D Y PATIL
INTERNATIONAL
UNIVERSITY
AKURDI PUNE

Project Synopsis

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

Dr. Jagadish S Jakati

Project Guide

Dr. Swapnil Waghmare
Project Coordinator

Dr. Maheshwari Biradar
HOD, BCA & MCA

School of Computer Science, Engineering and Applications

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

[Session 2024-2025]

TABLE OF CONTENTS

1. INTRODUCTION

1.1. Background

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.

1.2. Objectives

- To develop an automated system for predicting customer churn in the telecom industry using **Machine Learning (ML)** techniques.
- To collect and preprocess telecom customer data by handling missing values, encoding categorical variables, and scaling numerical features for better model performance.
- To implement and compare multiple ML models such as Decision Tree, Random Forest, Gradient Boosting, and K-Nearest Neighbors (KNN) to identify the best-performing model for churn prediction.
- To apply feature engineering techniques such as feature selection and transformation to enhance the dataset quality and model performance.
- To evaluate the performance of the models using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to ensure reliable predictions.
- To fine-tune the hyperparameters of the selected model using techniques like Grid Search and Random Search to optimize efficiency and accuracy.

- To develop a user-friendly tool or dashboard that allows telecom companies to input customer data and receive real-time churn predictions.
- To help telecom companies reduce customer churn by providing actionable insights and improving customer retention strategies through data-driven decision-making.

1.3. Purpose

The purpose of this project is to develop an efficient and accurate machine learning-based system for predicting customer churn in the telecom industry. Customer churn leads to significant revenue losses, and predicting it in advance enables telecom companies to take preventive measures, improve customer retention strategies, and enhance overall business performance.

This project aims to analyze customer behavior using historical data, identifying key factors that contribute to churn. By leveraging machine learning algorithms, the model will provide actionable insights to help telecom companies implement data-driven strategies to reduce churn and improve customer satisfaction.

The main goals of this project include:

- To build a predictive model that accurately classifies customers into churn and non-churn categories.
- To provide insights into key customer behaviors and attributes influencing churn.
- To assist telecom companies in taking proactive retention measures by identifying at-risk customers.
- To develop an automated system that helps decision-makers optimize customer engagement and reduce customer turnover.
- To enhance business profitability by minimizing churn and improving long-term customer relationships.

By integrating various machine learning models such as Decision Tree, Random Forest, Gradient Boosting, and K-Nearest Neighbors (KNN), this system will serve as a valuable tool for telecom operators to predict churn and develop effective customer retention strategies.

1.4. 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 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%). The study concludes that ADASYN slightly

outperforms SMOTE in handling imbalanced data.
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. To handle class imbalance, they used the Synthetic Minority Oversampling Technique (SMOTE). Feature selection was performed using Pearson correlation, and the model achieved an impressive 99% accuracy with Random Forest. The study also incorporated survival analysis and Cox Proportional Hazard Model to predict customer retention strategies.
Drawback: The extremely high accuracy (99%) raises concerns about potential overfitting, especially since the dataset was imbalanced. Additionally, the use of SMOTE may introduce synthetic samples that do not accurately represent real-world data, leading to distorted feature relationships. The study also lacks a detailed discussion on the computational complexity of the models, particularly for large-scale datasets.

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 machine learning techniques to predict customer churn in the telecom industry. The dataset consists of 2,666 rows and 20 features. After preprocessing and applying SMOTE for class balancing, multiple models were tested, including Bernoulli Naive Bayes, Gaussian Naive Bayes, SVM, KNN, Decision Tree, Random Forest, and XGBoost. The XGBoost classifier achieved the highest accuracy of 94%. The study highlights that international plans and customer service calls strongly influence churn.
Drawback: The study does not analyze potential overfitting issues with XGBoost, given the dataset size. Computational complexity and scalability for large datasets are not discussed.

