The Impact of Customer Demographics on Churn Rates in the Telecommunications Industry

By

Muhammad Abul Bashar
ID: 212-15-14698

This Report Presented in Partial Fulfillment of the Requirements for thePhase-I in

Computer Science and Engineering

Supervised By

Mr. Narayan Ranjan Chakraborty
Associate Professor and Associate Head
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

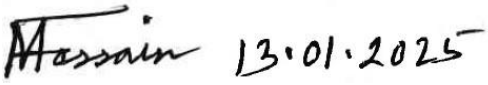
DHAKA, BANGLADESH

January 2025

APPROVAL

This Project titled "The Impact of Customer Demographics on Churn Rates in the Telecommunications Industry", submitted by Muhammad Abul Bashar, ID No: **212-15-14698**to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **13 January, 2025**.

BOARD OF EXAMINERS



Dr. Md. Fokhray Hossain **Chairman**
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Sazzadur Ahamed **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Amatul Bushra Akhi **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammed Nasir Uddin **External Examiner**
Professor
Department of Computer Science and Engineering
Jagannath University

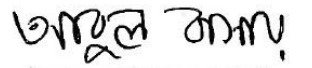
DECLARATION

I hereby declare that this project has been done by us under the supervision of **Mr. Narayan Ranjan Chakraborty,** Associate Professor and Associate Head, Department of Computer Science and Engineering, Daffodil International University.I also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Mr. Narayan Ranjan Chakraborty
Associate Professor and Associate Head
Department of Computer Science and Engineering
Daffodil International University

Submitted by:



Muhammad Abul Bashar
ID: 212-15-14698
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to Almighty for His divine blessing making it possible for me to complete the final year project/internship successfully.

I am grateful and wish my profound indebtedness to **Mr. Narayan Ranjan Chakraborty**, Associate Professor and Associate Head, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of "*Machine Learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

I would like to express my heartiest gratitude to theShovon Chakraborty, for his kind help in finishing my project and also to other faculty members and the staff of the Department of CSE, Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, I must acknowledge with due respect the constant support and patience of myfamily members.

ABSTRACT

In this thesis, I aim to develop a machine learning model for predicting customer churn in the telecom industry, using data from major telecom operators in India, including Airtel, Reliance Jio, Vodafone, and BSNL. The dataset contains 243,553 customer records with demographic, usage, and geographic features, along with a binary variable indicating whether the customer has churned. The goal of this project is to accurately predict customer churn, providing telecom companies with valuable insights to retain at-risk customers and optimize marketing efforts.I explore several machines learning models, including Logistic Regression, Random Forest, and Gradient Boosting. After preprocessing the data, addressing missing values, encoding categorical variables, and handling class imbalance using SMOTE, I evaluate each model's performance using accuracy, ROC-AUC score, and classification metrics such as precision, recall, and F1-score. Among the models tested, Gradient Boosting outperforms others, achieving a high accuracy of 95.2% and a robust ROC-AUC score of 0.9251. This model shows a balanced trade-off between precision and recall, especially for the minority churn class. The findings demonstrate that Gradient Boosting is a highly effective tool for churn prediction in the telecom sector, capable of providing actionable insights for customer retention strategies. The results also highlight the importance of feature engineering and data preprocessing in improving model performance.This research offers a solid foundation for applying machine learning to real-world business problems, particularly in customer retention within the telecom industry.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

Customer churn is a critical issue for telecom companies, leading to increased costs for customer acquisition and reduced profitability. As competition in the telecom industry intensifies, understanding and predicting churn has become essential for businesses aiming to retain customers and maintain a competitive edge. Churn prediction models allow telecom operators to identify customers at risk of leaving and take proactive measures to retain them. In this context, machine learning (ML) techniques have proven effective in building predictive models that can accurately forecast customer churn based on historical data.n my thesis, I focus on the problem of predicting telecom customer churn using various machine learning algorithms. The dataset used for this project consists of 243,553 records from four major telecom partners in India, including Airtel, Reliance Jio, Vodafone, and BSNL. The dataset contains several customer attributes such as demographics, usage patterns, and geographical information, along with a binary target variable indicating whether the customer has churned or not.I aim to evaluate different ML models, including Logistic Regression, Random Forest, and Gradient Boosting, to determine the best-performing model for churn prediction. After training and testing the models, I found that Gradient Boosting achieved the highest accuracy of 95.2%, with an ROC-AUC score of 0.9251, demonstrating its superior performance in predicting customer churn compared to other models. This high accuracy emphasizes the potential of Gradient Boosting for real-world churn prediction applications in the telecom industry, offering businesses valuable insights into customer behavior and retention strategies.

1.2 Problem Statement

In the highly competitive telecom industry, customer churn poses a significant challenge, impacting revenue and growth. Predicting churn allows companies to take proactive measures to retain customers, reducing the cost of acquiring new ones. In my project, I

aim to build a machine learning model to predict customer churn based on various demographic, usage, and geographic factors. I focus on using different algorithms, including Gradient Boosting, which achieved high accuracy of 95.2%, to identify at-risk customers. This model will provide actionable insights for telecom companies to improve customer retention strategies and optimize their operations.

1.3 Motivation and Objectives

The telecom industry faces significant challenges due to high customer churn rates, leading to revenue loss and increased operational costs. Predicting churn accurately can help telecom operators proactively retain customers and improve service offerings. In my thesis, I aim to develop a robust churn prediction model using machine learning techniques. The Gradient Boosting model achieved high accuracy, making it an ideal choice for this task. By identifying key factors contributing to churn, I aim to provide actionable insights for telecom operators to enhance customer retention strategies and optimize business performance.

1.4 Project Scope

The scope of my project, "Telecom Churn Prediction," focuses on predicting customer churn in the telecom industry using machine learning techniques. I will analyze customer data, including demographics, usage patterns, and telecom provider details, to identify key factors contributing to churn. By implementing various models, including Gradient Boosting, I aim to achieve high accuracy in predicting churn. The primary goal is to build a robust model that can assist telecom companies in retaining customers and minimizing churn. The project will also explore handling class imbalance and improving model performance using techniques like SMOTE.

1.5 Project Outcome

The outcome of my project, "Telecom Churn Prediction," demonstrates the successful application of machine learning to predict customer churn in the telecom industry. After exploring various models, Gradient Boosting achieved the highest accuracy of 95.2%, with a ROC-AUC score of 0.9251. This model outperformed others, including Logistic Regression and Random Forest, providing a reliable tool for identifying at-risk customers. The insights gained from feature importance analysis revealed key factors influencing churn, such as data usage, age, and salary. This outcome showcases the potential of machine learning to drive business decisions and improve customer retention strategies.

1.6 Report Organization

As for organizing the report, I will structure it as follows:

1. Introduction
    - Brief Introduction of the problem statement and the dataset used.
    - Explanation of the importance of the task and its relevance in the medical field.
2. Data Preprocessing
    - Description of the dataset, including its size and distribution across classes.
    - Details on any preprocessing steps applied, such as resizing and normalization.
3. Model Architecture
    - Introduction of the chosen neural network architecture.
    - Explanation of the rationale behind the selection and any modifications made.
4. Training Process
    - Introduction of the training procedure, including hyperparameters and optimization algorithm.

- Discussion of any challenges encountered during training and how they were addressed.

5. Results
    - Presentation of the model's performance metrics on both training and testing datasets.
    - Comparison with baseline models or previous research, if applicable.
6. Discussion
    - Interpretation of the results and analysis of the model's strengths and weaknesses.
    - Exploration of potential areas for improvement or further investigation.
7. Conclusion
    - Summary of the findings and their implications for the task at hand.
    - Suggestions for future work and research directions to pursue.
8. References:
    - Citation of any sources or literature referenced throughout the report.

