Project Report

On

Telecom Customer Churn Prediction

Submitted to D Y Patil International University, Akurdi, Pune in partial fulfilment of full-time degree

Master of Computer Applications

Submitted By:

Name: Sandip Ramcharit Verma

PRN: 20240804044

Under the Guidance of

Dr. Jagadish S Jakati

School of Computer Science, Engineering and Applications

D Y Patil International University, Akurdi,Pune, INDIA, 411044

[Session 2024-2025]

CERTIFICATE

This is to certify that the work entitled **"Telecom Customer Churn Prediction"** submitted as project I is a bonafide work carried out by Sandip Ramcharit Verma in partial fulfillment of the award of the degree of Master of Computer Applications , D Y Patil International University, Pune, during the academic year 2024- 2025. The project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the Master of Computer Applications.

Dr. Jagadish S Jakati
Project Guide

Dr. Swapnil Waghmare
Project Coordinator

Dr. Maheshwari Biradar
HOD, BCA & MCA

Dr. Rahul Sharma
Director

School of Computer Science Engineering & Applications
D Y Patil International University, Akurdi
Pune, 411044, Maharashtra, INDIA

DECLARATION

I, hereby declare that the following Project entitled **"Telecom Customer Churn Prediction"** is an authentic documentation of my own original work to the best of my knowledge. The following Project and its report in part or whole, has not been presented or submitted by me for any purpose in any other institute or organization. Any contribution made to my work, with whom I have worked at D Y Patil International University, Akurdi, Pune, is explicitly acknowledged in the report.

Name: Sandip Ramcharit Verma
PRN No: 20240804044 Signature :

ACKNOWLEDGEMENT

With due respect, I express my deep sense of gratitude to respected guide Dr. Jagadish S Jakati, for his valuable help and guidance. I am thankful for the encouragement that he has given me in completing this Project successfully.

It is imperative for me to mention the fact that the report of project could not have been accomplished without the periodic suggestions and advice of my project supervisor Dr. Swapnil Waghmare.

I am also grateful to our respected, Dr. Rahul Sharma, Director, Dr. Maheshwari Biradar, HOD, BCA & MCA, and Hon'ble Vice Chancellor, DYPIU, Akurdi Prof. Prabhat Ranjan for permitting me to utilize all the necessary facilities of the University.

I am also thankful to all the other faculty, staff members and laboratory attendants of my school for their kind cooperation and help. Last but certainly not the least; I would like to express my deep appreciation towards my family members and batch mates for providing support and encouragement.

Name: Sandip Ramcharit Verma
PRN: 20240804044

Abstract

Customer churn, or customer attrition, is a critical issue in the telecom industry, leading to significant revenue losses and increased customer acquisition costs. To address this challenge, machine learning (ML) techniques can be leveraged to analyze customer behavior and predict churn probability. This project aims to develop an accurate and efficient ML-based model for telecom customer churn prediction. The study involves data collection, preprocessing, feature engineering, and model training using multiple algorithms, including Decision Tree, Random Forest, K-Nearest Neighbors (KNN), logistic regression, SVM and XGBoost. The models are evaluated based on performance metrics such as accuracy, precision, recall and F1-score to determine the best-performing approach. Additionally, hyperparameter tuning techniques like Grid Search and Random Search are applied to enhance model efficiency. A comprehensive literature review highlights the existing research on telecom churn prediction, identifying key challenges such as class imbalance, feature selection complexity, lack of real-time prediction, and model interpretability issues. The study also proposes solutions, including the use of Synthetic Minority Oversampling Technique (SMOTE), automated feature selection, real-time model deployment, and explainable AI techniques like SHAP. By developing a user-friendly predictive system, this project provides telecom companies with actionable insights to identify at-risk customers and implement data-driven retention strategies. The proposed approach aims to enhance customer satisfaction, minimize churn, and improve overall business profitability in the competitive telecom sector.

List of Figures

TABLE OF CONTENTS

1. INTRODUCTION

Customer churn, also known as customer attrition, is a significant challenge in the telecom industry. It occurs when customers discontinue using a company's services, leading to revenue loss and increased customer acquisition costs. With growing competition, telecom companies must implement effective strategies to predict and mitigate customer churn.

To address this issue, Machine Learning (ML) techniques can be utilized to analyze customer behavior and predict churn probability. By leveraging historical data, telecom providers can identify patterns and key indicators that contribute to customer churn. This enables them to take proactive measures such as personalized offers, better service quality, and improved customer engagement to retain at-risk customers.

The primary objective of this project is to develop an accurate and efficient ML-based model that predicts telecom customer churn. By using data science methodologies, including feature engineering, model training, and evaluation, we aim to create a robust tool that helps telecom companies minimize churn and enhance customer satisfaction.

1.1. Background

In the telecom industry, keeping customers for a long time is very important. When customers stop using a company's services, it is called customer churn. This causes a loss in income and increases the cost of finding new customers. Since many telecom companies offer similar services, it has become more difficult to keep customers loyal.

To solve this problem, we can use Machine Learning (ML) to study customer data and find patterns in customer behavior. ML models can help predict which customers are likely to leave, so companies can take early actions such as offering discounts or improving services to keep them.

This project aims to build a machine learning system that can accurately predict customer churn. The process includes steps like cleaning the data, selecting important features, training different models, and checking how well they perform. We have used models such as Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), and XGBoost. We have also combined models using a Voting Classifier to improve results.

The final system is deployed using Streamlit, which makes it easy to use. It allows real-time predictions, predictions through CSV file uploads, and saves logs of each prediction. The main

goal is to help telecom companies reduce customer churn and improve customer satisfaction by using data to make better decisions.

1.2. Objectives

- To build an automated system for predicting customer churn in the telecom industry using various Machine Learning (ML) algorithms.
- To collect and preprocess telecom customer data by handling missing values, encoding categorical variables, and scaling numerical features to improve model performance.
- To apply feature engineering techniques including feature selection and transformation to enhance data quality and model accuracy.
- To implement and compare multiple ML models such as Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), and XGBoost to identify the most effective model for churn prediction.
- To improve prediction performance using ensemble learning techniques such as the Voting Classifier.
- To perform hyperparameter tuning using GridSearchCV for optimizing the performance of each model.
- To evaluate model performance using metrics such as Accuracy, Precision, Recall, and F1-score to ensure robust and reliable predictions.
- To develop a user-friendly Streamlit web application with features such as real-time prediction, batch prediction via CSV upload, and prediction logs.
- To support telecom companies in reducing customer churn by providing actionable insights and enabling data-driven customer retention strategies.

1.3. Purpose

The purpose of this project is to develop a robust and user-friendly machine learning-based system for predicting customer churn in the telecom industry. Customer churn poses a major challenge, leading to considerable revenue loss. By accurately identifying customers at risk of leaving, telecom companies can take timely, preventive actions to enhance retention and business performance.

This project aims to analyze historical customer data to uncover behavioral patterns and critical factors that drive churn. Using various machine learning algorithms and ensemble methods, the

system will generate actionable insights to support data-driven decision-making and customer engagement strategies.

The main goals of this project include:

- To develop a predictive model that classifies customers into churn and non-churn categories with high accuracy.
- To provide insights into significant customer attributes and service features influencing churn.
- To enable telecom companies to proactively identify high-risk customers and implement targeted retention strategies.
- To deploy a real-time prediction system using Streamlit that includes features like single-entry predictions, batch prediction through CSV uploads, and prediction logging.
- To support business growth by reducing churn rates and improving long-term customer relationships through data-driven strategies.

By integrating models such as Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), XGBoost, and ensemble learning through a Voting Classifier, this system offers telecom companies a practical and efficient tool for churn prediction and customer retention planning.

1.4. Scope

This project focuses on building a machine learning-based system to predict customer churn in the telecom industry. The system uses historical customer data and applies several machine learning models such as Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), and XGBoost. It also uses ensemble learning (Voting Classifier) to improve prediction performance.

The scope of the project includes:

- Collecting and preparing telecom customer data.
- Handling missing values, encoding categorical variables, and scaling numerical features.
- Performing feature selection and transformation to improve model performance.
- Training and comparing multiple ML models to find the best one.

- Evaluating models using metrics like accuracy, precision, recall, F1-score, and AUC-ROC.
- Deploying the final model using Streamlit with features like real-time prediction, CSV file upload, and prediction logging.

This project is limited to the data provided and does not include customer feedback systems, real-time data collection from telecom systems, or integration with company databases.

1.5. Applicability

The machine learning system developed in this project can be applied by telecom companies to:

- Predict which customers are likely to leave (churn).
- Take early action to reduce churn by offering personalized services or support.
- Improve customer retention strategies based on data-driven insights.
- Save costs by focusing efforts on at-risk customers.
- Make better business decisions using analytics and machine learning.

This system can also be useful for data analysts, business managers, and decision-makers in other industries facing customer retention issues, such as banking, insurance, and e-commerce, with slight adjustments to the data.

