where as A logistic regression model will try to guess the probability of belonging to one group or another. The logistic regression is essentially an extension of a linear regression, only the predicted outcome value is between [0, 1]. The model will identify relationships between our target feature, Churn, and our remaining features to apply probabilistic calculations for determining which class the customer should belong to.

1. Accuracy is high.

2. Efficient inference

3. Rate of hyper-parameter tuning (shorter the optimization time, the better)

### Advantage:

XGB boosting is an algorithm consisting of a group of weaker trees that add up their calculations to predict their target variable with better accuracy.

## 3.3 PURPOSE OF THE SYSTEM:

The globalization and advancements of telecommunication industry, exponentially raises the number of operators in the market that escalates the competition. In this competitive era, it has become mandatory to maximize the profits periodically, for that various strategies have been proposed, namely, acquiring new customers, up-selling the existing customers & increasing the retention period of existing customers. Among all the strategies, retention of existing customers is least expensive as compared to others. In order to adopt the third strategy, companies have to reduce the potential customer churn i.e., customer movement form the one service provider to other. The main reason of churn is the dissatisfaction of consumer service and support system. The key to unlock solutions to this problem is by forecasting the customers which are at risk of churning.

One of the main aim of Customer Churn prediction is to help in establishing strategies for customer retention. Along with growing competition in markets for providing services, the risk of customer churn also increases exponentially. Therefore, establishing strategies to keep track of loyal customers (non-churners) has become a necessity. The customer churn models aim to identify early churn signals and try to predict the customers that leave voluntarily. Thus many companies have realized that their existing database is one of their most valuable asset and according to Abbasdimehr, churn prediction is a useful tool to predict customers at risk

By knowing which clients are at the highest risk of leaving, we can better target our rescue efforts. For example, we can reach out to these clients with a marketing campaign, reminding them that they haven't purchased from us in a while, or even offering them a benefit. In addition to knowing which clients to target, we can use the churn model to calculate the maximum benefit price that is still worthwhile.

In our existing model, every involved weak classier is generally given a weight in conformity with the accuracy further training it using the re-weighted training data. XGBoost algorithm consisting of a group of weaker trees that add up their calculations to predict a target variable with better accuracy.

## 3.4 FEASIBILITY STUDY:

Feasibility Study in Software Engineering is a study to evaluate feasibility of propose project or system. Feasibility study in one of stage among important four stages of Software Project Management Process. Feasibility study is carried out based on many proposes to analyze whether software product will be right in terms of development, implantation, contribution of project to the organization etc. During system analysis the feasibility study of the proposed system is to be carried out. Feasibility study is so important stage of Software Project Management Process as after completion of feasibility study it gives a conclusion of whether to go ahead with proposed project as it is practically feasible or to 13 stop proposed project here as it is not right/feasible to develop or to think/analyze about proposed project again.
Three key considerations involved in the feasibility analysis are,

- TECHNICAL FEASIBILITY
- ECONOMICAL FEASIBILITY
- SOCIAL FEASIBILITY

### Technical Feasblity

In Technical Feasibility current resources both hardware software along with required technology are analyzed/assessed to develop project. This technical feasibility study gives report whether there exists correct required resources and technologies which will be usedfor project development. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to highdemands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

## Economical Feasbility

In Economic Feasibility study cost and benefit of the project is analyzed. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus, the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

## Social Feasbility

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level ofacceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar within it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed,as he is the final user of the system

# Chapter - 4
# SOFTWARE REQUIREMENTS SPECIFICATIONS

# CHAPTER – 4
# SOFTWARE REQUIREMENTS SPECIFICATIONS

## 4.1GENERAL

These are the requirements for doing the project. Without using these toolsand software's we can't do the project. So, we have two requirements to do the project. They are:

1. Hardware Requirements.
2. Software Requirements

## 4.2HARDWARE REQUIREMENTS

The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete and consistent specification of the wholesystem. They are used by software engineers as the starting point for the system design. It should what the system does and not how it should be implemented.

- Processor: Core i5
- RAM: 4 GB
- OS: Windows 7/8/10 (32 or 64 bit)

## 4.3 SOFTWARE REQUIREMENTS

The software requirements document is the specification of the system. It should includeboth a definition and a specification of requirements. It is a set of what the system shoulddo rather than how it should do it. The software requirements provide a basis for creatingthe software requirements specification. It is useful in estimating cost, planning team activities, performing tasks and tracking the teams and tracking the team's progress throughout the development activity.

- Python libraries such as pandas, NumPy.
- Sklearn library using with many subordinate models such as model selectionclassifiers, and metrics etc.
- SUBLIME TEXT for implementation purpose.
- Sublime Text is a shareware cross-platform source code editor.

## 4.4FUNCTIONAL REQUIREMENTS

A functional requirement defines a function of a software-system or its component. A function is described as a set of inputs, the behavior, in proposed system, XGBoost is anensemble learning algorithm, and is used for classification, regression and for ranking problems. XGBoost has the advantage over other algorithms that it is highly flexible, usesthe power of parallel processing, and supports Regularization, faster than Gradient Boosting. It has been found to provide a good estimate of generalization error and resistant to overfitting. In Patron Defection Data Analysis**,** Customer churn dataset is collected and divided it as 80% of data set for training and 20% of data for testing.Using XGBoost, user can run cross validation after each iteration. Finally, the optimization is done and the accuracy obtained by XGBoost is 80.103% and Logistic Regression is 80.38%.

# Chapter – 5

# SYSTEM DESIGN

# CHAPTER – 5

# SYSTEM DESIGN

## 5.1 WORK FLOW:

The data was provided by a Finnish insurance company that wanted to predict whether acustomer is going to stay or leave after the current period or not. Leaving or churning is defined as “Yes” and not leaving is defined as “No” in this dataset. The selection of a data set from a choice of separate or distinct implicitly applicable ones seem to be nothing but a numerical comparison between the essentials of the business with the practicable data quality enumerated in the metadata description.

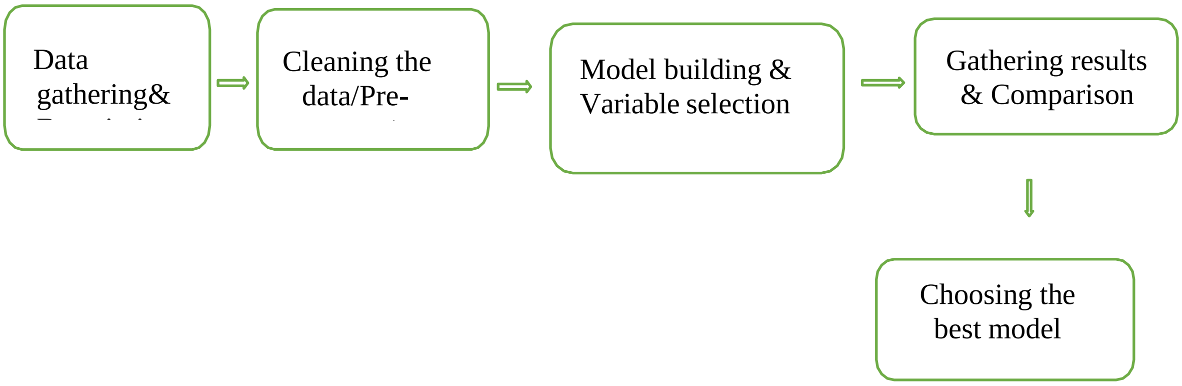



**Fig 5.1: Process of Choosing the Best Model**

Implementation of various models to predict churn on the data before and after oversampling the data, after splitting the data into test and train set. The train set includesthe values in the churn column which is used to train the model and same model will be applied on test set, to test the quality of the model.

## 5.2 MODEL DEVELOPMENTt:

### The Data

As mentioned above, the data is sourced from Kaggle. In our dataset, we have 7043 rows(each representing a unique customer) with 21 columns: 19 features, 1 target feature (Churn). The data is composed of both numerical and categorical features, so we will needto address each of the datatypes respectively.

**Target:**

- **Churn** — Whether the customer churned or not (Yes, No)

**Numeric Features:**

- **Tenure** — Number of months the customer has been with the company
- **MonthlyCharges** — The monthly amount charged to the customer
- **TotalCharges** — The total amount charged to the customer

**Categorical Features:**

- **CustomerID**
- **Gender** — M/F
- **SeniorCitizen** — Whether the customer is a senior citizen or not (1, 0)
- **Partner** — Whether customer has a partner or not (Yes, No)
- **Dependents** — Whether customer has dependents or not (Yes, No)
- **PhoneService** — Whether the customer has a phone service or not (Yes, No)
- **MulitpleLines** — Whether the customer has multiple lines or not (Yes, No, NoPhone Service)
- **InternetService** — Customer's internet service type (DSL, Fiber Optic, None)
- **OnlineSecurity** — Whether the customer has Online Security add-on (Yes, No,No Internet Service)
- **OnlineBackup** — Whether the customer has Online Backup add-on (Yes, No, NoInternet Service)
- **DeviceProtection** — Whether the customer has Device Protection add-on (Yes,No, No Internet Service)
- **TechSupport** — Whether the customer has Tech Support add-on (Yes, No, NoInternet Service)
- **StreamingTV** — Whether the customer has streaming TV or not (Yes, No, NoInternet Service)
- **StreamingMovies** — Whether the customer has streaming movies or not (Yes,No, No Internet Service)
- **Contract** — Term of the customer's contract (Monthly, 1-Year, 2-Year)
- **PaperlessBilling** — Whether the customer has paperless billing or not (Yes, No)
- **PaymentMethod** — The customer's payment method (E-Check, Mailed Check,Bank Transfer (Auto), Credit Card (Auto))

## Algorithms

Implementation of various models to predict churn on the data before and after oversampling the data, after splitting the data into test and train set. The train set includesthe values in the churn column which is used to train the model and same model will be applied on test set, to test the quality of the model. Following are the different models used to predict churn with its implementation considering ensemble modelling and boosting methods for best results:

**Random Forest:** Random Forest, like the name indicates, conforms a substantial numberof unique decision trees that emits a class prediction and administer as an ensemble. Finally, the model's prediction is decided to be the class with the maximum votes.

A great number of relevantly uncorrelated models is the reason that the random forest model strives so well i.e., the individual trees administering as a committee performs better than any of the individual component models.

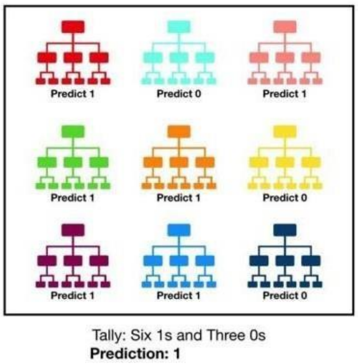



**Fig – 5.2**: **Random Forest Architecture**

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples andtakes 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 handlethe data set containing continuous variables as in the case of regression and categoricalvariables as in the case of classification. It performs better results for classification problems.

**KNN Classifier:** "The k-nearest neighbors (KNN) algorithm" is an absolute and easy toimplement machine learning algorithm. It makes an assumption that alike entities exist incloseness to each other.

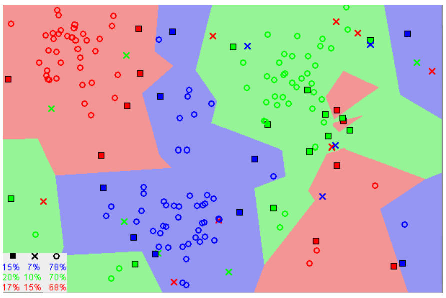


**Fig – 5.3: KNN Classifier**

In the Fig-5.3, majority of the time, the data points that are similar to each other are near each other. For this algorithm to be useful, it assumes this to be true. The KNN algorithmdetermines the proximity by calculating the distance between the points on the graph.

K Nearest Neighbor algorithm falls under the Supervised Learning category and is used for classification (most commonly) and regression. It is a versatile algorithm also used for imputing missing values and resampling datasets. As the name (K Nearest Neighbor)suggests it considers K Nearest Neighbors (Data points) to predict the class or continuousvalue for the new Datapoint.

