Comparative Analysis of Machine Learning Models for Predicting Customer Churn in a Subscription-Based Service

Farah Walid Abdelsalam

Software Engineer

Uneeq interns Machine learning task 1

***Abstract*— This report presents a comparative analysis of three machine learning models—Logistic Regression, Decision Tree Classifier, and k-Nearest Neighbors (k-NN)—for predicting customer churn in a subscription-based service. The dataset used for this study was preprocessed through various steps, including handling missing values and encoding categorical data. The models were evaluated based on their accuracy, recall, precision, and F1 scores. Among the models tested, Decision Tree showed the highest performance with Accuracy, precision, and F1-score while KNN performed the highest in Recall, indicating that these 2 could be the best 2 options for such datasets.**

Keywords— k-NN, customer churn, accuracy, precision, recall, f1-score, decision tree, logistic regression, classification models.

I. Introduction

Predicting client churn is an essential challenge for subscription-based organisations that want to reduce turnover and keep consumers. Businesses can prevent consumers from leaving by taking proactive measures if they can accurately forecast which ones are likely to do so. his report explores the performance of three widely used classification algorithms: Logistic Regression, Decision Tree Classifier, and k-NN—in predicting customer churn. The goal is to determine which algorithm offers the most reliable performance in this context.

II. Data Description and Preprocessing Steps

The dataset comprises two parts: a training set and a testing set, which were concatenated for easier preprocessing. It contains both numerical and categorical features, including customer demographics, subscription details, and contract length. The target variable is a binary indicator of customer churn.

Preprocessing Steps:

- Handling Missing Values: Missing values were identified and addressed by dropping rows with null values.
- Encoding Categorical Data: Categorical features such as Gender, Subscription Type, and Contract Length were manually encoded into numerical values for model compatibility.
- Feature Scaling: Numerical features were scaled using StandardScaler to ensure uniformity across features.

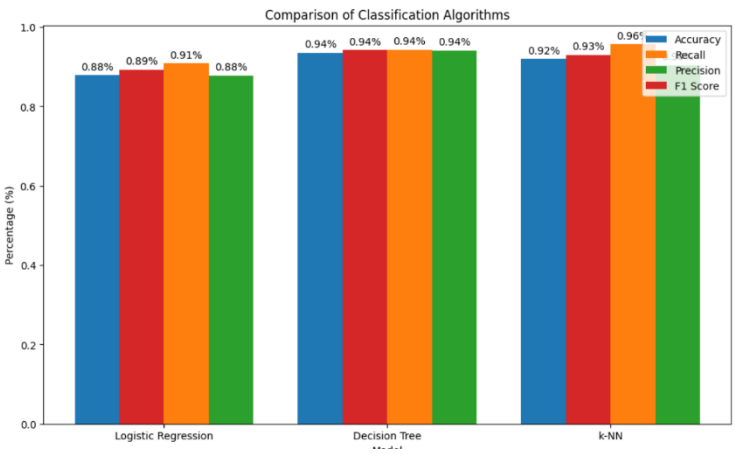
III. Using Methodology/Approach

Using the preprocessed dataset, the methodology entails training and assessing three machine learning models: k-NN, Decision Tree Classifier, and Logistic Regression. 30% of the data set was used for testing after the models had been trained on 70% of it. Accuracy, recall, precision, and F1 score were used to assess each model's performance, giving a thorough understanding of its advantages and disadvantages.

IV. Results

The results obtained from the three models are as follows:

Model Comparison			
	Logistic Regression	Decision Tree	KNN
Accuracy	87.84%	93.53%	91.95%
Precision	87.74%	94.10%	90.43%
Recall	90.83%	94.28%	95.67%
F1-Score	89.26%	94.19%	92.97%



Bar chart comparison of the models used on the dataset

V. Conclusion

As previously said, this study shows that Decision tree, which came in second to KNN, provides the most dependable performance for churn prediction of customers, as seen by its near 95% results in all evaluation criteria. The findings suggest that meticulous data preprocessing, when paired with a suitable machine learning model selection, can greatly influence the accuracy and efficiency of churn prediction. Future work could explore additional models, such as ensemble methods, or further refine the feature engineering process to enhance predictive performance.