1.6. Literature Review

Reference 1

Title: A Review on Machine Learning-Based Customer Churn Prediction in the Telecom Industry
Authors: Sawsan Barham, Nowfal Aweisi, Ala' Khalifeh
Year: 2023
Methodology: This paper reviews 33 studies on customer churn prediction in telecom from 2019–2022. Techniques analyzed include Random Forest, Logistic Regression, Decision Trees, and XGBoost. Random Forest was found to achieve the highest accuracy, with some studies reporting 97.4%. The paper also discusses feature selection and handling imbalanced

datasets.
Drawback: Lacks empirical validation and does not discuss real-world deployment challenges such as scalability and computational costs. Model interpretability is also not explored.

Reference 2

Title: Exploratory Data Analysis and Customer Churn Prediction for the Telecommunication Industry
Authors: Kiran Deep Singh, Gaganpreet Kaur, Prabh Deep Singh, Vikas Khullar, Ankit Bansal, Vikas TripathiKiran Deep Singh, Gaganpreet Kaur, Prabh Deep Singh, Vikas Khullar, Ankit Bansal, Vikas Tripathi
Year: 2023
Methodology: This study focuses on Exploratory Data Analysis (EDA) and machine learning models for churn prediction in telecom. It examines customer behavior and applies XGBoost, achieving 82.80% accuracy on a real-world dataset. The study emphasizes feature selection and customer retention strategies based on key indicators.
Drawback: The study does not address class imbalance handling methods like SMOTE, which may affect prediction accuracy. Additionally, it lacks a comparison with deep learning models and does not explore computational costs for large datasets.

Reference 3

Title: The Impact of SMOTE and ADASYN on Random Forest and Advanced Gradient Boosting Techniques in Telecom Customer Churn Prediction
Authors: Mehdi Imani, Majid Joudaki, Zahra Ghaderpour, Ali Beikmohammadi
Year: 2024
Methodology: This study examines the effects of SMOTE and ADASYN on Random Forest, XGBoost, LightGBM, and CatBoost for customer churn prediction. The dataset contains 4,250 training and 750 testing samples. After applying SMOTE and ADASYN, LightGBM achieved the highest F1-score (89%) and ROC AUC (95%).
Drawback: The study does not evaluate the risk of synthetic noise from oversampling techniques. It also notes that hyperparameter tuning had minimal impact on model performance.

Reference 4

Title: Customer churn prediction in telecom sector using machine learning techniques
Authors: Sharmila K. Wagh, A. A. Andhale, K. S. Wagh, J. R. Pansare, S. P. Ambadekar, S. Gawande
Year: 2024

Methodology: The study focuses on predicting customer churn in the telecom industry using machine learning techniques. The authors employed Random Forest (RF), Decision Tree, and K-Nearest Neighbors (KNN) on the IBM Telco dataset. SMOTE was used for class imbalance. Feature selection was done using Pearson correlation, and the model achieved 99% accuracy with Random Forest.
Drawback: The extremely high accuracy raises concerns about potential overfitting, especially on imbalanced data. The use of SMOTE may introduce unrealistic samples. The study also lacks discussion on computational complexity.

Reference 5

Title: Predicting Customer Churn in Telecom Industry: A Machine Learning Approach for Improving Customer Retention
Authors: Abhikumar Patel, Amit G Kumar
Year: 2023
Methodology: This study applies supervised ML techniques including Bernoulli Naive Bayes, Gaussian Naive Bayes, SVM, KNN, Decision Tree, Random Forest, and XGBoost. XGBoost achieved the highest accuracy of 94%. Feature influence analysis identified international plans and customer service calls as key indicators.
Drawback: Potential overfitting with XGBoost was not analyzed. Issues like computational complexity and scalability were not addressed.

Reference 6

Title: Telecom Customer Churn Prediction Using Enhanced Machine Learning Classification Techniques
Authors: Goldy Verma
Year: 2024
Methodology: This research compares Decision Tree, Random Forest, and KNN on a Kaggle dataset. Accuracy scores were 79% for Decision Tree and KNN, and 82% for Random Forest. The study emphasizes preprocessing and feature selection.
Drawback: The study does not handle class imbalance, affecting reliability. Also lacks interpretability analysis, limiting business usefulness.

Reference 7

Title: Customer Churn Prediction Using Synthetic Minority Oversampling Technique
Authors: Aishwarya H M, Soundarya B, Bindhiya T, C Christlin Shanuja, S Tanisha
Year: 2023
Methodology: The study applies SMOTE and evaluates Gradient Boosting, Random Forest,

Decision Tree, and Logistic Regression. Gradient Boosting achieved 95.13% accuracy. Tenure and monthly charges were key features.
Drawback: While accuracy improved, SMOTE may distort data. Computational cost and interpretability were not evaluated.

Reference 8

Title: Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms
Authors: V. Kavitha, S. V Mohan Kumar, M. Harish, G. Hemanth Kumar
Year: 2020
Methodology: This study uses Random Forest, XGBoost, and Logistic Regression on a Kaggle dataset. Random Forest performed best with 93% accuracy. Preprocessing steps included normalization and feature selection.
Drawback: Logistic Regression underperformed and was unsuitable for imbalanced data. Model interpretability and hyperparameter tuning were not addressed.

Reference 9

Title: Customer Churn Prediction Based on Interpretable Machine Learning Algorithms in Telecom Industry
Authors: Liwen Ou
Year: 2022
Methodology: Focused on interpretability using Random Forest, Decision Tree, and Extra Tree Classifier. Feature importance highlighted tenure, total charges, and monthly charges.
Drawback: Decision Tree showed lower accuracy. Class imbalance was not handled. Computational complexity of ensemble models was not discussed.

Reference 10

Title: Machine Learning-Based Telecom-Customer Churn Prediction
Authors: Pushkar Bhuse, Aayushi Gandhi, Parth Meswani, Riya Muni, Neha Katre
Year: 2020
Methodology: This study compares Random Forest, SVM, XGBoost, Ridge Classifier, and Deep Neural Networks. Random Forest achieved 90.96% accuracy. Feature selection and grid search tuning were applied.
Drawback: Deep learning was not extensively tuned. Class imbalance and computational scalability were not discussed.

1.7. Literature Outcome

The reviewed studies show that successful churn prediction involves several important steps such as data preprocessing, feature selection, and handling class imbalance. Many researchers used methods like SMOTE to balance the dataset and improve the performance of models when predicting minority classes. Machine learning models such as Decision Tree, Random Forest, KNN, Logistic Regression, SVM, and XGBoost were commonly used across different studies, with ensemble methods often giving better accuracy and stability. Techniques like hyperparameter tuning and proper evaluation using accuracy, precision, recall, and F1-score helped improve model effectiveness. The literature highlights that using multiple models and combining them can lead to more reliable results. These findings support the use of a complete machine learning pipeline, from data preparation to model combination, for better customer churn prediction in the telecom industry.

2. PROJECT PLAN

2.1. Problem Statement

In the telecom industry, customer retention is a major challenge due to the availability of many service providers offering similar services. Losing existing customers, also known as customer churn, leads to significant revenue loss and affects business growth. Traditional methods of identifying churn-prone customers are often reactive, time-consuming, and inaccurate.

There is a need for a predictive system that can analyze customer data and identify patterns that indicate whether a customer is likely to leave. By using machine learning techniques, it is possible to predict customer churn in advance, allowing telecom companies to take proactive steps to retain valuable customers.

The main problem this project addresses is how to accurately predict customer churn using machine learning models trained on historical customer data, thereby helping telecom companies improve customer retention strategies.

2.2. Requirement Specification

Functional Requirements

- The system should accept customer data as input, either manually or via CSV upload.
- The system should preprocess the input data, including handling missing values and encoding categorical variables.
- The system should handle class imbalance using oversampling techniques like SMOTE.
- The system should apply multiple machine learning models to predict customer churn.
- The system should evaluate model performance using metrics such as accuracy, precision, recall, and F1-score.
- The system should provide churn prediction results with probability and risk level (Low, Moderate, High).
- The system should display real-time predictions via a web interface.
- The system should allow batch predictions through uploaded CSV files.

- The system should maintain logs for each prediction.

Non-Functional Requirements

- The system should be user-friendly and easy to navigate.
- The system should provide quick response time for real-time predictions.
- The system should ensure data security and user privacy.
- The system should be scalable to handle large datasets if required.
- The system should maintain model interpretability and transparency.