**Logistic Regression:** Logistic regression predicts the probability of a classification by calculating the relationship between one or more independent features with a dependent variable. Definitively, logistic regression predicts the chances of a data point that belongsto the default bracket. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes. It is an extensively employed algorithm for classification in industry. The logistic regression model, like theAdaline and perceptron, is a statistical method for binary classification that can be generalized to multiclass classification.

The below curve is known as the Sigmoid Curve. The sigmoid function is represented bythe symbol sigma. Its graphical behavior has been described in the below figure.

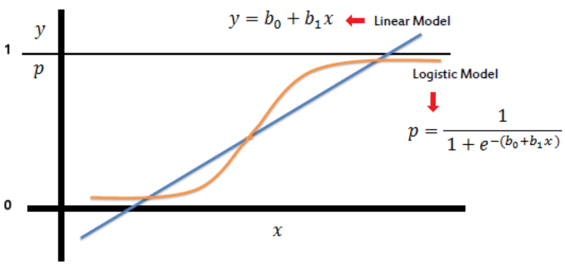



**Fig – 5.4: Logistic Regression Graph**

Logistic Regression, a method for classifying the data in Machine Learning. Logistic regression is generally used where we have to classify the data into two or more classes. One is binary and the other is multi-class logistic regression. As the name suggests, the binary class has 2 classes that are Yes/No, True/False, 0/1, etc. In multi-class classification, there are more than 2 classes for classifying data.

**Decision Tree Classifier:** The decision tree classifier creates theclassification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute. Each leaf represents class labels associated with the instance. Instances in the training set are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. Starting from the rootnode of the tree, each node splits the instance space into two or more sub-spaces according to an attribute test condition. Then moving down the tree branch corresponding to the value of the attribute, a new node is created. This process is then repeated for the subtreerooted at new node, until all records in the training set have been classified.

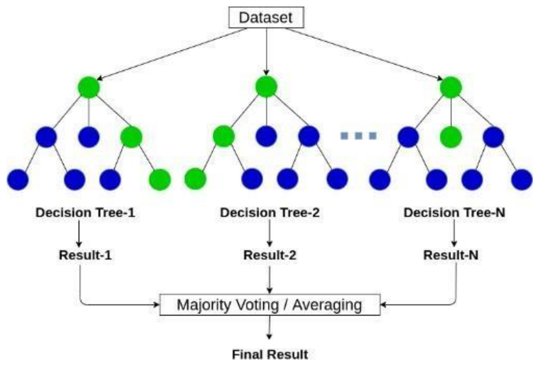



**Fig – 5.5: Decision Tree Classifier**

Extra Tree Classifier: Extremely Randomized Trees Classifier (Extra Trees Classifier) is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a "forest" to output its classification result. In concept, it is very similar to a Random Forest Classifier and only differs fromit in the manner of construction of the decision trees in the forest.

Extra Trees Forest is constructed from the original training sample. Then, at each test node, each tree is provided with a random sample of k featuresfrom the feature-set from which each decision tree must select the best feature to split thedata based on some mathematical criteria (typically the Gini Index).

**Gradient Boost Classifier:** Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees;it usually outperforms random forest. A gradient-boosted trees model is built in a stage- wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

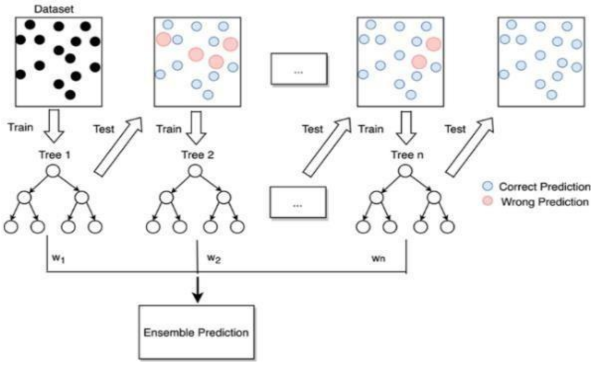



**Fig –5.6: Gradient Boosting Classifier**

**Ridge Classifier:** Ridge regression is a model tuning method that is used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. When issue of multicollinearity occurs, least-squares are unbiased,and variances are large, this results in predicted values being far away from the actual values.

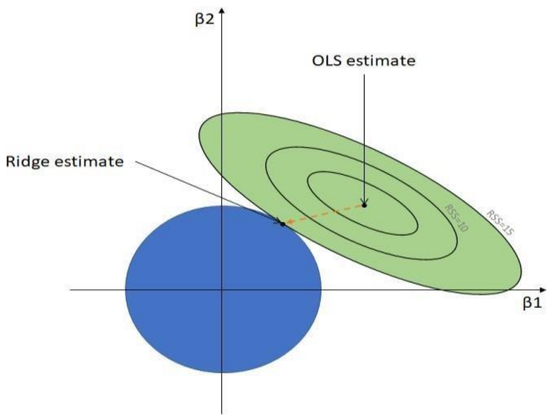



**Fig –5.7: Ridge Classifier Graph**

The Ridge Classifier, based on Ridge regression method, converts the label data into [-1,1] and solves the problem with regression method. The highest value in prediction is accepted as a target class and for multiclass data multi-output regression is applied.

**Bagging Ridge Classifier:** A Bagging regressor is an ensemble meta-estimator that fits base regressors each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Sucha meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its constructionprocedure and then making an ensemble out of it.

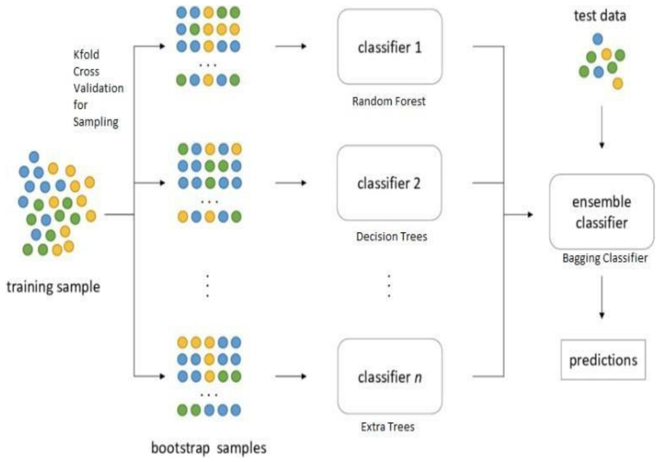



**Fig – 5.8: Bagging Ridge Classifier**

When random subsets of the dataset are drawn as random subsets of the samples, then this algorithm is known as Pasting. If samples are drawn with replacement, then the method is known as Bagging. When random subsets of the dataset are drawn as random subsets of the features, then the method is known as Random Subspaces. Finally, when base estimators are built on subsets of both samples and features, then the method is known as Random Patches.

**XGBoost Classifier:** XGBoost is the implementation of the gradient boosted tree algorithms that is commonly used for classification and regression problems. Gradient boosting is an algorithm consisting of a group of weaker trees that add up their calculations to predict a target variable with better accuracy.

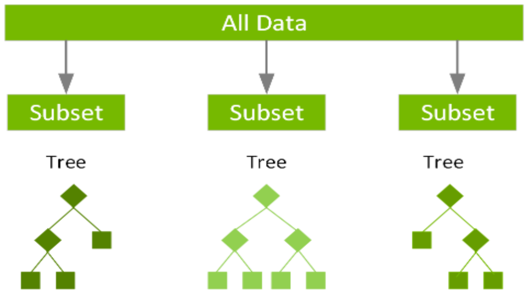



**Fig –5.9: XGBoost Classifier**

In this algorithm, decision trees are created in sequential form. Weights play an importantrole in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrongby the tree is increased and these variables are then fed to the second decision tree. Theseindividual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, rank and user-defined prediction problems.

## 5.2 UML DIAGRAMS:

### General

Configuration Engineering manages the different UML [Unified Modeling language] graphs for the execution of task. Configuration is a significant designing portrayal of a thing that willbe constructed. Programming configuration is an interaction through whichthe prerequisites are converted into portrayal of the product. Configuration is where quality is delivered in programming. Configuration is the way to precisely make an interpretation of client necessities into completed item.

Diagrams in UML can be broadly classified as:

**Structural Diagrams** – Capture static aspects or structure of a system. Structural Diagrams include: Component Diagrams, Object Diagrams, Class Diagrams and Deployment Diagrams.

**Behavior Diagrams** – Capture dynamic aspects or behavior of the system. Behavior diagrams include: Use Case Diagrams, State Diagrams, Activity Diagrams and Interaction Diagrams.

### 5.3.1 Class Diagram:

In software engineering, a class diagram in the Unified Modeling 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.
The class shape itself consists of a rectangle with three rows. The top row contains the name of the class, the middle row contains the attributes of the class, and the bottom section expresses the methods or operations that the class may use. Classes and subclassesare grouped together to show the static relationship between each object

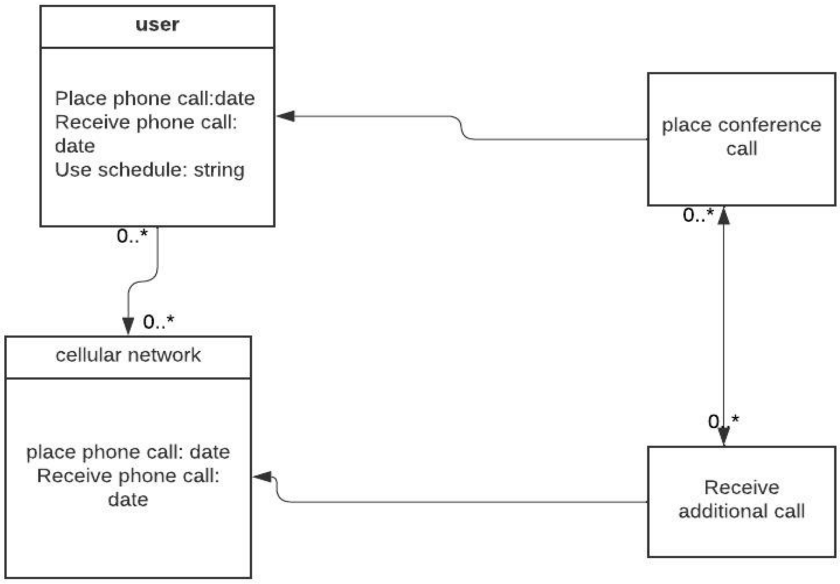



**Fig – 5.10: Class Diagram**

### 5.3.2 Use case Diagram:

The purpose of a use case diagram in UML is to demonstrate the different ways that a user might interact with a system. In the Unified Modeling Language (UML), a use casediagram can summarize the details of your system's users (also known as actors) and theirinteractions with the system. To build one, you'll use a set of specialized symbols and connectors.

UML is the modelling toolkit that you can use to build your diagrams. Use cases are represented with a labelled oval shape. Stick figures represent actors in the process, and the actor's participation in the system is modelled with a line between the actor and use case. To depict the system boundary, draw a box around the use case itself.

