PROJECT REPORT: TELECOM CUSTOMER CHURN PREDICTION

SUBMITTED BY: KRISHNENDHU P A

1. Introduction

In the highly competitive telecommunications sector, retaining existing customers is significantly more cost-effective than acquiring new ones. This project focuses on analysing customer behaviour and service usage logs to predict the likelihood of churn. By identifying at-risk customers before they leave, the business can deploy targeted retention strategies to safeguard recurring revenue.

2. Abstract

This project implements a comprehensive data science pipeline to solve the churn problem using a real-world dataset of telecom subscribers. The approach integrates SQL for robust data engineering, XGBoost for high-performance predictive modeling, and SHAP for model explainability. Key findings revealed that customer complaints are the strongest predictors of churn, while high usage frequency—contrary to initial hypotheses—actually correlates with higher loyalty unless paired with service failures. The final model achieved an F1-Score of 0.80, providing a reliable foundation for automated retention triggers.

3. Tools Used

The following tools and technologies were utilized to build the end-to-end solution:

SQL (SQLite): Used for data ingestion, schema normalization, and complex feature engineering.

Python: The primary programming language for the analysis and modelling pipeline.

Pandas & NumPy: For data manipulation and numerical processing.

XGBoost: The machine learning algorithm used for binary classification.

Scikit-learn: For model evaluation, metrics, and dataset splitting.

SMOTE (Imbalanced-Learn): To handle the significant class imbalance between churned and loyal customers.

SHAP: For interpretability and identifying feature importance.

4. Steps Involved in Building the Project

The project was executed in four distinct phases:

Phase 1: Data Engineering & SQL Integration

The raw dataset was loaded into a local SQLite database to demonstrate production-level data handling.

Columns were normalized to snake case, and a specialized "High Value at Risk" feature was created using SQL logic to flag top-tier users with recent complaints.

Phase 2: Exploratory Data Analysis (EDA) & Preprocessing

Performed analysis on usage metrics such as seconds_of_use and call_failure.

Addressed class imbalance using SMOTE, ensuring the model could learn patterns from the minority "Churn" class effectively.

Phase 3: Predictive Modeling

Trained an XGBoost Classifier on the enriched dataset.

Evaluated performance using a confusion matrix, achieving 85% accuracy and balanced precision/recall metrics.

Phase 4: Explainability & Strategy Development

Applied SHAP to decode the model's internal logic.

Segmented users into "At Risk," "Loyal," and "Dormant" categories to provide actionable business recommendations.

5. Conclusion

The analysis successfully moved the churn problem from a "black box" prediction to an actionable business strategy. We concluded that while high usage volume indicates loyalty, it does not immunize a customer against churn if service quality (measured by complaints and call failures) is poor. By implementing the proposed "Service Recovery Protocol" for high-value complainers, the organization can proactively reduce churn and protect its most valuable revenue streams.