Final Project: Churn Prediction & Data Drift (Offline MLOps)

Date of Submission:

February 26, 2026

Course:

Data Analysis &Engineering

Instructor:

Fabien LIONTI

Group 5:

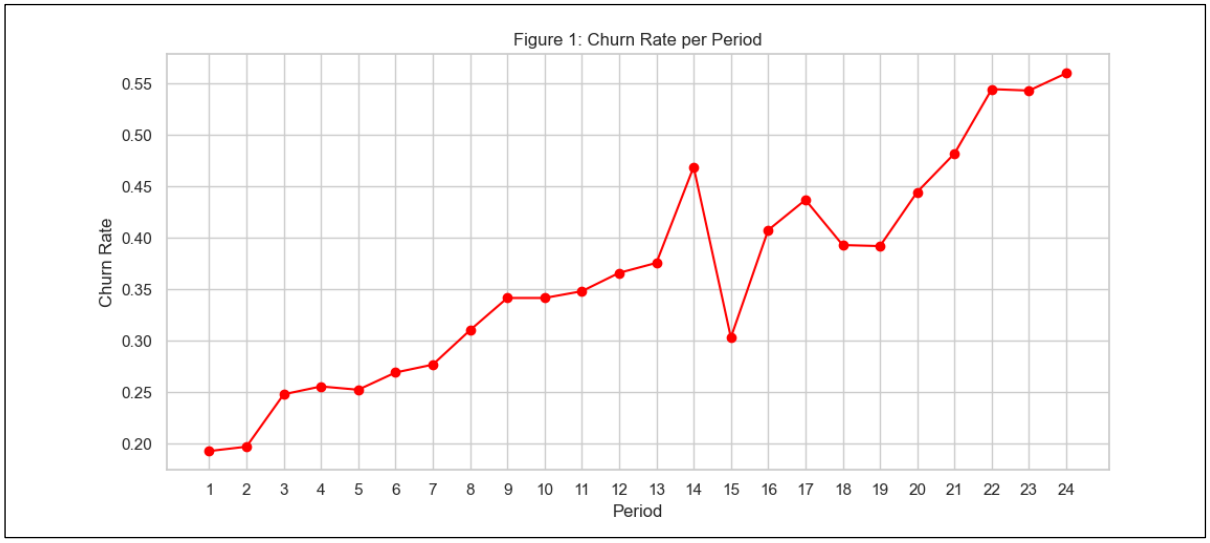
- **Lareinne KOUONANG**
- **Merveille KAMDEM**
- **Simo BORIS**
- **Shashidhar MARIKUKULA**