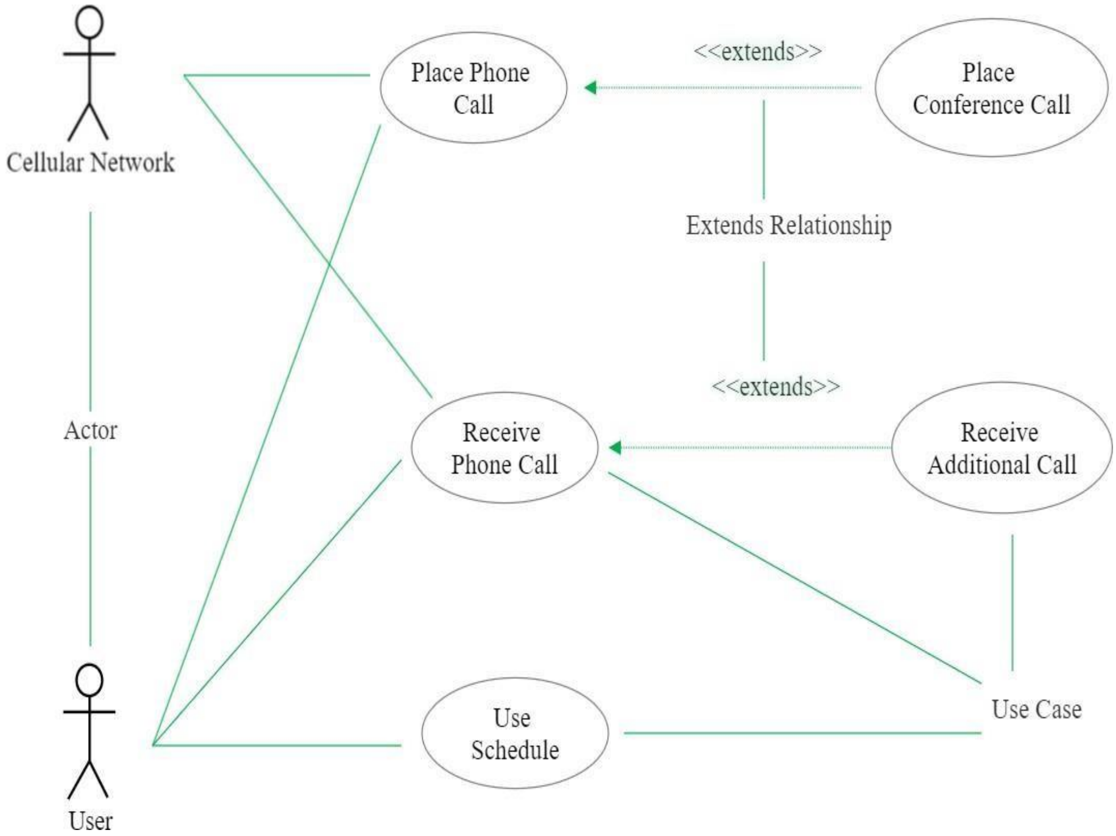



**Fig – 5.11: Use Case Diagram**

### 5.3.3 Activity Diagram:

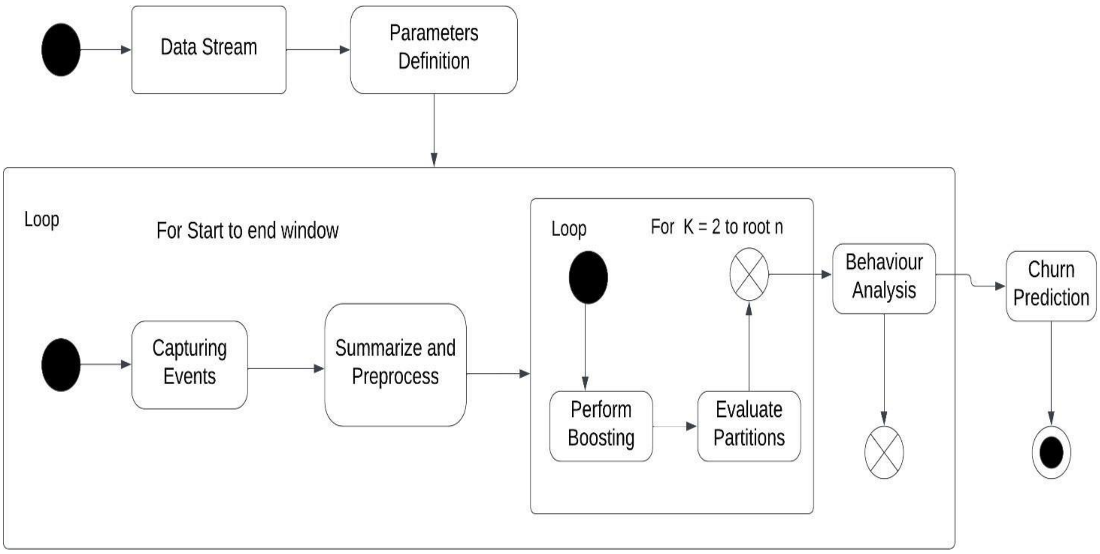



**Fig –5.12: Activity diagram for the Proposed System**

## Explanation:

We use **Activity Diagrams** to illustrate the flow of control in a system and refer to the steps involved in the execution of a use case. We model sequential and concurrent activities using activity diagrams. So, we basically depict workflows visually using an activity diagram. An activity diagram focuses on condition of flow and the sequence in which it happens. We describe or depict what causes a particular event using an activity diagram. UML models basically three types of diagrams, namely, structure diagrams, interaction diagrams, and behavior diagrams. An activity diagram is a **behavioral diagram** i.e., it depicts the behavior of a system. An activity diagram portrays the control flow from a start point to a finish point showing the various decision paths that exist while the activity is being executed. We can depict both sequential processing and concurrent processing of activities using an activity diagram. They are used in business and process modelling where their primary use is to depict the dynamic aspects of a system. An activity diagram is very **similar to a flowchart**.

### 5.3.4 Data Flow Diagram:

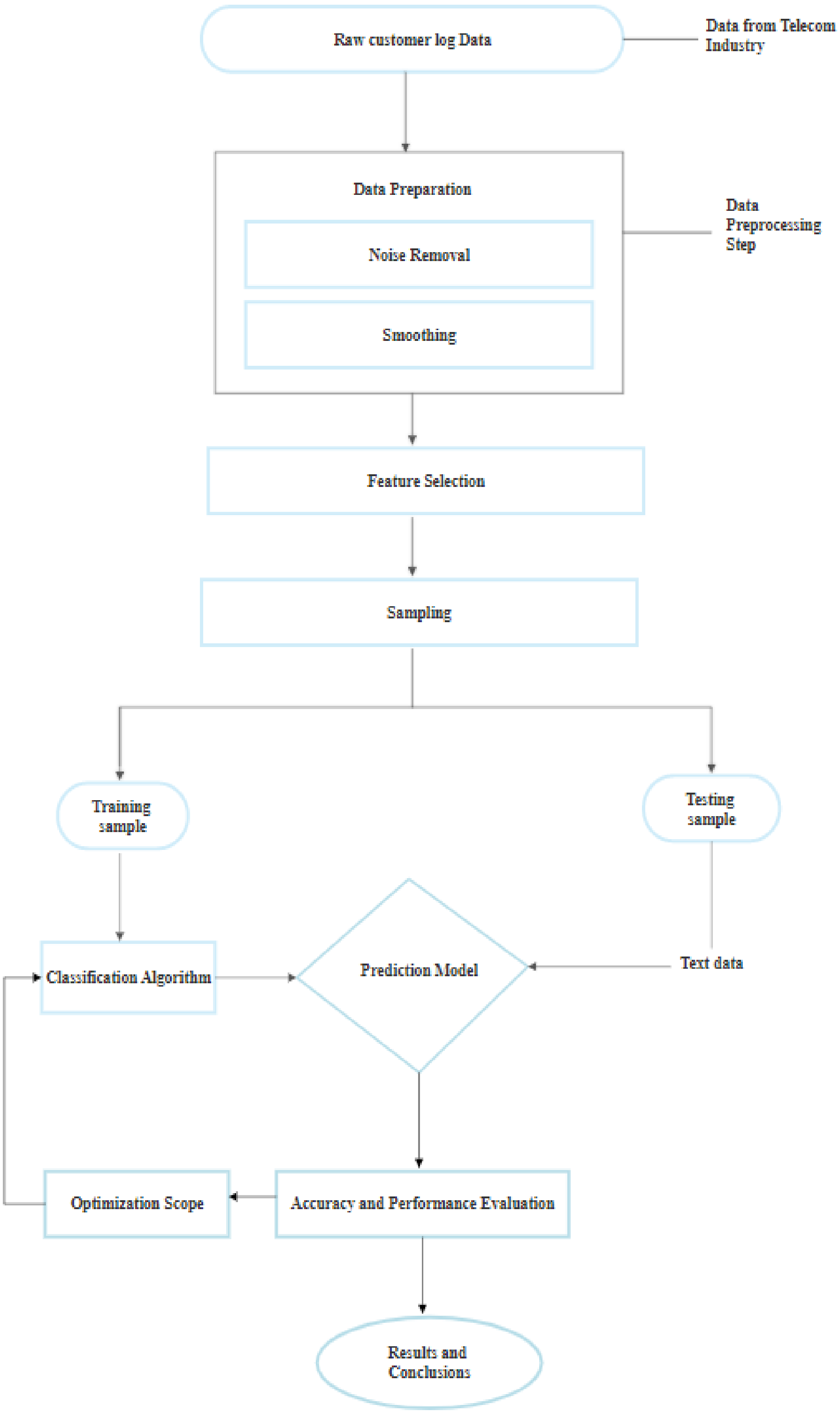



**Fig: 5.13 Data Flow Diagram**

## Explanation:

A data-flow diagram is **a way of representing a flow of data through a process or a system** (usually an information system). The DFD also provides information about the outputs and inputs of each entity and the process itself.

Data Flow Diagrams provide a straightforward, efficient way for organizations to understand, perfect, and implement new processes or systems. They're visual representations of your process or system, so they make it easy to understand and prune.You can either use logical or physical diagrams to describe the same flow of informationor you can use them in conjunction to understand a process or system on a more granularlevel. But, before you can use a DFD to understand your system or process' flow of information, you need to know the standard notations or symbols used to describe it.

## 5.3 System Architecture:

The original features with customers' consumption and behavior are equidistantly grouped to construct new features. The stacking model consists of two levels with four algorithms: Xgboost (XGB), Logistic regression (LR), Decision tree (DT) and Naive Bayes classifier (NBC) to achieve better prediction accuracy. The third step consists of asoft voting. The results of the stacking model are input to the soft voting.

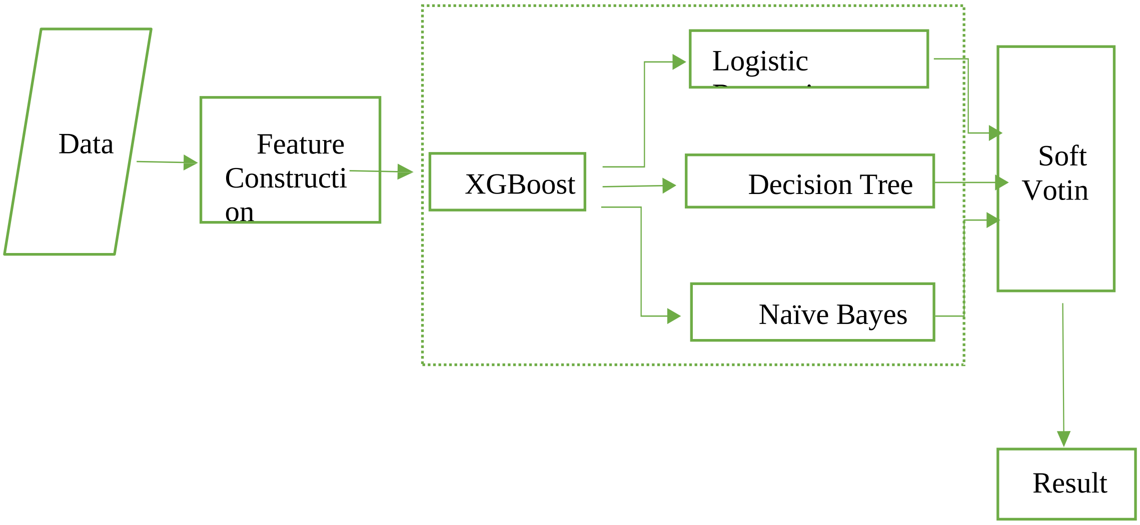



**Fig: 5.14 System Architecture**

The stacking model is a general method of using a higher-accuracy algorithm to combinelower-accuracy algorithms to achieve greater predictive accuracy. The best results are obtained when the higher-accuracy algorithm is in a first level and lower-accuracy algorithms are in a second level. In this study, the stacking model consists of two layers as shown in, level 1 and level 2. A higher accuracy model is used in the first layer (level 1), while lower-accuracy models are used in the second layer (level 2).

Soft voting estimates the class probability with different algorithms having contrasting approaches to improve the accuracy of prediction. It assigns a larger weight to the important classifier, and the highest category is selected by summing the probabilities predicted by the model. This proposed system can determine important factors affecting customer purchasing behavior in the telecommunications industry.