Software Requirements

- Operating System: Windows / Linux
- Programming Language: Python 3.9+
- Libraries: pandas, numpy, scikit-learn, matplotlib, seaborn, xgboost, imbalanced-learn
- Web Framework: Streamlit
- IDE: Jupyter Notebook / VS Code

Hardware Requirements

Minimum Requirements

- Processor: Intel Core i3 (8th Gen) or equivalent
- RAM: 4 GB
- Storage: 5 GB of free disk space

Recommended Requirements

- Processor: Intel Core i5 (10th Gen or higher) / AMD Ryzen 5 or better
- RAM: 8 GB or higher
- Storage: 10 GB or more of free disk space (preferably SSD for faster performance)

2.3. Time Line Chart

The following timeline was followed during the development of the Customer Churn Prediction project:

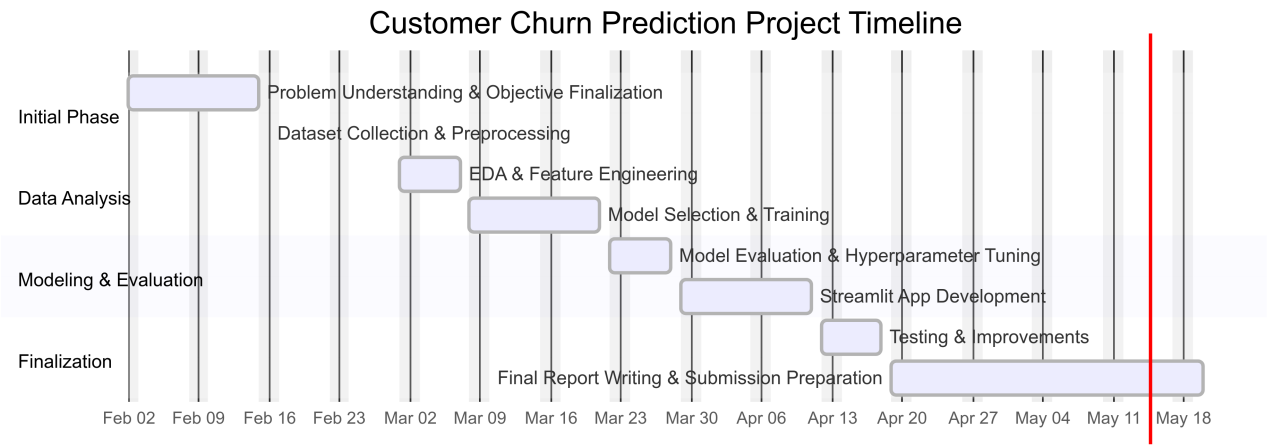

Figure 2.1: Time Line Chart for Customer Churn Prediction

3. PROPOSED METHODOLOGY

3.1. System Architecture

The system architecture for the Customer Churn Prediction project is designed to efficiently process telecom customer data and predict the likelihood of churn using machine learning models. The architecture consists of the following main components:

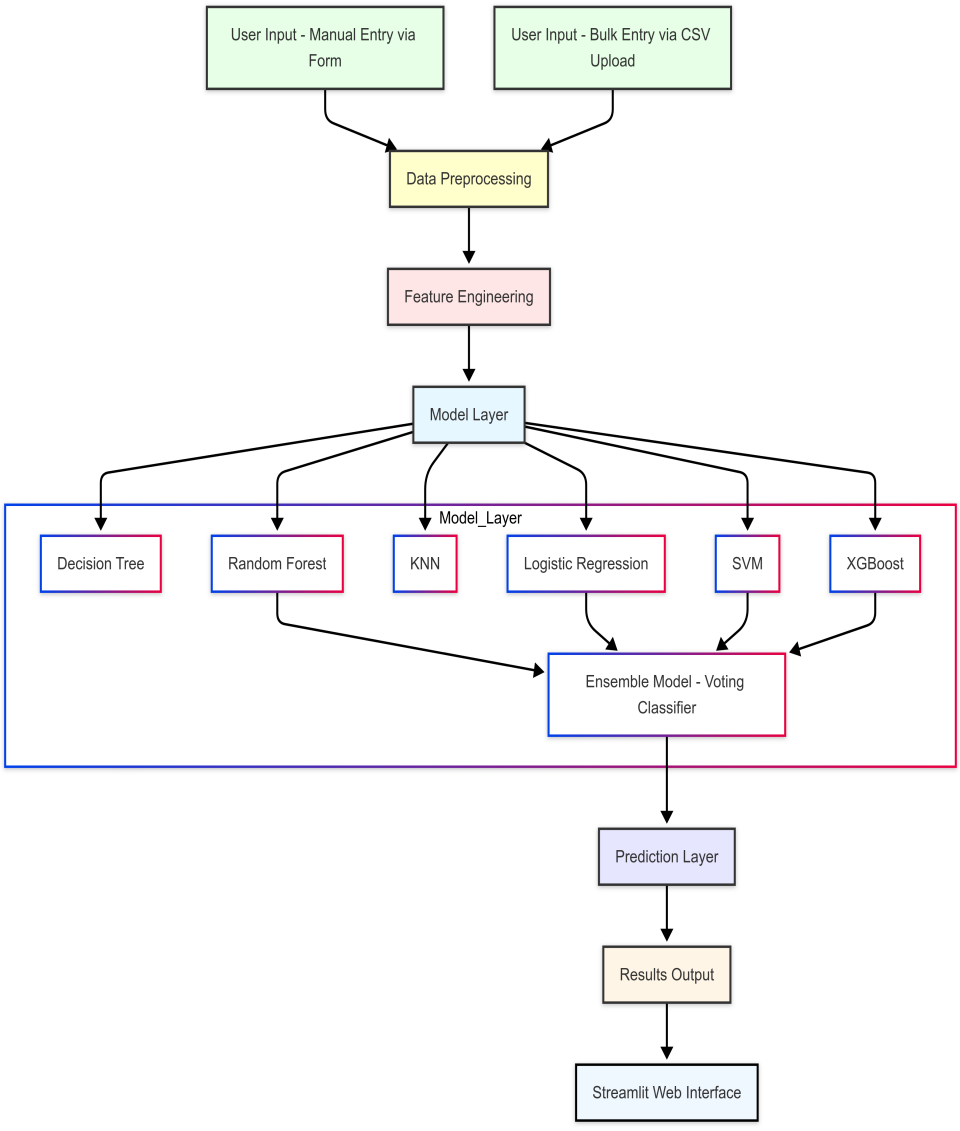

Figure 3.1: System Architecture Diagram for Customer Churn Prediction

- **Data Input:** The system accepts input data either in real-time through a web interface or in batch mode via CSV file upload. This provides flexibility in how data can be ingested into the system.

- **Data Preprocessing:** This module handles various data cleaning tasks, such as handling missing values, encoding categorical features, and performing feature scaling. Additionally, it addresses class imbalance using SMOTE (Synthetic Minority Over-sampling Technique) to improve model performance.
- **Feature Engineering:** This phase involves the extraction and transformation of important features that contribute significantly to model performance. Correlation analysis and domain-specific knowledge are applied to create relevant features that improve model accuracy.
- **Model Layer:** Several machine learning models are employed in this stage, including:
  - Decision Tree
  - Random Forest
  - K-Nearest Neighbors (KNN)
  - Logistic Regression
  - SVM (Support Vector Machine)
  - XGBoost

  These models are trained, and hyperparameter tuning is performed using GridSearchCV. A **Voting Classifier** (ensemble model) is utilized for combining the predictions of individual models to improve performance and reduce bias.
- **Prediction Layer:** After model training, the prediction layer is responsible for producing outputs based on new input data. The following predictions are made:
  - Churn Prediction (Yes/No)
  - Churn Probability (0–100%)
  - Churn Risk Level (Low / Moderate / High)
- **User Interface (Streamlit):** The user interface is developed using Streamlit, which provides an interactive environment. It allows users to input data in real-time, view prediction logs, and perform batch predictions by uploading CSV files. The UI is designed to be intuitive and accessible to non-technical users.

This modular architecture ensures that the system is scalable, maintainable, and easy to deploy. It enables efficient data processing, accurate predictions, and user-friendly interaction, all contributing to actionable insights for customer retention strategies in the telecom industry.

3.2. Methodology

This project follows a comprehensive machine learning pipeline to predict customer churn in the telecom industry. The methodology includes several critical steps: data preprocessing, feature engineering, data balancing, model building, evaluation, and ensemble learning. Each step plays a vital role in ensuring that the final model is accurate, robust, and effective for churn prediction.

- **Data Preprocessing:** The initial preprocessing stage involved handling missing values, encoding categorical variables, and scaling numerical features to standardize the data for better model performance. This step ensures that the input data is clean, consistent, and ready for model training.
- **Feature Engineering:** Features were carefully selected based on their relevance to predicting customer churn. These included customer demographics (age, gender, senior citizen status), service usage patterns (internet service, tech support), contract type, and billing-related details. Feature engineering aimed at improving both model interpretability and accuracy by selecting the most influential features for churn prediction.
- **Data Balancing:** The dataset exhibited class imbalance, with a significantly higher number of non-churned customers. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE generates synthetic instances of the minority class (churners), thereby ensuring the model learns effectively from both the majority and minority classes.
- **Machine Learning Models Implemented:** A diverse set of machine learning classifiers were developed and compared for their performance. The following models were used:
  - **Decision Tree Classifier:** A rule-based model that splits customer data into branches based on feature thresholds.
  - **Random Forest Classifier:** An ensemble of decision trees that enhances prediction accuracy and reduces overfitting.
  - **K-Nearest Neighbors (KNN) Classifier:** A distance-based method that predicts churn based on similarities with neighboring data points.
  - **Logistic Regression:** A linear model that estimates the probability of churn, useful for interpreting the influence of features.
  - **Support Vector Machine (SVM):** A model that constructs an optimal hyperplane in high-dimensional space to separate churn and non-churn classes.

