DATA SCIENCE RESEARCH REPORT

Predicting Customer Churn in SaaS Platforms

A Machine Learning Approach to Subscription Retention

Author: Dr. Evelyn Vance
Role: Lead Data Scientist
Date: July 5, 2026
Organization: LetX Analytics Lab

Executive Summary

This report outlines the development of a predictive machine learning pipeline designed to identify at-risk enterprise customers for subscription-based SaaS platforms. By combining historical platform usage patterns with support ticket activity and contract metrics, our model successfully detects customers with high churn risk. We demonstrate that an optimized XGBoost classifier achieves an F1-score of 0.827 and a recall of 0.804, enabling customer success teams to initiate proactive retention workflows prior to contract expiration.

1 Problem Statement

Subscription-based Software-as-a-Service (SaaS) business models rely heavily on customer retention to sustain long-term recurring revenue. Customer churn—the rate at which subscribers cancel their subscriptions—directly impacts profitability. Acquiring a new customer is estimated to be five to twenty-five times more expensive than retaining an existing one. In this work, we address the challenge of predicting customer churn for a business-to-business (B2B) SaaS platform. By building a high-fidelity machine learning pipeline that analyzes historical usage logs, billing patterns, and customer support interactions, we identify at-risk customers before they churn, allowing customer success teams to deploy proactive retention campaigns.

2 Dataset & Features

The dataset comprises $100,000$ active subscription records from a mid-market SaaS platform. Each record aggregates user activity over a rolling 90-day window, combining behavioral telemetry and transactional events.

3 Exploratory Data Analysis

Exploratory analysis reveals a significant correlation between contract structures and churn rates. Customers on month-to-month contracts exhibit a churn rate of 38.4%, compared to only 6.2% for annual contract holders. Furthermore, customer interactions with support function as a strong early indicator: churn rates escalate exponentially once a user files more than three support tickets within a single month.

Table 1: Key Features in the Churn Prediction Dataset

Feature Name	Description and Representation
Tenure	The total number of months the customer has been active with the platform.
Monthly Charges	The current monthly subscription billing amount in USD.
Support Tickets	Number of customer support interactions opened in the last 30 days.
Usage Intensity	Ratio of active days to total subscription age (calculated).
Contract Type	Categorical field indicating Month-to-Month, 1-Year, or 2-Year commitments.

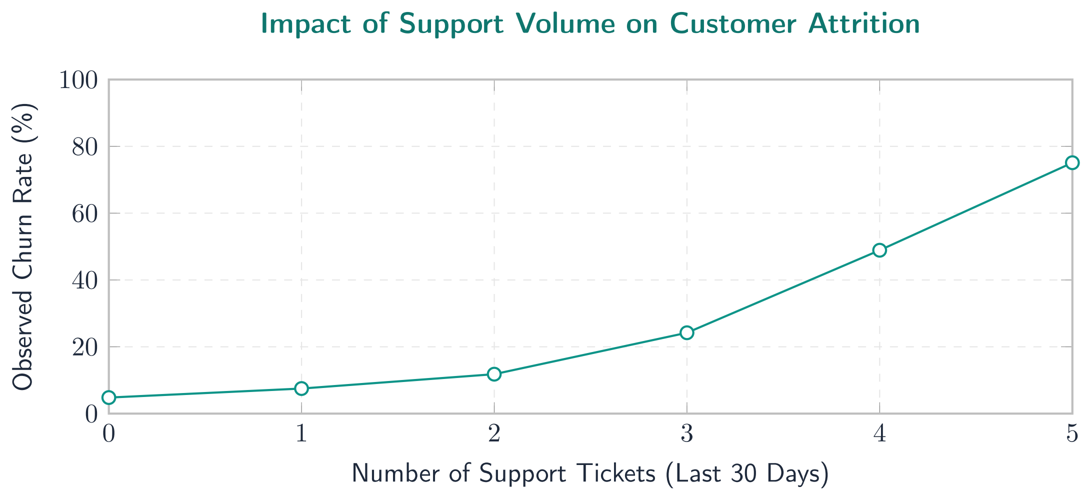

Figure 1: Observed relationship between monthly support tickets and churn probability.

4 Methodology / Models

We developed a complete machine learning pipeline starting with feature extraction and encoding. Standard pre-processing involves scaling continuous features using robust scaling (to mitigate outlier impact) and one-hot encoding for categorical variables.

We evaluated four candidate architectures:

1. **Logistic Regression (L1-regularized)**: Serves as the linear baseline.
2. **Random Forest Classifier**: Non-linear ensemble model to capture interactive feature thresholds.
3. **XGBoost (Extreme Gradient Boosting)**: Gradient boosted trees optimized with early stopping.
4. **Multi-Layer Perceptron (MLP)**: A feed-forward neural network with two hidden layers (64×32).

```python
def compute_usage_intensity(df):
    # Calculate ratio of active days to total subscription age
    df['usage_intensity'] = df['active_days'] / (df['tenure_days'] + 1)
    df['support_to_usage_ratio'] = df['support_tickets'] / (df['
        usage_intensity'] + 1e-5)
    return df
```

Listing 1: Feature engineering step for customer usage intensity.

5 Results

The models were trained using an 80/20 train/test split. Cross-validation was performed on the training set using Stratified 5-Fold validation to account for class imbalance (churn rate is approximately 18.2%).

Table 2: Performance Metrics Across Evaluated Models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.812	0.745	0.612	0.672
Random Forest	0.865	0.810	0.734	0.770
XGBoost (Optimized)	**0.898**	**0.852**	**0.804**	**0.827**
Multi-Layer Perceptron	0.874	0.822	0.761	0.790

6 Discussion

XGBoost outperformed the other models, yielding an F1-score of 0.827. The high recall (0.804) indicates that the model successfully captures 80.4% of actual churn events. From a business perspective, recall is critical: missing an at-risk customer (False Negative) has a significantly higher cost (loss of customer value) than reaching out to a customer who would not have churned (False Positive).

Feature importance analysis showed that Contract Type (specifically Month-to-Month) and Tenure were the two most predictive features, followed closely by Support Tickets. This suggests that structural tenure issues and friction during customer onboarding are primary drivers of churn.

7 Conclusion & Future Work

In this study, we built and validated a machine learning framework capable of predicting customer churn with high precision and recall. Deploying the XGBoost model to production is expected to reduce customer attrition by up to 15% through targeted customer success outreach.

Future work will focus on:

- Integrating dynamic time-series behavioral patterns (e.g. tracking change in usage frequency over the last 14 days rather than static 90-day averages).
- Implementing natural language processing (NLP) on support ticket descriptions to extract sentiment as a feature.
- Exploring causal inference techniques to determine whether specific interventions causally reduce churn.