Telco Customer Churn (Option 1)

PROBLEM STATEMENT (TELCO)

Telecommunications companies lose significant revenue when existing customers discontinue their service ("churn"). Acquiring new customers is typically more expensive than retaining current ones, so proactively identifying customers at high risk of churn is a key business priority.

In this project, we use the "Telco Customer Churn" dataset from Kaggle, where each row represents a telecom customer with information about demographics, account details, services subscribed, and monthly charges, along with a binary label Churn indicating whether the customer left in the last month. Our goal is to build a supervised classification model that predicts the probability of churn for each customer and to understand which factors most strongly drive churn.

The final model and analysis should help business stakeholders design targeted retention strategies (e.g., discounts, contract changes, or service upgrades) focused on segments with high churn risk.

MODELING PLAN

1. **Problem Framing & Track Justification**
    - Type: Supervised Learning – Classification (target Churn: Yes/No).
    - Justification: We have labeled historical outcomes, and the business question is to predict a categorical outcome (churn vs no churn) and analyze drivers of that outcome.

2. **Data understanding & preprocessing**
    - Load the Kaggle CSV and perform dataset overview: number of rows, columns, data types, basic statistics.
    - Handle missing values (e.g., blank TotalCharges in certain rows), encode categorical variables (e.g., one-hot), and scale numerical features if needed for Logistic Regression.
    - Split data into train/test sets (e.g., 70/30 or 80/20) using stratification on Churn to preserve class balance.

3. **Exploratory Data Analysis (EDA)**

- Univariate:
    - Histograms and box plots for numerical features (e.g., tenure, MonthlyCharges, TotalCharges).
    - Bar plots and value counts for categorical features (e.g., Contract, InternetService, PaymentMethod).

- Missing values analysis: heatmap or bar chart to show which columns have missing values and how many.
- Bivariate:
    - Churn rate by contract type, tenure buckets, payment method, and monthly charges.
    - Correlation matrix for numeric features; stacked bar plots/box plots of features vs Churn.
- Target analysis:
    - Plot class distribution of Churn and quantify class imbalance (e.g., % of customers that churned).

4. Baseline model: Logistic Regression

- Train a Logistic Regression classifier using scaled numeric features and encoded categorical variables.
- Evaluate on the test set using Accuracy, Precision, Recall, F1-score, and Confusion Matrix (no RMSE).
- Interpret coefficients (e.g., odds ratios) to understand which features increase or decrease churn risk.

5. Advanced model: Random Forest (or Decision Tree) classifier

- Train a tree-based model on the same feature set.
- Tune key hyperparameters (e.g., max_depth, n_estimators for Random Forest) via cross-validation.
- Evaluate with the same metrics and confusion matrix; extract feature importance scores to see which variables the model relies on most.

6. Class imbalance handling & error analysis

- Inspect class balance; if churners are a minority, experiment with:
    - Class weights in models, or
    - Oversampling methods (e.g., simple random oversampling) on the training set.
- Compare metrics and confusion matrices before and after imbalance handling; discuss impact on false positives vs false negatives.
- Perform error analysis: which customer segments are most frequently misclassified? Is the model missing “silent churners” or over-flagging low-risk customers?

7. Model comparison & business interpretation

- Compare Logistic Regression vs Random Forest on accuracy, precision, recall, F1, and interpretability.
- Justify the chosen “production” model for the capstone (e.g., Random Forest for performance, Logistic Regression for interpretability, or a trade-off).

- Translate technical insights into business actions: e.g., “month-to-month contracts, fiber optic internet, and high monthly charges are associated with higher churn risk; retention strategies should focus on customers in these segments.”