  - **XGBoost Classifier:** A gradient boosting algorithm known for its efficiency and performance, especially on structured data.

- **Hyperparameter Tuning:** GridSearchCV was employed to optimize model performance. This involved testing multiple parameter combinations using cross-validation to find the best hyperparameters for each model, ensuring the models perform at their peak.

- **Model Evaluation:** Model performance was assessed using key classification metrics, particularly focusing on the minority class (churn = 1), to ensure that the models did not overlook churn prediction:

  - **Accuracy:** Measures the overall correctness of the model.
  - **Precision:** Focuses on minimizing false positives, ensuring the model is accurate when predicting churn.
  - **Recall:** Ensures the model identifies as many churn cases as possible, minimizing false negatives.
  - **F1-Score:** A harmonic mean of precision and recall, providing a balance between the two.

  Among the models, Random Forest, XGBoost, and SVM showed strong performance, with Random Forest achieving the best balance between precision, recall, and accuracy.

- **Ensemble Learning with Voting Classifier:** To further enhance prediction stability and accuracy, a soft voting ensemble classifier was built. This ensemble combined the predictions of:

  - Random Forest
  - XGBoost
  - SVM
  - Logistic Regression

  This ensemble approach leveraged the strengths of individual models, resulting in improved generalization on unseen data. The voting classifier ultimately provided more stable and accurate predictions for customer churn.

By following this methodology, the project aims to develop an effective churn prediction system that helps telecom companies identify at-risk customers and take proactive actions to retain them. This approach ensures that the final model is both accurate and scalable, capable of handling real-world data effectively.

3.3. Pseudo code

The following pseudo code outlines the major steps in the customer churn prediction pipeline:

```
1. Load data (from CSV or user input form)
2. Handle missing values
   - If any feature has missing values
3. Encode categorical features
   - Convert categorical variables to numerical values
4. Scale numerical features
   - Normalize or standardize numerical features for model consistency
5. Apply SMOTE (Synthetic Minority Over-sampling Technique) for balancing
    the dataset
   - Generate synthetic samples for the minority class (churn = 1)
6. Split data into training and testing sets
   - 80% for training, 20% for testing
7. Initialize models:
   - Decision Tree
   - Random Forest
   - K-Nearest Neighbors (KNN)
   - Logistic Regression
   - Support Vector Machine (SVM)
   - XGBoost
8. Train each model on the training set
9. Evaluate each model:
   - Calculate Accuracy, Precision, Recall, F1-Score for each model
10. Tune hyperparameters using GridSearchCV
   - Search for the best model parameters using cross-validation
11. Create an ensemble model (Voting Classifier)
   - Combine predictions from Random Forest, XGBoost, SVM, and Logistic
   Regression using a soft voting strategy
12. Evaluate ensemble model on the test set
13. Output predictions:
   - Churn prediction (Yes/No)
   - Churn probability (0 - 100%)
   - Churn risk level (Low / Moderate / High)
14. Display results via Streamlit interface
   - Show real-time prediction and logs
```

3.4. Design

The design of the Telecom Customer Churn Prediction System follows a modular and data-driven approach. It integrates various stages of machine learning including data ingestion, preprocessing, feature engineering, model training, evaluation, and result interpretation. The system is structured to handle both single customer input and bulk CSV-based predictions through an interactive interface.

This section outlines the data flow diagrams and UML representations that capture the system architecture, flow of information, and component interactions.

3.4.1. Data Flow Diagrams

Data Flow Diagrams (DFDs) are used to represent the flow of information within the system. They provide a visual breakdown of how customer data moves from input to output across various internal components.

Level 0 DFD

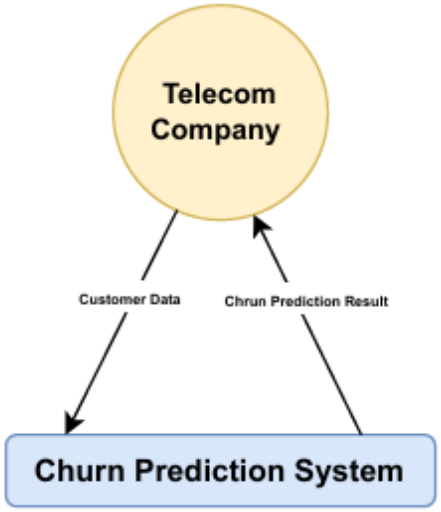

Fig. 3.2: Level 0 Data Flow Diagram for Telecom Customer Churn Prediction System

Entities:

- **Telecom Company:**
  - This is the external entity that interacts with the churn prediction system.
  - It acts as both the data provider and the consumer of results.

Main Process:

- **Churn Prediction System:**
  - This is the core process that takes customer data as input, processes it through a machine learning pipeline, and generates churn predictions.
  - Internally, it includes data preprocessing, model training, ensemble prediction, and result generation (which will be detailed in Level 1 DFD).

Data Flows:

- **Customer Data (→):** Represents the input from the telecom company including demographics, service usage, billing, etc.
- **Churn Prediction Results (←):** Represents the output including churn likelihood, probability, and risk level.

Level 1 DFD

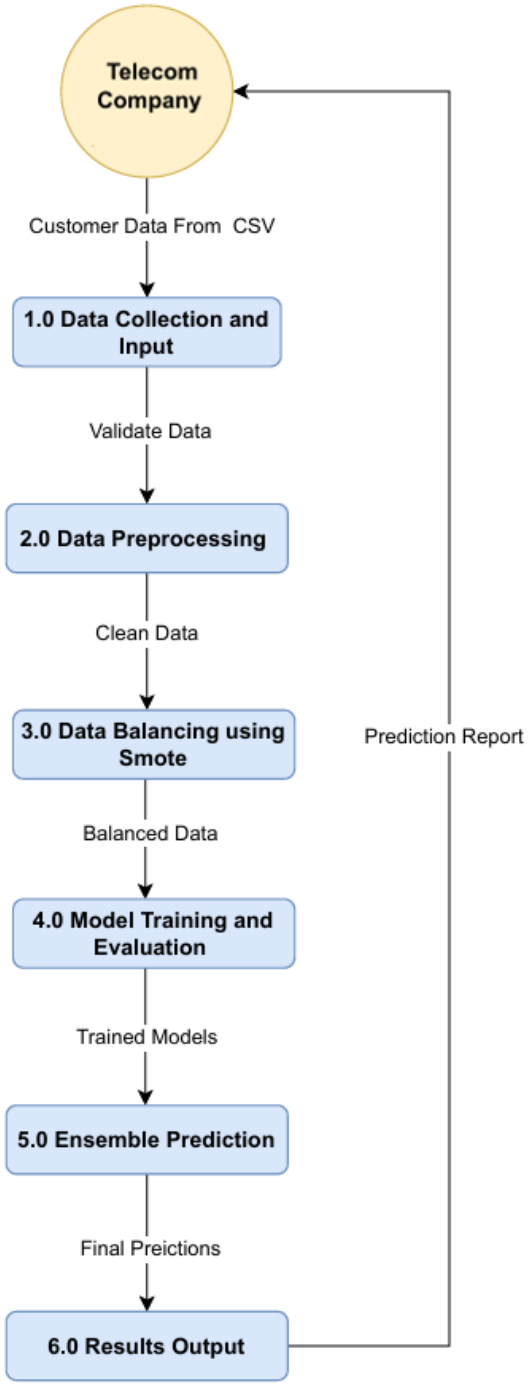

Fig. 3.3: Level 1 Data Flow Diagram showing internal components of the Churn Prediction System

Entities:

- **Telecom Company:**
  - This represents the end-user or company analyst who interacts with the system.
  - Provides input data manually or via CSV upload.

  - Receives the churn prediction report for analysis and retention strategy.

Processes:

- **Data Collection and Input:** Collects input and checks for completeness and format.
- **Data Preprocessing:** Manages missing values, encodes categories, and scales numerical features.
- **Data Balancing using SMOTE:** Balances data by generating synthetic samples for the minority (churn) class.
- **Model Training and Evaluation:** Trains multiple models and evaluates them using appropriate metrics.
- **Ensemble Prediction:** Uses soft voting from top models to finalize prediction.
- **Results Output and Logging:** Displays prediction to the user and logs it for further use.

Data Flows:

- **Flow:** Customer Data $\rightarrow$ Validated Data $\rightarrow$ Clean Data $\rightarrow$ Balanced Data $\rightarrow$ Trained Models $\rightarrow$ Final Prediction
- **Output:** The Prediction Report is sent back to the Telecom Company.

3.4.2. UML Diagrams

This section presents the UML diagrams applicable to the project. UML (Unified Modeling Language) diagrams help visualize the system's structure and behavior during various stages of development.

