**Problem Formulation: Customer Churn Prediction Pipeline**

**1. Business Problem Definition**

**Objective:**

Customer churn is a critical concern for businesses, particularly in subscription-based industries such as telecommunications, banking, and e-commerce. High customer churn rates lead to revenue loss and increased costs for acquiring new customers. The goal of this project is to build an end-to-end machine learning pipeline that predicts customer churn based on historical customer data.

**Key Business Objectives:**

* Reduce customer churn by identifying at-risk customers early.
* Improve customer retention strategies through targeted interventions.
* Optimize marketing and customer service efforts to enhance customer experience.
* Increase revenue by reducing customer acquisition costs.

**2. Data Sources and Attributes**

**Primary Data Sources:**

1. **Customer Churn Dataset (Local CSV File)**
   * Attributes:
     + CustomerId: Unique identifier for customers.
     + Surname: Customer's last name (not used in prediction).
     + CreditScore: Customer’s creditworthiness score.
     + Geography: Customer’s country.
     + Gender: Male/Female.
     + Age: Customer’s age.
     + Tenure: Number of years as a customer.
     + Balance: Account balance.
     + NumOfProducts: Number of products used.
     + HasCrCard: Whether the customer has a credit card (0/1).
     + IsActiveMember: Whether the customer is an active user (0/1).
     + EstimatedSalary: Estimated yearly salary.
     + Exited: Whether the customer churned (1) or stayed (0).
2. **Kaggle Customer Churn Dataset (Fetched via API)**
   * Attributes:
     + customerID: Unique identifier for customers.
     + SeniorCitizen: Indicates if the customer is a senior.
     + tenure: Length of stay in months.
     + MonthlyCharges: Monthly service charges.
     + TotalCharges: Total amount billed.
     + InternetService, Contract, PaymentMethod: Service-related details.
     + Churn: Whether the customer churned (‘Yes’/‘No’).

**3. Expected Outputs**

**Pipeline Deliverables:**

1. **Clean datasets** suitable for exploratory data analysis (EDA) and feature engineering.
2. **Transformed features** optimized for machine learning models.
3. **A trained, deployable model** that predicts customer churn based on customer attributes.
4. **Versioned datasets** at different processing stages for reproducibility.
5. **Automated pipeline orchestration** with Apache Airflow for data ingestion, preprocessing, training, and deployment.

**4. Evaluation Metrics**

To measure the performance of the churn prediction model, the following metrics will be used:

* **Accuracy:** Measures the proportion of correctly predicted churn labels.
* **Precision:** Ensures that when a customer is predicted to churn, they are likely to actually churn.
* **Recall (Sensitivity):** Measures how many actual churned customers are correctly identified.
* **F1 Score:** A balance between precision and recall.
* **ROC-AUC Score:** Evaluates the model’s ability to distinguish between churned and non-churned customers.

**5. Deliverables**

* **Problem formulation document (this document) in PDF/Markdown format.**
* **Data pipeline documentation** outlining data ingestion, validation, feature engineering, model training, and deployment.
* **Performance reports and visualization of model evaluation metrics.**
* **Source code for all pipeline components, stored in a version-controlled repository.**