Bangladesh Army University of Science and Technology (BAUST), Saidpur

Department of Computer Science and Engineering (CSE)

PROJECT REPORT

Customer Churn Prediction Using Machine Learning

Course Code: CSE 4140

Course Title: Machine Learning Sessional

Submitted By:

Sumiya Naznin Haque (220201020)

Md. Abdullah (220201034)

Submitted To:

Engr. Rohul Amin
Lecturer, Department of CSE

Nadim Reza
Lecturer, Department of CSE

Submission Date: 25-11-2025

Customer Churn Prediction Using Machine Learning

Sumiya Naznin Haque, Md. Abdullah

***Abstract*—Customer churn prediction has become a critical task for telecom service providers aiming to retain subscribers and reduce operational losses. This study presents a machine learning–based framework for predicting customer churn using a publicly available telecom dataset. The dataset was preprocessed through data cleaning, feature encoding, and normalization to ensure high-quality model input. Three supervised learning algorithms—Logistic Regression, K-Nearest Neighbors (KNN), and Decision Tree—were trained and evaluated to identify the most effective model for churn detection. Model performance was assessed using accuracy, precision, recall, F1-score, ROC–AUC, and confusion matrix analysis. Experimental results indicate that all three models demonstrate reasonable predictive capability, with the best performance achieved by the model obtaining the highest AUC score. The findings highlight the potential of machine learning techniques in supporting telecom operators with early churn identification and data-driven customer retention strategies.**

***Index Terms*—Customer Churn, Machine Learning, Logistic Regression, KNN, Decision Tree, ROC-AUC, Confusion Matrix.**

I. Introduction

Customer churn refers to the phenomenon where existing customers discontinue a service or switch to a competing provider. Customer Churn Prediction (CCP) aims to identify such customers in advance by analyzing behavioral, demographic, and service-usage patterns using machine learning models. Early identification enables organizations—especially telecom operators—to implement targeted retention strategies and minimize revenue loss. Customer churn has become a major challenge for the global telecom industry. According to a recent study published in the Journal of Big Data, more than 20–40% of telecom customers exhibit churn tendencies annually, depending on the region and service quality 1. Another large-scale analysis by the International Journal of Computer Applications reported that telecom companies worldwide lose billions of dollars each year due to customer churn and the high cost of acquiring new customers relative to retaining existing ones 2. For example, a report by Ahmad et al. (2019) found that approximately 26% of users in their telecom dataset churned within a single year, demonstrating the severity of the issue 1. These statistics highlight the necessity of churn prediction systems that can proactively identify at-risk customers. Machine learning techniques have been widely applied to churn prediction due to their ability to learn complex patterns from large-scale customer datasets. By analyzing historical records such as service usage, billing information, customer support interactions, and contract details, predictive models can classify customers into churners and non-churners with high accuracy. Such insights enable telecom companies to improve customer satisfaction, reduce switching behavior, and maintain long-term profitability.

A. Contribution

Sumiya Naznin Haque

- Dataset Collection
- Exploratory Data Analysis (EDA)
- Data Preprocessing

Md. Abdullah

- Feature Engineering
- Model Training
- Model Evaluation

II. Literature Review

Several previous studies have investigated customer churn prediction using different machine learning approaches. Ahmad et al. [1] utilized a real telecom operator dataset and proposed a big-data–based machine learning framework for churn prediction. Their study employed Logistic Regression, Random Forest, and Gradient Boosting using the Apache Spark platform, achieving an AUC of approximately 93%. Although their model demonstrated strong scalability, the work lacked deeper analysis on model interpretability and feature importance, which are essential for business decision-making. Similarly, Vafeiadis et al. [2] conducted a comparative study using various machine learning techniques—including Support Vector Machines (SVM), Artificial Neural Networks (ANN), Decision Trees, and Logistic Regression—for predicting customer churn. Their findings showed that SVM achieved the highest accuracy, whereas ANN required longer training time and Decision Trees suffered from overfitting. However, the study mainly relied on accuracy as the primary evaluation metric, without using more comprehensive metrics such as ROC, AUC, Precision, Recall, and F1-Score.

Despite the strengths of these studies, several limitations remain, including:

- Limited set of evaluation metrics

- Minimal emphasis on feature engineering
- Lack of comprehensive comparison among simple baseline models
- Limited interpretability and analysis of influencing factors
- Insufficient multi-metric performance evaluation

In this work, we address these limitations by applying Logistic Regression, K-Nearest Neighbors (KNN), and Decision Tree models on the Telco Customer Churn dataset with structured feature engineering. We further conduct an extensive performance evaluation using Confusion Matrix, Precision, Recall, F1-Score, ROC Curve, and AUC, providing a more comprehensive comparative analysis than the prior studies.