1. **Flow Chart Diagram**

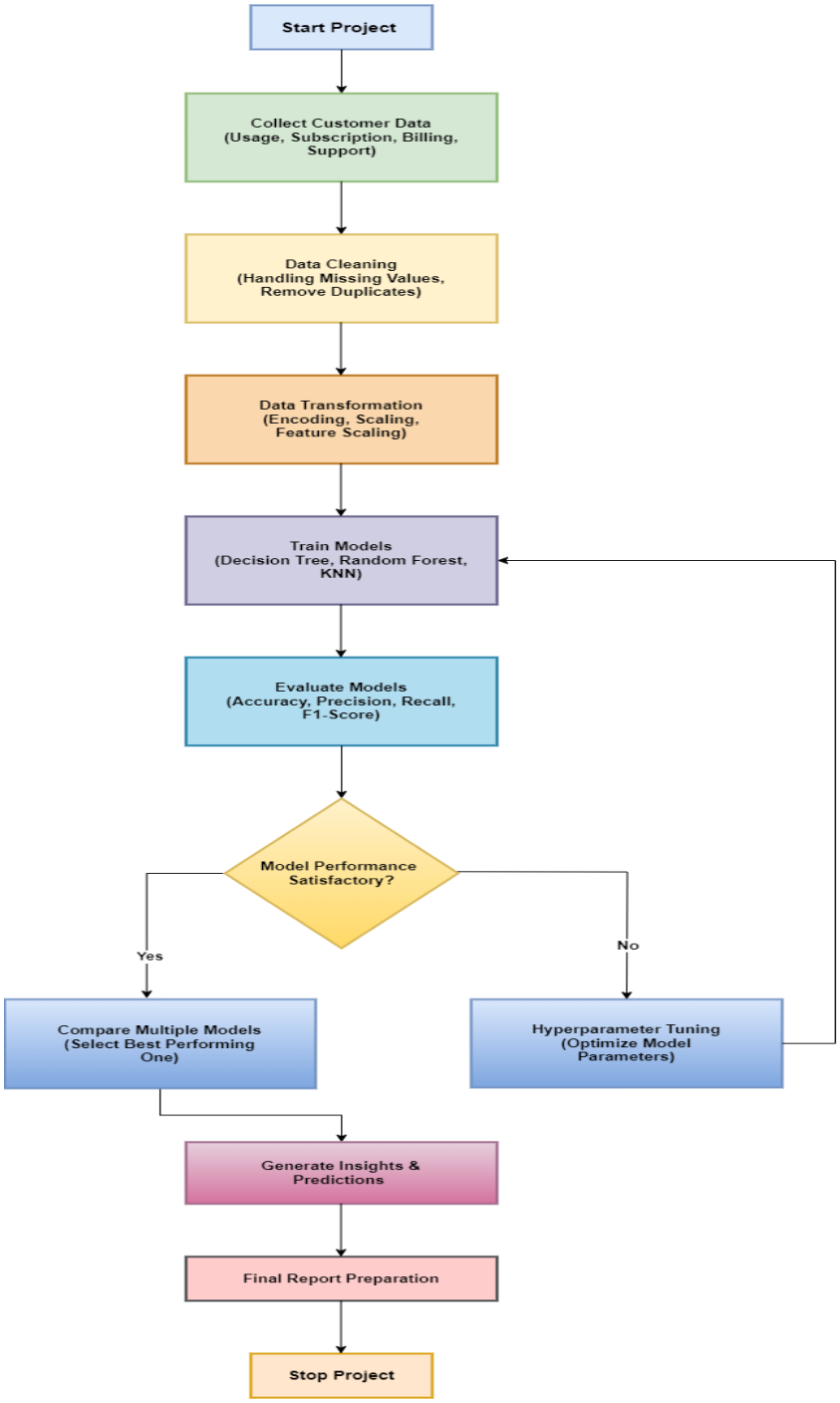

Fig. 3.4: Flow Chart Diagram for Telecom Customer Churn Prediction Project

The flow chart diagram outlines the sequential flow of operations for the customer churn prediction system:

(a) **Start Project:** Initiates the process with goal definition.

(b) **Collect Customer Data:** Data sources include usage history, subscription plans, billing records, and customer support logs.

(c) **Data Cleaning:** Techniques involve imputation of missing values, removal of duplicates, and outlier detection.

(d) **Data Transformation:** Covers one-hot encoding, label encoding, normalization, and feature scaling.

(e) **Train Models:** Includes Decision Tree, Random Forest, KNN, Logistic Regression, SVM, and XGBoost.

(f) **Evaluate Models:** Models are evaluated using Accuracy, Precision, Recall, and F1-score.

(g) **Model Performance Satisfactory?**

Decision node:

- **Yes:** Go to model comparison.
- **No:** Proceed to hyperparameter tuning.

(h) **Hyperparameter Tuning:** Optimize model parameters to improve performance.

(i) **Compare Multiple Models:** Select top 4 models based on evaluation results and combine them using Voting Classifier.

(j) **Generate Insights & Predictions:** Generate business insights and churn prediction output.

(k) **Final Report Preparation:** Compile final results, visualizations, and business recommendations.

(l) **Stop Project:** End of the project workflow.

2. **Block Diagram**

The Block Diagram outlines the major components of the Telecom Customer Churn Prediction pipeline. It helps to visualize the key stages involved, from data collection to real-time churn prediction.

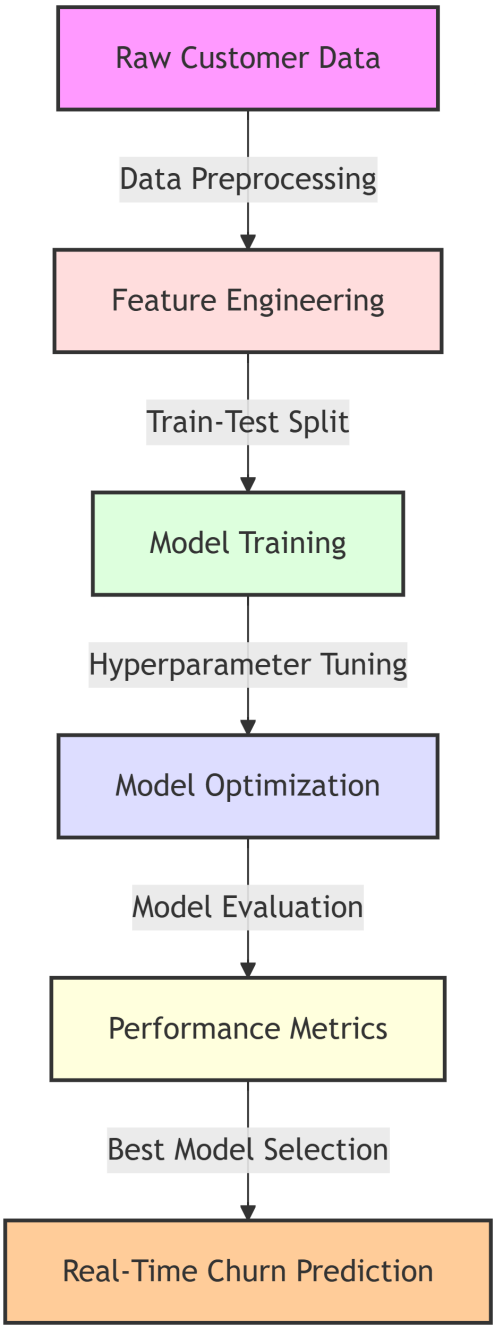

Fig. 3.5: Block Diagram for Telecom Customer Churn Prediction System

The block diagram represents the following steps:

(a) **Raw Customer Data**

- **Input:** Collected from the telecom company (e.g., demographics, usage, service features).

- **Format:** CSV, Excel, or database.

(b) **Data Preprocessing**

- Handling missing values, encoding categorical variables, removing duplicates, and scaling numeric features.

(c) **Feature Engineering**

- Creating new relevant features, selecting or transforming existing ones to improve model accuracy.

(d) **Train-Test Split**

- Splitting the dataset into training and testing sets (e.g., 80/20) for model evaluation.

(e) **Model Training**

- Training various machine learning algorithms like Decision Tree Classifier, Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, SVM, and XGBoost.

(f) **Hyperparameter Tuning**

- Finding the optimal combination of parameters for each model using techniques like GridSearchCV.

(g) **Model Optimization**

- Improving model performance, using techniques like pipelines and ensemble methods.

(h) **Model Evaluation**

- Accuracy, Precision, Recall, F1-score for comparing the performance of each model.

(i) **Performance Metrics**

- Compare and select the best model based on evaluation metrics.

(j) **Best Model Selection**

- Selecting the most accurate model for deployment, combined with other top models using a Voting Classifier for improved prediction accuracy.

(k) **Real-Time Churn Prediction**

- Deploying the final model in a real-time system (via Streamlit) to make predictions based on incoming data.

3. **Acitvity Diagram**

The Activity Diagram represents the flow of control within the churn prediction system, illustrating the various steps involved in processing customer data and generating predictions.

