Customer Churn Prediction Report

Mohamed Kasm Mohamed Saqr Zeyad

August 28, 2025, 06:50 PM EEST

Abstract

This report presents a comprehensive data science project to predict customer churn in a telecom company using the Telco Customer Churn dataset from Kaggle. The analysis follows the full data science life cycle, including data collection, exploration, cleaning, preprocessing, visualization, feature engineering, model building, optimization, and interpretation. We evaluated Logistic Regression, Random Forest, and XGBoost models, with the tuned XGBoost achieving the best performance (0.85 ROC-AUC). Key findings indicate that specific contract types, internet service options, and payment methods drive churn. Recommendations include targeting short-term contract customers and enhancing service bundles to reduce churn.

1 Executive Summary

This project aimed to predict customer churn for a telecom company to enable proactive retention strategies. Using the Telco Customer Churn dataset (7,043 records, 21 features), we followed a structured data science pipeline:

- **Dataset**: 7,043 customers, 26.5% churn rate (1,869 churned, 5,174 retained).
- **Models**: Logistic Regression, Random Forest, XGBoost (default and tuned).
- **Best Model**: Tuned XGBoost (0.85 ROC-AUC, 81% accuracy).
- **Key Features**: Contract types, internet service, payment methods.
- **Recommendations**: Focus retention on short-term contract customers, offer bundled services.

The tuned XGBoost model provides a reliable tool for identifying at-risk customers, potentially saving millions in acquisition costs by reducing churn.

2 Introduction and Goal

The goal was to predict whether a telecom customer will churn (leave the service) based on demographic, service, and billing data. Churn prediction enables targeted retention strategies, as acquiring new customers costs 5–6 times more than retaining existing ones. The project followed the data science life cycle:

1. Data source & collection
2. Data exploration (EDA)
3. Data cleaning & preprocessing
4. Data visualization
5. Data processing & feature engineering

6. Machine learning (model building & evaluation)
7. Model optimization (hyperparameter tuning)
8. Results & conclusions

3 Data Source and Collection

- **Dataset**: Telco Customer Churn (Kaggle).
- **Source**: `https://www.kaggle.com/datasets/blastchar/telco-customer-churn`.
- **Size**: 7,043 rows, 21 columns.
- **Features**:
  - *Demographics*: gender, SeniorCitizen, Partner, Dependents.
  - *Services*: PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies.
  - *Billing*: tenure, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges.
  - *Target*: Churn (Yes/No).

A sample of the dataset is shown in Table 1.

Table 1: Sample of Telco Customer Churn Dataset (First 3 Rows)

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	Inter
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	
5575-GNVDE	Male	0	No	No	34	Yes	No	
3668-QPYBK	Male	0	No	No	2	Yes	No	

4 Data Exploration (EDA)

- **Data Types**: Mix of categorical (e.g., gender, InternetService) and numeric (e.g., tenure, MonthlyCharges).
- **Statistics**: Tenure (0–72 months, mean 32), MonthlyCharges ($18–$118, mean $64), TotalCharges (some blanks).
- **Target Distribution**: Imbalanced – 73.5% No Churn (5,174), 26.5% Yes Churn (1,869).
- **Issues**: TotalCharges stored as string with 11 missing values; no other major missing data.

5 Data Cleaning and Preprocessing

- Converted TotalCharges to numeric, imputed missing values with median (1,395 USD).
- Dropped irrelevant customerID.
- Separated features: 3 numeric (tenure, MonthlyCharges, TotalCharges), 16 categorical.
- Mapped Churn to binary (1=Yes, 0=No).

No significant outliers or duplicates were identified.

6 Data Visualization

Visualizations provided insights into churn patterns (details omitted for brevity; refer to original notebook).

7 Data Processing and Feature Engineering

- **Pipeline**:
  - Numeric: StandardScaler for normalization.
  - Categorical: OneHotEncoder (drop first to avoid multicollinearity).

- **Feature Engineering**: Created tenure$_{g}roupbins(0--12, 12--24, etc.)$.**Train-Test Split** $: 80/20stratifiedsplittomo$

8 Machine Learning (Model Building and Evaluation)

Three models were evaluated. Metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC (details omitted for brevity; refer to original report).

9 Model Optimization (Hyperparameter Tuning)

XGBoost was optimized using RandomizedSearchCV (details omitted; refer to original report).

Tuned XGBoost Performance:

- Accuracy: 81%
- Precision: 67%
- Recall: 54%
- F1-Score: 60%
- ROC-AUC: 0.85

10 Results and Conclusions

- **Model Comparison**: Tuned XGBoost outperformed others (0.85 ROC-AUC).
- **Feature Importance** (Tuned XGBoost, Table 2):
  - Top features reflect contract stability, service type, and payment preferences.
- **Business Insights**:
  - Target customers on short-term contracts with retention offers.
  - Enhance service bundles for specific internet types.
  - Monitor payment method trends.
- **Limitations and Next Steps**:
  - Explore imbalance handling (e.g., SMOTE).
  - Add temporal features.
  - Deploy with threshold tuning.

Table 2: Top 5 Feature Importances (Tuned XGBoost)

Rank	Feature	Importance
1	Internet Service No	0.20
2	Contract Two Year	0.18
3	Internet Service Fiber Optic	0.15
4	Contract One Year	0.13
5	Payment Method Electronic Check	0.11

11 Acknowledgments

This project was conducted using Python, scikit-learn, XGBoost, and visualization libraries. The dataset was sourced from Kaggle.