## Problem Statement

Customer churn is a major challenge for telecom companies. Retaining existing customers is often more cost-effective than acquiring new ones, but to do so effectively, companies must identify the factors that drive customers to leave. You are provided with the IBM Telco Customer Churn Dataset, which contains customer demographics, subscription details, account information, and churn status. Your task: 1. Perform a thorough Exploratory Data Analysis (EDA) to understand churn patterns. 2. Create meaningful visualizations to communicate findings clearly. 3. Train a Machine Learning classification model to predict customer churn and interpret the results for business action. **Dataset Link:** IBM Telco Customer Churn CSV

## EDA Questions

### 1. Dataset Overview

- What is the total number of rows and columns in the dataset?
- What are the data types of each column?
- Which columns are categorical, numerical, or binary?
- Are there irrelevant columns (e.g., customerID) that should be dropped?
- How many unique values are there in each categorical variable?

### 2. Data Quality Checks

- Are there missing values in any column?
- Are there blank spaces stored as empty strings?
- Does TotalCharges contain non-numeric values?
- Are there duplicate customerID entries?
- Is TotalCharges approximately equal to MonthlyCharges × tenure?

### 3. Target Variable Analysis

- What is the churn rate (%)?
- Is the target variable (Churn) balanced or imbalanced?
- Which churn value is more frequent?

### 4. Univariate Analysis

- What is the distribution of tenure, MonthlyCharges, and TotalCharges?
- Which internet service type is most common?
- Which contract type is most popular?
- Which payment method is most used?
- Which service add-ons are most/least subscribed?

### 5. Bivariate Analysis

- What is the churn rate by Contract type?
- How does PaymentMethod affect churn?
- Do customers with PaperlessBilling churn more?
- How does churn vary across InternetService types?
- Do customers with OnlineSecurity or TechSupport churn less?
- How does MonthlyCharges differ between churned and retained customers?
- How does tenure differ between churned and retained customers?
- Does senior citizen status affect churn?
- Does gender affect churn?
- Do customers with dependents churn less?

### 6. Multivariate Analysis

- How does churn vary across Contract type and InternetService?
- How does churn vary by PaymentMethod and PaperlessBilling together?

- Are short-tenure, high-monthly-charge customers more likely to churn?
- How does churn change with the number of subscribed services?
- Do bundled services reduce churn more than single services?

### 7. Business Insight Questions

- Which three features are most strongly related to churn?
- Which customer segment has the highest churn risk?
- Which segment is high-value but high-risk?
- What patterns indicate early-stage churn risk?
- Which features should be targeted in retention strategies?

## Visualization Questions

- Churn Distribution – Bar chart of churn counts and percentages.
- Contract Type & Churn – Stacked bar or 100% stacked bar.
- Payment Method & Churn – Grouped bar chart.
- Internet Service & Churn – Bar chart.
- Tenure Distribution – Histogram + KDE.
- Monthly Charges vs Churn – Boxplot or violin plot.
- Tenure vs Monthly Charges – Scatter or hexbin plot with churn color coding.
- Bundle Depth vs Churn – Bar chart of churn rate by number of subscribed services.
- Payment Friction – 100% stacked bar for PaymentMethod × PaperlessBilling.
- Customer Value at Risk – KDE or histogram of TotalCharges for churned vs retained.

## ML Model Training Task

- Target & Features – Encode Churn as binary (1 = Yes, 0 = No).
- Data Cleaning – Convert TotalCharges to numeric, handle errors, impute missing values, drop customerID.
- Encoding – One-hot encode categorical variables.
- Scaling – Standardize numerical variables.
- Train/Test Split – Stratified split (80/20).
- Handle Class Imbalance – Use class_weight='balanced' or SMOTE.
- Baseline Models – Train Logistic Regression and Random Forest.
- Evaluation Metrics – Accuracy, Precision, Recall, F1-score, ROC-AUC, PR-AUC.
- Feature Importance – Use coefficients (Logistic) and permutation importance (RF).
- Business Recommendations – Translate findings into actionable strategies.