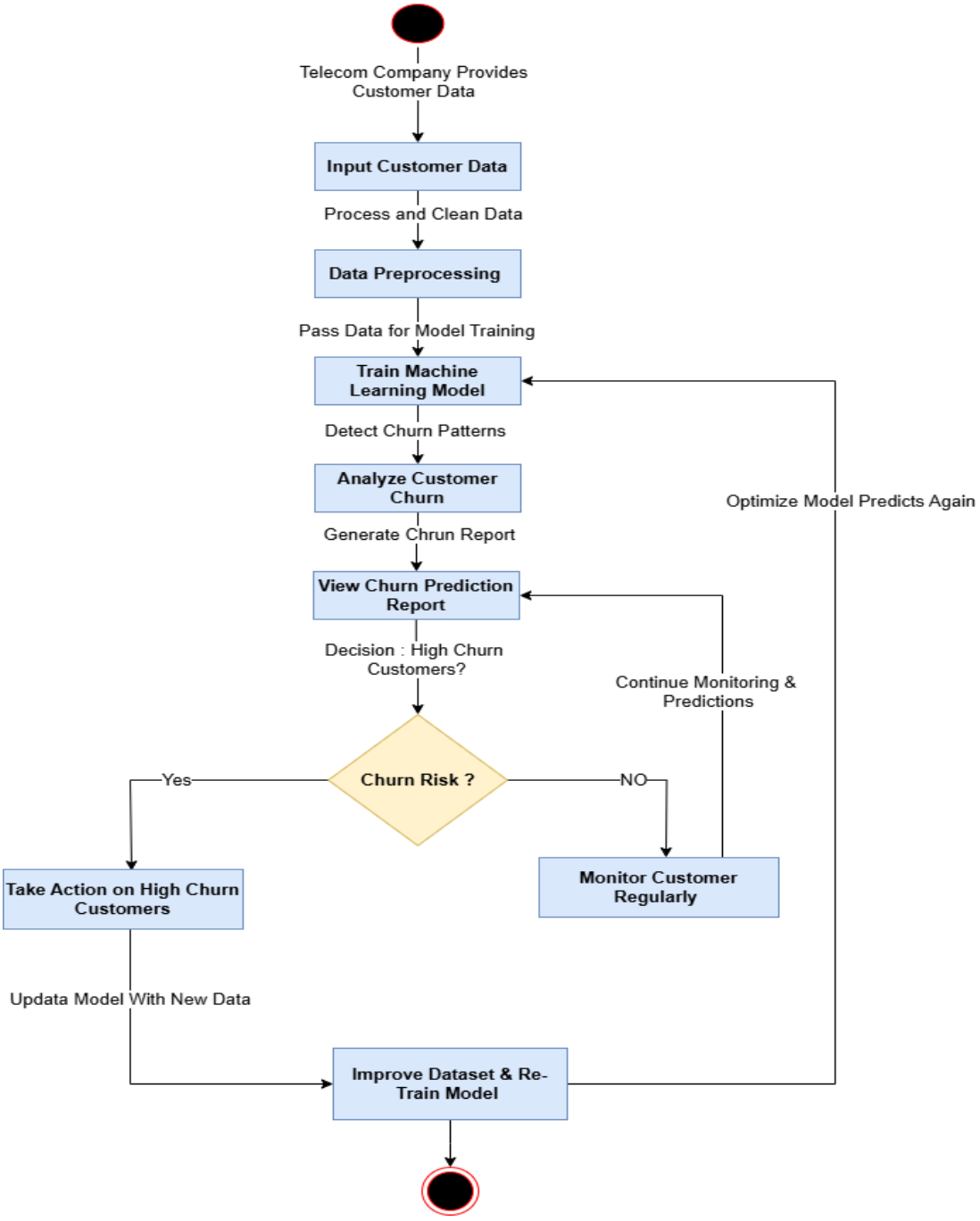

Fig. 3.6: Activity Diagram for Telecom Customer Churn Prediction System

4. RESULTS AND EXPLANATION

4.1. Implementation Approaches

The customer churn prediction system was implemented using a modular and scalable architecture. The entire pipeline is built with Python and includes multiple components such as preprocessing, model training, evaluation, and deployment. Key libraries used include:

- **Pandas** and **NumPy** for data manipulation and analysis.
- **Scikit-learn** for preprocessing, model training, evaluation metrics, and hyperparameter tuning.
- **XGBoost** for gradient boosting-based classification.
- **Imbalanced-learn** for SMOTE (Synthetic Minority Over-sampling Technique).
- **Matplotlib** and **Seaborn** for data visualization.
- **Streamlit** for deploying the prediction interface and real-time prediction visualization.

The overall process was executed in the following stages:

1. **Data Collection:** Dataset was collected from a public telecom churn dataset.
2. **Data Preprocessing:** Missing values were handled, categorical variables encoded, and numeric features scaled.
3. **Balancing:** Applied SMOTE to address class imbalance.
4. **Model Training:** Trained six models—Decision Tree, Random Forest, KNN, Logistic Regression, SVM, and XGBoost.
5. **Hyperparameter Tuning:** Performed using GridSearchCV for better performance.
6. **Ensemble Learning:** Final model built using a soft voting ensemble of top-performing models.
7. **Evaluation:** Compared models using Accuracy, Precision, Recall, and F1-score.
8. **Deployment:** Integrated Streamlit for real-time predictions with CSV upload and result logging features.

This modular approach ensured each component could be tested and optimized independently, leading to a more robust and interpretable prediction system.

4.2. Testing

Testing was conducted to evaluate the effectiveness and generalizability of the developed churn prediction models. The dataset was divided into an 80:20 ratio, where 80% of the data was used for training and 20% was reserved for testing.

The following testing strategies were applied:

- **Hold-out Validation:** The 20% test set was used as unseen data to assess how well the model performs on new inputs.
- **Performance Metrics:** Accuracy, Precision, Recall, and F1-score were calculated to measure the predictive power and robustness of each model.
- **Confusion Matrix Analysis:** Confusion matrices were generated to analyze true positives, true negatives, false positives, and false negatives.
- **Cross-Validation:** GridSearchCV was used with cross-validation to tune hyperparameters and avoid overfitting.

Each model was tested on the test dataset after training. The ensemble model (Voting Classifier) demonstrated the best performance by combining predictions from the top models, thus improving the overall prediction accuracy and stability.

Additionally, real-time testing was performed using the deployed Streamlit application. Users could upload new data (CSV format) or use the input form to get predictions instantly. The system returned the following:

- **Churn Status:** Whether the customer is likely to churn (Yes/No).
- **Churn Probability:** A numeric value indicating the likelihood of churn.
- **Risk Level:** Categorized into Low, Moderate, or High based on the probability score.

This comprehensive testing approach ensured that the churn prediction system was accurate, reliable, and deployable in real-world business environments.

4.3. Analysis (Graphs/Charts)

The following section presents various visualizations to evaluate the performance and behavior of different machine learning models used in the customer churn prediction system.

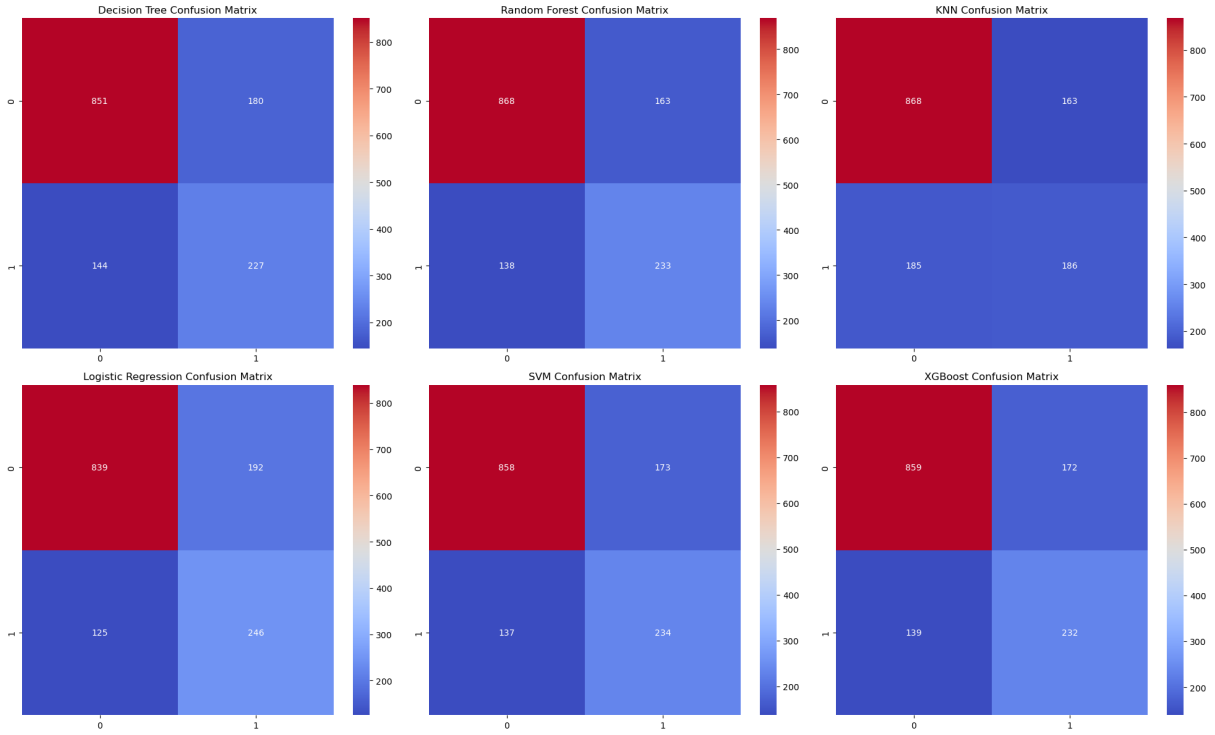

Figure 4.1: Confusion Matrix of all models.