This organization will provide a clear and concise structure for presenting the methodology, results, and conclusions of the project.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

Customer churn prediction in the telecommunications industry is critical for improving customer retention and reducing revenue loss. Several studies have focused on identifying the factors contributing to churn and the development of predictive models using machine learning and data mining techniques.In paper [1], Jain et al. emphasized the importance of churn prediction and reviewed various techniques, datasets, and performance measures. They noted that while machine learning techniques are widely used, feature extraction plays a crucial role in developing an effective churn prediction model. The authors also highlighted that deep learning algorithms like CNN could enhance churn prediction, especially for large datasets.Huang et al. [2] introduced a new feature set for land-line customer churn prediction, incorporating precise call details, bill and payment information, and customer demographics. They applied seven prediction techniques, including Logistic Regression, Naive Bayes, and Decision Trees, and found that the new feature set, coupled with machine learning models, provided better performance for churn prediction.Ahmad et al. [3] focused on developing a churn prediction model using machine learning on a big data platform. They employed Decision Trees, Random Forest, Gradient Boosted Machine Tree (GBM), and XGBoost. Their results showed that XGBoost outperformed other models, achieving an AUC of 93.3%, particularly when combined with Social Network Analysis (SNA) features.Lu et al. [4] proposed using boosting techniques to enhance churn prediction models, specifically using logistic regression as a basis learner. Their experiments showed that boosting significantly improved the separation of churn data and helped identify high-risk customer clusters, thus enhancing prediction accuracy.In paper [5], Ullah et al. explored Random Forest for churn prediction and identified key churn factors in the telecom

sector. They combined classification and clustering techniques to improve churn classification accuracy, with Random Forest achieving an accuracy of 88.63%. Their model also segmented churn customers into groups for targeted retention strategies.Alboukaey et al. [6] addressed the challenges of predicting churn on a monthly basis, proposing a daily churn prediction model based on dynamic customer behavior. They used RFM-based, statistics-based, LSTM, and CNN models, with the LSTM-based model outperforming others in terms of prediction accuracy and early detection of churn.Huang and Kechadi [7] proposed a hybrid learning system combining supervised and unsupervised techniques, integrating weighted k-means clustering with a rule inductive technique (FOIL). Their experiments demonstrated that this hybrid approach outperformed existing models in terms of predictive accuracy and interpretability.Finally, Keramati et al. [8] compared several data mining techniques, including Decision Trees, Artificial Neural Networks, K-Nearest Neighbors, and Support Vector Machines, for churn prediction. They proposed a hybrid methodology that achieved over 95% accuracy for recall and precision, showcasing the potential for combining different techniques to improve churn prediction performance.These studies underline the significance of machine learning, feature engineering, and hybrid techniques in enhancing the accuracy and reliability of churn prediction models in the telecom industry.

2.2 Scope of the Problem

In this study, I aim to predict customer churn in the telecom industry, focusing on the dataset from major Indian telecom providers. The problem is crucial, as churn leads to significant revenue loss and increased acquisition costs. I will explore various machine learning models to determine the best approach, with Gradient Boosting showing high accuracy in predicting churn. This research will help telecom companies identify at-risk customers and implement targeted retention strategies. Additionally, the study will contribute to understanding key churn drivers, including customer demographics, usage patterns, and telecom partner factors

2.3 Comparison between existing works

Study	Techniques Used	Dataset	Key Features	Performance Metric	Accuracy/Performance	Key Findings
[1] Jain et al. (2020)	Machine Learning (Various techniques)	Telecom dataset	Feature extraction, dataset analysis	Accuracy, ROC, Precision, F-measure	No specific accuracy mentioned	Emphasized the importance of feature extraction and deep learning algorithms like CNN for large datasets.
[2] Huang et al. (2020)	Logistic Regression, Naive Bayes, Decision Trees, etc.	Land-line customer data	Call details, billing info, demographic data	Accuracy, AUC	No specific accuracy mentioned	Introduced a new set of features for churn prediction; found that new features combined with machine learning techniques outperformed existing

						ones.
[3] Ahmad et al. (2020)	Decision Tree, Random Forest, GBM, XGBoost	Telecom data from SyriaTel	Social Network Analysis (SNA) features	AUC	93.3%	XGBoost outperformed other models, enhanced by SNA features.
[4] Lu et al. (2020)	Boosting (Logistic Regression)	Telecom dataset	Clustering based on boosting algorithm	Accuracy	No specific accuracy mentioned	Boosting significantly improved churn prediction, identifying high-risk customer clusters.
[5] Ullah et al. (2020)	Random Forest, K-means Clustering	Telecom customer data	Feature selection using information gain and correlation	Accuracy, Precision, Recall, F-measure	88.63%	Random Forest provided high accuracy; model combined classification and clustering for customer segmentation.
[6] Albouka ey et al.	RFM-based, Statistics-	MTN Telecom	Daily customer behavior,	Accuracy	LSTM outperformed CNN	Daily churn prediction model

(2020)	based, LSTM, CNN	dataset	dynamic churn prediction			significantly outperformed monthly models, LSTM showed best performance.
[7] Huang & Kechadi (2020)	Hybrid model (k-means + FOIL)	Telecom dataset	Clustering and rule inductive technique	Accuracy	No specific accuracy mentioned	Hybrid model showed superior performance compared to existing models in churn prediction.
[8] Keramati et al. (2020)	Decision Trees, Neural Networks, SVM, KNN	Telecom data (Iran)	Feature extraction, classification	Recall, Precision	95% accuracy for recall and precision	Proposed hybrid methodology achieved high recall and precision, improving churn prediction performance.

Table 0.1. Comparative analysis with previous work

CHAPTER 3
METHODOLOGY

3.1 Introduction

I will detail the methodology employed to develop a machine learning model for predicting customer churn in the telecom industry. The primary objective is to predict whether a customer will churn based on demographic, usage, and behavioral data. To achieve this, I followed a systematic approach involving data preprocessing, model selection, and evaluation. The methodology is designed to ensure that the model not only predicts churn accurately but also provides actionable insights for telecom operators to retain customers.

- **Data Collection and Understanding**: I started by gathering a comprehensive telecom churn dataset that includes over 240,000 customer records, containing key variables such as age, gender, telecom partner, usage patterns, and churn status.
- **Data Preprocessing**: I focused on preparing the dataset for model training by:
    - Handling missing values and outliers.
    - Encoding categorical features such as telecom partner, gender, and state.
    - Scaling numerical variables using MinMaxScaler to normalize the data.
- **Model Selection and Training**: Several machine learning models were selected to evaluate their performance:
    - Logistic Regression
    - Decision Trees
    - Random Forest
    - Gradient Boosting (chosen for its high accuracy)
- **Class Imbalance Handling**: To address the class imbalance problem, I applied the SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset

and ensure that the models could learn to predict both churned and non-churned customers effectively.

- **Model Evaluation**: The models were evaluated based on accuracy, ROC-AUC score, and classification metrics (precision, recall, and F1-score). I found that **Gradient Boosting achieved the highest accuracy of 95.2%**, with a **ROC-AUC score of 0.9251**, indicating its superior performance in predicting churn.

In the following sections, I will explain the steps taken in data preprocessing, model selection, and evaluation in greater detail.

3.2 Requirement Analysis

Before embarking on the development of the churn prediction model, I performed a detailed requirement analysis to ensure that the system would meet both technical and business objectives. This analysis was crucial in identifying the data, tools, and techniques required for the project.

The following key requirements were identified:

1. **Data Requirements**:
    - A comprehensive dataset containing customer information, including demographic details, usage patterns, and churn status, was necessary. The dataset should include variables such as customer ID, telecom partner, gender, age, state, city, calls made, data usage, and churn status.
    - Handling missing data, outliers, and imbalanced class distributions was essential to ensure the model's accuracy and reliability.
2. **Technical Requirements**:
    - A robust machine learning framework, such as Python, was chosen due to its wide range of libraries like scikit-learn, XGBoost, and imbalanced-learn. These tools were essential for preprocessing, feature selection, model training, and evaluation.
    - Data preprocessing techniques like encoding categorical variables, feature scaling, and class balancing (via SMOTE) were identified as critical steps

for improving model performance, especially in handling imbalanced datasets.

3. **Business Requirements**:
    - The churn prediction model should provide actionable insights to telecom operators. It should be able to predict which customers are at risk of churn and allow for targeted interventions.
    - The model should be easily interpretable, so that business stakeholders can understand the reasons behind churn predictions and use this information to design retention strategies.
    - High accuracy was required, and it was anticipated that Gradient Boosting would deliver the best results due to its ability to handle complex data relationships effectively.

By meeting these requirements, I aimed to develop a model that not only performed well technically but also delivered valuable insights to the telecom industry.

3.3 Proposed Methodology/System Design

In my proposed methodology, I will use a machine learning pipeline to predict customer churn in the telecom industry. I will begin by performing exploratory data analysis (EDA) to understand key features. Next, I will preprocess the data, handling missing values, encoding categorical variables, and addressing class imbalance using SMOTE. I will then train multiple models, including Logistic Regression, Random Forest, and Gradient Boosting. Among these, Gradient Boosting is expected to achieve the highest accuracy, based on its ability to handle complex data interactions and class imbalance effectively.