# Chapter - 6

# IMPLEMENTATION

# CHAPTER – 6
# IMPLEMENTATION

## 6.1 MODEL IMPLEMENTATION:

Implementation of various models to predict churn on the data before and after oversampling the data, after splitting the data into test and train set. The train set includesthe values in the churn column which is used to train the model and same model will be applied on test set, to test the quality of the model. Following are the different models used to predict churn with its implementation considering ensemble modeling andboosting methods for best results:

### Performance Evaluation Parameters

The following evaluation parameters used

**False Positive (FP):** refers to the number of negative tuples that were incorrectly labelledas positive.

**False Negative (FN):** refers to the number of positive tuples that were incorrectly labelledas negative.

**True Positive (TP):** refers to the positive tuples that were labelled correctly as positive. **True negative (TN):** refers to the number of negative tuples that were labelled correctlyby the classifier.

**Precision:**

Precision is the number of correct results divided by the number of all returned results.

Precision = tp / (tp + fn) (1)

**Recall:**

Recall is the number of correct results divided by the number of results that shouldhasbeen returned.

Recall = tp / (tp + fn) (2)

**F1-Score:**

A measure that combines precision and recall is the harmonic mean of precision and recall.

F = (2 * precision ∗ recall) / (precision + recall) (3)

**Measure of evaluation:**

The probability of success in recognizing the right class of an instance.

Accuracy = (tp + tn) / (tp + tn + fp + fn) (4)

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.96 | 0.97 | 0.96 | 530 |
| 1 | 0.97 | 0.96 | 0.97 | 652 |
| accuracy | | | 0.96 | 1182 |
| macro avg | 0.96 | 0.96 | 0.96 | 1182 |
| weighted avg | 0.96 | 0.96 | 0.96 | 1182 |

## 6.2 Modules:

### NumPy

This article will help you get acquainted with the widely used array-processing libraryin Python, NumPy. **What is NumPy?** NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It is open-source software. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

### Pandas

pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labelled" data both easy and intuitive. Itaims to be the fundamental high-level building block for doing practical, real-world dataanalysis in Python.

Additionally, it has the broader goal of becoming the most powerful and flexible open- source data analysis/manipulation tool available in any language**.**

Here are just a few of the things that pandas do well:

- Easy handling of **missing data** (represented as NaN) in floating point as well as non-floating-point data.
- Size mutability: columns can be **inserted and deleted** from DataFrame and higher dimensional objects.
- Automatic and explicit **data alignment**: objects can be explicitly aligned to a set of labels, or the user can simply ignore the labels and let Series, DataFrame, etc. automatically align the data for you in computations.
- Powerful, flexible **group by** functionality to perform split-apply-combine operations on data sets, for both aggregating and transforming data.
- Make it **easy to convert** ragged, differently-indexed data in other Python and NumPy data structures into DataFrame objects.
- Intelligent label-based **slicing**, **fancy indexing**, and **subsetting** of large data sets.
- Intuitive **merging** and **joining** data sets.
- Flexible **reshaping** and pivoting of data sets.
- **Hierarchical** labelling of axes (possible to have multiple labels per tick).
- Robust IO tools for loading data from **flat files** (CSV and delimited), Excel files,databases, and saving / loading data from the ultrafast **HDF5 format**.
- **Time series**-specific functionality: date range generation and frequency conversion, moving window statistics, date shifting, and lagging.

## Seaborn

Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

## Matplotlib

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.It provides an object-oriented API for embedding plots into applications using general- purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.

## Plotly

The plotly Python library is an interactive, open-source plotting library that supports over40 unique chart types covering a wide range of statistical, financial, geographic, scientific,and 3-dimensional use-cases.

Built on top of the Plotly JavaScript library (plotly.js), plotly enables Python users to create beautiful interactive web-based visualizations that can be displayed in Jupyter notebooks, saved to standalone HTML files, or served as part of pure Python-built web applications using Dash. The plotly Python library is sometimes referred to as "plotly.py"to differentiate it from the JavaScript library.

## Scikit-Learn

**Scikit-learn** (formerly **scikits.learn** and also known as **sklearn**) is a freesoftware machine learning library for the Python programming language.

It features various classification, regression and clustering algorithms which including support vector machines, random forests, gradient boosting, *k*-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Scikit-learn is a library in Python that provides many unsupervised and supervised learning algorithms. It’s built upon some of the technology you might already be familiarwith, like NumPy, pandas, and Matplotlib!

The functionality that scikit-learn provides include:

- **Regression**, including Linear and Logistic Regression
- **Classification**, including K-Nearest Neighbors
- **Clustering**, including K-Means and K-Means++
- **Model selection**
- **Preprocessing**, including Min-Max Normalization

Scikit-learn is largely written in Python, and uses NumPy extensively for high-performance linear algebra and array operations. Furthermore, some core algorithms are written in Cython to improve performance. Support vector machines are implemented bya Cython wrapper around LIBSVM; logistic regression and linear support vector machines by a similar wrapper around LIBLINEAR. In such cases, extending these methods with Python may not be possible.

Scikit-learn integrates well with many other Python libraries, such as Matplotlib and plotly for plotting, NumPy for array vectoroization, Pandas dataframes, SciPy, and many more.

## XGBoost

The XGBoost python module is able to load data from many different types of data format, including:

- NumPy 2D array
- SciPy 2D sparse array
- Pandas data frame
- cuDF DataFrame
- cupy 2D array
- dlpack
- datatable
- XGBoost binary buffer file.
- LIBSVM text format file
- Comma-separated values (CSV) file

Xgboost is a supervised learning library that is used for classification as well as regeneration. It implements ML algorithms under the Gradient Boosting framework,and it provides a parallel tree boosting (also known as GBDT, GBM) which solves manydata science problems in a fast and accurate way. It can run on both single and distributed frameworks like Apache Hadoop, Apache Spark, etc. In this article, we are going to see how to install Xgboost in Anaconda Python.

## 6.3 Code:

```
import pandas as pd
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.ticker as mtick
import matplotlib.pyplot as plt
%matplotlib inlinedf=pd.read_csv('/content/WA_Fn-UseC_-Telco-Customer-
Churn.csv')
df.head()
df.info()
df.dropna(inplace=True)
# drop customerID
df.drop('customerID', axis=1, inplace=True)
df.info()
df_cat = df.select_dtypes(include=[object])
df_cat.head()
for i in df_cat.columns:
      print(i, df[i].unique())
from sklearn.preprocessing import OneHotEncoder
ohe = OneHotEncoder(handle_unknown='ignore',
sparse_output=False).set_output(transform='pandas')
ohetransform = ohe.fit_transform(df[df_cat.columns])
ohetransform.head()
df = pd.concat([df, ohetransform], axis=1)
df.head()
for i in df_cat.columns:
     df.drop([i], axis=1, inplace=True)
df.drop('Churn_No', axis=1, inplace=True)
df.rename(columns={'Churn_Yes':'Churn'}, inplace=True)
df.head(10)
df.describe()
import matplotlib.pyplot as plt
fig, axes = plt.subplots(nrows=1, ncols=3, figsize=(15, 5))
columns = ['tenure', 'MonthlyCharges', 'TotalCharges']
for i, column in enumerate(columns):
    axes[i].hist(df[column], bins=50, color='skyblue', edgecolor='black')
    axes[i].set_title(f'Distribution of {column}')
```

```
    axes[i].set_xlabel(column)
    axes[i].set_ylabel('Frequency')
plt.tight_layout()
plt.show()
df['Churn'].value_counts().plot(kind='barh', figsize=(8, 6))
plt.xlabel('Count')
plt.ylabel('Churn')
for i, predictor in enumerate(df.drop(columns=['Churn', 'TotalCharges',
'MonthlyCharges', 'tenure'])):
    plt.figure(i, figsize=(5, 3))
    sns.countplot(data=df, x=predictor, hue='Churn')
df[['MonthlyCharges', 'Churn']]
plt.figure(figsize=(10, 5))
sns.kdeplot(data=df, x="MonthlyCharges", hue="Churn", fill=True, alpha=0.5)
plt.title('Density Plot of Monthly Charges by Churn Status')
plt.xlabel('Monthly Charges')
plt.ylabel('Density')
plt.show()
plt.figure(figsize=(20,8))
     df.corr()['Churn'].sort_values(ascending = False).plot(kind='bar')
import pandas as pd
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.metrics import recall_score
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.tree import DecisionTreeClassifier
from imblearn.over_sampling import SMOTE
df.head()
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression(max_iter=10000)
logreg.fit(X_train, y_train)
y_pred = logreg.predict(X_test)
y_pred
# confusion_matrix
from sklearn import metrics
cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
cnf_matrix
class_names=[0,1] # name  of classes
fig, ax = plt.subplots()
tick_marks = np.arange(len(class_names))
plt.xticks(tick_marks, class_names)
plt.yticks(tick_marks, class_names)
# create heatmap
sns.heatmap(pd.DataFrame(cnf_matrix), annot=True, cmap="YlGnBu" ,fmt='g')
ax.xaxis.set_label_position("top")
plt.tight_layout()
plt.title('Confusion matrix', y=1.1)
plt.ylabel('Actual label')
plt.xlabel('Predicted label')
```

```
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred, labels=[0, 1]))
# UpSampling
from imblearn.combine import SMOTEENN
sm = SMOTEENN()
X_res, y_res = sm.fit_resample(X,y)
Y_res.value_counts()
Xr_train, Xr_test, yr_train, yr_test = train_test_split(X_res, y_res,test_size=0.2)
logreg.fit(Xr_train, yr_train)
yr_pred = logreg.predict(Xr_test)
print(classification_report(yr_test, yr_pred, labels=[0, 1]))
model_dt=DecisionTreeClassifier(criterion = "gini",random_state =
100,max_depth=6, min_samples_leaf=8)
model_dt.fit(Xr_train,yr_train)
y_pred=model_dt.predict(Xr_test)
print(classification_report(yr_test, y_pred, labels=[0,1]))
from sklearn.ensemble import RandomForestClassifier
model_rf=RandomForestClassifier(n_estimators=100, criterion='gini', random_state =
100,max_depth=6, min_samples_leaf=8)
model_rf.fit(Xr_train,yr_train)
y_pred=model_rf.predict(Xr_test)
print(classification_report(yr_test, y_pred, labels=[0,1]))
from xgboost import XGBClassifier
model_xg = XGBClassifier()
model_xg.fit(Xr_train, yr_train)
y_pred=model_rf.predict(Xr_test)
print(classification_report(yr_test, y_pred, labels=[0,1]))
from sklearn.model_selection import RandomizedSearchCV
from xgboost import XGBClassifier
from scipy.stats import uniform, randint
# Define the model
model_xg = XGBClassifier()
param_distributions = {
    'max_depth': randint(3, 6),
    'learning_rate': uniform(0.01, 0.2),
    'n_estimators': randint(100, 300),
    'subsample': uniform(0.8, 0.2)
}
random_search = RandomizedSearchCV(estimator=model_xg,
                                   param_distributions=param_distributions,
                                   n_iter=100, cv=3, verbose=2, random_state=42, n_jobs=-1)
random_search.fit(Xr_train, yr_train)
print(random_search.best_params_)
best_model = random_search.best_estimator_
y_pred = best_model.predict(Xr_test)
print(classification_report(yr_test, y_pred, labels=[0, 1]))
```

# Chapter – 7
# SYSTEM TESTING

# CHAPTER – 7
# SYSTEM TESTING

## 7.1 Testing:

This section describes the different types of testing that may be used to test a software.

### Manual Testing

Manual testing includes testing a software manually, i.e., without using any automated tool orany script. In this type, the tester takes over the role of an end-user and tests the software to identify any unexpected behavior or bug. There are different stages for manual testing such asunit testing, integration testing, system testing, and user acceptance testing.

### Automation Testing

Automation testing, which is also known as Test Automation, is when the tester writes scriptsand uses another software to test the product. This process involves automation of a manual process. Automation Testing is used to re-run the test scenarios that were performed manually,quickly, and repeatedly.

### What to Automate?

It is not possible to automate everything in a software. The areas at which a user can make transactions such as the login form or registration forms, any area where large number of userscan access the software simultaneously should be automated.

