Customer Churn Prediction – Summary Report

Author: Kidima Medy Masuka

Aspiring Data Scientist | Data Analyst

Date: 2026

1. Problem Statement

Customer churn poses a significant financial risk for subscription-based businesses. The objective of this project was to develop a predictive model capable of identifying customers at risk of churning, enabling proactive retention strategies.

2. Data Overview

The project uses a simulated customer churn dataset sourced from Kaggle. Although not based on real company data, the dataset reflects realistic customer behaviour patterns commonly observed in subscription-based services.

3. Methodology

- A structured data science workflow was followed:
- Data cleaning and exploratory analysis
- Feature engineering and preprocessing
- Train-test split to avoid data leakage
- Scaling and encoding using Column Transformer
- Model training and optimisation using GridSearchCV
- ROC-AUC was used for hyperparameter optimisation due to class imbalance, while recall and F1-score were prioritised during evaluation.

4. Model Evaluation

Three models were evaluated:

- Logistic Regression
- Random Forest
- XGBoost

XGBoost achieved the strongest overall performance, with the highest recall, F1-score, and ROC-AUC. Random Forest achieved the highest precision, while Logistic Regression showed limited churn detection capability in comparison.

5. Business Impact

Assuming a customer base of 5,000 with a 20% churn rate:

Approximately 1,000 customers are expected to churn.

The XGBoost model correctly identifies around 430 of these customers.

Retaining 50% of identified churners could save approximately 215 customers.

At an estimated value of R2,000 per customer per year, this results in potential annual savings of R430,000.

6. Limitations and Future Work

The dataset is simulated and does not incorporate customer lifetime value. Future enhancements could include CLV-based prioritisation, decision-threshold optimisation, and deployment within a real-time decision-support system.

7. Conclusion

This project demonstrates an end-to-end churn prediction pipeline with a strong emphasis on correct methodology, transparent evaluation, and business-driven decision-making. Rather than optimising for accuracy alone, model selection was aligned to minimise customer loss.

Usage & Attribution

This project is shared for educational and portfolio purposes.

If reused or adapted, appropriate credit must be given to the author.

📰This project is part of my personal data science portfolio ✅