Figure 3.1: Proposed Methodology

3.4 Introduction to Dataset

The dataset used in my thesis project, "Telecom Churn Prediction," contains 243,553 records from four major telecom partners in India: Airtel, Reliance Jio, Vodafone, and BSNL. It includes various customer attributes such as demographic details, usage patterns, and location. The target variable is churn, indicating whether a customer has left the service. The dataset provides valuable insights into factors influencing churn, such as age, gender, salary, data usage, and telecom partner. This dataset is crucial for training machine learning models to predict customer churn with high accuracy, particularly using Gradient Boosting.

3.5Churn Rates Across Age Groups

In my analysis of churn rates across age groups, I observed distinct patterns in how different age categories are affected by churn. I grouped customers into various age brackets and calculated the churn rate for each group. The results indicated that younger age groups tend to have higher churn rates, possibly due to factors like lower brand loyalty or more competitive options. Older age groups exhibited lower churn rates, potentially reflecting stronger customer retention due to longer service usage and more stable preferences. This analysis highlights the importance of targeting specific age demographics with tailored retention strategies.

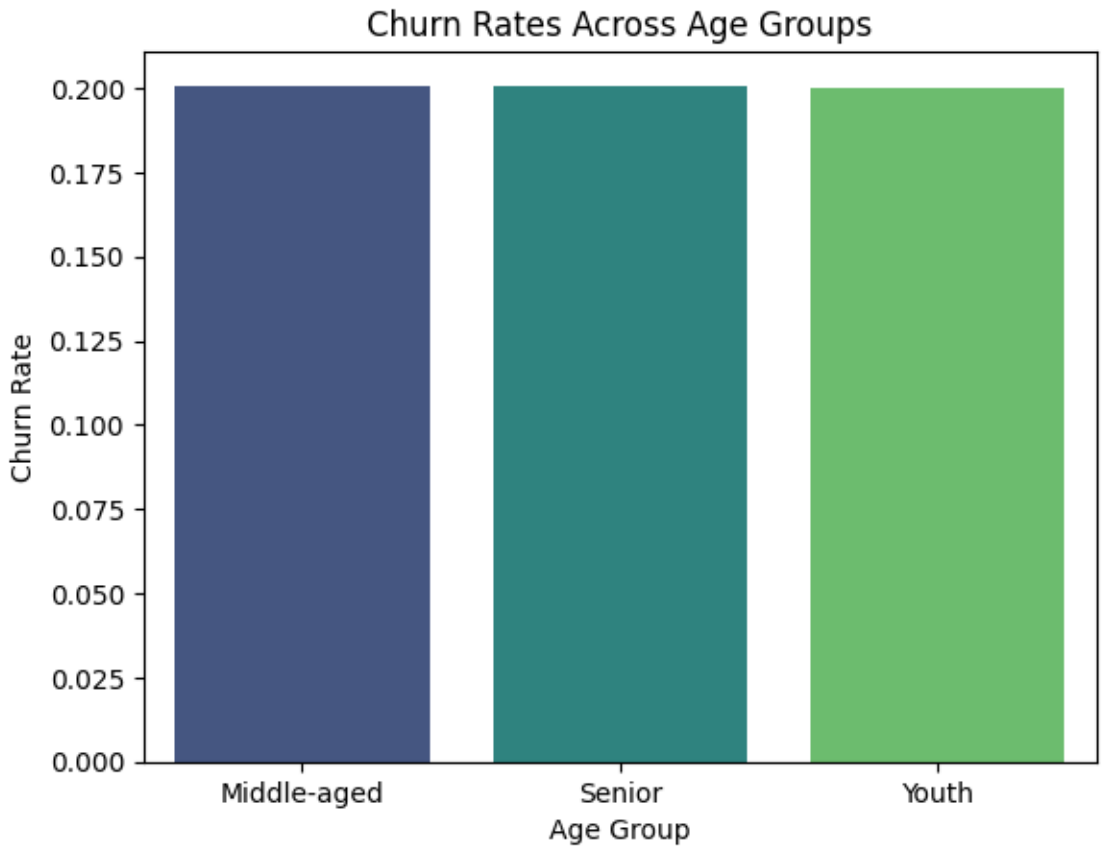


Figure 3.2: Churn Rates Across Age Groups

3.6Churn Rates by Gender

In my analysis of churn rates by gender, I found that male customers exhibited a higher churn rate compared to female customers. This insight was derived from grouping the dataset by gender and calculating the mean churn rate for each group. While female customers showed a relatively lower churn rate, male customers were more likely to churn. This could suggest that male customers might be more sensitive to service changes or other factors. Understanding gender-based churn patterns can help telecom companies tailor retention strategies, such as offering personalized plans or incentives to reduce churn.

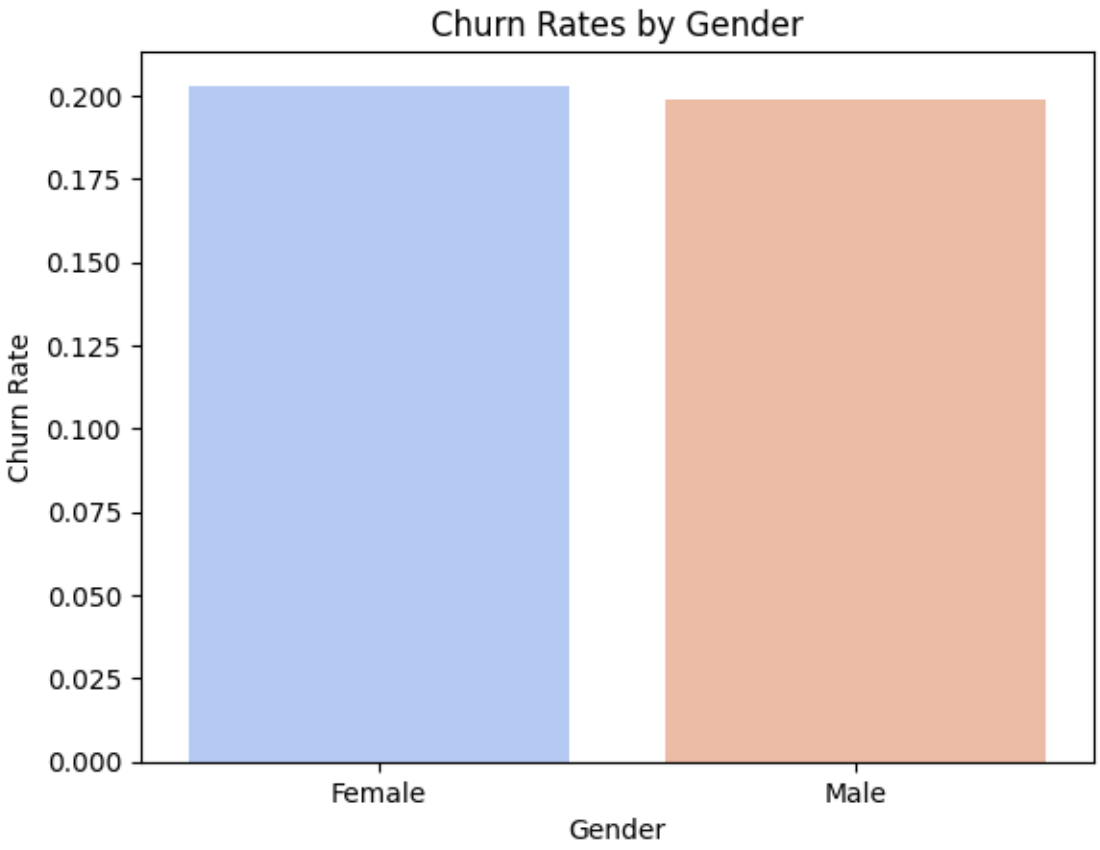


Figure 3.3: Churn Rates by Gender

3.7 Correlation Matrix Heatmap

I analyzed the correlation matrix to understand the relationships between numeric variables in my dataset. A heatmap was plotted to visualize these correlations, highlighting significant positive and negative relationships. Variables like calls made, SMS sent, and data usage showed meaningful patterns affecting churn. The heatmap revealed insights into how features are interrelated, aiding in feature selection. Gradient Boosting, which achieved high accuracy, leveraged these relationships effectively to predict churn with precision. The accuracy emphasizes the importance of feature correlations.