III. Methodology

Our pipeline follows these steps:

1) **Data Collection and EDA:** Inspect distributions, missing values, and correlations to guide preprocessing.
2) **Feature Engineering:** Create derived features, encode categorical variables (one-hot/label), and select informative attributes.
3) **Preprocessing:** Handle missing values, encode categorical features, and normalize numerical features.
4) **Model Training:** Train Logistic Regression, KNN, and Decision Tree on stratified 80/20 train/test split.
5) **Evaluation:** Compute accuracy, precision, recall, F1-score, confusion matrices, ROC and AUC values.

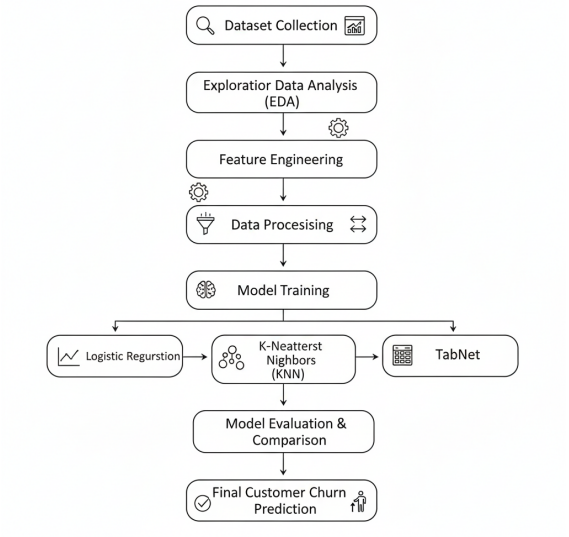

Fig. 1: Methodology flowchart.

Our project followed a structured machine-learning pipeline for Customer Churn Prediction. First, the dataset was collected and explored through Exploratory Data Analysis (EDA) to understand patterns, distributions, and missing values. Next, feature engineering was performed to convert categorical variables, handle missing data, and prepare the dataset for modeling. The processed data was split into 80 percent training and 20 percent testing. Multiple machine-learning models were then trained on the training set.

IV. Results and Analysis

A. Logistic Regression

The Logistic Regression model achieved a 79% test accuracy. For the non-churn class (0), the model performed well with 0.83 precision and 0.90 recall, meaning it correctly identified most non-churn customers. For the churn class (1), performance was moderate, with 0.64 precision and 0.49 recall, showing that the model missed many actual churn cases.

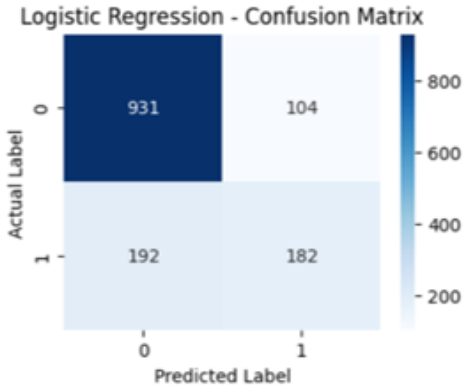

Fig. 2: Confusion Matrix for Logistic Regression.

The confusion matrix shows:

- 931 non-churn customers were correctly predicted.
- 182 churn customers were correctly predicted.
- A large number (192) of churn customers were misclassified as non-churn.

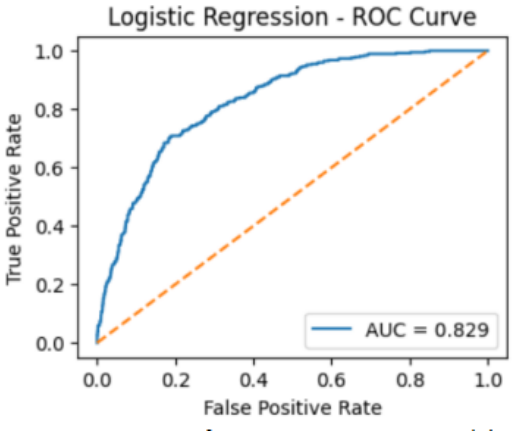

Fig. 3: ROC Curve for Logistic Regression.

Overall, Logistic Regression performs well for the majority class but struggles to detect churn due to class imbalance.

B. K-Nearest Neighbors

The K-Nearest Neighbors (KNN) model achieved a 78% test accuracy, which is slightly lower than Logistic Regression. For the non-churn class (0), the model performed well, with 0.83 precision and 0.87 recall, meaning it correctly identified most non-churn customers. For the churn class (1), the performance was moderate, with 0.59 precision and 0.52 recall. This shows that the model still struggled to detect churn customers, although slightly better recall compared to Logistic Regression. Overall, KNN provides good performance for the majority class but continues to face difficulty in identifying churn cases due to class imbalance.

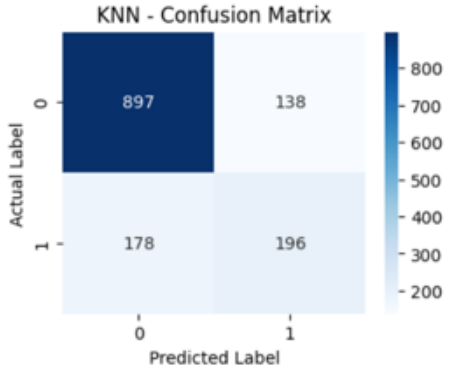

Fig. 4: Confusion Matrix for KNN.