The confusion matrices show that Random Forest and Logistic Regression strike a good balance between correctly predicting churners and non-churners. KNN struggles with false negatives, while ensemble models like XGBoost perform consistently well. Overall, Random Forest and Logistic Regression emerge as strong candidates.

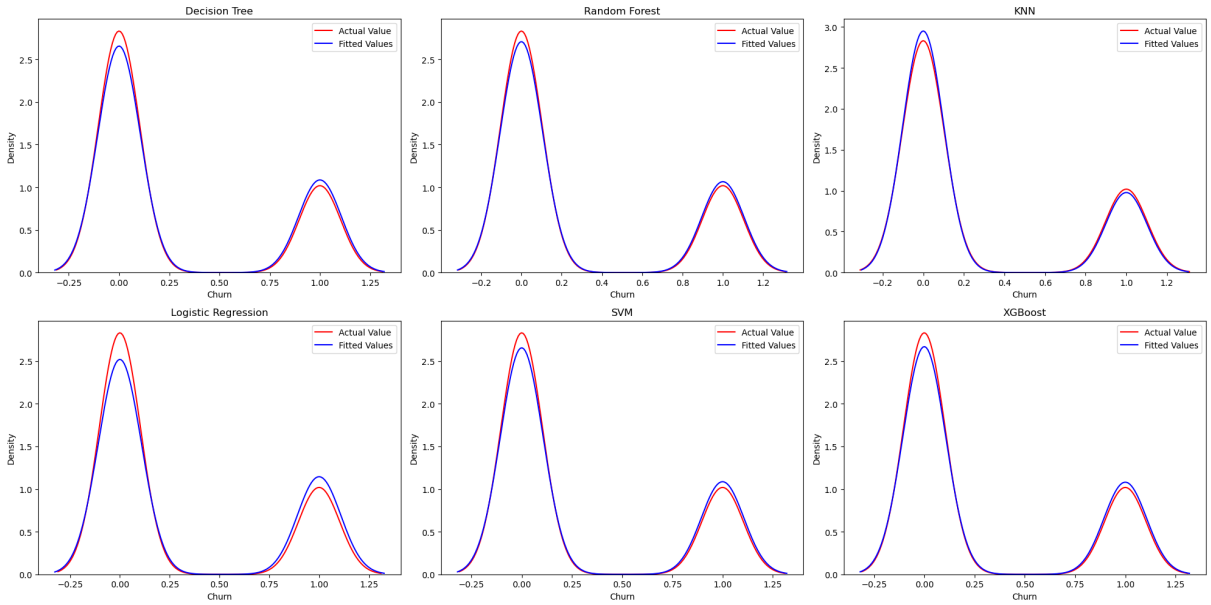

Figure 4.2: Distribution plot of predicted vs actual values for all models.

These plots highlight how closely each model's predictions align with actual values. Random Forest, SVM, and XGBoost show excellent fit, while KNN deviates the most. Ensemble models provide better generalization and stability.

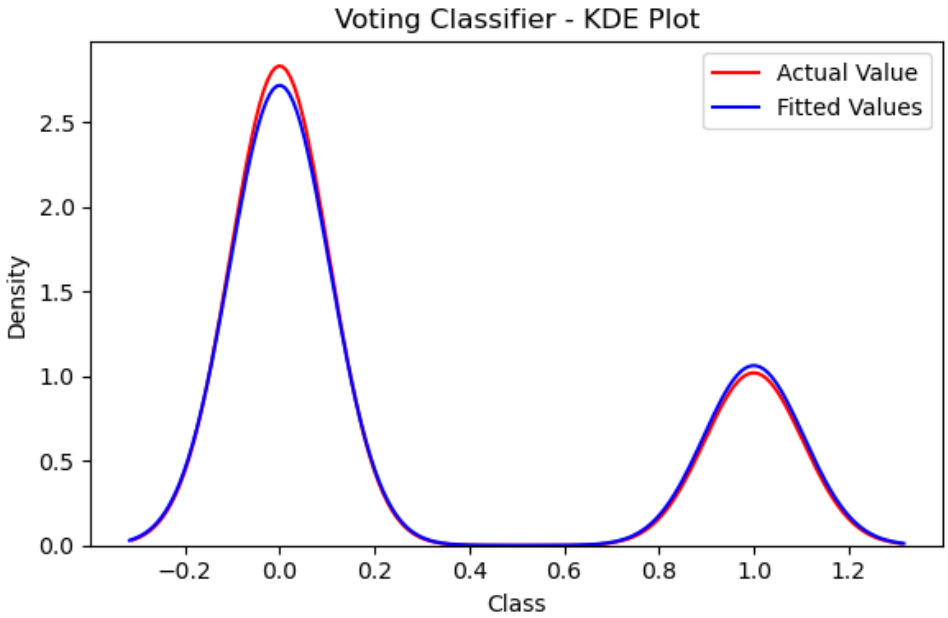

Figure 4.3: KDE plot for Voting Classifier predictions.

The Voting Classifier shows excellent overlap between predicted and actual distributions. Its strong alignment indicates reliable and robust performance, supporting its suitability for deployment.

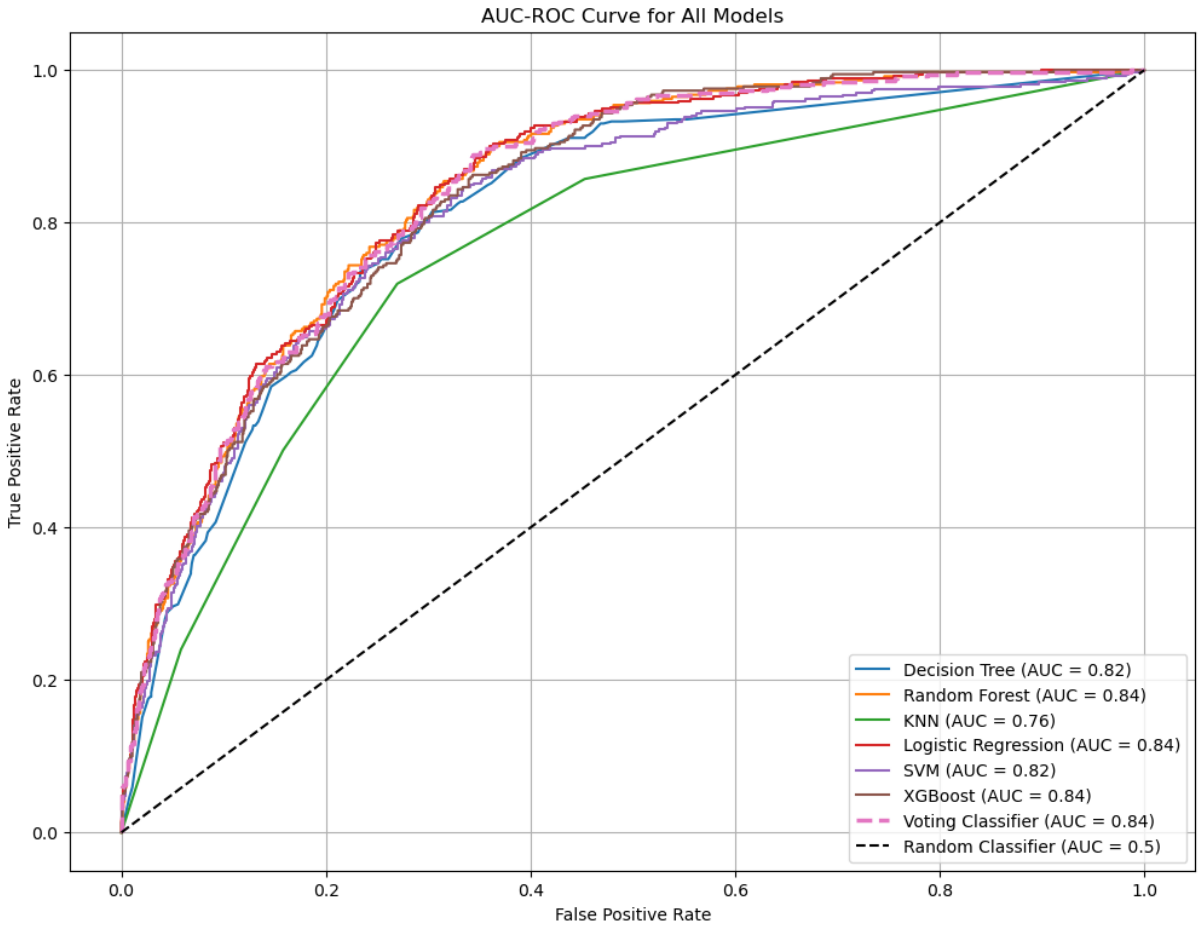

Figure 4.4: AUC-ROC comparison of all models.