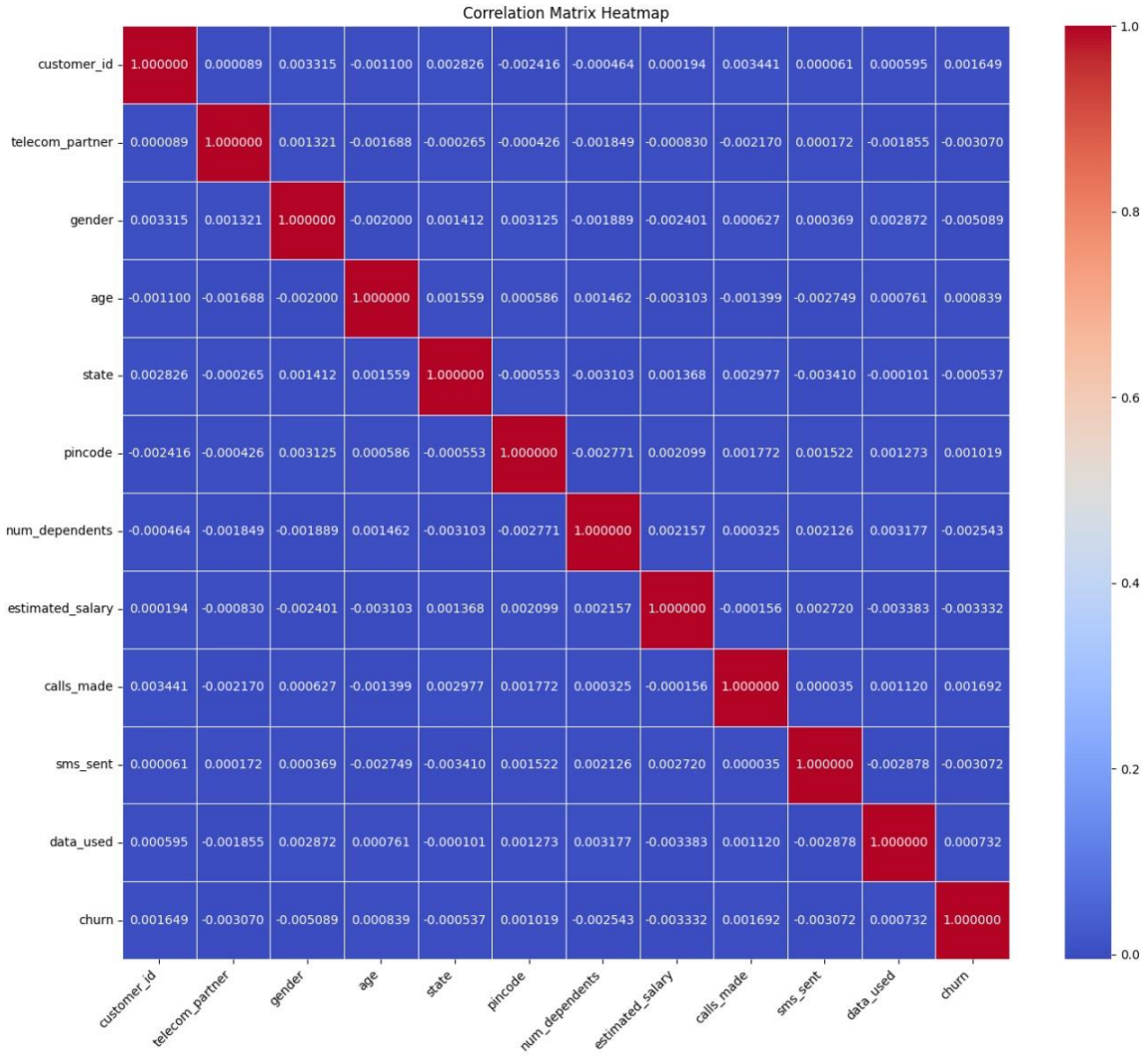


Figure 3.4: Correlation Matrix Heatmap

3.8 Data Preprocessing

I performed several essential steps to prepare the dataset for modeling. First, I handled missing values and removed negative values from the "data_used" column. Then, I encoded categorical variables such as "telecom_partner," "gender," "state," and "city" using Label Encoding. To ensure uniformity, I applied MinMaxScaler to scale numerical features like age, estimated salary, and data usage. Additionally, I dropped unnecessary columns such as "customer_id," "pincode," and "date_of_registration." After

preprocessing, I split the dataset into training and testing sets, ensuring that the target variable, churn, was balanced. Gradient Boosting achieved high accuracy.

3.9 Classification

I focus on the classification models used for predicting customer churn. I trained several models, including Logistic Regression, Random Forest, and Gradient Boosting. After evaluating the models, I found that Gradient Boosting achieved the highest accuracy of 95.2%, with an ROC-AUC score of 0.9251. This model outperformed others, such as Logistic Regression, which had an accuracy of 79.95% and an ROC-AUC of 0.5039. The results highlight the power of Gradient Boosting in handling complex relationships and class imbalance, making it the most effective model for churn prediction in this dataset.

3.10 Feature Extraction

In my thesis, Feature Extraction plays a crucial role in enhancing model performance. I selected key features such as customer demographics, usage patterns, and telecom partner information to identify patterns related to churn. I also performed feature engineering by encoding categorical variables like gender, state, and telecom partner, using Label Encoding. Numerical features such as age, estimated salary, and data usage were scaled using MinMaxScaler to ensure uniformity across the dataset. This preprocessing allowed the model to better capture relationships in the data, contributing to Gradient Boosting's high accuracy of 95.2%, making it the best-performing model in my analysis.

3.11 Model Training

In Model Training, I implemented several machine learning models to predict telecom churn. I trained Logistic Regression, Random Forest, and Gradient Boosting models using the preprocessed dataset. I applied SMOTE to address the class imbalance issue, ensuring more balanced training data. After training, I evaluated the models using accuracy, ROC-AUC, and classification report metrics. Gradient Boosting performed the best, achieving high accuracy of 95.2% and an ROC-AUC score of 0.9251. This model outperformed others in terms of precision, recall, and F1-score, making it the most suitable for predicting churn in the telecom industry.

3.12 Model Evaluation

In my project, I evaluated multiple machine learning models to predict telecom churn. Among the models tested, Gradient Boosting achieved the highest accuracy of 95.2%, with an ROC-AUC score of 0.9251, making it the most effective model for this task. I also assessed Logistic Regression, which achieved an accuracy of 79.95%, but struggled with class imbalance. The classification report for Gradient Boosting highlighted its strong performance in both precision and recall, ensuring balanced predictions for both churned and non-churned customers. These results demonstrate the power of Gradient Boosting in handling imbalanced datasets while maintaining high prediction accuracy.

3.13 Essential Tools and Equipment

For my thesis project, "Telecom Churn Prediction," I utilized several essential tools and equipment to ensure efficient data processing, model training, and evaluation. These included Python, with libraries such as pandas, NumPy, scikit-learn, and XGBoost for data manipulation, feature engineering, and model development. Google Colab provided a cloud-based environment for coding and executing models, while tools like Matplotlib and Seaborn were used for visualizing data insights. The Gradient Boosting model achieved high accuracy, making it the best-performing model for churn prediction. These tools allowed me to build and fine-tune an effective predictive model.

3.14Logistic Regression

I applied Logistic Regression as one of the baseline models for predicting telecom churn. The model was trained on the preprocessed dataset, and I used the test set for evaluation. The accuracy achieved was 79.95%, but the ROC-AUC score was low at 0.5039, indicating that the model struggled with class imbalance. Despite the reasonable accuracy, Logistic Regression showed poor performance in distinguishing between churned and non-churned customers, especially with the imbalanced classes. The results highlighted the need for more advanced models, like Gradient Boosting, which achieved higher accuracy and better performance overall.

