DATA ENGINEERING ANALYSIS

FINAL PROJECT

CHURN PREDICTION & DATA DRIFT

(OFFLINE MLOps)

TEAM MEMBERS

Hariprasath SOUPRAMANIANE

Rahul BALAKRISHNAN

Shamini DJEASSEGAR

Contents

1.EXECUTIVE SUMMARY

This report details the methodology and findings of the Churn Prediction project. The primary objective is to build a robust predictive model that anticipates customer churn while proactively addressing temporal data drift. By simulating an offline MLOps environment, we established leakage-free preprocessing pipelines, dynamically evaluated retraining strategies (Fixed vs. Rolling), and finalized the model decision threshold based on business-specific costs.

2.DRIFT-ORIENTED EXPLORATORY DATA (EDA)

Before modelling, we performed a continuous monitoring-style baseline EDA to identify non -stationarity in the dataset:

2.1 Target Drift (Concept Drift):

We analysed the churn rate over successive periods, observing notable fluctuations that confirm the environment is dynamic.

2.2 Covariate Drift:

Visualizations established shifts in the distributions of input features over time. For example, the “Monthly Amount” continuously evolved, and the `Tenure` KDE distribution demonstrated noticeable differences between early and recent periods.

2.3 Missing Value Analysis:

The missingness of the “Support Calls” feature was tracked per period, revealing that missing rates fluctuated. Furthermore, missingness was strongly correlated with the `Churn` variable (indicating Missing Not at Random -MNAR), meaning the absence of data itself acts as a predictive signal.

3. PREPROCESSING PIPELINE & DATA LEAKAGE PREVENTION

To emulate a true production environment, random train/test splits were strictly avoided. Instead, the data was sequentially processed respecting the Period variable. A Scikit-Learn Pipeline and Column Transformer were utilized to apply transformations independently within each training window.

3.1 Numerical Features:

Handled using median imputation followed by standard scaling.

3.2 Categorical Features:

Processed using constant imputation (filled with the value “MISSING”) followed by One-Hot Encoding.

By fitting this pipeline strictly within each temporal training loop, we ensured a leakage-free process, where statistics from future periods never contaminated historical training batches.

4. RETRAINING POLICIES & OFFLINE VALIDATION

Two retraining strategies were evaluated using a Logistic Regression baseline to handle the observed data drift.

4.1 Fixed Strategy:

A model trained once on an initial static set of historical periods and then used to predict all future periods without retraining.

4.2 Expanding Strategy:

A model retrained at every new period using all available historical data up to that period (train on 1..t-1, test on t).

4.3 Findings:

Performance (tracked via AUC) showed that the Expanding strategy maintained slightly more stable performance over time, while the Fixed strategy gradually degraded due to data drift.

5. MODEL COMPARISON: LOGISTIC REGRESSION VS RANDOM FOREST

After identifying the optimal history length via offline back testing, we selected a 3-month rolling window strategy for adaptive retraining.

Two models were compared:

1. Logistic Regression (Linear Baseline)

2. Random Forest Classifier (Non-linear Model)

Both models were trained using the selected 3-month rolling window. The evaluation was conducted period-by-period using AUC.

Results showed:

- Logistic Regression Avg AUC ≈ 0.8707

- Random Forest Avg AUC ≈ 0.8608

Logistic Regression demonstrated slightly better consistency and was therefore selected as the final model. PREPROCESSING PIPELINE & DATA LEAKAGE PREVENTION

6.BUSINESS COST- BASED THRESHOLD OPTIMIZATION

Standard machine learning metrics (like AUC or Accuracy) do not reflect the true financial impact of churn. We optimized the final decision boundary based on asymmetric business costs:

1.False Negative Cost (C_{FN}): Cost of missing a churner = 10 units.

2. False Positive Cost (C_{FP}): Cost of proactively reaching out to a retained customer unnecessarily = 2units.

Using the final out-of-time test period, we simulated decision thresholds from `0.01` to `0.99`. By mapping the confusion matrix to the business cost formula ($Total Cost = FN \times 10 + FP \times 2$), we identified the threshold that minimized the financial impact.

- Optimal Decision Threshold: ~0.23
- Minimized Business Cost: 118

units

7. CONCLUSION

This project successfully implemented a robust offline MLOps pipeline. By respecting the temporal nature of the data, proactively handling drift through optimal retraining windows, and translating probabilistic predictions into concrete business value via cost-threshold optimization, the resulting model offers a practical and financially viable solution to customer churn prediction.