### When to Automate?

Test Automation should be used by considering the following aspects of a software –

- Large and critical projects
- Projects that require testing the same areas frequently
- Requirements not changing frequently
- Accessing the application for load and performance with many virtual users
- Stable software with respect to manual testing
- Availability of time

### How to Automate?

Automation is done by using a supportive computer language like VB scripting and an automated software application. There are many tools available that can be used to write automation scripts. Before mentioning the tools, let us identify the process that can be used toautomate the testing process –

- Identifying areas within a software for automation
- Selection of appropriate tool for test automation
- Writing test scripts
- Development of test suits
- Execution of scripts
- Create result reports
- Identify any potential bug or performance issues

## 7.2 Black-Box Testing

The technique of testing without having any knowledge of the interior workings of the application is called black-box testing. The tester is oblivious to the system architecture and does not have access to the source code. Typically, while performing a black-box test, a testerwill interact with the system's user interface by providing inputs and examining outputs withoutknowing how and where the inputs are worked upon.

The following table lists the advantages and disadvantages of black-box testing.

| **Advantages** | **Disadvantages** |
|---|---|
| Well suited and efficient for large code segments. | Limited coverage, since only A selectednumber of test scenariosis actually performed. |
| Code access is not required. | Inefficient testing, due to the fact That the tester only has limited knowledge about an application. |
| Clearly separates user's perspective from the developer's perspective through visibly definedroles. | Blind coverage, since the tester cannot target specific code segments or error prone areas. |

| | |
|---|---|
| Large numbers of moderately skilled testers can test the application with no knowledge of implementation, programming language, or operating systems. | The test cases are difficult to design. |

## 7.3 hite-Box Testing

The technique of testing without having any knowledge of the interior workings of the application is called black-box testing. The tester is oblivious to the system architecture and does not have access to the source code. Typically, while performing a black-box test, a testerwill interact with the system's user interface by providing inputs and examining outputs without knowing how and where the inputs are worked upon.
The following table lists the advantages and disadvantages of black-box testing.

| **Advantages** | **Disadvantages** |
|---|---|
| Well suited and efficient for large codesegments. | Limited coverage, since only a selected number of test scenarios is Actually performed. |
| Code access is not required. | Inefficient testing, due to the fact that The tester only has limited knowledge about an application. |
| Clearly separates user's perspective from the Developer’s perspective through visibly defined roles. | Blind coverage, since the tester Cannot target specific code segments or errorprone areas. |
| Large numbers of moderately skilled testers can testthe application with no knowledge of implementation, programming language, or Operating systems. | The test cases are difficult todesign. |

# Chapter - 8

# RESULTS

# CHAPTER – 8

# RESULTS

# 8. Results

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

5 rows × 21 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7032 non-null   float64
 20  Churn             7043 non-null   object
dtypes: float64(2), int64(2), object(17)
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 7032 entries, 0 to 7042
Data columns (total 20 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   gender            7032 non-null   object
 1   SeniorCitizen     7032 non-null   int64
 2   Partner           7032 non-null   object
 3   Dependents        7032 non-null   object
 4   tenure            7032 non-null   int64
 5   PhoneService      7032 non-null   object
 6   MultipleLines     7032 non-null   object
 7   InternetService   7032 non-null   object
 8   OnlineSecurity    7032 non-null   object
 9   OnlineBackup      7032 non-null   object
 10  DeviceProtection  7032 non-null   object
 11  TechSupport       7032 non-null   object
 12  StreamingTV       7032 non-null   object
 13  StreamingMovies   7032 non-null   object
 14  Contract          7032 non-null   object
 15  PaperlessBilling  7032 non-null   object
 16  PaymentMethod     7032 non-null   object
 17  MonthlyCharges    7032 non-null   float64
 18  TotalCharges      7032 non-null   float64
 19  Churn             7032 non-null   object
dtypes: float64(2), int64(2), object(16)
```

| | gender | Partner | Dependents | PhoneService | MultipleLines | InternetService | OnlineSecurity | OnlineBackup | DeviceProtection | TechSupport | StreamingTV | StreamingMovies | Contract | PaperlessBilling | PaymentMet |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Female | Yes | No | No | No phone service | DSL | No | Yes | No | No | No | No | Month-to-month | Yes | Electro ch |
| 1 | Male | No | No | Yes | No | DSL | Yes | No | Yes | No | No | No | One year | No | Mailed ch |
| 2 | Male | No | No | Yes | No | DSL | Yes | Yes | No | No | No | No | Month-to-month | Yes | Mailed ch |
| 3 | Male | No | No | No | No phone service | DSL | Yes | No | Yes | Yes | No | No | One year | No | Bank trans (automa |
| 4 | Female | No | No | Yes | No | Fiber optic | No | No | No | No | No | No | Month-to-month | Yes | Electro ch |