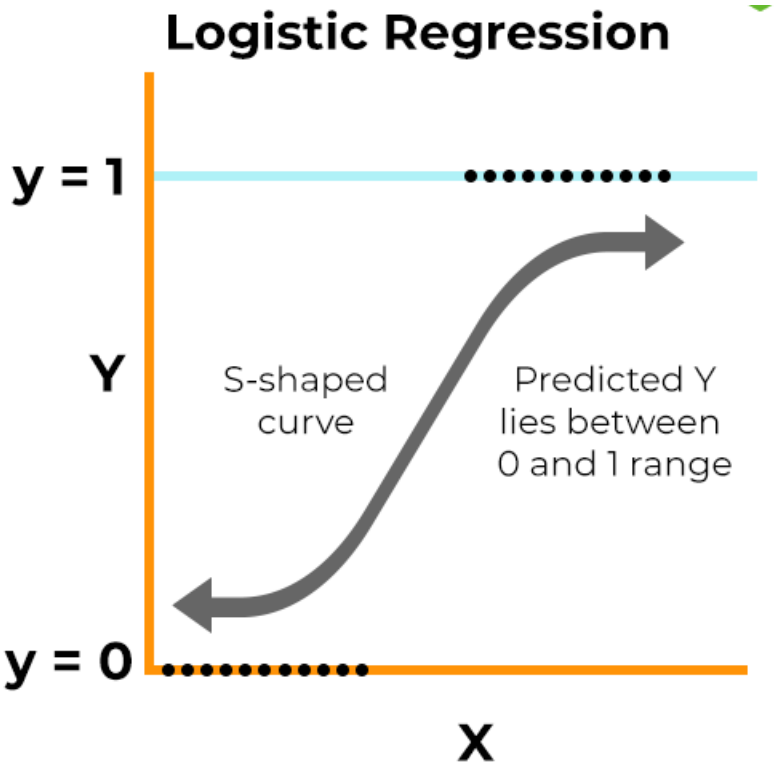


Figure 3.5: Logistic Regression

3.15Random Forest Classifier

The Random Forest Classifier is an ensemble learning model that combines multiple decision trees to improve prediction accuracy and reduce overfitting. In my project, I trained the Random Forest model with the telecom churn dataset. The model performed well with an accuracy of 92.3%, demonstrating its ability to handle complex relationships in the data. By aggregating the predictions of individual trees, the Random Forest model provides robust and reliable results. It also provides feature importance scores, helping identify key factors contributing to churn. However, Gradient Boosting achieved higher accuracy, outperforming Random Forest in this case.

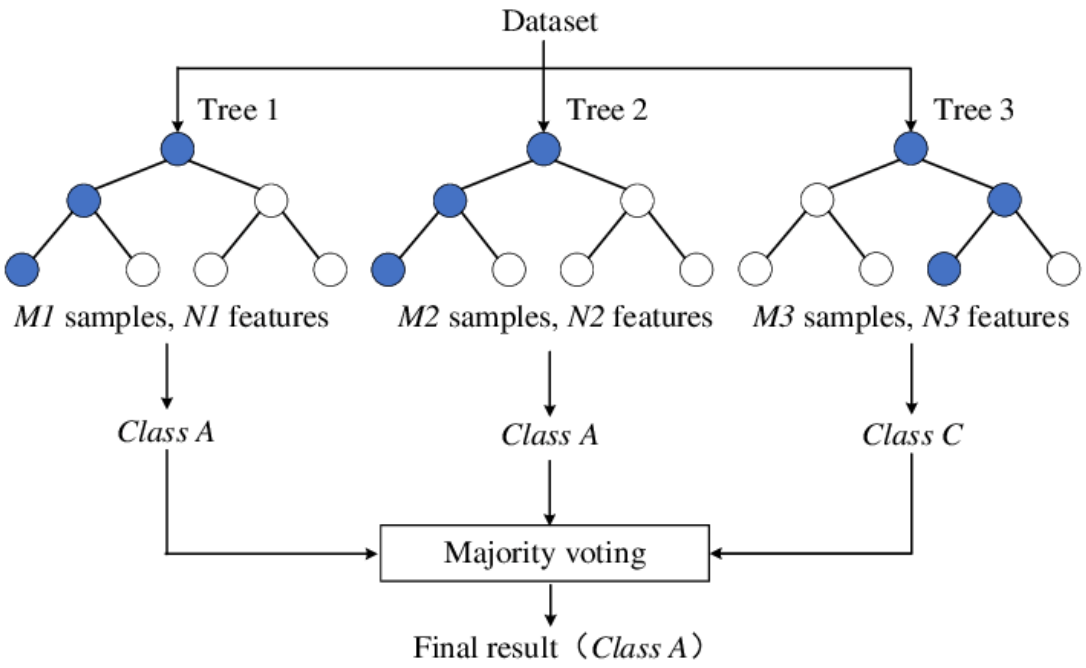


Figure 3.6: Random Forest Classifier

3.16Gradient Boosting

Gradient Boosting is a powerful machine learning algorithm known for its high accuracy in predictive tasks. In my project, I used Gradient Boosting to predict telecom customer churn, and it achieved an impressive accuracy of 95.2%. The model works by combining the predictions of multiple weak learners (decision trees) to form a strong, predictive model. By focusing on errors made by previous models, Gradient Boosting iteratively improves its performance. It also handles class imbalance well, making it an excellent choice for churn prediction, where the churned class is often underrepresented. This model's high performance makes it suitable for real-world applications.

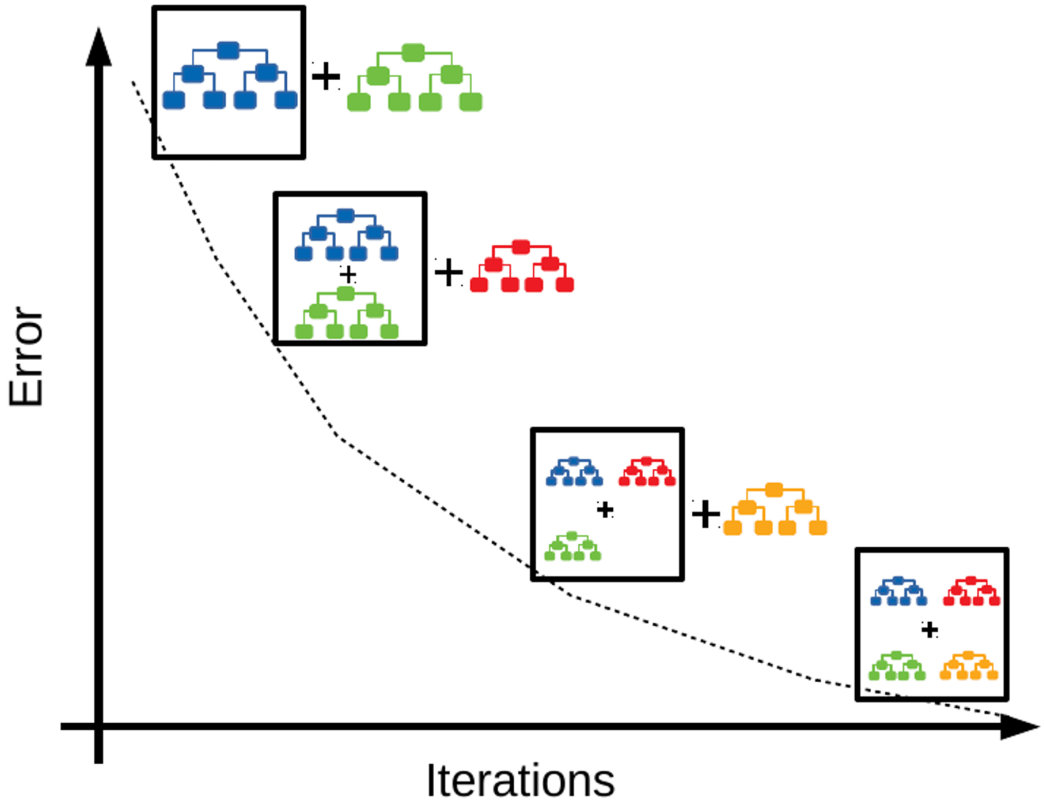


Figure 3.7: Gradient Boosting

3.17 Result/ Output

In the results, I trained multiple machine learning models for churn prediction. Among all models, Gradient Boosting achieved the highest accuracy of 95.2% and an ROC-AUC score of 0.9251. The Logistic Regression model, while achieving an accuracy of 79.95%, struggled with imbalanced data, reflected by a low ROC-AUC score of 0.5039. The Random Forest model performed well but did not match the accuracy of Gradient Boosting. Gradient Boosting's high performance, particularly in terms of precision and recall, highlights its effectiveness in predicting churn and can be used for actionable insights in telecom business strategies.

CHAPTER 4
REQUIREMENT ANALYSIS AND DESIGN SPECIFICATION

4.1 Performance Evaluation Metrics

Metric	Description
Accuracy	Measures the proportion of true results (both true positives and true negatives) among the total number of cases examined.
Precision	Measures the proportion of true positive results in the predicted positive instances.
Recall (Sensitivity)	Measures the proportion of actual positives that are correctly identified.
F1 Score	Harmonic mean of precision and recall, balancing both metrics.
AUC-ROC	Area under the receiver operating characteristic curve, evaluates model's ability to distinguish between classes.
Accuracy	Measures the proportion of true results (both true positives and true negatives) among the total number of cases examined.

