Telecom Customer Churn Analysis – Project Report

Abstract

This project develops a data-driven approach to predict telecom customer churn, explain key drivers, and recommend targeted retention strategies. Using six months of synthetic customer interactions (usage, recharges, complaints, and payments), we trained classification models and produced actionable insights and segments to guide win-back and loyalty initiatives.

Introduction

Customer churn directly impacts revenue and acquisition costs. The objective is to (1) predict churn propensity, (2) understand the drivers behind churn risk, and (3) translate findings into practical retention playbooks aligned to distinct customer segments.

Tools Used

- Languages & Libraries: Python (pandas, numpy, scikit-learn, xgboost, seaborn, matplotlib, shap, eli5)

- Notebooks & Reporting: JupyterLab, python-pptx

- Artifacts: metrics_summary.json, scored CSV with churn_prob & segment, visuals (ROC/CM/SHAP), PPT report

Steps Involved in Building the Project

1. Data Generation & Preparation

- Created a synthetic yet realistic dataset for 10,000 customers with 6 months of interactions (calls, recharges, complaints, payments) and demographics.

- Engineered features: recency (days since last call/recharge/complaint/bill), usage_trend, complaint_intensity, recharge_consistency, roaming/weekend patterns.

2. Modeling & Validation

- Trained RandomForest, LogisticRegression (best), and optional XGBoost with Stratified K-fold and GridSearchCV.

- Selected the best model on ROC-AUC; saved metrics and artifacts for reporting.

3. Explainability & Insights

- Used SHAP/ELI5 to identify key drivers (recency, recharge frequency/value, complaint intensity, on-time pay ratio) and provide interpretable explanations.

- Generated evaluation visuals: ROC curve and confusion matrix.

4. Segmentation & Targeting

- Scored all customers to compute churn_prob.

- Applied business rules to form segments (At Risk, Loyal, Dormant, Needs Attention) for tailored interventions.

5. Reporting & Deliverables

- Produced a PowerPoint (executive-ready), a scored CSV, and this concise report.

Results (Key Numbers)

- Best Model: LogisticRegression

- ROC-AUC: 0.655 | F1: 0.081 | Accuracy: 0.636
- Segment distribution (N = 10,000): At Risk 257 | Loyal 2,328 | Dormant 0 | Needs Attention 7,415

Conclusion

The solution prioritizes early identification of churn via recency and behavior features, provides clear drivers for stakeholder trust, and translates insights into targeted retention playbooks. Operationalizing this pipeline with real labels and CRM integration can reduce churn, improve ARPU, and enhance customer experience. Future iterations should (1) replace synthetic labels with production churn outcomes, (2) evaluate threshold tuning to optimize precision/recall for campaigns, and (3) add model monitoring for drift and calibration.