Predictive Analytics for Customer Churn: Dataset

Team Members: Valerie Garcia (vlg170000@utdallas.edu), Patrick Ninan (pgn230000@utdallas.edu), Albin Poulose (axp220239@utdallas.edu), Srivani Kakumani (sxk230164@utdallas.edu), Ashiq Mohammed Al Ameen (axa220171@utdallas.edu)

Target and Motivation

The project aims to predict customer churn in a subscription-based service using anonymized data. By analyzing subscription types, payment methods, and customer interactions, the model seeks to identify customers at risk of canceling their subscriptions. The motivation lies in retaining revenue, enhancing customer satisfaction, and ensuring business sustainability. Additionally, predicting churn provides a competitive advantage and optimizes resource allocation. Ultimately, proactive churn management enables targeted retention strategies and fosters stronger customer relationships.

Target:
Churn — Whether the customer churned or not (Yes, No)

Numeric Features:
Tenure — Number of months the customer has been with the company
MonthlyCharges — The monthly amount charged to the customer
TotalCharges — The total amount charged to the customer

Categorical Features:
CustomerID, Gender — M/F
SeniorCitizen — Whether the customer is a senior citizen or not (1, 0)
Partner — Whether customer has a partner or not (Yes, No)
Dependents — Whether customer has dependents or not (Yes, No)
PhoneService — Whether the customer has a phone service or not (Yes, No)
Multiple Lines — Whether the customer has multiple lines or not (Yes, No, No Phone Service)
InternetService — Customer's internet service type (DSL, Fiber Optic, None)
OnlineSecurity — Whether the customer has Online Security add-on (Yes, No, No Internet Service)
OnlineBackup — Whether the customer has Online Backup add-on (Yes, No, No Internet Service)

DeviceProtection — Whether the customer has Device Protection add-on (Yes, No, No Internet Service)
TechSupport — Whether the customer has Tech Support add-on (Yes, No, No Internet Service)
StreamingTV — Whether the customer has streaming TV or not (Yes, No, No Internet Service)
StreamingMovies — Whether the customer has streaming movies or not (Yes, No, No Internet Service)
Contract — Term of the customer's contract (Monthly, 1-Year, 2-Year)
PaperlessBilling — Whether the customer has paperless billing or not (Yes, No)
PaymentMethod — The customer's payment method (E-Check, Mailed Check, Bank Transfer (Auto), Credit Card (Auto))

Data

We are using a Kaggle Dataset denoting Predictive Analytics for Customer Churn: Dataset (kaggle.com).
The provided datasets consist of customer data for a streaming service, with one dataset for training containing 243,787 entries and another for testing with 104,480 entries.

Each entry contains 21 columns in the training set and 20 columns in the test set, including features such as account age, monthly charges, total charges, subscription type, payment method, and customer preferences. Key features include account age, viewing habits (viewing hours per week, average viewing duration), content engagement metrics (content downloads per month, watchlist size), and customer demographics (gender).

Both datasets have columns of various data types including integers, floats, and objects (likely representing categorical variables), with no missing values observed. The target variable 'Churn', present in the training set, likely indicates whether a customer has discontinued their subscription, crucial for churn prediction modeling. These datasets provide rich information for customer behavior analysis and predictive modeling aimed at understanding and potentially reducing churn within the streaming service platform.

Plan for the Project

We plan to use logistic regression, ROC curve and AUC and a decision tree to predict whether the customer churned or not (Yes, No).