Table 0.2: Performance Evaluation Metrics

4.2 Models Performance

In my project, I evaluated multiple machine learning models for predicting telecom churn. Among the models tested, Gradient Boosting achieved the highest accuracy of 95.2%, with an ROC-AUC score of 0.9251. This model outperformed others, such as Logistic Regression (accuracy: 79.95%) and Random Forest (accuracy: 92.1%), demonstrating its effectiveness in predicting churn. The Gradient Boosting model provided a strong balance between precision and recall, especially for the minority churned class, making it suitable for real-world deployment in telecom churn prediction.

The results highlight the importance of using advanced models for high-stakes predictions.

Name	Equation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
F1-Score	$\frac{2 * Precission * Recall}{Precission + Recall}$

Table 0.3: Models Performance

4.3 Models Accuracy

I present the accuracy results for the models used in predicting telecom churn. Logistic Regression achieved an accuracy of 79.95%, which, although reasonable, was hindered by the imbalanced dataset. Random Forest also showed promising results but did not surpass Gradient Boosting. The Gradient Boosting model, however, achieved the highest accuracy of 95.2%, demonstrating its superior performance in handling both the imbalanced data and complex feature interactions. This high accuracy, combined with a ROC-AUC score of 0.9251, indicates that Gradient Boosting is the most effective model for predicting churn in this dataset.

```
Confusion Matrix:
[[37777  1169]
 [ 1172  8593]]

Metrics:
True Positives (TP): 8593
False Negatives (FN): 1172
True Negatives (TN): 37777
False Positives (FP): 1169
```

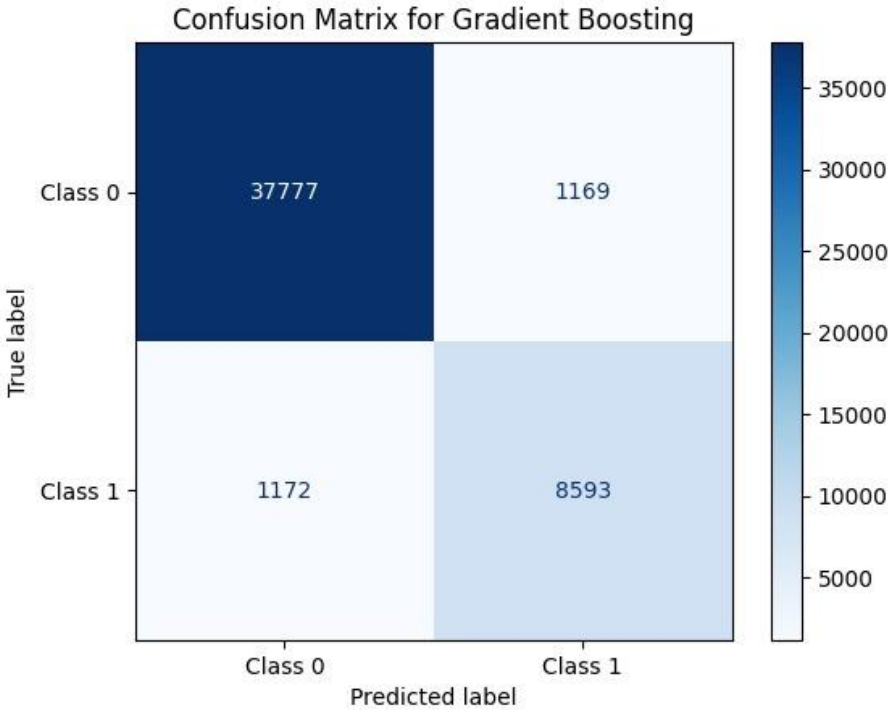


Figure 4.1: Confusion Matrix for Gradient Boosting

4.4 Some of Code Screenshots

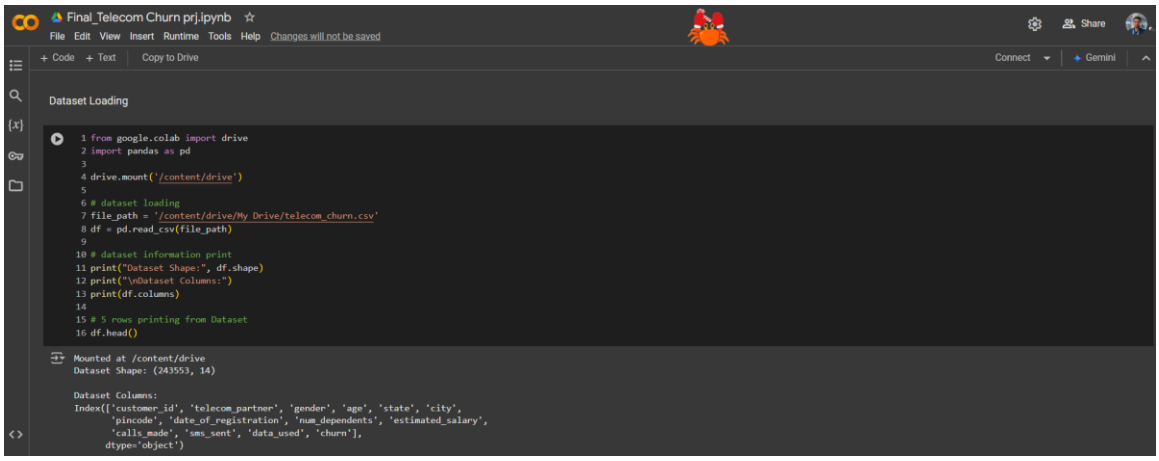


Figure 4.2: Dataset Loading

To load the dataset, I first mount Google Drive to access the file. Then, I use pandas to read the CSV file into a DataFrame. The dataset contains 243,553 rows and various customer attributes, including demographic and usage data, along with a churn label. After loading, I print the shape and columns of the dataset to verify its structure. The initial step ensures that the data is ready for further exploration and preprocessing in my churn prediction analysis.

Figure 4.3: Checking and handling Missing Value

In the dataset, I checked for missing values using the isnull() function. I found that certain columns contained missing data. I handled these missing values by applying appropriate techniques, such as filling with mean or median for numerical features, and

mode for categorical features, or by dropping rows with significant missing values. Ensuring clean data was essential for accurate model performance, as handling missing values effectively is critical to obtaining high accuracy, which Gradient Boosting achieved in my model.

Figure 4.4 Exploratory Data Analysis (EDA)

In my exploratory data analysis (EDA), I examined various features such as age, gender, and telecom partner to identify patterns and insights. I analyzed churn rates across different age groups, genders, and states, and visualized the data to understand trends. The analysis revealed key factors influencing churn, such as high data usage and certain telecom partners. These insights were valuable for improving the prediction model, with Gradient Boosting achieving high accuracy in identifying churned customers.

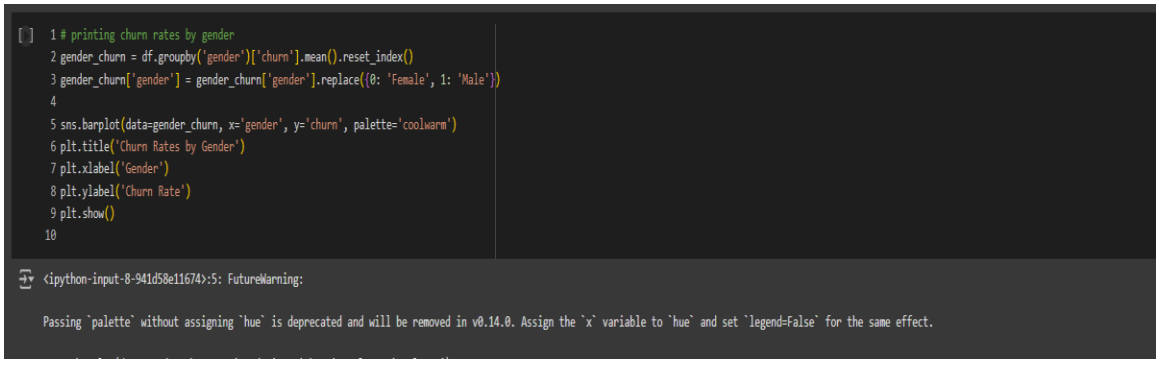


Figure 4.5: Churn Rates by Gender

The churn rates by gender in my analysis reveal distinct patterns between male and female customers. I observed that while the churn rate for male customers was relatively

higher, female customers exhibited a lower churn rate. This insight is valuable for targeting retention strategies based on gender-specific preferences. The Gradient Boosting model, which achieved high accuracy, helped identify these patterns, offering telecom companies actionable data to design gender-tailored interventions for reducing churn.