```
gender ['Female' 'Male']
Partner ['Yes' 'No']
Dependents ['No' 'Yes']
PhoneService ['No' 'Yes']
MultipleLines ['No phone service' 'No' 'Yes']
InternetService ['DSL' 'Fiber optic' 'No']
OnlineSecurity ['No' 'Yes' 'No internet service']
OnlineBackup ['Yes' 'No' 'No internet service']
DeviceProtection ['No' 'Yes' 'No internet service']
TechSupport ['No' 'Yes' 'No internet service']
StreamingTV ['No' 'Yes' 'No internet service']
StreamingMovies ['No' 'Yes' 'No internet service']
Contract ['Month-to-month' 'One year' 'Two year']
PaperlessBilling ['Yes' 'No']
PaymentMethod ['Electronic check' 'Mailed check' 'Bank transfer (automatic)'
 'Credit card (automatic)']
Churn ['No' 'Yes']
```

| | gender_Female | gender_Male | Partner_No | Partner_Yes | Dependents_No | Dependents_Yes | PhoneService_No | PhoneService_Yes | MultipleLines_No | MultipleLines_No phone service | ... | Contract_One year | Contract_Two year | PaperlessBilling_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | ... | 0.0 | 0.0 | |
| 1 | 0.0 | 1.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | ... | 1.0 | 0.0 | |
| 2 | 0.0 | 1.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | ... | 0.0 | 0.0 | |
| 3 | 0.0 | 1.0 | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | ... | 1.0 | 0.0 | |
| 4 | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | ... | 0.0 | 0.0 | |

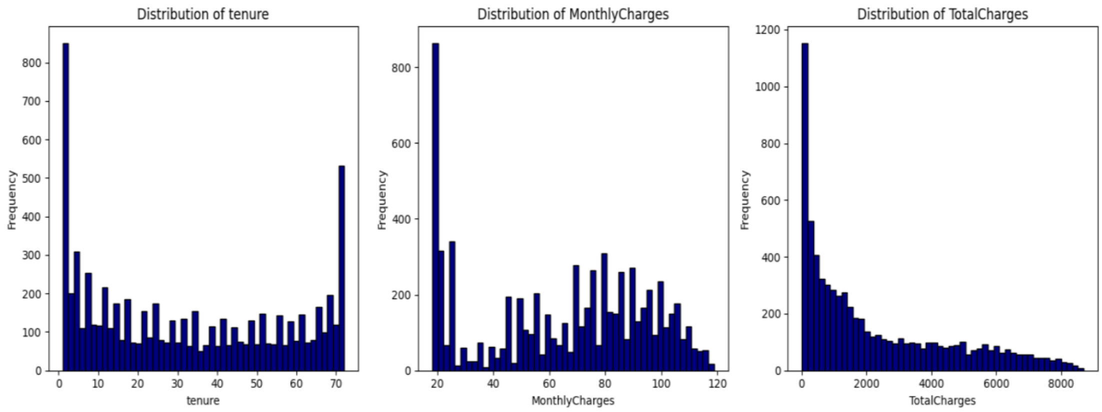
5 rows × 43 columns

| | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | MultipleLines | InternetService | OnlineSecurity | OnlineBackup | ... | Contract_One year | Contract_Two year | PaperlessBilling_No | PaperlessBilling_Yes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Female | 0 | Yes | No | 1 | No | No phone service | DSL | No | Yes | ... | 0.0 | 0.0 | 0.0 | 1.0 |
| 1 | Male | 0 | No | No | 34 | Yes | No | DSL | Yes | No | ... | 1.0 | 0.0 | 1.0 | 0.0 |
| 2 | Male | 0 | No | No | 2 | Yes | No | DSL | Yes | Yes | ... | 0.0 | 0.0 | 0.0 | 1.0 |
| 3 | Male | 0 | No | No | 45 | No | No phone service | DSL | Yes | No | ... | 1.0 | 0.0 | 1.0 | 0.0 |
| 4 | Female | 0 | No | No | 2 | Yes | No | Fiber optic | No | No | ... | 0.0 | 0.0 | 0.0 | 1.0 |

5 rows × 63 columns

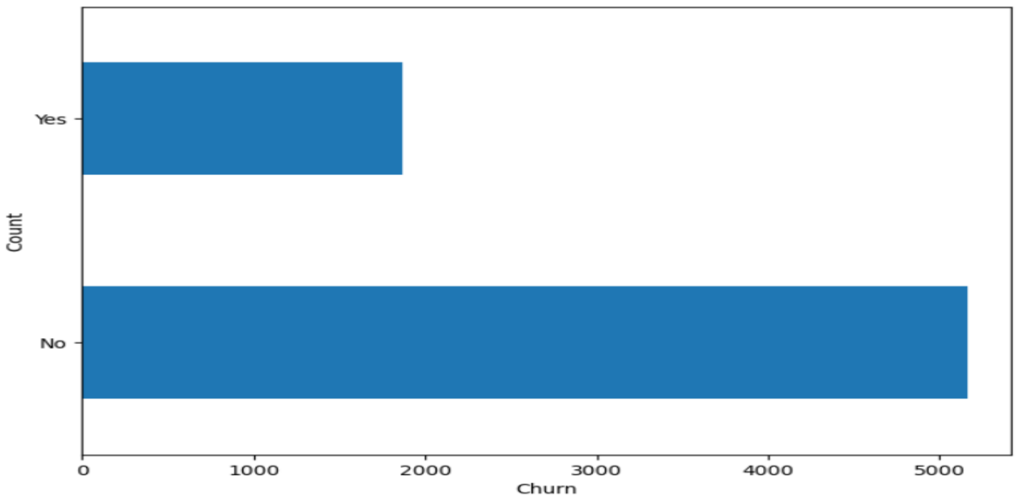
| | SeniorCitizen | tenure | MonthlyCharges | TotalCharges | gender_Female | gender_Male | Partner_No | Partner_Yes | Dependents_No | Dependents_Yes | ... | Contract_Month-to-month | Contract_One year | Contract_Two year | PaperlessBilling_I |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 29.85 | 29.85 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | ... | 1.0 | 0.0 | 0.0 | 0 |
| 1 | 0 | 34 | 56.95 | 1889.50 | 0.0 | 1.0 | 1.0 | 0.0 | 1.0 | 0.0 | ... | 0.0 | 1.0 | 0.0 | 1 |
| 2 | 0 | 2 | 53.85 | 108.15 | 0.0 | 1.0 | 1.0 | 0.0 | 1.0 | 0.0 | ... | 1.0 | 0.0 | 0.0 | 0 |
| 3 | 0 | 45 | 42.30 | 1840.75 | 0.0 | 1.0 | 1.0 | 0.0 | 1.0 | 0.0 | ... | 0.0 | 1.0 | 0.0 | 1 |
| 4 | 0 | 2 | 70.70 | 151.65 | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | ... | 1.0 | 0.0 | 0.0 | 0 |
| 5 | 0 | 8 | 99.65 | 820.50 | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | ... | 1.0 | 0.0 | 0.0 | 0 |
| 6 | 0 | 22 | 89.10 | 1949.40 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | ... | 1.0 | 0.0 | 0.0 | 0 |
| 7 | 0 | 10 | 29.75 | 301.90 | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | ... | 1.0 | 0.0 | 0.0 | 1 |
| 8 | 0 | 28 | 104.80 | 3046.05 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | ... | 1.0 | 0.0 | 0.0 | 0 |
| 9 | 0 | 62 | 56.15 | 3487.95 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | ... | 0.0 | 1.0 | 0.0 | 1 |

10 rows × 46 columns



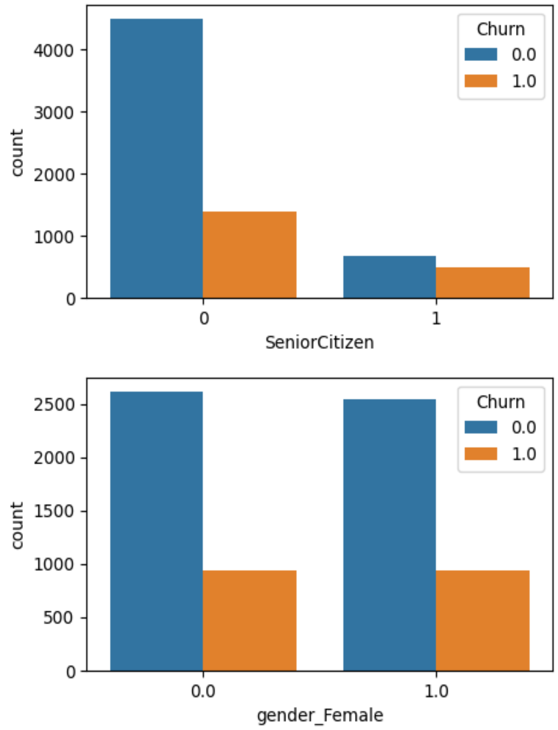

The distribution of tenure,monthly charges and total charges.

Text(0, 0.5, 'Count')

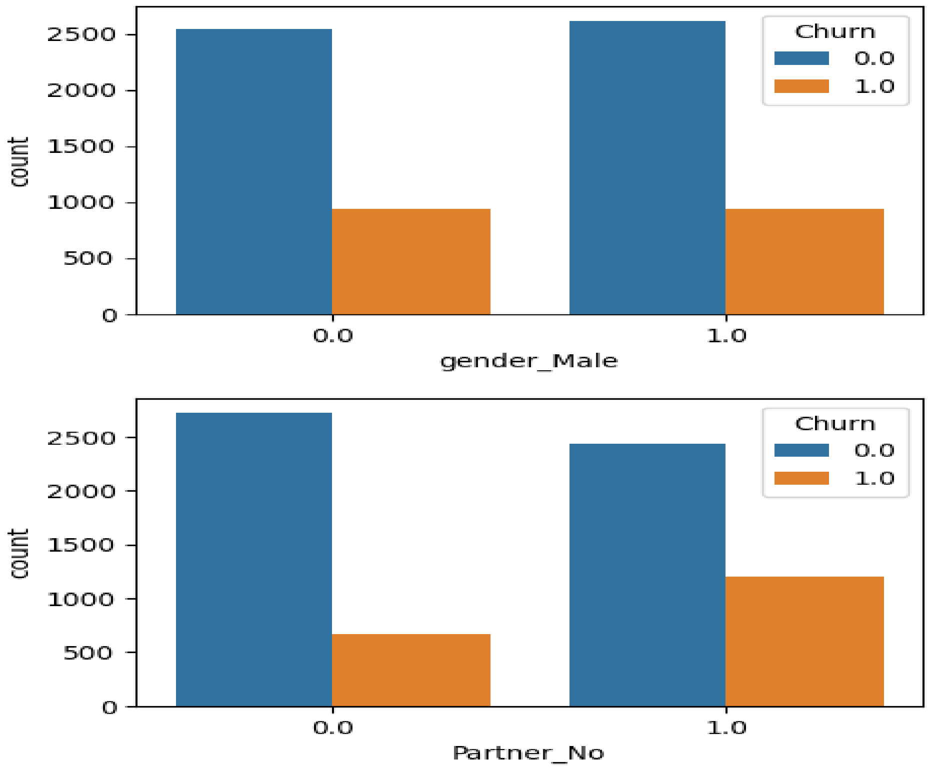



The count of churn and non- churn in the format of bar graph.

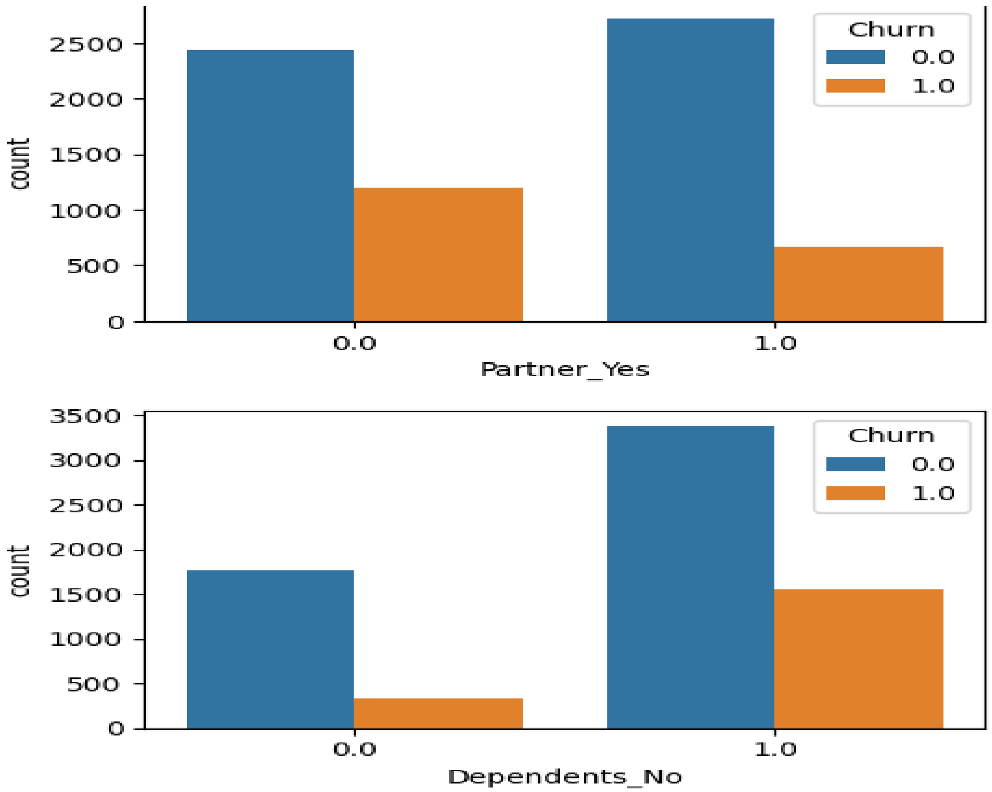
```
<ipython-input-29-cc9fda9d625f>:2: RuntimeWarning: More than 20 figures have been opened. Figures created through the pypl
  plt.figure(i, figsize=(5, 3))
```




The above bar graph shows the count of senior citizen and gender_female of churn.




The above bar graph shows the count of gender_male and partner_No of churn.

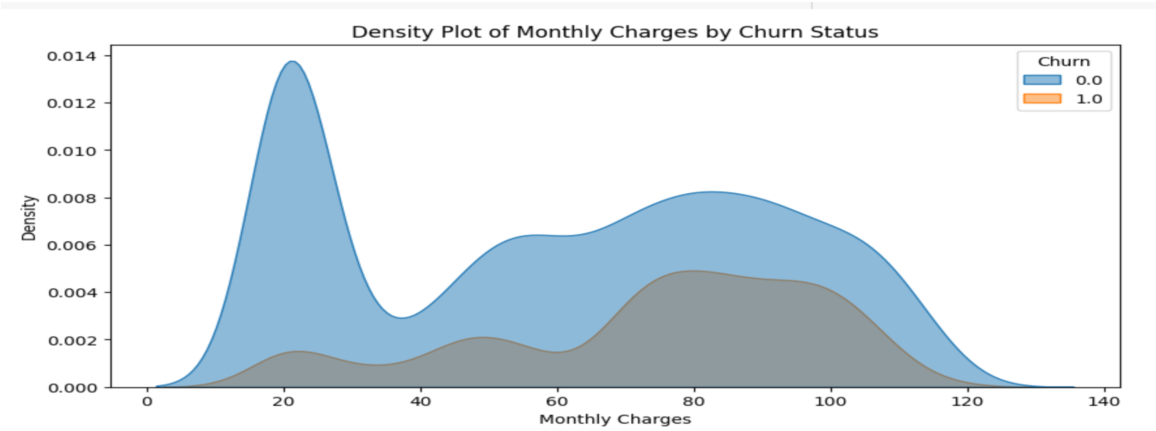



The above bar graph shows the count of partner_Yes and Dependents_No of churn.

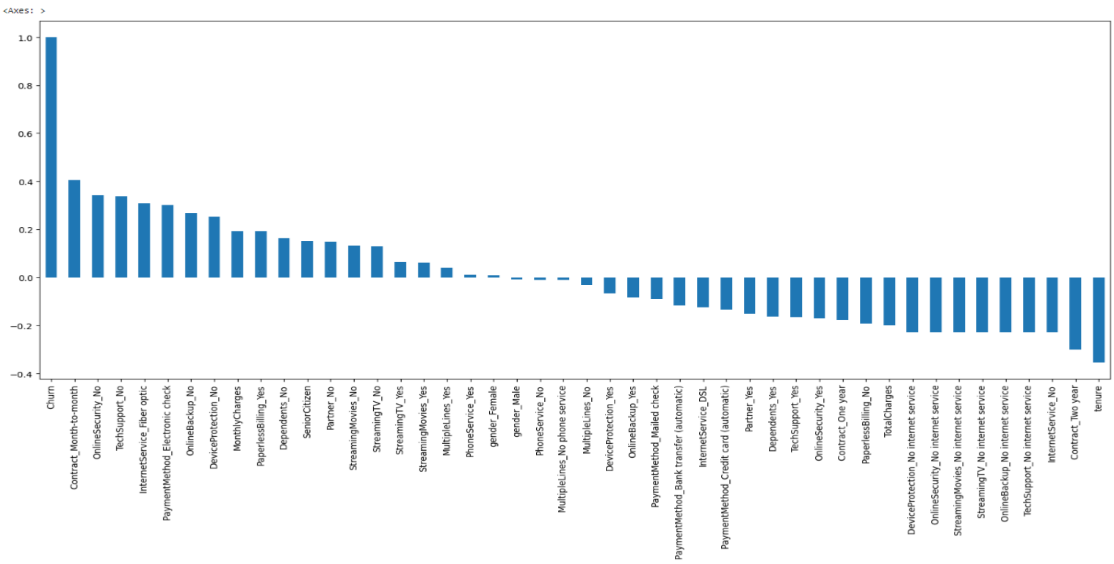
| | MonthlyCharges | Churn |
|---|---|---|
| 0 | 29.85 | 0.0 |
| 1 | 56.95 | 0.0 |
| 2 | 53.85 | 1.0 |
| 3 | 42.30 | 0.0 |
| 4 | 70.70 | 1.0 |
| ... | ... | ... |
| 7038 | 84.80 | 0.0 |
| 7039 | 103.20 | 0.0 |
| 7040 | 29.60 | 0.0 |
| 7041 | 74.40 | 1.0 |
| 7042 | 105.65 | 0.0 |

7032 rows × 2 columns

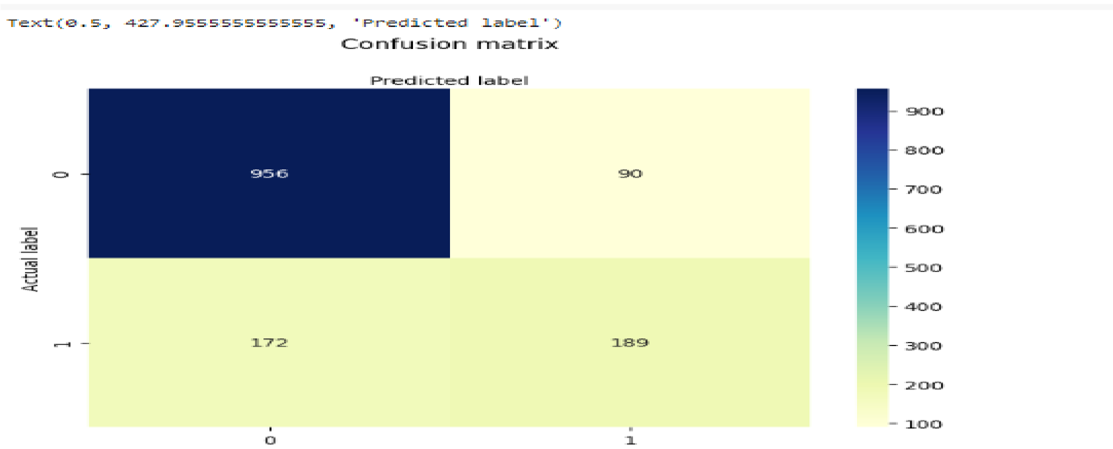
The above shows the monthly charges of patron.




Density Plot of Monthly Charges by Churn Status




Water fall graph




Confusion Matix of actual and predicted label.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.85 | 0.91 | 0.88 | 1046 |
| 1 | 0.68 | 0.52 | 0.59 | 361 |
| accuracy | | | 0.81 | 1407 |
| macro avg | 0.76 | 0.72 | 0.74 | 1407 |
| weighted avg | 0.80 | 0.81 | 0.81 | 1407 |

Classification report of existing model.

