Phase 1: Project and Data Overview

1. Introduction

Customer churn represents a significant challenge for the telecommunications industry, where retaining existing subscribers is substantially more cost-effective than acquiring new ones. A high churn rate directly undermines revenue growth and poses long-term risks to business sustainability. Without early detection mechanisms, firms risk losing valuable customers who could otherwise be retained through timely interventions.

This study seeks to address this challenge by developing a predictive model to identify customers at high risk of churn. Through statistical analysis and machine learning, the project will generate actionable insights to support managerial decision-making and enhance customer retention strategies.

2. Problem Definition and Project Objectives

Problem Statement

The central problem concerns the inability to anticipate which customers are most likely to discontinue services. Churn erodes customer lifetime value and reduces operational efficiency, especially in highly competitive telecom markets.

Project Objectives

The project aims to:

1. Develop a predictive model capable of accurately classifying customers into churn and non-churn groups.
2. Identify and quantify the primary factors influencing customer attrition.
3. Provide business-oriented recommendations to inform targeted retention campaigns.

Expected Outcomes

The expected deliverables include:

- A machine learning model optimized for recall to minimize false negatives in churn detection.
- Empirical insights on the behavioral and service-related drivers of churn.
- Visualization dashboards and a business report for managerial interpretation.
- A complete project repository (technical notebooks, documentation, and GitHub codebase) suitable for both deployment and portfolio demonstration.

3. Data Sources Overview

Dataset Description

The analysis is based on the telecom customer churn dataset, provided as a CSV file (telecom_churn_dirty.csv). It contains **3,343 customer records** and **11 variables**. The dataset primarily captures contractual details, service usage, and billing behavior.

Data Coverage

- **Target Variable:** Churn (binary: 0 = retained, 1 = churned).
- **Service Features:** DataPlan, DataUsage, CustServCalls, DayMins, DayCalls, RoamMins.
- **Billing Features:** MonthlyCharge, OverageFee.
- **Contract Features:** AccountWeeks, ContractRenewal.

Limitations

Demographic factors (e.g., age, gender, geographic location) are absent, restricting the potential for customer segmentation and profiling. Nevertheless, the dataset provides adequate behavioral and service attributes for predictive modeling.

4. Exploratory Data Overview

Dataset Structure

The dataset includes 3,343 records with one binary target variable (Churn) and ten predictors, spanning contract, service, and billing attributes. Data types are primarily numeric or binary, making the dataset highly amenable to classification modeling.

Descriptive Statistics

- Churn prevalence is **14.5%**, indicating **class imbalance**.
- Average service tenure: **101 weeks**, with a maximum of 243.
- Data plan adoption: only **27%** of customers; mean usage = **0.82 GB/month**.
- Customer service calls: median = 1, but extreme cases reach 50.
- Billing: average monthly charge ≈ **56 CAD**; roaming minutes ≈ **10 minutes** per user.

Preliminary Observations

- **Missing Values:** Present in DataUsage (5.3%), DayMins (5.1%), and MonthlyCharge (5.0%).

- **Outliers:** Detected in CustServCalls (up to 50), DayMins (up to 1000 minutes), MonthlyCharge (>100 CAD), and RoamMins (≈20 minutes).
- **Implication:** These anomalies require rigorous cleaning and validation in the next phase.

5. Data Quality Assessment

Missing Values

- DataUsage: 177 missing (5.3%)
- DayMins: 169 missing (5.1%)
- MonthlyCharge: 168 missing (5.0%)
  → To be handled via imputation (median/regression) or row removal depending on data integrity.

Outliers

Using the IQR rule and distribution plots:

- CustServCalls: 273 outliers (8.17%) — may contain **business-relevant signals** (e.g., dissatisfaction).
- DayMins: 30 outliers (0.9%) — extreme usage up to 1000 minutes.
- MonthlyCharge: 35 outliers (1.05%) — charges >100 CAD.
- RoamMins: 46 outliers (1.38%) — unusually high roaming activity.

Class Imbalance

- Churn: 14.5%
- Non-Churn: 85.5%
  → Indicates potential **bias toward majority class**; requires techniques such as SMOTE, undersampling, or class weighting.

6. Next Steps Planning

Data Preprocessing

- **Missing Values:** Imputation strategies (median, regression, or domain-guided).
- **Outliers:** Retain business-significant anomalies (e.g., CustServCalls) but cap/transform extreme noise.

- **Encoding:** No further action required as all categorical variables are binary.

Feature Engineering

- Derived features (e.g., **TotalCharges = MonthlyCharge × AccountWeeks**).
- Interaction terms (e.g., DataPlan × DataUsage).
- Scaling and correlation analysis for redundancy reduction.

Analytical Framework

- Data splitting: training, validation, and test sets.
- Modeling: Logistic Regression (baseline) and advanced classifiers (Random Forest, XGBoost).
- Evaluation: Emphasis on **Recall, Precision, F1-score, and ROC-AUC**, beyond simple accuracy.

7. Conclusion

Phase 1 established the groundwork for predictive modeling of customer churn. The dataset is sufficiently comprehensive, though it presents notable challenges: missing values, outliers, and class imbalance. These issues will be systematically addressed in next phases through preprocessing, feature engineering, and robust model design.

By articulating the business problem, outlining the dataset, and identifying quality concerns, Phase 1 ensures a structured transition into subsequent analytical phases.