Predictive Modeling for Customer Retention

Herik Patel
Jenish Patel
Rutgers University

October 31, 2025

Abstract

For subscription-based businesses, customer churn represents a significant threat to revenue and profitability. This project aims to predict which customers are likely to cancel their subscriptions using predictive modeling and interpretable machine learning methods. We develop and evaluate multiple classification models, focusing on achieving both high predictive accuracy (AUC-ROC > 0.85) and actionable insights that can guide retention strategies.

1 Background / Problem Statement

What problem are you trying to solve? Customer churn—the rate at which subscribers discontinue a service—is a critical and costly problem for any subscription-based business. Losing a customer not only eliminates future revenue but also incurs additional marketing expenses for acquiring replacements. The central challenge lies in identifying customers at high risk of churning before they take action. Current retention approaches tend to be reactive, intervening only after dissatisfaction is expressed.

Why is it important or interesting? Retention is substantially more cost-effective than acquisition. By accurately predicting churn risk, companies can proactively implement targeted interventions such as discounts, personalized support, or product recommendations. This project is particularly interesting because it transitions from descriptive analytics to predictive analytics, directly improving customer lifetime value and business efficiency.

2 Introduction / Goal

What are you planning to build or study? We plan to build a high-accuracy classification model that assigns each active customer a probability of churn within the next 30 days.

Main Objective

- **Prediction:** Develop a Gradient Boosting Machine (XGBoost) that achieves an AUC-ROC above 0.85 on unseen data.
- **Actionability:** Identify and rank key features (e.g., tenure, service usage, payment method) to provide actionable insights for business teams.

3 Dataset

Source and Description We will use the publicly available *Telco Customer Churn Dataset* from Kaggle, which contains roughly 7,000 customer records. Each record includes demographic, account, and service-related features.

Attributes

- **Demographics:** Gender, Senior Citizen, Partner, Dependent
- **Account Information:** Tenure, Contract type, Payment method, Monthly Charges, Total Charges
- **Services:** Internet type, Security, Streaming, Tech Support
- **Target:** Binary variable `Churn`

4 Method / Approach

Modeling Strategy We will implement a three-phase pipeline: Preprocessing, Modeling, and Tuning/Interpretation.

4.1 Phase 1: Baseline Modeling

A simple Logistic Regression will serve as the baseline for performance comparison.

4.2 Phase 2: Advanced Modeling

We will train ensemble models including Random Forest and XGBoost, as boosting methods often perform best for structured data.

4.3 Implementation Details

- **Data Cleaning:** Impute or remove missing Total Charges values; one-hot encode categorical variables.
- **Feature Engineering:** Create behavioral ratios (e.g., Monthly Charges / Tenure) to capture spending trends.
- **Training:** Use an 80/20 train-validation split with K-Fold cross-validation for robustness.
- **Hyperparameter Tuning:** Perform Grid or Randomized Search to optimize performance metrics.

5 Evaluation

Performance Measurement Model performance will be compared against the baseline Logistic Regression and the target AUC-ROC threshold of 0.85.

Metrics

- **AUC-ROC:** Primary metric for discrimination performance.
- **Recall:** Sensitivity in detecting actual churners.
- **Precision:** Accuracy of churn predictions.
- **F1-Score:** Balanced metric combining precision and recall.
- **Confusion Matrix:** Visual breakdown of predictions.

6 Related Work

Previous studies have used several approaches for churn prediction:

- **Survival Analysis:** Modeling time-to-event (customer lifetime).
- **Statistical Classifiers:** SVMs, logistic regression, and simple neural networks.
- **Ensemble Methods:** Random Forest and Gradient Boosting for state-of-the-art accuracy.

Novelty and Contribution Our approach focuses on interpretability and business applicability, not just accuracy. Using SHAP values, we provide transparent, human-readable explanations such as: "Customer X is high risk due to short tenure, month-to-month contract, and high charges." This bridges the gap between prediction and actionable strategy.

7 Conclusion

By combining interpretable machine learning with robust predictive modeling, this project aims to create a practical and accurate churn prediction tool. The ultimate goal is to support proactive, data-driven retention strategies that improve both customer experience and business profitability.