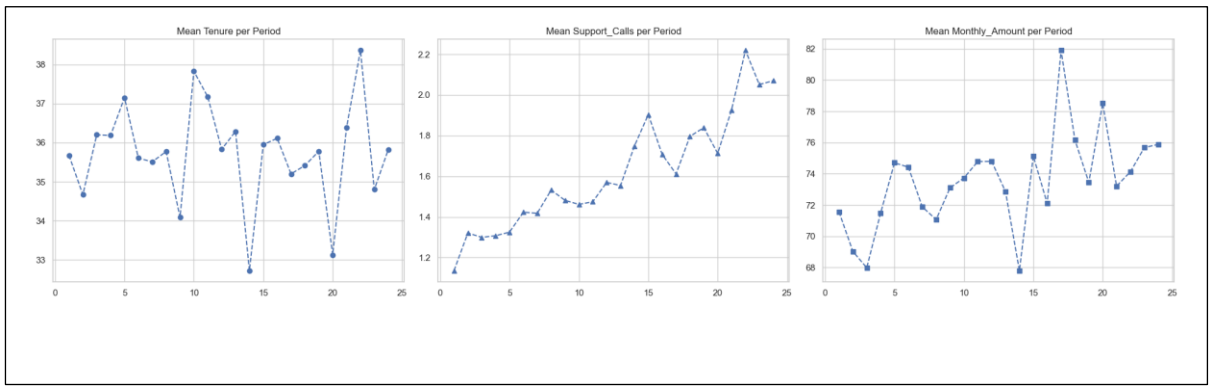
1. Introduction

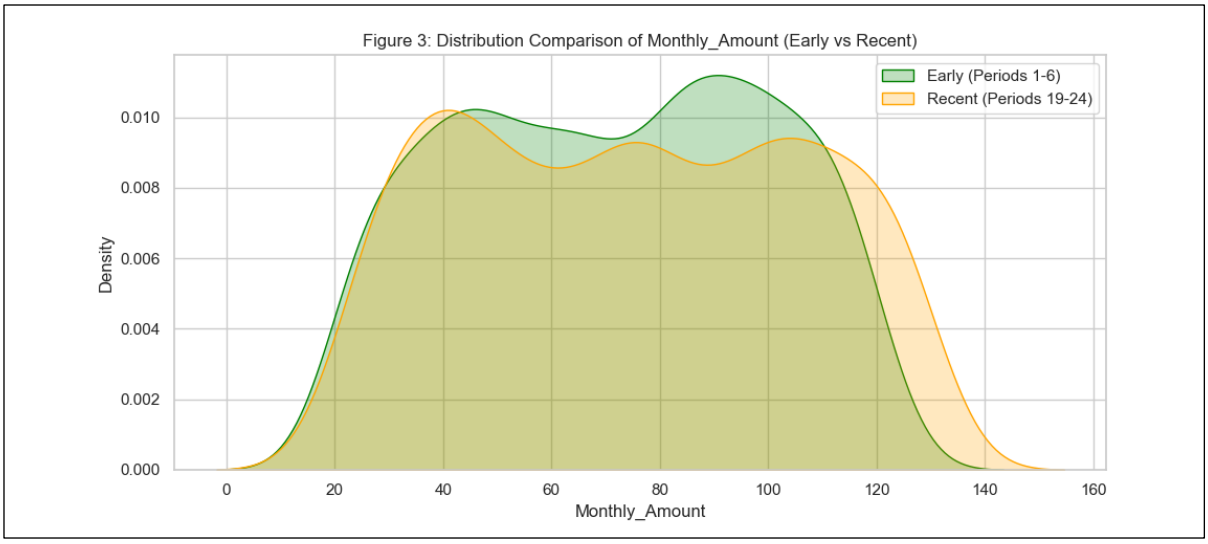
Attracting a new customer **costs five times** as much as keeping an existing. **Customer churn**, also known as customer retention, is defined as when customers or subscribers discontinue doing business with a firm or service. Customer churn is a critical metric because it is much less expensive to retain existing customers than it is to acquire new customers. To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.

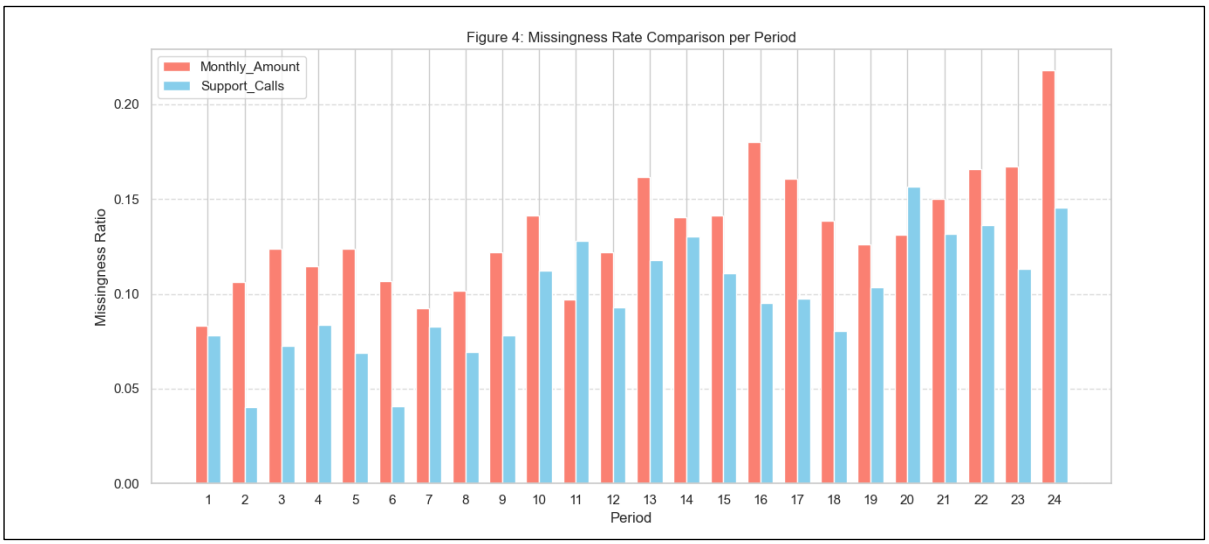
2. EDA

The objective of this step is to highlight the non-stationarity of the data. In a churn context, non-stationarity means that the data-generating process (both the features and the target) changes over time.