```
Churn
1.0     3201
0.0     2626
Name: count, dtype: int64
```

The count of churn and non-churn.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.92 | 0.94 | 0.93 | 512 |
| 1 | 0.95 | 0.93 | 0.94 | 654 |
| accuracy | | | 0.94 | 1166 |
| macro avg | 0.94 | 0.94 | 0.94 | 1166 |
| weighted avg | 0.94 | 0.94 | 0.94 | 1166 |

The above screenshot shows metrics of Random forest.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.94 | 0.91 | 0.92 | 512 |
| 1 | 0.93 | 0.96 | 0.94 | 654 |
| accuracy | | | 0.93 | 1166 |
| macro avg | 0.94 | 0.93 | 0.93 | 1166 |
| weighted avg | 0.93 | 0.93 | 0.93 | 1166 |

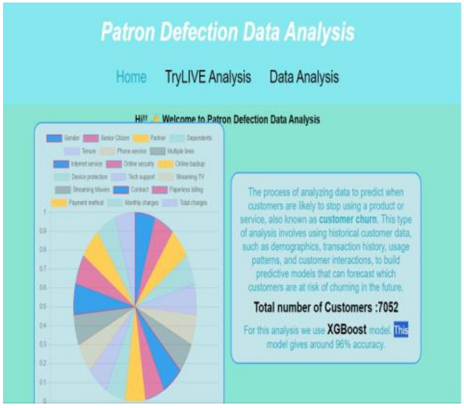
The above screenshot shows metrics of Decision tree.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.94 | 0.91 | 0.92 | 512 |
| 1 | 0.93 | 0.96 | 0.94 | 654 |
| accuracy | | | 0.93 | 1166 |
| macro avg | 0.94 | 0.93 | 0.93 | 1166 |
| weighted avg | 0.93 | 0.93 | 0.93 | 1166 |

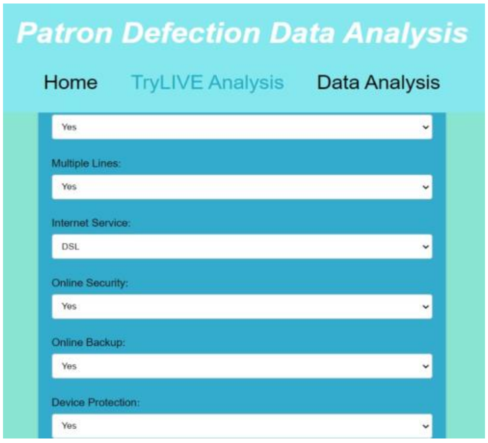
The above screenshot shows metrics of XGBoost.

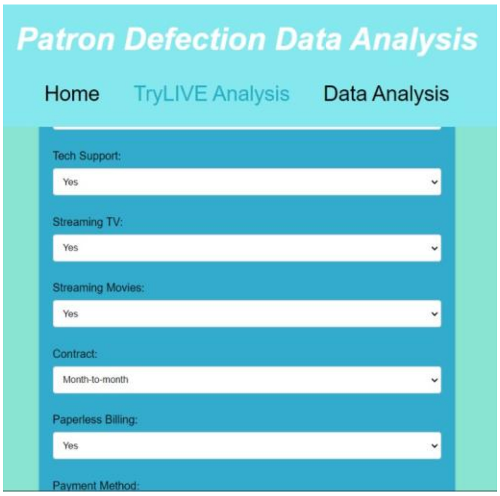
```
Fitting 3 folds for each of 100 candidates, totalling 300 fits
 ▸ RandomizedSearchCV
 ▸ estimator: XGBClassifier
     ▸ XGBClassifier
```

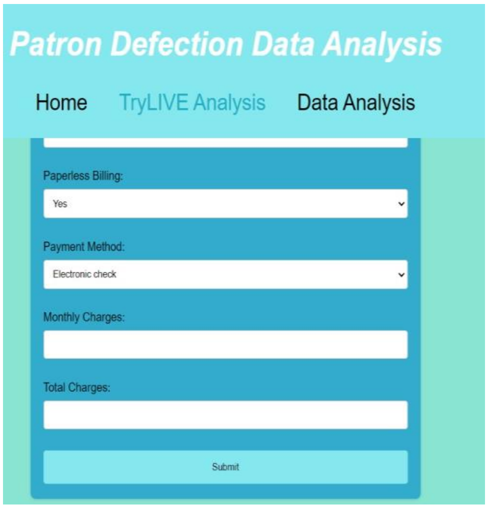
## Screenshots of Frontend output:



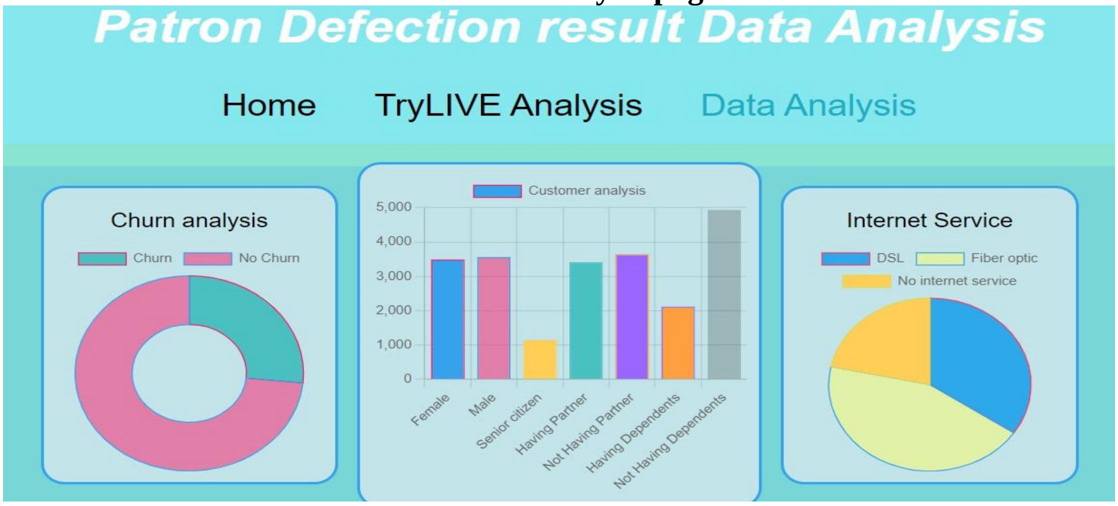

**Home Page**

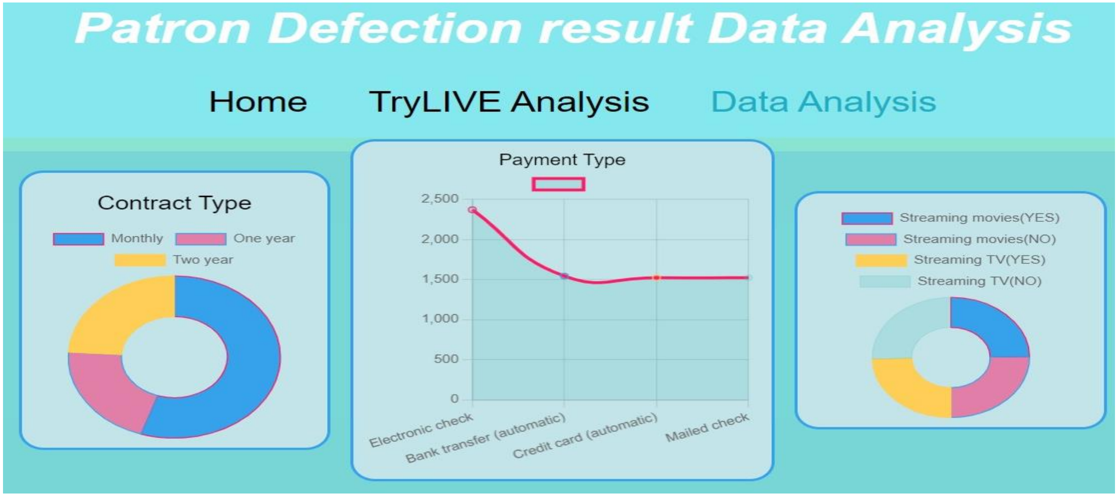



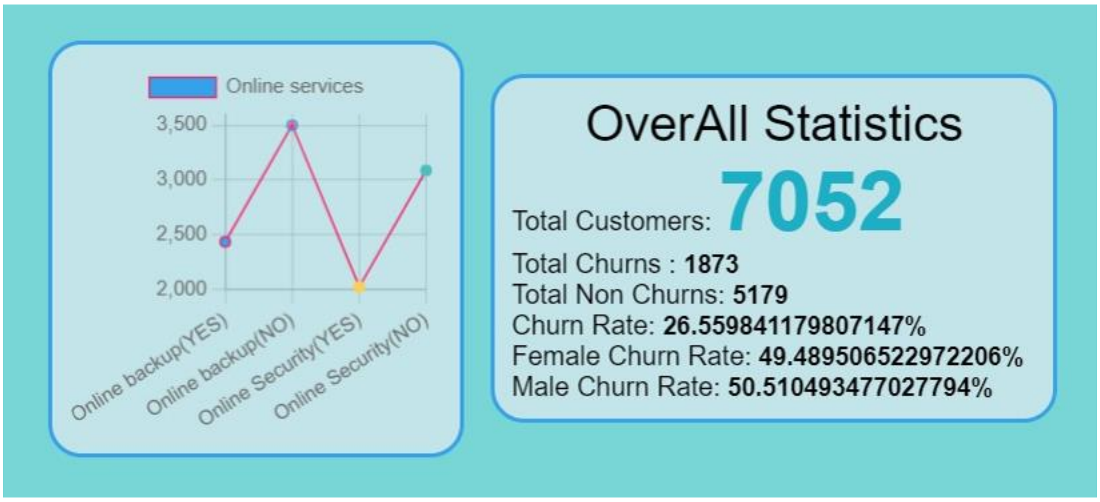



**Live Analysis page**




**Data Analysis Page**

# Chapter – 9

# CONCLUSION

# CHAPTER – 9
# CONCLUSION

## 8. CONCLUSION:

The Patron Defection Data Analysis project provided valuable insights into the patterns and trends surrounding patron defection within the organization. Through rigorous data analysis, several key findings emerged. Firstly, it became evident that certain fluctuations in patron behavior preceded defection, indicating potential warning signs that could be monitored and addressed proactively. These patterns varied across different segments of patrons, highlighting the importance of segmentation in understanding defection dynamics.
Delving deeper into the root causes of defection, it became apparent that dissatisfaction with service quality, changes in personal circumstances, and increased competition were among the primary drivers. Understanding these underlying reasons is crucial for devising effective retention strategies tailored to address the specific needs and concerns of at-risk patrons.
Segmentation analysis further reinforced the notion that not all patrons are equally prone to defection.By leveraging historical data and identifying key predictors, such as frequency of interactions, satisfaction ratings, and tenure, the models demonstrated promising results in anticipating future churn. However, it's essential to continuously refine these models and incorporate additional variables to enhance predictive accuracy over time.
In conclusion, the Patron Defection Data Analysis project has provided valuable insights into the factors driving patron defection and laid the groundwork for implementing proactive retention strategies. By leveraging these insights and recommendations, the organization can strengthen its relationships with patrons, foster loyalty, and ultimately drive sustainable growth and success.

### Future Scope

In the future, the Patron Defection Data Analysis project could focus on real-time monitoring for early defection signs, advanced predictive modeling with machine learning, optimizing loyalty programs through experimentation, and integrating data from multiple channels for a holistic view of customer behavior. Additionally, efforts could be directed towards comprehensive customer journey mapping and implementing closed-loop feedback systems. These advancements could result in a more proactive and personalized approach to retaining patrons, driving sustained growth and fostering strong customer relationships.

# Chapter – 10

# BIBLIOGRAPHY

# CHAPTER – 10
# BIBLIOGRAPHY


1. Anderson, E. W., Fornell, C., & Lehmann, D. R. (1994). Customer Satisfaction, Market Share, and Profitability: Findings from Sweden. Journal of Marketing, 58(3), 53-66.

2. Reichheld, F. F. (1996). The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Value. Harvard Business Review Press.

3. Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society. Series B (Methodological), 36(2), 111-147.

4. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

5. Kim, Y. (2017). Predictive modeling for customer churn: A case study of an insurance company. Journal of Big Data, 4(1), 1-12.

6. Verhoef, P. C., & Lemon, K. N. (2013). Successful Customer Relationship Marketing: New Thinking, New Strategies, New Tools for Winning the Customer's Heart. Springer Science & Business Media.

7. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate Data Analysis: A Global Perspective. Pearson Education.

8. Kandampully, J., & Suhartanto, D. (2000). Customer Loyalty in the Hotel Industry: The Role of Customer Satisfaction and Image. International Journal of Contemporary Hospitality Management, 12(6), 346-351.