The confusion matrix shows:

- 897 non-churn customers were correctly predicted.
- 196 churn customers were correctly predicted.
- A large number (178) of churn customers were misclassified as non-churn.

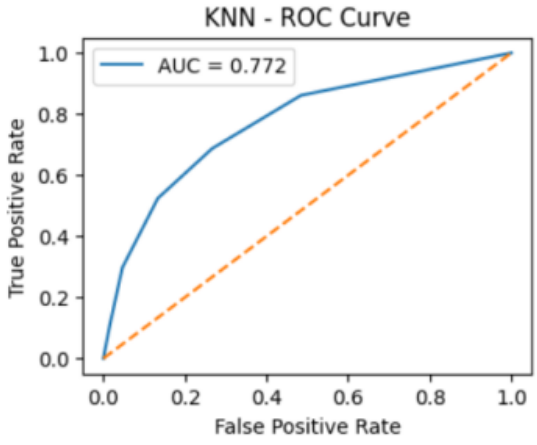

Fig. 5: ROC Curve for KNN.

C. Tebnet

The Decision Tree (Tebnet) model achieved a 72% test accuracy, which is lower compared to Logistic Regression and KNN. The overall macro average precision, recall, and F1-score are around 0.63–0.64, showing that the model performs moderately across both classes

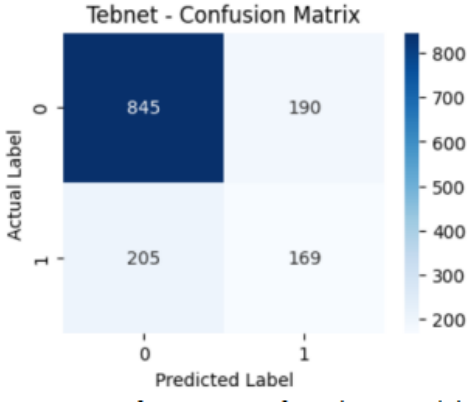

Fig. 6: Confusion Matrix for Decision Tree.

The confusion matrix shows:

- 845 non-churn customers were correctly predicted.
- 169 churn customers were correctly predicted.
- A large number (205) of churn customers were misclassified as non-churn.

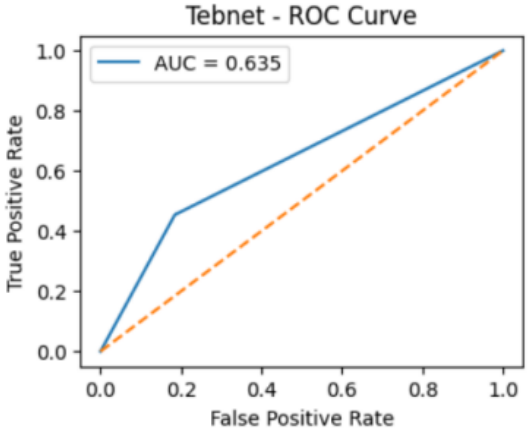

Fig. 7: ROC Curve for Tebnet Model.

D. Summary Table

TABLE I: Model Comparison

Model	Precision	Recall	F1-Score	Accuracy
Logistic Regression	0.78	0.79	0.78	0.79
KNN	0.77	0.78	0.77	0.78
Decision Tree	0.72	0.72	0.72	0.72

V. Discussion

The performance of all three models shows that customer churn prediction is a challenging binary classification task due to class imbalance and overlapping features. Logistic Regression achieved the highest accuracy and provided a good balance between precision and recall for the churn class, indicating that it can generalize well on this dataset. KNN performed moderately, but its lower recall on the churn class suggests difficulty capturing minority patterns.The Decision Tree model showed the lowest performance, likely because it overfits noisy patterns in the data. Overall, the results indicate

that simpler, linear models are more effective for this dataset than non-parametric models.

VI. Conclusion

This study applied Logistic Regression, K-Nearest Neighbors, and Decision Tree models on the Kaggle Telco Customer Churn dataset to predict churn. Among the three, Logistic Regression delivered the best overall performance with the highest test accuracy and balanced classification metrics. KNN performed reasonably but struggled with recall for the churn class, while the Decision Tree model showed lower generalization ability. These findings highlight that churn prediction benefits from models that can handle linear relationships and avoid overfitting. The study demonstrates that machine learning can effectively support telecom companies in identifying at-risk customers and taking proactive retention strategies.

References

[1] 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. 28, 2019.

[2] S. Vafeiadis et al., "A comparison of machine learning techniques for customer churn prediction," *International Journal of Computer Applications*, vol. 127, no. 8, 2015.

[3] Customer Churn Dataset, Kaggle. Available at: https://www.kaggle.com/datasets/rashadrmammadov/customer-churn-dataset