

Customer Churn Prediction – Final Report

This report details the steps taken to build a supervised machine learning model to predict customer churn using a telecom dataset. It includes data cleaning, preprocessing, model training, evaluation, and recommendations for future improvements.

Dataset Overview

The dataset contains 7,043 customer records with features like tenure, contract type, internet service, and charges. The target variable is 'Churn', indicating whether a customer has left the service.

Data Cleaning and Preparation

We converted the 'TotalCharges' column to numeric and dropped missing values. The 'customerID' column was removed as it doesn't provide predictive value. Categorical columns were encoded using Label Encoding.

Handling Class Imbalance

Since churners made up only 27% of the dataset, we used SMOTE to oversample the minority class in the training set. This helped balance the model's sensitivity.

Model Training and Tuning

A Random Forest Classifier was used with `class_weight='balanced'`. Hyperparameter tuning was performed using GridSearchCV with cross-validation to improve performance.

Evaluation Results

The final model achieved the following on the test set:

- Accuracy: ~76%
- Precision for Churn: ~54%
- Recall for Churn: ~65%
- F1-Score for Churn: ~0.59

These metrics indicate a solid model with good recall for identifying actual churners.

Visual Analysis

A ROC Curve demonstrated the model's discriminative power. A feature importance plot revealed that Contract type, Tenure, and Monthly Charges were the top contributors to churn prediction.

Recommendations for Improvement

To further enhance this model:

- Integrate SHAP for local interpretability
- Experiment with XGBoost or LightGBM

- Deploy using Streamlit for real-time use
- Add a precision-recall curve and incorporate business cost analysis to fine-tune decision thresholds.