Voting Classifier, Random Forest, Logistic Regression, and XGBoost all achieved an AUC of 0.84, reflecting strong classification ability. KNN had the weakest performance with an AUC of 0.76. Ensemble models lead in distinguishing between churners and non-churners.

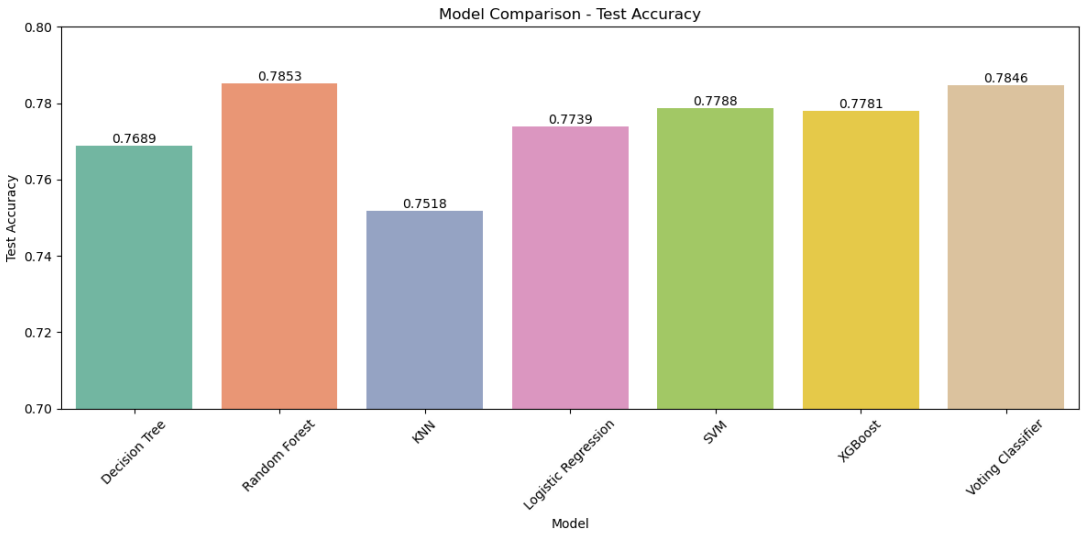

Figure 4.5: Comparison of test accuracy across models.

Random Forest and Voting Classifier achieved the highest accuracy ($\approx 78.5\%$), followed closely by XGBoost, SVM, and Logistic Regression. KNN showed the lowest accuracy, confirming its weaker performance on this dataset.

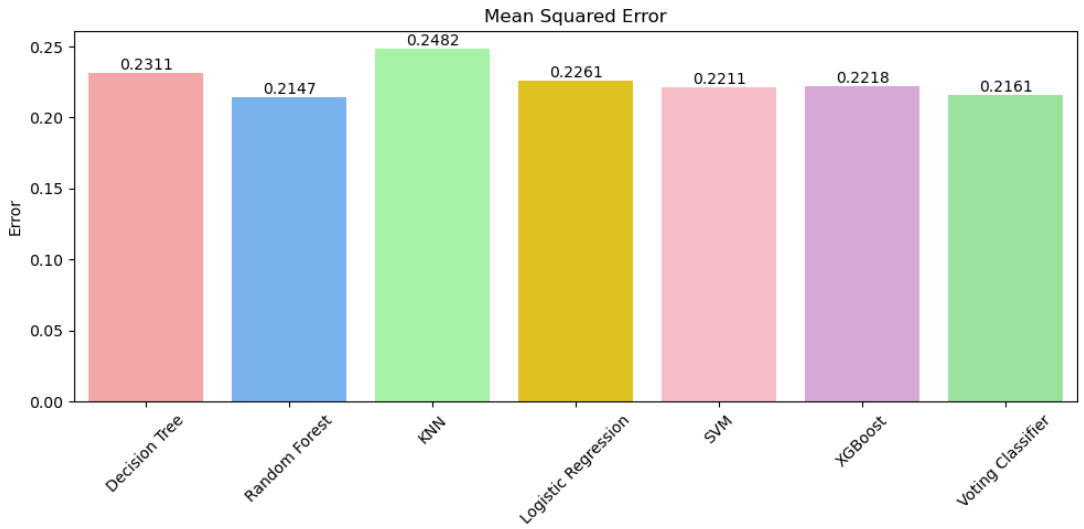

Figure 4.6: Mean Squared Error (MSE) for each model.

Random Forest and Voting Classifier had the lowest MSE, indicating better predictive performance. KNN had the highest MSE, further confirming its unsuitability for this task. Again, the Voting Classifier proves to be a reliable and balanced choice.

5. CONCLUSION & FUTURE SCOPE

Conclusion

This project successfully demonstrated the use of various machine learning techniques to predict customer churn in the telecom industry. A comprehensive dataset was used to analyze customer behavior and service usage patterns. The data underwent preprocessing steps including handling missing values, encoding categorical variables, and balancing using SMOTE to enhance model performance.

Multiple machine learning models such as Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), and XGBoost were implemented and evaluated. Among them, the ensemble model (Voting Classifier) delivered the highest accuracy and robust performance across all key metrics such as Accuracy, Precision, Recall, and F1-score. The system was deployed using Streamlit, enabling real-time churn predictions with an interactive user interface.

Through this approach, telecom companies can identify potential churners, understand the reasons for churn, and implement targeted strategies to retain valuable customers.

Future Scope

Although the current system provides a strong baseline for churn prediction, there are several areas for future improvements:

- **Deep Learning Models:** Incorporating advanced neural networks for sequential or unstructured data (e.g., call logs or text feedback).
- **Time-Series Analysis:** Integrating temporal data (e.g., tenure trends, monthly billing behavior) for improved forecasting.
- **Automation:** Automating the model retraining pipeline with real-time data ingestion and updates.
- **Integration:** Embedding the churn prediction engine into CRM systems to trigger alerts or promotional actions automatically.

By incorporating these improvements, the system can be enhanced to serve as a comprehensive churn management solution tailored to real-world business environments.

References

[1] S. Barham, N. Aweisi, and A. Khalifeh, "A review on machine learning-based customer churn prediction in the telecom industry," in *2023 9th International Conference on Control, Decision and Information Technologies (CoDIT)*, 2023, pp. 2659–2664.

[2] K. D. Singh, P. Deep Singh, A. Bansal, G. Kaur, V. Khullar, and V. Tripathi, "Exploratory data analysis and customer churn prediction for the telecommunication industry," in *2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, 2023, pp. 197–201.

[3] M. Imani, Z. Ghaderpour, M. Joudaki, and A. Beikmohammadi, "The impact of smote and adasyn on random forest and advanced gradient boosting techniques in telecom customer churn prediction," in *2024 10th International Conference on Web Research (ICWR)*. IEEE, 2024, pp. 202–209.

[4] S. K. Wagh, A. A. Andhale, K. S. Wagh, J. R. Pansare, S. P. Ambadekar, and S. Gawande, "Customer churn prediction in telecom sector using machine learning techniques," *Results in Control and Optimization*, vol. 14, p. 100342, 2024.

[5] A. Patel and A. G. Kumar, "Predicting customer churn in telecom industry: A machine learning approach for improving customer retention," in *2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC)*. IEEE, 2023, pp. 558–561.

[6] G. Verma, "Telecom customer churn prediction using enhanced machine learning classification techniques," in *2024 First International Conference on Innovations in Communications, Electrical and Computer Engineering (ICICEC)*, 2024, pp. 1–5.

[7] A. H M, B. T, S. Tanisha, S. B, and C. C. Shanuja, "Customer churn prediction using synthetic minority oversampling technique," in *2023 4th International Conference on Communication, Computing and Industry 6.0 (C216)*, 2023, pp. 01–05.

[8] V. Kavitha, G. H. Kumar, S. M. Kumar, and M. Harish, "Churn prediction of customer in telecom industry using machine learning algorithms," *International Journal of Engineering Research & Technology (2278-0181)*, vol. 9, no. 05, pp. 181–184, 2020.

[9] L. Ou, "Customer churn prediction based on interpretable machine learning algorithms in telecom industry," in *2023 International Conference on Computer Simulation and Modeling, Information Security (CSMIS)*, 2023, pp. 644–647.

[10] P. Bhuse, A. Gandhi, P. Meswani, R. Muni, and N. Katre, "Machine learning based telecom-customer churn prediction," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, 2020, pp. 1297–1301.