Telecom Customer Churn Prediction

Machine Learning Project Report

Author: Victor Prefa

Credentials: Medical Doctor | MSc Data Science & Business Analytics

GitHub: github.com/23396784/telecom-customer-churn-prediction

Executive Summary

This project develops machine learning models to predict customer churn in the telecommunications industry. Using a dataset of 3,333 customer records with 21 features, we trained and evaluated three classification models: Decision Tree, Random Forest, and AdaBoost. The Random Forest model achieved the highest accuracy of 85.71%, while AdaBoost demonstrated better recall for identifying at-risk customers. Key predictors of churn include Day Minutes, Customer Service Calls, and Day Charges.

1. Business Problem

Customer churn is a critical challenge for telecom companies, as acquiring new customers costs 5-25 times more than retaining existing ones. This project aims to:

• Identify customers at high risk of churning

• Understand the key factors that drive customer churn

• Enable proactive retention strategies through predictive modeling

2. Dataset Overview

The dataset contains 3,333 observations and 21 variables representing customer behavior and account information.

Category	Features

Usage Metrics	DayMins, EveMins, NightMins, IntlMins, DayCalls, EveCalls, NightCalls, IntlCalls
Billing Features	DayCharge, EveCharge, NightCharge, IntlCharge
Account Info	AccountLength, VMailMessage, IntlPlan, VMailPlan, CustServCalls
Identifiers	State, AreaCode, Phone
Target Variable	Churn (Yes/No)

Churn Distribution: 85.5% Retained | 14.5% Churned

3. Methodology

Data Processing Pipeline:

1. **Data Loading:** Loaded dataset using PySpark DataFrame for scalable processing

2. **Preprocessing:** Handled missing values, encoded categorical variables

3. **EDA:** Generated visualizations including histograms, count plots, and correlation matrix

4. **Model Training:** Applied Decision Tree, Random Forest, and AdaBoost classifiers

5. **Evaluation:** Assessed models using accuracy, precision, and recall metrics

Tools Used: PySpark 4.0.1, Python 3.x, Scikit-learn, Pandas, Matplotlib, Seaborn

Data Split: 80% Training (2,666 samples) | 20% Testing (667 samples)

4. Key Findings from Exploratory Data Analysis

High Churn Indicators:

• **Customer Service Calls:** Customers with 4+ service calls have significantly higher churn rates, indicating unresolved issues leading to dissatisfaction.

• **Day Minutes Usage:** High daytime usage correlates with higher charges and increased churn probability.

• **International Plan:** Subscribers to international plans show higher churn rates, suggesting pricing or service quality concerns.

5. Model Performance Results

Model	Accuracy	Precision	Recall
Decision Tree	78.57%	40%	40%
Random Forest	85.71%	100%	20%
AdaBoost	82.14%	50%	40%

Model Selection Insights:

• **Random Forest:** Achieved highest overall accuracy (85.71%) with perfect precision. Best for minimizing false positives in retention campaigns.

• **AdaBoost:** Better recall means catching more actual churners. Recommended when the cost of missing a churner is high.

6. Top Predictors of Churn

Rank	Feature	Importance	Business Insight
1	Day Minutes	HIGH	High usage leads to higher bills and churn
2	Customer Service Calls	HIGH	4+ calls indicate unresolved issues
3	Day Charge	MEDIUM	Directly correlated with day minutes
4	International Plan	MEDIUM	Plan subscribers show higher churn
5	Evening Minutes	LOW	Less impact on churn decisions

7. Business Recommendations

Priority 1 - Proactive Service Intervention: Flag customers after their 3rd service call for manager callback and issue resolution before reaching the critical 4+ call threshold.

Priority 2 - High-Usage Loyalty Programs: Offer discounted day-time rates or unlimited plans to heavy users before they seek alternatives.

Priority 3 - International Plan Review: Audit international plan pricing and quality. Higher churn among international subscribers indicates competitive disadvantage.

Priority 4 - Predictive Retention System: Deploy Random Forest model in production to score customers monthly and trigger automated retention workflows.

8. Conclusion

This project successfully developed machine learning models to predict telecom customer churn with up to 85.71% accuracy. The analysis identified key predictors including Day Minutes, Customer Service Calls, and Day Charges. By implementing the recommended retention strategies and deploying the predictive model, telecom companies can significantly reduce churn rates and improve customer lifetime value. Reducing churn by just 1% can save millions annually in customer acquisition costs.

Project Repository: github.com/23396784/telecom-customer-churn-prediction

Contact: Victor Prefa | LinkedIn: linkedin.com/in/victor-prefa-99810929