# Churn Prediction – Baseline Analysis

## 1. Objective

The objective was to develop a baseline churn prediction model using the dataset 'Test 1 – Tabela 1.xlsx' containing 5000 telecom subscribers and 12 attributes. The goal was to predict whether a customer churns (binary variable 'CHURN').

## 2. Data Overview

Records: 5000  
Initial features: 12 (11 numeric, 1 categorical)  
Excluded columns: SUBSCRIBER\_ID (identifier)  
Dropped columns (missing >40%): PROD\_CNT\_MACRO, REV\_BUN\_MAC, TOPUP\_AMT  
Target balance: 78.3% non-churn, 21.7% churn

## 3. Data Profiling (YData Profiling)

Missing data visualized via count/matrix/heatmap plots. Detected several incomplete columns with low completeness (<60%). No duplicate rows detected.  
Moderate correlations with churn:  
BNUM\_IN (0.26), BNUM\_OUT (0.22), PROD\_CNT\_MACRO (0.21), REV\_BUN\_MAC (0.18), TOPUP\_CNT (0.15).  
Potential temporal leakage suspected — features describe user activity possibly after churn event.  
Profiling report generated: churn\_profile\_report.html

## 4. Model Pipeline Setup

Preprocessing with:  
• Numerical: median imputation + standard scaling.  
• Categorical: most frequent imputation + one-hot encoding.  
Models used: Logistic Regression (max\_iter=2000) and Gradient Boosting Classifier.

## 5. Evaluation Strategy

Train/test split: 75/25, stratified.  
Metrics: Accuracy, Precision, Recall, F1, ROC AUC, Average Precision (AP).  
Visualizations: ROC Curves and Precision–Recall Curves on holdout set.

## 6. Results Summary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 | ROC AUC |
| Logistic Regression | 0.9992 | 1.0000 | 0.9963 | 0.9982 | 1.0000 |
| Gradient Boosting | 0.9992 | 1.0000 | 0.9963 | 0.9982 | 0.9999 |

Both models achieve near-perfect performance, indicating data leakage — predictors likely encode post-churn behavior.

## 7. Diagnostics and Leakage Check

Correlation inspection confirms strong behavioral predictors (calls, revenue, top-ups). Next steps:  
1. Limit features to those available before churn month.  
2. Rebuild dataset with a rolling window (features from month N predicting churn in month N+1).  
3. Add k-fold validation.  
4. Consider threshold tuning for business recall/precision trade-off.

## 8. Output Artifacts

Profiling report: churn\_profile\_report.html  
ROC plot: roc\_curves.png  
Precision–Recall plot: pr\_curves.png  
Holdout predictions: predictions\_holdout.csv

## 9. Conclusion

The baseline pipeline functions correctly (EDA → preprocessing → model → evaluation). However, extremely high metrics confirm that the data contains leakage or post-event features. The next iteration should focus on temporal validation and feature filtering to ensure predictive reliability.