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 for customer churn prediction using a Kaggle dataset. The models were trained and tuned, achieving 79% accuracy for Decision Tree and KNN, while Random Forest performed best at 82% accuracy. Data preprocessing and feature selection were emphasized, showing that shorter tenures and higher monthly charges increase churn risk.

Drawback: The study does not address class imbalance explicitly, which may affect model performance. It also lacks an interpretability analysis, which is crucial for business decision-making.

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: This study applies SMOTE to handle class imbalance in telecom churn prediction. It compares Gradient Boosting, Random Forest, Decision Tree, and Logistic Regression, with Gradient Boosting achieving 95.13% accuracy, followed by Random Forest and Decision Tree. Factor analysis highlights tenure and monthly charges as key churn indicators, emphasizing the importance of feature selection in model performance.

Drawback: While SMOTE improves accuracy, it may introduce synthetic data that distorts real-world patterns, leading to overfitting. The study does not explore the computational cost of SMOTE, which can be significant for large datasets. Additionally, it lacks interpretability analysis for Gradient Boosting, which is crucial for business decision-making.

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: The study compared the performance of Random Forest, XGBoost, and Logistic Regression for predicting customer churn in the telecom industry. The authors used a dataset from Kaggle and performed data preprocessing, including data filtering, feature selection, and data normalization. The Random Forest model achieved the highest accuracy of 93%, followed by XGBoost and Logistic Regression. The study also visualized the results using Confusion Matrices and discussed the importance of customer retention strategies.

Drawback: The Logistic Regression model underperformed, achieving only 73% accuracy, which indicates that it may not be suitable for imbalanced datasets. The study also did not address the interpretability of the models, which is crucial for understanding the factors driving customer churn. Additionally, the authors did not explore the impact of hyperparameter tuning on model performance.

Reference 9

Title: Customer Churn Prediction Based on Interpretable Machine Learning Algorithms in Telecom Industry
Authors: Liwen Ou
Year: 2022
Methodology: This study focuses on model interpretability for churn prediction using Random Forest, Decision Tree, and Extra Tree Classifier on the IBM Telco dataset. Feature importance analysis identified tenure, total charges, and monthly charges as key factors. Random Forest and Extra Tree Classifier achieved 82.3% and 82.5% accuracy, while Decision Tree performed slightly worse at 77.5%.
Drawback: The Decision Tree model's lower accuracy suggests it may struggle with high-dimensional datasets. The study does not address class imbalance, which is crucial for accurate churn prediction. Additionally, it lacks discussion on the computational complexity of ensemble models, which could limit large-scale deployment.

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, Support Vector Machines (SVM), XGBoost, Ridge Classifier, and Deep Neural Networks for customer churn prediction in the telecom industry. A grid search was used for hyperparameter tuning. The Random Forest model achieved the highest accuracy (90.96%) before tuning. Feature selection was performed to refine the dataset, and various classification techniques were evaluated to balance efficiency and accuracy.
Drawback: While the study explores multiple models, deep learning techniques were not optimized extensively, and no discussion on class imbalance handling methods like SMOTE was included. Additionally, computational complexity for larger datasets was not addressed.

2. GAP IDENTIFICATION

Despite advancements in machine learning and predictive analytics, telecom customer churn prediction still faces several challenges. Existing churn prediction models often lack efficiency in real-time deployment and struggle with various limitations that impact their effectiveness. The major gaps identified in this domain include:

- **Data Imbalance:** Churn cases are significantly fewer than non-churn cases, leading to biased model performance. Techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or class weighting can be used to address this issue.
- **Feature Selection Complexity:** Identifying the most relevant factors affecting churn remains difficult due to multiple influencing variables. Automated feature selection techniques and domain expertise are essential for improving model accuracy.
- **Lack of Real-Time Prediction:** Many existing models operate on batch processing, delaying actionable insights. Deploying ML models as APIs or using real-time streaming services like Apache Kafka can help achieve real-time predictions.
- **Interpretability Issues:** Some models, such as ensemble learning, act as black-box algorithms, making it difficult to understand their predictions. Techniques like SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) can enhance model interpretability.
- **Customer Behavior Changes:** Dynamic customer preferences make static models less effective over time. Regular model retraining with fresh data ensures adaptability and improved predictive performance.

