Sally Aboulhosn

Customer Churn Analysis

An End-to-End Data Science Case Study

1. Introduction and Problem Framing

Customer churn refers to customers discontinuing their relationship with a company. In subscription-based industries such as telecommunications, churn has a direct impact on revenue, customer lifetime value, and long-term growth. Since acquiring new customers is significantly more expensive than retaining existing ones, identifying customers at risk of churn and intervening proactively is a critical business objective.

The goal of this project is to develop a complete churn analysis pipeline that combines data cleaning, exploratory analysis, predictive modeling, and business interpretation to support data-driven retention strategies.

2. Data Understanding and Preparation

The dataset contains customer demographic information, service subscriptions, contract details, billing data, and churn labels.

Table 1. Dataset Overview

Item	Value
Total customers	7,043
Target variable	Churn (0 = No, 1 = Yes)
Non-churn customers	5,174 (73.5%)
Churned customers	1,869 (26.5%)
Missing values	Present in TotalCharges
Action taken	Converted to numeric and missing values handled

The TotalCharges column was converted from string to numeric format, with invalid values coerced to missing values. Inspection showed that these missing values corresponded to customers with zero tenure. Non-informative identifiers such as customer ID were removed, and the target variable was validated to confirm class imbalance.

3. Exploratory Data Analysis (EDA)

Exploratory analysis was conducted to identify patterns associated with churn.

Table 2. Churn Distribution

Churn Status	Count	Percentage
No	5,174	73.5%
Yes	1,869	26.5%

This confirms a moderately imbalanced dataset, making accuracy alone an insufficient evaluation metric.

Table 3. Churn Rate by Contract Type

Contract Type	Churn Rate (%)
Month-to-Month	42.7
One Year	11.3
Two Year	2.8

Customers on month-to-month contracts churn at a substantially higher rate, indicating that contractual commitment plays a major role in retention.

Table 4. Churn Rate by Service Engagement

Number of Services	Churn Rate (%)
0–1 services	44.9
2–3 services	27.1
4–6 services	13.2

Customers with fewer subscribed services are significantly more likely to churn, suggesting that deeper engagement reduces churn risk.