Analysis Pillar	Statistical Observation	Detected Phenomenon	Impact on Modeling Strategy
Figure 1: Target	Churn climbs from 19.3% to 56.0% over 24 months.	**Target Drift** $P(Y\|t)$	Static models will fail. Use **Time-Series Split** for validation and monitor baseline drift.
Figure 2: Features	Monthly_Montant shows non-linear fluctuations and growth.	**Covariate Drift** $E[X\|t]$	Requires **Period-relative scaling** (Z-score per month) to stabilize input feature range.
Figure 3: Density	Distribution in months 19-24 is shifted right vs months 1-6.	**Input Shift (OOD)** $P(X\|t)$	Older training data is "Out-of-Distribution." **Weight on recent samples** increases during training.

Figure 4: Quality	Missing data rates doubled, peaking at 21.8%.	Data Quality Drift	Reliability drops over time. Do not use mean imputation.
MNAR Analysis	Churn is +22.3% higher when data is missing.	Missing Not At Random	Create **"Missing Indicators"** to capture this signal.

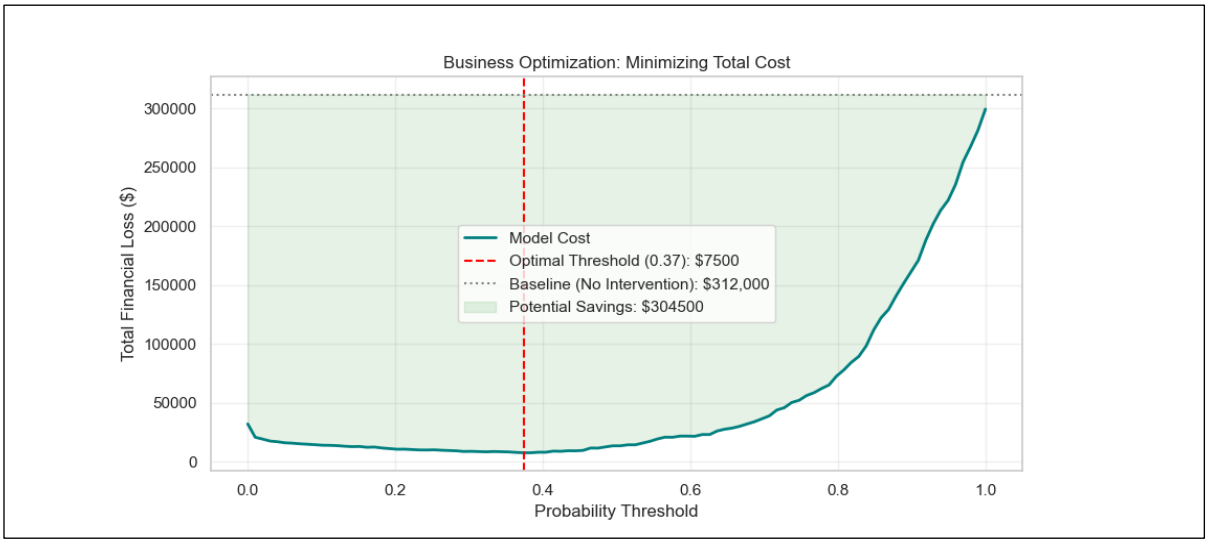
3. Business Optimization

We define a Total Economic Loss Function $L(\tau)$

$$L(\tau) = (C_{FN} \times FN_{\tau}) + (C_{FP} \times FP_{\tau})$$

Where:

- τ : The decision threshold.
- FN: A customer churns, and we did nothing (False Negative).
- FP: We offer a discount to a loyal customer (False Positive).
- C_{FN}:The cost of a False Negative.
- C_{FP}: The cost of a False Positive.



Business Decision:

- **The Strategy:** The company should target any customer with a churn probability higher than **37%**.
- **The Trade-off:** We will accept many False Alarms (customers who get a discount but aren't leaving) because catching most actual churners saves the company over **304,500$** compared to doing nothing. This optimization compensates for drift by adjusting the decision boundary toward risk-aversion.