Addressing these challenges will help in developing a more robust and efficient churn prediction system. By improving feature selection, handling data imbalance, incorporating real-time prediction capabilities, and enhancing model interpretability, this project aims to create an enhanced model that provides better accuracy and actionable insights for telecom companies.

3. METHODOLOGY

3.1. Methodology

This project follows a structured machine learning pipeline to predict customer churn in the telecom industry. The key steps are:

- **Data Preprocessing:** Handling missing values, encoding categorical variables, and scaling numerical features for better model performance.
- **Feature Engineering:** Selecting relevant features such as customer demographics, service usage, contract details, and billing information to enhance model accuracy.
- **Machine Learning Models Used:**
  - **Decision Tree Classifier:** A rule-based model that splits customer data into different classes to predict churn.
  - **Random Forest Classifier:** An ensemble of decision trees that improves accuracy and reduces overfitting.
  - **K-Nearest Neighbors (KNN) Classifier:** A distance-based algorithm that classifies customers based on similarity to other customers.
- **Model Evaluation:** Assessing model performance using metrics such as accuracy, precision, recall, and F1-score.
- **Hyperparameter Tuning:** Optimizing model parameters using techniques like Grid Search to improve predictive accuracy.

By following this methodology, the project aims to develop an effective churn prediction system that helps telecom companies identify and retain at-risk customers.