9. R Development Core Team. (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

10. Nguyen, T., & Huang, Q. (2019). Predictive analytics in customer relationship management: a systematic literature review and future research directions. Expert Systems with Applications, 115, 408-422.
11. Fader, P. S., & Hardie, B. G. (2010). Customer-Base Valuation in a Contractual Setting: The Perils of Ignoring Heterogeneity. Marketing Science, 29(1), 85-93.

12. Mittal, V., & Kamakura, W. A. (2001). Satisfaction, Repurchase Intent, and Repurchase Behavior: Investigating the Moderating Effect of Customer Characteristics. Journal of Marketing Research, 38(1), 131-142.

13. Berry, L. L. (2000). Cultivating Service Brand Equity. Journal of the Academy of Marketing Science, 28(1), 128-137.

14. Gupta, S., Hanssens, D. M., Hardie, B. G., Kahn, W., Kumar, V., Lin, N. & Sriram, S. (2006). Modeling Customer Lifetime Value. Journal of Service Research, 9(2), 139-155.
15. Churn Management Solutions. (2003). Reducing Customer Churn: A Guide to Identifying, Analyzing, and Managing Customer Defections. John Wiley & Sons.

16. Lacey, R., & Sneath, J. Z. (2006). Customer Perceptions of Relationship Investment: A Study in the UK Financial Services Sector. Journal of Financial Services Marketing, 11(4), 283-293.

17. Wang, X., Yu, J., & Wei, Y. (2020). Predictive Analysis of Customer Churn Based on Deep Learning. IEEE Access, 8, 152727-152739.

18. Rust, R. T., & Oliver, R. L. (1994). Service Quality: Insights and Managerial Implications from the Frontier. Thousand Oaks, CA: Sage Publications.

19. Gupta, S., & Lehmann, D. R. (2006). Managing Customers as Investments: The Strategic Value of Customers in the Long Run. Upper Saddle River, NJ: Pearson Prentice Hall.

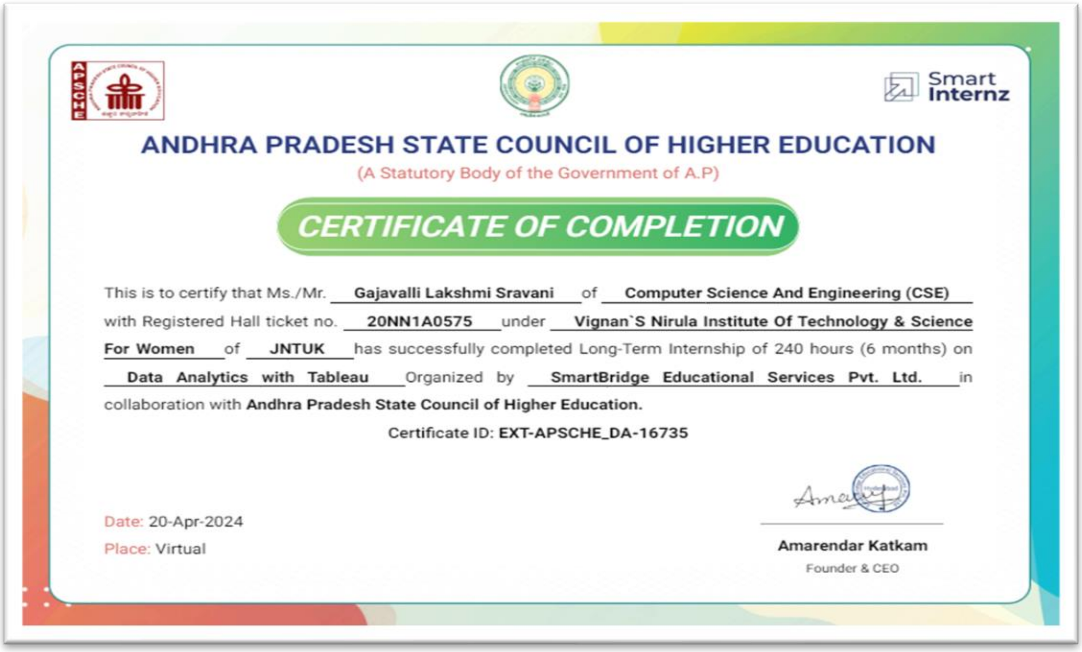
20. Hayes, B. E., & Jindal, P. (2018). Predictive Modeling of Customer Churn in the Wireless Telecommunication Industry. Journal of Retailing and Consumer Services, 42, 43-54.

# CHAPTER – 11

# INTERNSHIP CERTIFICATES

# CHAPTER - 11

# INTERNSHIP CERTIFICATES



ANDHRA PRADESH STATE COUNCIL OF HIGHER EDUCATION
(A Statutory Body of the Government of A.P)

CERTIFICATE OF COMPLETION

This is to certify that Ms./Mr. Gajavalli Lakshmi Sravani of Computer Science And Engineering (CSE) with Registered Hall ticket no. 20NN1A0575 under Vignan`S Nirula Institute Of Technology & Science For Women of JNTUK has successfully completed Long-Term Internship of 240 hours (6 months) on Data Analytics with Tableau Organized by SmartBridge Educational Services Pvt. Ltd. in collaboration with Andhra Pradesh State Council of Higher Education.

Certificate ID: EXT-APSCHE_DA-16735

Date: 20-Apr-2024
Place: Virtual

Amarendar Katkam
Founder & CEO

**Fig: 10.1 Internship Certificate of Member 1**

ANDHRA PRADESH STATE COUNCIL OF HIGHER EDUCATION
(A Statutory Body of the Government of A.P)

CERTIFICATE OF COMPLETION

This is to certify that Ms./Mr. Setty Lavanya of Computer Science And Engineering (CSE) with Registered Hall ticket no. 20NN1A05A4 under Vignan`S Nirula Institute Of Technology & Science For Women of JNTUK has successfully completed Long-Term Internship of 240 hours (6 months) on Data Analytics with Tableau Organized by SmartBridge Educational Services Pvt. Ltd. in collaboration with Andhra Pradesh State Council of Higher Education.

Certificate ID: EXT-APSCHE_DA-16909

Date: 20-Apr-2024
Place: Virtual

Amarendar Katkam
Founder & CEO

**Fig: 10.2 Internship Certificate of Member 2**

ANDHRA PRADESH STATE COUNCIL OF HIGHER EDUCATION

(A Statutory Body of the Government of A.P)

Smart Internz

CERTIFICATE OF COMPLETION

This is to certify that Ms./Mr. Singh Vijayalakshmi of Computer Science And Engineering with Registered Hall ticket no. 20NN1A05B1 under Vignan`S Nirula Institute Of Technology & Science For Women of JNTUK has successfully completed Long-Term Internship of 240 hours (6 months) on Data Analytics with Tableau Organized by SmartBridge Educational Services Pvt. Ltd. in collaboration with Andhra Pradesh State Council of Higher Education.

Certificate ID: EXT-APSCHE_DA-16928

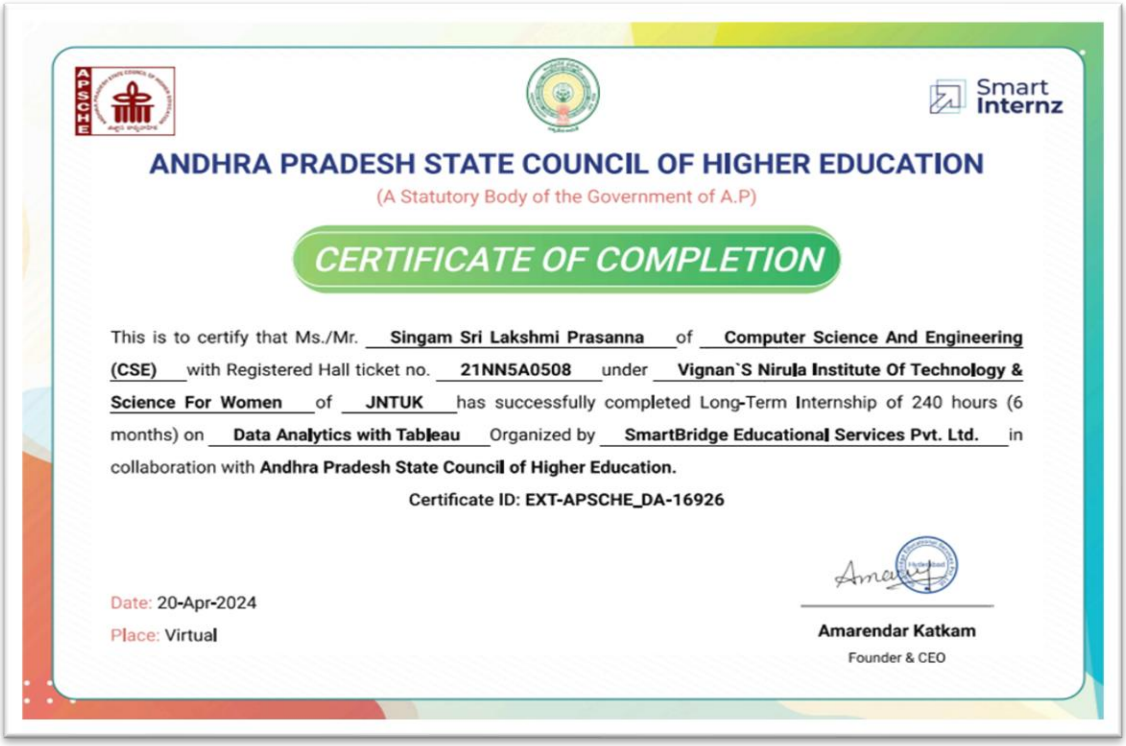
Date: 20-Apr-2024

Place: Virtual

Amarendar Katkam

Founder & CEO

Fig: 10.3 Internship Certificate of Member 3



ANDHRA PRADESH STATE COUNCIL OF HIGHER EDUCATION

(A Statutory Body of the Government of A.P)

Smart Internz

CERTIFICATE OF COMPLETION

This is to certify that Ms./Mr. Singam Sri Lakshmi Prasanna of Computer Science And Engineering (CSE) with Registered Hall ticket no. 21NN5A0508 under Vignan`S Nirula Institute Of Technology & Science For Women of JNTUK has successfully completed Long-Term Internship of 240 hours (6 months) on Data Analytics with Tableau Organized by SmartBridge Educational Services Pvt. Ltd. in collaboration with Andhra Pradesh State Council of Higher Education.

Certificate ID: EXT-APSCHE_DA-16926

Date: 20-Apr-2024

Place: Virtual

Amarendar Katkam

Founder & CEO

Fig: 10.4 Internship Certificate of Member 4