Final Report

Netflix Customer Churn:

Identifying the Behavioral Drivers of User Retention

Executive summary

This project analyzed Netflix customer behavior to identify drivers of churn and develop predictive models. While baseline models struggled, Random Forest achieved the highest PR AUC (0.241), followed closely by Logistic Regression (0.218). XGBoost underperformed compared to Random Forest and Logistic Regression (PR AUC = 0.185). Across all models, recall and precision remain limited, underscoring the need for richer behavioral features.

Engagement, discovery, satisfaction and early tenure emerged as the strongest behavioral drivers of churn. Based on these insights, we recommend strengthening onboarding, monitoring satisfaction proactively and focusing retention spend on high-value at-risk users.

1. Introduction

Netflix faces a persistent challenge with customer churn that directly erodes recurring revenue and limits upselling opportunities. Nowadays, the streaming industry has become increasingly competitive, driving up acquisition costs while consumers experience growing subscription fatigue. In this environment, retaining existing subscribers (particularly those with high lifetime value) proves far more cost-effective than constantly acquiring new customers.

Without clear insights into which behaviors distinguish loyal users from those likely to cancel, retention campaigns risk being overly broad and ineffective. Fortunately, Netflix's comprehensive behavioral dataset, encompassing watch history, search patterns, reviews, and recommendation interactions, presents a valuable opportunity to conduct an in-depth churn analysis.

This project pursues three primary objectives: developing a churn prediction model able to identify at-risk users despite significant class imbalance, uncovering the key behavioral drivers that differentiate retention from departure and delivering actionable recommendations to implement targeted strategies.

2. Data wrangling

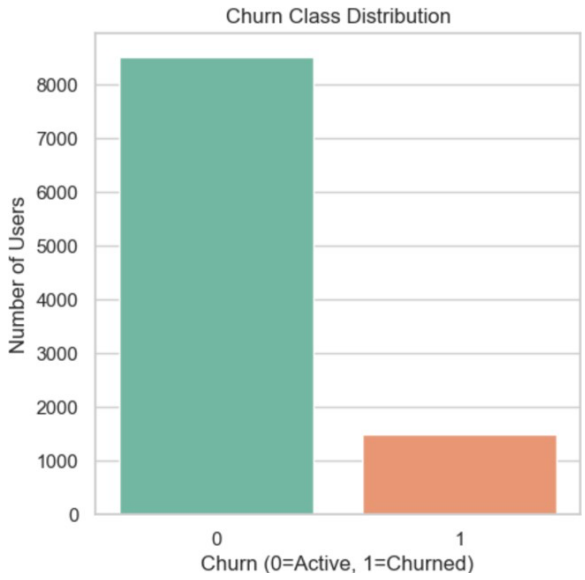
Our analysis leveraged the Netflix 2025 User Behavior Dataset, containing approximately 210 000 records distributed across six interconnected tables: users, watch_history, search_logs, reviews, movies and recommendation_logs.

First, we began by preparing and comprehensively cleaning the data, including removing duplicates across all fields and systematically handling missing values through median imputation for numeric features and “Unknown” categorization for sparse fields. Highly incomplete metrics were filled with zeros to maintain data integrity.

We then engineered sophisticated user-level features, including total and average watch time, completion rates, search activity metrics, review sentiment scores and recommendation click-through rates. A binary churn label was created where inactive users (is_active=False) were coded as churned (churn=1) and user tenure was calculated in days from subscription start.

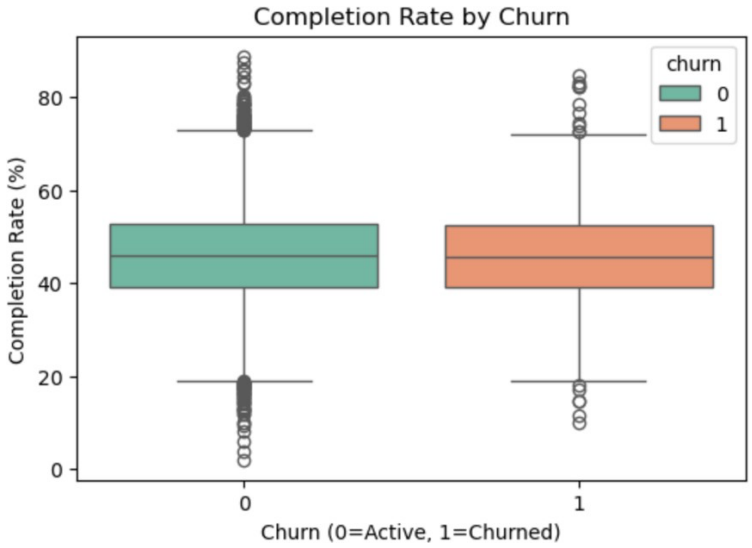
3. Exploratory Data Analysis

The initial analysis revealed a significant class imbalance, with 85% of users remaining active while only 15% had churned (a pattern that would heavily influence our modeling approach).



