Customer Churn – Model Development Report

# 1. Objective

Develop and validate a machine learning model to predict customer churn. The model is trained on the preprocessed dataset (`customer\_churn\_cleaned.csv`) and evaluated with metrics suited for imbalanced data (ROC-AUC, Precision, Recall, F1) with decision-threshold tuning.

# 2. Algorithm Selection & Rationale

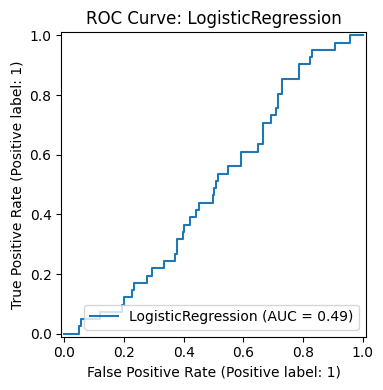
We trained two complementary models:  
• Logistic Regression (class\_weight='balanced') – interpretable, calibrated probabilities.  
• Random Forest (class\_weight='balanced') – captures non-linearities and interactions; strong baseline.  
Hyperparameters were tuned with cross-validation in the training script. Thresholds were optimized for F1.

# 3. Training & Validation

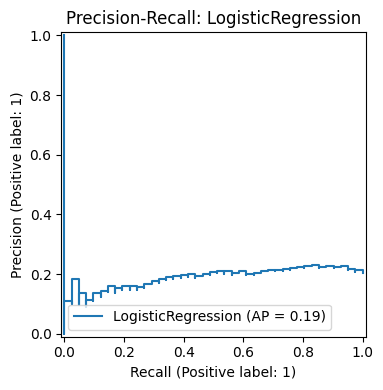
• Data split: stratified 80/20 train/test.  
• Cross-validation: 5-fold StratifiedKFold during hyperparameter search.  
• Primary metric: ROC-AUC (robust to class imbalance). Secondary: Precision, Recall, F1, Accuracy.  
• Imbalance handling: class\_weight='balanced'.  
• Threshold optimization: sweep 0.10–0.90 to maximize F1.

# 4. Performance & Evaluation

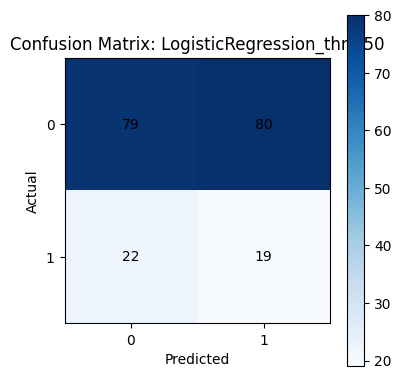
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | threshold | roc\_auc | precision | recall | f1 | accuracy |
| LogisticRegression | 0.5000 | 0.4875 | 0.1919 | 0.4634 | 0.2714 | 0.4900 |
| RandomForest | 0.5000 | 0.5145 | 0.2222 | 0.0488 | 0.0800 | 0.7700 |
| LogisticRegression\*bestT | 0.4600 | 0.4875 | 0.2229 | 0.9512 | 0.3611 | 0.3100 |
| RandomForest\*bestT | 0.2000 | 0.5145 | 0.2135 | 1.0000 | 0.3519 | 0.2450 |



ROC Curve



Precision–Recall Curve



Confusion Matrix (best threshold)

# 5. Business Utilisation

• Rank customers by churn probability; action top deciles.  
• Tailor interventions: low login frequency → engagement nudges; unresolved cases → service outreach; high-value at risk → targeted offers.  
• Monitor outcomes and iterate thresholds using campaign response data.

# 6. Potential Improvements

• Evaluate Gradient Boosting (XGBoost/LightGBM) for possible ROC-AUC gains.  
• Use cost-sensitive thresholding based on offer cost vs. retained value.  
• Add temporal features (tenure, trend of spend/engagement), probability calibration, and drift monitoring.

# 7. Artifacts & Reproducibility

**Trained models:**

model\_outputs/\*.pkl (latest RandomForest or LogisticRegression).

**Metrics & plots:**

model\_outputs/metrics.csv, metrics.json, roc\_\*.png, pr\_\*.png, cm\_\*.png

**Data:**

customer\_churn\_cleaned.csv