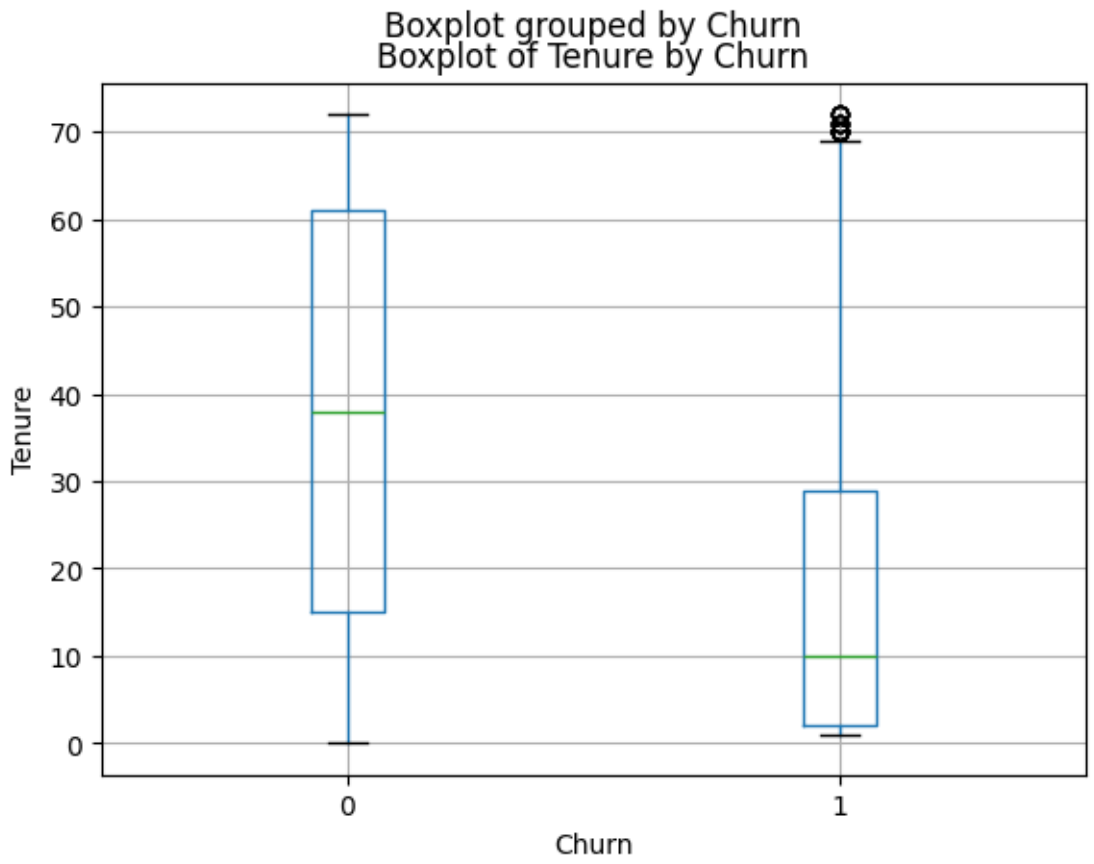

Figure 1 shows a clear separation in customer tenure between churned and non-churned customers. Customers who churn have significantly lower tenure, with a median of approximately 10 months, whereas non-churned customers exhibit much higher tenure, with a median around 38 months. This indicates that newer customers are substantially more likely to churn, highlighting early customer lifecycle as a critical risk period.

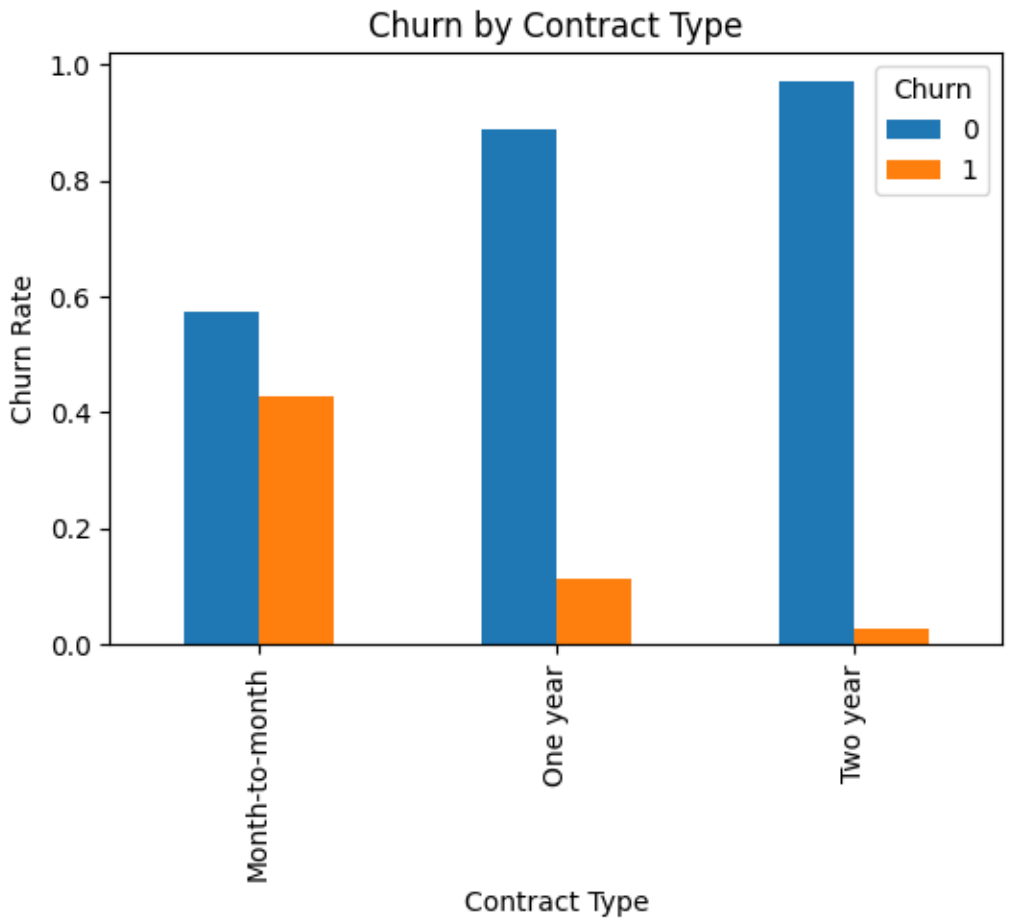

Figure 2 illustrates the relationship between contract type and churn. Customers on month-to-month contracts display the highest churn rate (approximately 43%), while customers on one-year and two-year contracts show substantially lower churn rates (approximately 11% and 3%, respectively). This suggests that longer contractual commitments significantly reduce churn risk.