The exploratory phase uncovered behavioral patterns that clearly distinguished churned from retained users. Churn risk peaked dramatically among new subscribers, with the majority of cancellations occurring within the first 90 days of subscription. This finding indicates the importance of the onboarding period.

Engagement emerged as the most powerful predictor of retention: users who have higher watch times, more frequent viewing sessions and greater content completion rates demonstrate much lower churn rates.



Similarly, the active use of discovery features (including search activity and recommendation interactions) strongly correlates with user loyalty. On the other hand, churned users largely ignore recommendations and show much lower search frequencies.

User satisfaction provided another critical dimension: subscribers who consistently left lower ratings or wrote reviews with negative sentiment showed higher churn probabilities. On the contrary, demographic variables and subscription preferences had minimal predictive power: churn rates only vary by 14-15% across different plans.

These patterns demonstrate that behavioral indicators prove to be effective in predicting churn risk.

4. Modeling

Our modeling framework treated churn as the target variable while carefully removing potential data leakage sources such as user identifiers and raw timestamps. The preprocessing pipeline included one-hot encoding for categorical variables, median imputation for missing values and standardization for numerical features. We applied an 80/20 training-testing split while preserving the original churn distribution.

Because the dataset is highly imbalanced (~15% churners), we prioritized Precision–Recall AUC (PR AUC) as the main evaluation metric, while also reporting ROC AUC, precision, recall, and F1 score.

	Model	ROC AUC	PR AUC	Precision	Recall	F1
0	RandomForest(1000)	0.5677	0.2407	1.0000	0.0623	0.1173
1	LogisticRegression	0.5480	0.2179	0.2384	0.1344	0.1719
2	XGBoost(2000, ES=50)	0.5423	0.1852	0.1983	0.3016	0.2393
3	Dummy(majority)	0.5000	0.1481	0.0000	0.0000	0.0000

Baseline performance

The Dummy Classifier, which always predicts the majority class (non-churn),

established our baseline with PR AUC = 0.148.

Logistic Regression

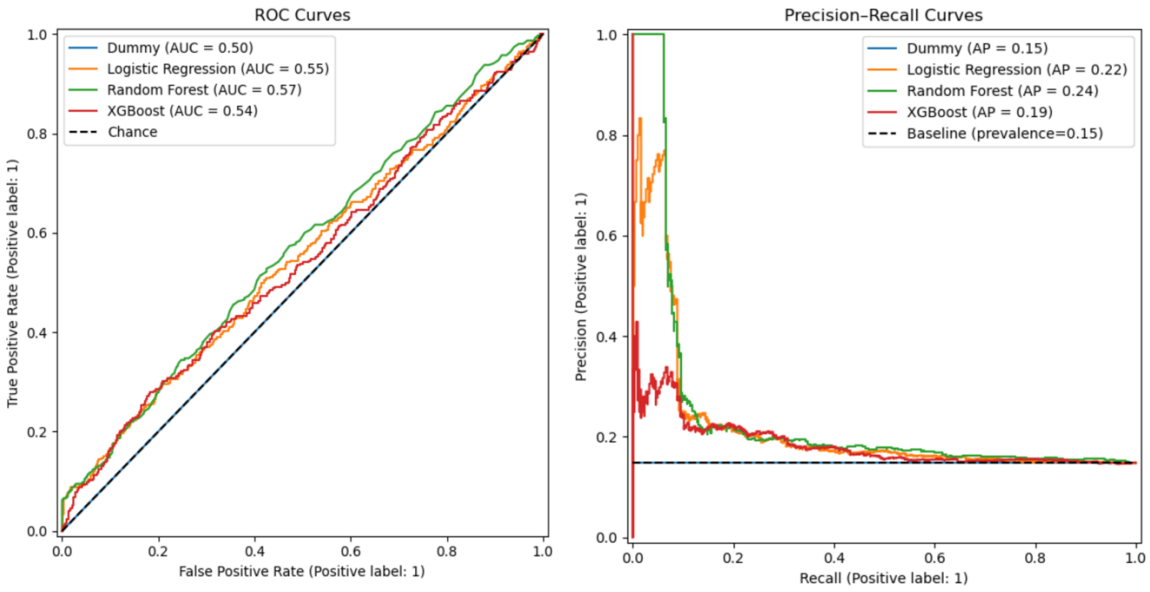
With balanced class weights, Logistic Regression achieved ROC AUC = 0.548 and PR AUC = 0.218, recovering about 13% of churners with 24% precision at its best F1 threshold (F1 = 0.172). It provides modest recall but generates many false positives.

