Customer Churn Prediction Project Report

Introduction

Customer churn, or customer attrition, refers to when a customer ceases their relationship with a company or service provider. In today's competitive business environment, understanding and reducing customer churn is critical for long-term success. This report documents the steps taken to build a machine learning model to predict customer churn, evaluates the model's performance, and provides actionable insights for business strategy.

Data Preprocessing

Data preprocessing is a critical step in machine learning that ensures data quality and compatibility with the model. The following steps were taken during preprocessing:

1. Handling Missing Values: The 'TotalCharges' column contained missing values which were replaced with the median value.

2. Encoding Categorical Data: Binary columns with 'Yes/No' values (e.g., 'Partner', 'Dependents') were encoded as 1/0. Multi-class categorical columns (e.g., 'InternetService', 'Contract') were one-hot encoded.

3. Scaling Numerical Data: Continuous variables such as 'tenure', 'MonthlyCharges', and 'TotalCharges' were scaled using StandardScaler to standardize the feature ranges.

Modeling

A Random Forest Classifier was selected as the machine learning model due to its ability to handle high-dimensional data, handle categorical features, and provide feature importance scores. The dataset was split into training and testing sets, with 80% of the data used for training and 20% for testing.

In the dataset, the number of non-churners (Class 0) is significantly higher than the number of churners (Class 1). The model tends to be biased toward the majority class (non-churners), leading to poor recall for the minority class (churners).

class_weight=balanced - misclassifying a churner is penalized **1.89 times more** than misclassifying a non-churner, making the model more sensitive to churners.

Model Evaluation

The model's performance was evaluated using precision, recall, F1-score, and ROC-AUC metrics. The key evaluation metrics are summarized below:

1. Precision and Recall: The model performed well in predicting non-churners (Class 0) with a recall of 0.91 and precision of 0.82. However, it struggled with churners (Class 1), achieving a recall of 0.44 and precision of 0.64.

2. Accuracy: The overall accuracy of the model was 79%, indicating a reasonable level of correct predictions.

3. ROC-AUC Score: The **Receiver Operating Characteristic - Area Under Curve (ROC-AUC)** measures the model's ability to distinguish between the two classes across all probability thresholds.

The ROC-AUC score of 0.832 indicates a strong ability to distinguish between churners and non-churners.

Insights and Business Implications

The model provided the following insights into customer churn:

1. Key Features: Features such as 'tenure', 'MonthlyCharges', 'TotalCharges', and 'InternetService_Fiber optic' were identified as the most important factors influencing churn.

2. Business Implications: Customers with short tenure, high monthly charges, and fiber optic internet service are more likely to churn. Retention strategies should focus on these customers by offering personalized discounts, improving service quality, and building customer loyalty programs.

Actionable Insights

Based on the model's outputs, the following actions are recommended to reduce customer churn:

1. Retention Programs: Offer discounts or loyalty programs to customers identified as high-risk (e.g., those with short tenure and high charges).

2. Customer Support: Improve customer support for high-risk customers, especially those with fiber optic internet service.

3. Targeted Marketing: Use the model to target at-risk customers with personalized marketing campaigns.

Conclusion

This project successfully developed a machine learning model to predict customer churn with reasonable accuracy and interpretability. The model's insights can help businesses take proactive steps to retain customers and minimize revenue loss. Future work should focus on addressing class imbalance and exploring advanced machine learning models for further improvement.