## **Customer Churn Prediction for a Telecom Company**

### **Overview**

#### **Business Need**

You are a data scientist at **Synergy Telecom Inc.**, a leading telecommunications company. The company is facing a significant challenge with customer churn, which is customers leaving for competitors. Acquiring a new customer is five times more expensive than retaining an existing one, so reducing churn is a top business priority.

The marketing department wants to launch a targeted retention campaign, offering special discounts and plan upgrades to customers who are at high risk of churning. However, they don't know who to target. A broad campaign would be too costly, while no campaign means continued revenue loss.

Your task is to analyze customer data and build a machine learning model that can accurately predict which customers are likely to churn in the near future. More importantly, you must identify the **key drivers** of churn. Understanding *why* customers leave is critical for Synergy Telecom to address root causes and improve its services.

This project will involve:

* Analyzing and preprocessing a real-world dataset with mixed data types.
* Building and comparing a simple baseline model with a more powerful one.
* Evaluating model performance with a focus on identifying at-risk customers.
* Interpreting the model's decisions to provide actionable insights to the business.

### **Data and Features**

You will be using the **Telco Customer Churn** dataset, a popular and realistic dataset provided by IBM.

* **Source:** [Kaggle: Telco Customer Churn](https://www.kaggle.com/datasets/blastchar/telco-customer-churn)
* **Description:** The dataset contains information about 7,043 customers of a fictional telecom company. Each row represents a unique customer.
* **Key Features:**
  + **Customer Demographics:** gender, SeniorCitizen, Partner, Dependents.
  + **Account Information:** tenure (how long they've been a customer), Contract, PaymentMethod, MonthlyCharges, TotalCharges.
  + **Subscribed Services:** PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies.
  + **Target Variable:** Churn (A binary feature: 'Yes' or 'No').
* **Critical Challenge: Data Preprocessing.** The dataset contains a mix of numerical and categorical features. The TotalCharges column is stored as an object and contains missing values that must be handled carefully. You will need a robust preprocessing strategy to prepare the data for modeling.

### **Learning Outcomes**

* **Skills:**
  + Perform a comprehensive Exploratory Data Analysis (EDA) to uncover initial insights.
  + Develop a robust data cleaning and preprocessing pipeline using scikit-learn.
  + Effectively encode categorical variables and scale numerical features.
  + Train and evaluate classification models like **Logistic Regression** and **XGBoost**.
  + Interpret model predictions using **SHAP** to identify the main drivers of churn.
  + Translate model findings into clear, actionable business recommendations.
* **Knowledge:**
  + Grasp the business impact of customer churn and the role of predictive analytics in retention strategies.
  + Understand how to evaluate a classification model beyond simple accuracy (e.g., using AUC, F1-Score).
  + Justify model selection based on both performance and interpretability.

### **Instructions**

#### **Phase 1 - Exploratory Data Analysis and Preprocessing**

- **Objective**: Define the precise goal, understand, and clean the data.

**- Your Tasks:**

1. **Data Loading and Initial Cleaning:**
   * Load the Telco-Customer-Churn.csv dataset.
   * Inspect the data types (.info()). Notice that TotalCharges is an object type. Investigate why and convert it to a numeric type, handling any errors or missing values that arise. This is a critical first step.
2. **Exploratory Data Analysis (EDA):**
   * Analyze the target variable Churn. Is the dataset balanced?
   * Create visualizations to understand the relationship between key features and churn. For example:
     + Does tenure affect churn? (e.g., a histogram of tenure for churned vs. non-churned customers).
     + How does the Contract type (Month-to-month, One year, Two year) relate to churn? (e.g., a count plot).
     + Are customers with higher MonthlyCharges more likely to leave?
3. **Data Preprocessing Pipeline:**
   * Separate your data into features (X) and the target (y).
   * Perform a train-test split.
   * Use scikit-learn's ColumnTransformer and Pipeline to build a preprocessing workflow that:
     + **Scales** numerical features (like tenure, MonthlyCharges) using StandardScaler.
     + **Encodes** categorical features using OneHotEncoder.

#### **Phase 2 - Model Building and Evaluation**

- **Objective**: Build and evaluate a machine learning model.

- **Your Tasks**:

1. **Baseline Model:**
   * Build a Pipeline that includes your preprocessing steps and a **Logistic Regression** model.
   * Train the model on your training data.
2. **Advanced Model:**
   * Build a second Pipeline that includes your preprocessing steps and a more powerful model like **XGBoost** (XGBClassifier).
   * Train this model on the same training data.
3. **Model Evaluation:**
   * For both models, make predictions on your test set.
   * Evaluate their performance using the following:
     + **Confusion Matrix:** To see the trade-off between correctly identifying churners and non-churners.
     + **Classification Report:** To see precision, recall, and F1-score.
     + **AUC Score:** To measure the model's overall ability to distinguish between the two classes.
   * Compare the two models. Which one performs better? Why?

#### **Phase 3 - Model Interpretation and Business Recommendations**

- **Objective**: Model interpretation and recommendation.

**- Your Tasks:**

1. **Identify Churn Drivers:**
   * Using your best-performing model (likely XGBoost), use the **SHAP** library to understand which features are most important for predicting churn.
   * Generate a **SHAP summary plot**. What are the top 5 factors that influence the model's predictions?
2. **Create a "Profile" of a High-Risk Customer:**
   * Based on your EDA and SHAP analysis, describe the typical customer who is likely to churn. (e.g., "Customers on a month-to-month contract with high monthly charges and no tech support are at the highest risk.").
3. **Formulate Actionable Recommendations:**
   * Write a short report (5-10 pages) for the "Head of Marketing."
   * **Your report must include:**
     + A brief summary of the project goal.
     + The key factors driving customer churn (with a visualization like the SHAP plot).
     + At least **three concrete, data-driven recommendations** for the marketing team. (e.g., "Offer a 15% discount to high-risk customers if they switch from a monthly to a one-year contract.").