Random Forest

The Random Forest model performed slightly better, with ROC AUC = 0.568 and PR AUC = 0.241. While precision was perfect (1.00) at very high thresholds, recall collapsed to near zero. This reflects the class imbalance: at strict thresholds, the model only flags a handful of churners (high precision, no recall). At more permissive thresholds, it achieved much higher recall (88%) but at the cost of very low precision (16%).

XGBoost

XGBoost reached ROC AUC = 0.542 and PR AUC = 0.185, delivering recall = 30% and precision = 20% at its best F1 threshold (F1 = 0.239). Although performance was weaker than Random Forest, it showed a more balanced trade-off between recall and precision.



At low thresholds, models achieve high recall but very low precision (many false alarms). At high thresholds, precision improves but recall collapses. This trade-off shows that before threshold tuning can produce actionable predictions, richer and more time-sensitive behavioral features (recency and drop-off patterns) are required.

5. Key findings

Our analysis revealed several patterns in Netflix churn behavior:

- **High engagement lowers churn.**

  Users with greater watch time and higher completion rates are less likely to cancel.

- **Exploration matters.**

  Customers who search frequently and click on recommendations tend to remain subscribed.

- **User satisfaction is predictive.**

  Low ratings or negative reviews often precede churn, providing early warning signals.

- **Onboarding is critical.**

  The first 90 days carry the highest churn risk, making it the key window for retention efforts.

- **Demographics add little value.**

  Subscription plans or personal attributes do not explain churn as strongly as behavioral signals.

Despite including recency-based features, the models show limited ability to separate churners from loyal users. This suggests that richer time-series signals (weekly activity changes or inactivity streaks) would be needed to uncover stronger behavioral drivers.

6. Recommendations

Based on these insights, we recommend Netflix implement the three following

initiatives:

- **Strengthen the onboarding experience**.
  Implement personalized welcome content, completion reminders and early exposure to recommendation features during the critical first 90 days to reduce new-user churn.

- **Monitor satisfaction proactively**.
  Create automated systems to flag users leaving low ratings or negative reviews, then trigger personalized outreach with tailored content suggestions or retention offers.

- **Focus on valuable customers**.
  Calculate Customer Lifetime Value using subscription spend and tenure data, then deploy retention resources strategically toward high-value and at-risk users.

- **Develop time-sensitive behavioral features** such as decline in watch time, search inactivity, and days since last session. These have the potential to improve model discrimination and provide actionable recall.

7. Limitations

Several factors limited our analysis. Indeed, some churn labels may reflect temporary inactivity rather than permanent cancellations, which introduces label noise and likely lowers achievable model performance. The absence of detailed revenue data required us to approximate customer value using basic subscription metrics. Additionally, the class imbalance fundamentally constrained our models' ability to achieve both high precision and recall.

8. Future work

Future improvements should focus on richer behavioral features that capture how user activity changes over time, rather than just totals or averages. In particular:

- Recency and decay patterns (e.g., sudden drops in watch time, inactivity streaks).
- Session-level dynamics (frequency, diversity, volatility in engagement).
- Cross-signal interactions (e.g., combining search activity, recommendations, and reviews).

Beyond feature engineering, model-driven interventions must be validated through A/B testing to measure real impact on retention and revenue. This ensures that predictive insights translate into business outcomes, guiding where to focus proactive retention efforts.

9. Conclusion

Netflix churn prediction remains a challenging task due to weak behavioral signals and class imbalance. Still, this project successfully identified engagement, content discovery, user satisfaction and early tenure as the most important drivers of retention.

While baseline models struggled, Random Forest modestly outperformed other approaches. However, none achieved strong predictive power, underscoring the limits of using only aggregate and recency-based features.

For Netflix leadership, the strategic implications are clear:

- Strengthen onboarding to reduce early churn risk.
- Monitor satisfaction signals (ratings, reviews, complaints) to act before cancellations.
- Prioritize high-value users for proactive retention, as broad campaigns are less cost-effective.

Overall, the models confirm that engagement, discovery, and satisfaction matter, but predictive performance remains modest. Future work should focus on richer temporal features and A/B testing of retention strategies to transform predictive insights into measurable impact.