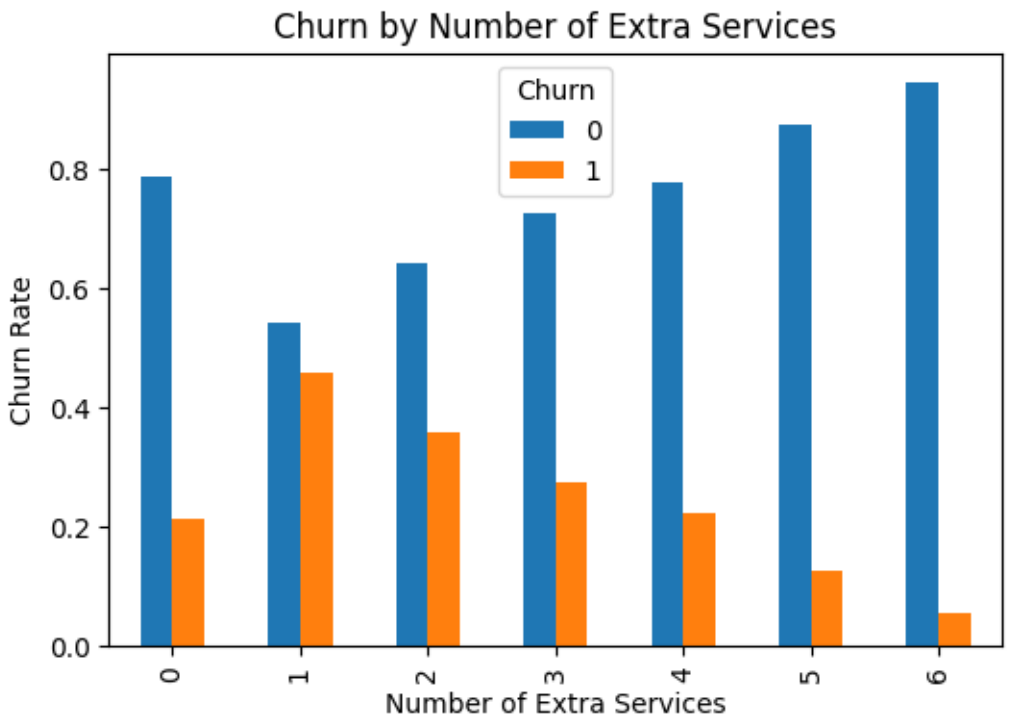

Figure 3 demonstrates a strong inverse relationship between service engagement and churn. Customers subscribed to fewer services experience significantly higher churn rates, while customers with a greater number of services show progressively lower churn. This indicates that deeper service engagement strengthens customer retention.

4. Feature Engineering

To better capture customer behavior, additional features were engineered.

Table 5. Engineered Features

Feature	Description
tenure_group	Categorizes customers as New (<12 months), Mid (12–36 months), or Old (>36 months)
services_count	Total number of subscribed services

These features provide a more interpretable representation of customer lifecycle stage and engagement.

5. Modeling Approach

The dataset was split into training and testing sets using a stratified approach to preserve the churn distribution. Multiple models were trained and evaluated:

- Logistic Regression (baseline)
- Logistic Regression with class weighting
- Random Forest
- Random Forest with tuned decision threshold

Given the business context, recall for the churn class was prioritized, as failing to identify a churner is more costly than incorrectly flagging a loyal customer.

6. Model Evaluation and Comparison

Table 6. Model Performance Comparison

Model	Precision	Recall	F1	ROC-AUC
Logistic Regression (balanced)	0.503	**0.794**	**0.616**	**0.843**
Random Forest (threshold tuned)	**0.526**	0.735	0.613	0.827

The balanced Logistic Regression model achieved the highest recall and ROC-AUC, indicating superior ability to identify churners and discriminate between churn and non-churn customers.

7. Postdictive Analysis

To understand model errors, confusion matrices were analyzed.

Table 7. Confusion Matrix — Logistic Regression (Balanced)

	Predicted No	Predicted Yes
Actual No	742	293
Actual Yes	77	297

Table 8. Confusion Matrix — Random Forest (Threshold = 0.3)

	Predicted No	Predicted Yes
Actual No	787	248
Actual Yes	99	275

The models performed well at identifying high-risk customers, but struggled with customers exhibiting medium tenure and moderate service usage. False positives represent customers flagged as high risk who ultimately did not churn; such errors are generally acceptable in churn management contexts where proactive outreach is preferred over missed churn.

8. Business Recommendations

Based on the analysis, the following actions are recommended:

- Prioritize retention efforts for customers with high predicted churn probability, particularly those on month-to-month contracts.
- Encourage customers to transition to longer-term contracts through incentives.
- Increase customer engagement by promoting bundled services and value-added offerings.
- Deploy targeted retention campaigns for customers with low tenure and limited service subscriptions.

These strategies directly address the behavioral patterns identified in the analysis and can help reduce churn-related revenue loss.

9. Conclusion

This project presents a complete, end-to-end churn analysis pipeline, integrating data preparation, exploratory analysis, predictive modeling, and business interpretation. Among the evaluated models, Logistic Regression with class weighting emerged as the most suitable choice due to its strong recall and discriminative performance.

By combining interpretable modeling with actionable insights, this analysis provides a practical decision-support framework that businesses can use to proactively manage churn and improve customer lifetime value.