3.2. Block Diagram

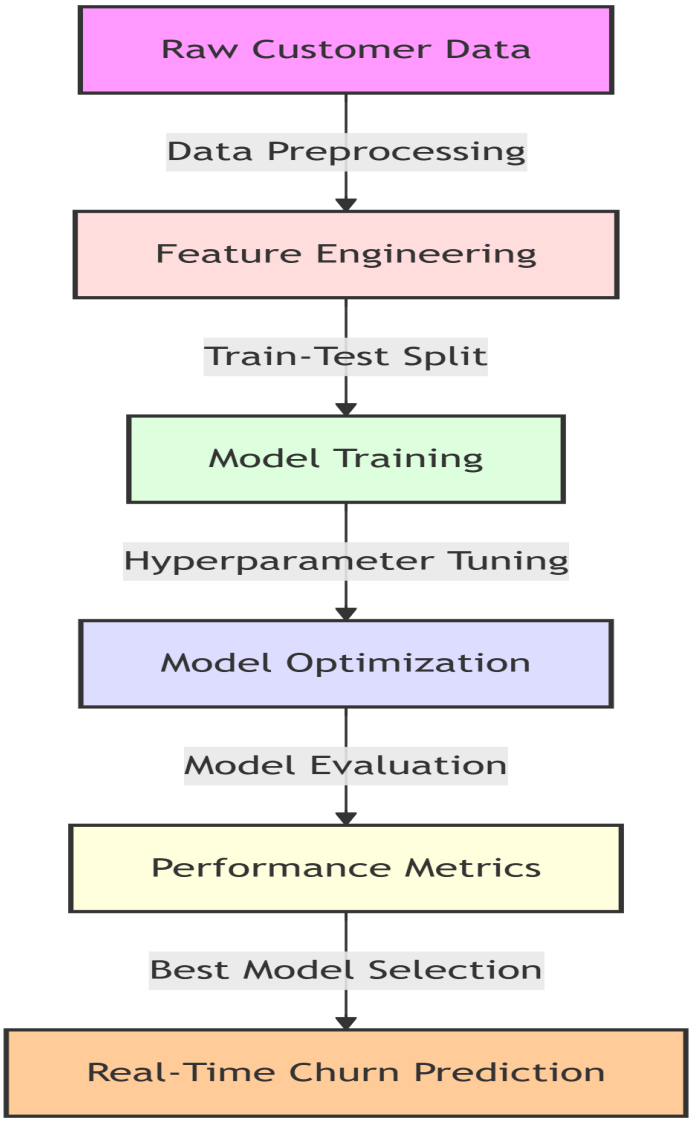

Figure 3.1: Block Diagram of Telecom Customer Churn Prediction

4. PROJECT FLOW DIAGRAMS

4.1. Gantt Chart

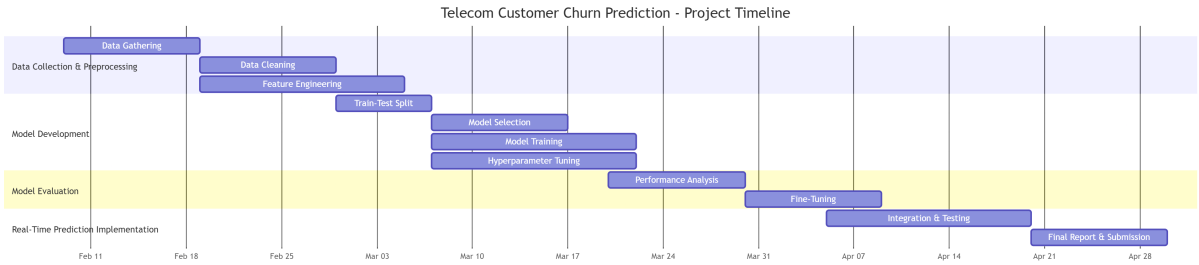

Figure 4.1: Gantt Chart for Telecom Customer Churn Prediction Project

4.2. Flowchart

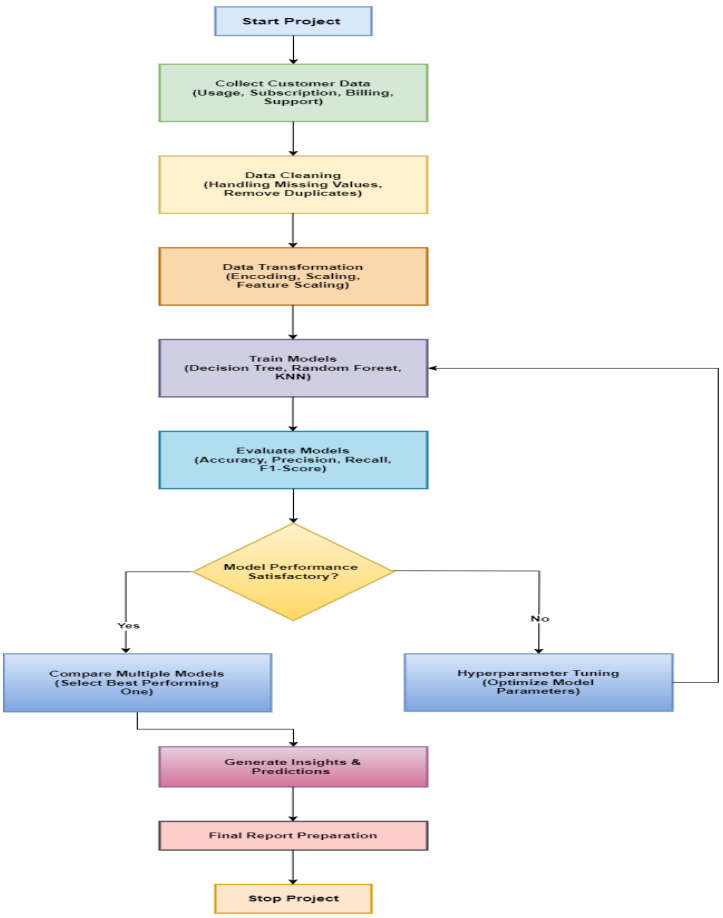

Figure 4.2: Flowchart for Telecom Customer Churn Prediction

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.