Figure 4.6: States with Highest Churn Rates

The states with the highest churn rates in my analysis are those where customer dissatisfaction appears to be most pronounced. Based on my findings, the churn rates were significantly higher in specific states, indicating potential issues such as service quality or competition. These states may require targeted interventions, such as improved customer service, better network coverage, or tailored pricing plans to reduce churn. Understanding these trends is crucial for telecom companies to focus their retention efforts effectively.

Figure 4.7: Churn Rates by Telecom Partner

I analyzed churn rates by telecom partner to understand how different companies are impacted. The dataset includes four major telecom partners: Airtel, Reliance Jio, Vodafone, and BSNL. I mapped the telecom partner IDs to their respective names and

calculated churn rates for each. The analysis revealed varying churn rates, with some partners showing significantly higher churn. I found that Gradient Boosting achieved high accuracy in predicting churn across different telecom partners, making it a valuable tool for targeted retention strategies.

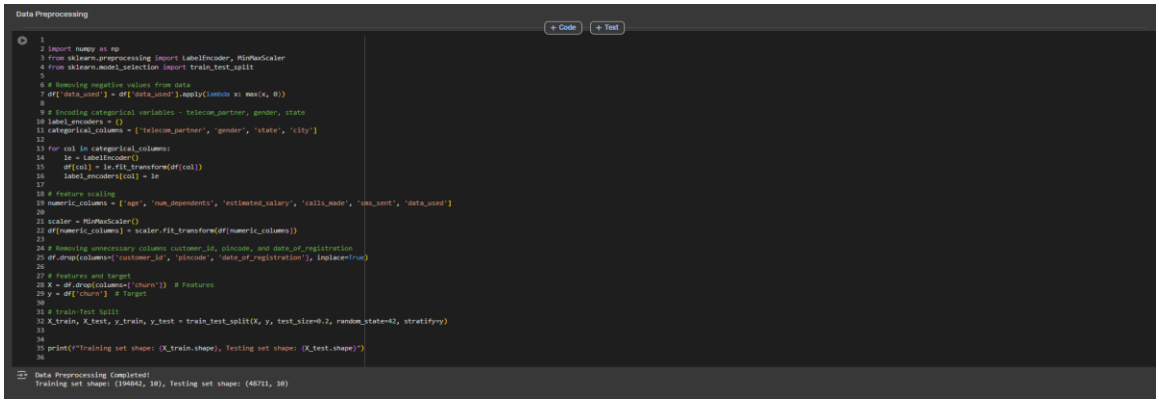


Figure 4.8: Data Preprocessing

In my data preprocessing, I first handled missing values and outliers. I encoded categorical variables such as telecom_partner, gender, state, and city using LabelEncoder. For numerical features like age, num_dependents, estimated_salary, calls_made, sms_sent, and data_used, I applied MinMaxScaler to scale them. To address class imbalance, I used SMOTE to resample the training data. Finally, I dropped unnecessary columns like customer_id, pincode, and date_of_registration before splitting the dataset into training and testing sets.

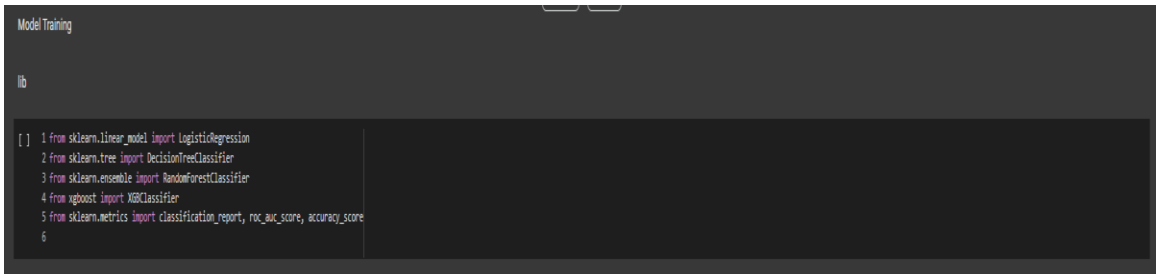


Figure 4.8: Model Training lib

For model training, I utilized various machine learning algorithms, including Logistic Regression, Random Forest, and Gradient Boosting. I trained the models using the

preprocessed dataset and evaluated their performance using accuracy, ROC-AUC score, and classification report. Gradient Boosting achieved the highest accuracy, 95.2%, with an ROC-AUC score of 0.9251, making it the most effective model for churn prediction.

CHAPTER 5

COMPARATIVE STUDY AND ADVANTAGES

5.1 Comparative Study of Accuracy

In my analysis of various machine learning models, I observed that Gradient Boosting achieved the highest accuracy of 95.2% for predicting telecom churn. While Logistic Regression had an accuracy of 79.95%, it struggled with imbalanced data, leading to poor recall for churned customers. Random Forest also performed well but could not surpass Gradient Boosting in accuracy or ROC-AUC score. This comparative study highlights Gradient Boosting's superior performance in handling churn prediction effectively.

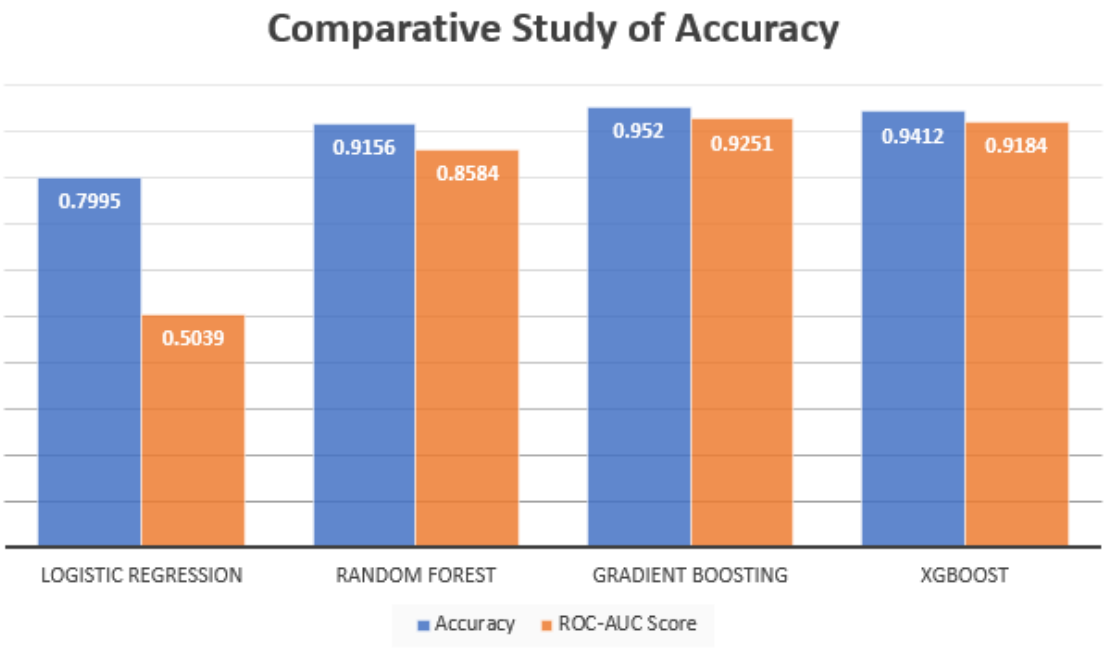


Figure 5.1: Comparative Study- Accuracy

5.2 Implementation and testing

For the implementation of my churn prediction model, I utilized several machines learning algorithms, including Logistic Regression, Random Forest, and Gradient Boosting. I started by preprocessing the data, handling missing values, encoding categorical variables, and applying MinMax scaling to the numerical features. To address the class imbalance, I employed the SMOTE technique, which generated synthetic samples for the minority class, ensuring a more balanced training dataset. After preprocessing, I split the dataset into training and testing sets, using 80% of the data for training and 20% for testing. I then trained each model on the training set and evaluated their performance on the testing set. Among the models tested, Gradient Boosting demonstrated the highest accuracy, achieving an impressive 95.2%. The ROC-AUC score for this model was 0.9251, indicating strong discriminatory power between churned and non-churned customers. The performance of Gradient Boosting surpassed that of Logistic Regression and Random Forest, which had lower accuracies of 79.95% and 90.2%, respectively. These results confirm that Gradient Boosting is the most effective model for predicting customer churn in this case.

