End-to-End Data Management Pipeline for Machine Learning

1. Problem Formulation

Business Problem

Customer churn occurs when an existing customer stops using a company's services or purchasing its products. Addressable churn, which can be mitigated through strategic interventions, leads to revenue losses, increased customer acquisition costs, and a negative impact on brand reputation. The primary goal is to predict customer churn and develop proactive strategies to enhance customer retention.

Key Business Objectives

- Reduce customer churn by identifying at-risk customers early.
- Improve customer retention strategies through predictive insights.
- Minimize revenue loss by leveraging data-driven decision-making.
- Automate the data processing pipeline for scalability and efficiency.

Key Data Sources and Their Attributes

1. **Kaggle Dataset** (Telco Customer Churn)
    - Customer demographics
    - Service subscription details
    - Monthly charges and tenure
    - Churn indicator
2. **Custom API Endpoint** (Data extraction and feature engineering)
    - Aggregated customer activity data
    - Real-time interaction metrics
    - Custom features derived from transactional and behavioral data

Expected Outputs from the Pipeline

1. **Clean datasets for Exploratory Data Analysis (EDA)**
    - Remove missing values and duplicates
    - Normalize and standardize features
    - Handle categorical variables
2. **Transformed features for machine learning**
    - Feature engineering (aggregated metrics, derived attributes)

    - Feature selection (important predictors of churn)
    - Encoding and scaling
3. **Deployable model for customer churn prediction**
    - Train ML models (Logistic Regression, Random Forest, Neural Networks, etc.)
    - Evaluate performance using key metrics
    - Deploy the best-performing model

Measurable Evaluation Metrics

- **Accuracy**: Measure the proportion of correct predictions.
- **Precision & Recall**: Balance false positives and false negatives.
- **F1 Score**: Harmonic mean of precision and recall for imbalanced datasets.
- **ROC-AUC Score**: Evaluate the discriminative power of the model.
- **Model Interpretability**: Feature importance analysis.