Customer Churn Prediction

Analysis and Preprocessing Report

Prepared for: Telecommunications Leadership Team Prepared

by: **MIHIR MILIND UGHADE**

Data Science & Analytics

Table of Contents

Executive Summary

This report documents the foundational phases of a customer churn prediction initiative for a telecommunications provider. It consolidates data understanding, exploratory analysis, preprocessing, and feature engineering to enable a robust modeling stage. Early signals indicate that contract type, tenure, internet service, and monthly charges are key churn drivers. The project now transitions to model development and validation, with an emphasis on fair and actionable predictions to support retention strategies.

Business impact opportunities include: reducing churn among month-to-month customers through targeted retention offers; reassessing pricing and experience for fiber-optic customers; and implementing early-life engagement programs for short-tenure segments. The modeling phase will quantify uplift from potential interventions and inform ROI-positive actions.

1. Introduction

The objective of this project is to identify customers at risk of churn and to surface the primary drivers that explain this risk. Using a structured data pipeline and reproducible analysis, the team aims to produce a reliable predictive model that can power retention workflows, proactive outreach, and pricing decisions.

Scope of this report includes project setup, exploratory data analysis (EDA), data preprocessing, and feature engineering. Modeling, evaluation, and deployment will be addressed in subsequent phases.

2. Data Overview

Dataset: Telco Customer Churn (publicly available). The working corpus is split into train, validation, and test files. It captures customer demographics, subscribed services, account and billing attributes, and the target variable (Churn).

2.1 Demographics

Representative attributes include age, gender, partner status, and dependents. These fields contextualize service usage and inform segmentation strategies.

2.2 Services Subscribed

Service-related attributes include phone service, internet service (DSL, Fiber Optic, Cable), and add-ons such as online security, online backup, device protection, tech support, and streaming. Service combinations often delineate risk profiles and price sensitivity.

2.3 Account & Billing

Key fields include tenure (months), contract type, payment method, paperless billing, monthly charges, and total charges. These variables frequently exhibit strong associations with churn risk and lifetime value.

2.4 Target Variable

Churn is a binary label indicating whether a customer churned (1) or remained (0). Class imbalance is anticipated and addressed during modeling and evaluation.

3. Exploratory Data Analysis (EDA)

EDA focused on assessing target distribution, bivariate relationships, and early hypotheses regarding drivers of churn. Observed patterns align with telecommunications benchmarks and indicate several actionable levers for retention.

Key insights and business implications include:

- Customers on month-to-month contracts are substantially more likely to churn than those on annual or biennial contracts. Implication: consider incentives to promote longer-term contracts at renewal.
- Fiber Optic subscribers exhibit higher churn rates relative to DSL, suggesting potential sensitivity to pricing or experience. Implication: audit fiber plans for price-value alignment and service reliability.
- Short-tenure customers are at elevated risk, particularly within the first 12 months. Implication: deploy early-life cycle engagement and onboarding programs to consolidate loyalty.
- Higher monthly charges correlate with increased churn propensity, especially among fiber customers. Implication: explore tiered pricing, bundling, or targeted discounts to mitigate price-driven attrition.

4. Data Preprocessing

The preprocessing pipeline ensures data quality, leakage prevention, and model compatibility. Key decisions and justifications are outlined below.

Column Pruning

Removed Customer ID, Churn Category, and Churn Reason. The former is non-predictive; the latter two are directly tied to the label and would introduce leakage.

Missing Value Treatment

Filled missing values in Internet Type as 'No Internet Service' to maintain category integrity. Converted Total Charges from object to numeric and imputed blanks as zero to reflect new accounts.

Categorical Encoding

Applied one-hot encoding to represent categories as binary indicators, enabling compatibility with linear and tree-based classifiers.

5. Feature Engineering

To capture non-linear tenure effects and interaction patterns, engineered features complement the raw variables.

Tenure Binning: segmented tenure into bins (0–12, 12–24, 24–48, 48–60, 60–72 months) with one-hot indicators to model distinct risk regimes.

Interaction Readiness: the encoded service and contract variables facilitate interactions during modeling (e.g., trees) without manual cross-features at this stage.

6. Next Steps

Model Development: train baseline and advanced classifiers (Logistic Regression, Random Forest, XGBoost).

Evaluation: assess on validation data with AUC, F1-score, precision, recall, and calibration; monitor performance across customer segments to ensure fairness.

Tuning: apply hyperparameter optimization (cross-validation, early stopping) and compare uplift vs. complexity.

Final Test: lock the best model and evaluate on the test set to obtain an unbiased performance estimate.

Deployment Readiness: prepare inference pipeline, monitoring plan, and threshold strategy aligned with business goals.

7. Conclusion and Recommendations

The analysis confirms that contract structure, tenure, internet service type, and pricing are central to churn risk. With a clean and engineered dataset, the team is positioned to develop a high-precision model to guide targeted retention actions.

Contract Strategy: incentivize transitions from month-to-month to annual commitments at renewal, prioritizing high-risk cohorts.

Pricing and Value: review fiber-optic plans for pricing elasticity and perceived value; consider bundle discounts or service guarantees.

Early-Life Engagement: implement onboarding journeys for customers within 12 months of tenure, emphasizing service education and proactive support.

Risk-Based Outreach: use model scores to trigger retention offers and service checks, measuring ROI through controlled experiments.

8. Risks and Assumptions

Data Leakage: features directly reflective of churn events must remain excluded throughout modeling and deployment.

Class Imbalance: mitigation via metrics choice, resampling, and calibrated thresholds is required to avoid biased performance.

Generalizability: model drift risk due to pricing changes, market dynamics, or network performance variations; monitoring essential.

9. Reproducibility and Governance

All steps are traceable through notebooks and scripts. Versioning, environment capture, and data validation checks are recommended to ensure auditability. Prior to deployment, establish governance over model updates, performance thresholds, fairness safeguards, and incident response.