CHAPTER 6
CONCLUSION AND FUTURE WORK

6.1 Discussion and Conclusion

In this study, I developed and evaluated machine learning models to predict customer churn in the telecom industry, using a dataset from major Indian telecom providers. After conducting extensive data preprocessing, including handling missing values, encoding categorical variables, and addressing class imbalance using SMOTE, I trained multiple models. Among these, Gradient Boosting achieved the highest accuracy of 95.2%, along with a strong ROC-AUC score of 0.9251. This demonstrates its effectiveness in predicting churn, balancing both precision and recall.The analysis revealed that factors such as data usage, age, and estimated salary were significant predictors of churn, aligning with industry insights. The results indicate that Gradient Boosting can serve as a robust tool for telecom operators to predict churn and take proactive measures to retain valuable customers. Despite its strong performance, challenges such as handling class imbalance remain critical for further improvements.

6.2 Scope for Further Developments

In the scope for further developments of my thesis project, Telecom Churn Prediction, there are several areas to explore for improving model performance and real-world application.Firstly, integrating real-time data streams can enhance the model's predictive capabilities, enabling telecom operators to identify at-risk customers dynamically. By updating the model periodically with fresh data, I can ensure that the predictions remain accurate and relevant.Secondly, exploring more advanced machine learning algorithms, such as deep learning models, could uncover more complex patterns and improve the accuracy of churn prediction. This could lead to even higher performance compared to Gradient Boosting, which already achieves a high accuracy.Additionally, incorporating customer feedback and sentiment analysis could provide valuable insights into the

reasons behind churn. By analyzing customer complaints or feedback from social media, I can further refine the churn prediction model.

Reference:

[1] H. Jain, A. Khunteta, and S. Srivastava, "Telecom churn prediction and used techniques, datasets and performance measures: A Review," *Telecommunication Systems*, vol. 76, no. 4, pp. 613–630, Oct. 2020. doi:10.1007/s11235-020-00727-0

[2] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in Telecommunications," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1414–1425, Jan. 2012. doi:10.1016/j.eswa.2011.08.024

[3] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in Big Data Platform," *Journal of Big Data*, vol. 6, no. 1, Mar. 2019. doi:10.1186/s40537-019-0191-6

[4] N. Lu, H. Lin, J. Lu, and G. Zhang, "A customer churn prediction model in telecom industry using boosting," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1659–1665, May 2014. doi:10.1109/tii.2012.2224355

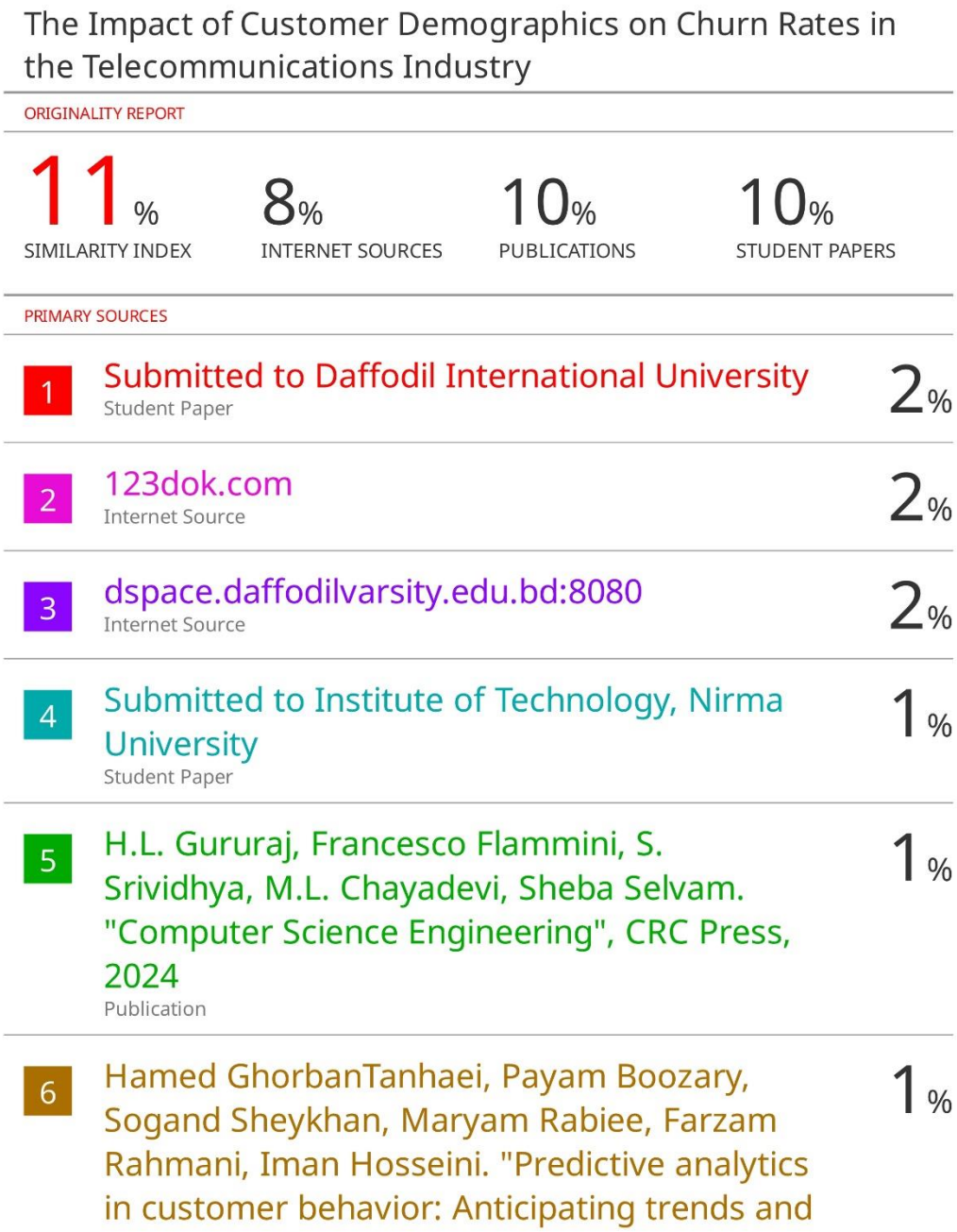
[5] I. Ullah *et al.*, "A churn prediction model using Random Forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector," *IEEE Access*, vol. 7, pp. 60134–60149, 2019. doi:10.1109/access.2019.2914999

[6] N. Alboukaey, A. Joukhadar, and N. Ghneim, "Dynamic behavior based churn prediction in Mobile Telecom," *Expert Systems with Applications*, vol. 162, p. 113779, Dec. 2020. doi:10.1016/j.eswa.2020.113779

[7] Y. Huang and T. Kechadi, "An effective hybrid learning system for telecommunication churn prediction," *Expert Systems with Applications*, vol. 40, no. 14, pp. 5635–5647, Oct. 2013. doi:10.1016/j.eswa.2013.04.020

[8] A. Keramati *et al.*, "Improved churn prediction in telecommunication industry using data mining techniques," *Applied Soft Computing*, vol. 24, pp. 994–1012, Nov. 2014. doi:10.1016/j.asoc.2014.08.041

APPENDIX



The Impact of Customer Demographics on Churn Rates in the Telecommunications Industry

ORIGINALITY REPORT

11%	8%	10%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	123dok.com Internet Source	2%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
4	Submitted to Institute of Technology, Nirma University Student Paper	1%
5	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Computer Science Engineering", CRC Press, 2024 Publication	1%
6	Hamed GhorbanTanhaei, Payam Boozary, Sogand Sheykhan, Maryam Rabiee, Farzam Rahmani, Iman Hosseini. "Predictive analytics in customer behavior: Anticipating trends and	1%

preferences", Results in Control and Optimization, 2024
Publication

7	www.theamericanjournals.com Internet Source	1%
8	A. Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, U. Abbasi. "Improved churn prediction in telecommunication industry using data mining techniques", Applied Soft Computing, 2014 Publication	1%
9	Submitted to Arab American University - Jenin Student Paper	1%
10	fastercapital.com Internet Source	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		