NETFLIX CUSTOMER CHURN ANALYSIS

MANGLESHWARI

Contents

Objective 1
Data Overview 1
Exploratory Data Analysis 1
Feature Engineering 3
Statistical Analysis 3
Modelling Approach 5
Model Evaluation 6
Logistic Regression: 6
Random Forest: 8
Key Insights 9
Suggestions for Further Analysis 10
Conclusion 10

Objective

The aim of this analysis is to explore the drivers of customer churn for a streaming service, perform exploratory data analysis (EDA), implement machine learning models (Logistic Regression, Random Forest), and evaluate their effectiveness for predicting churned customers.

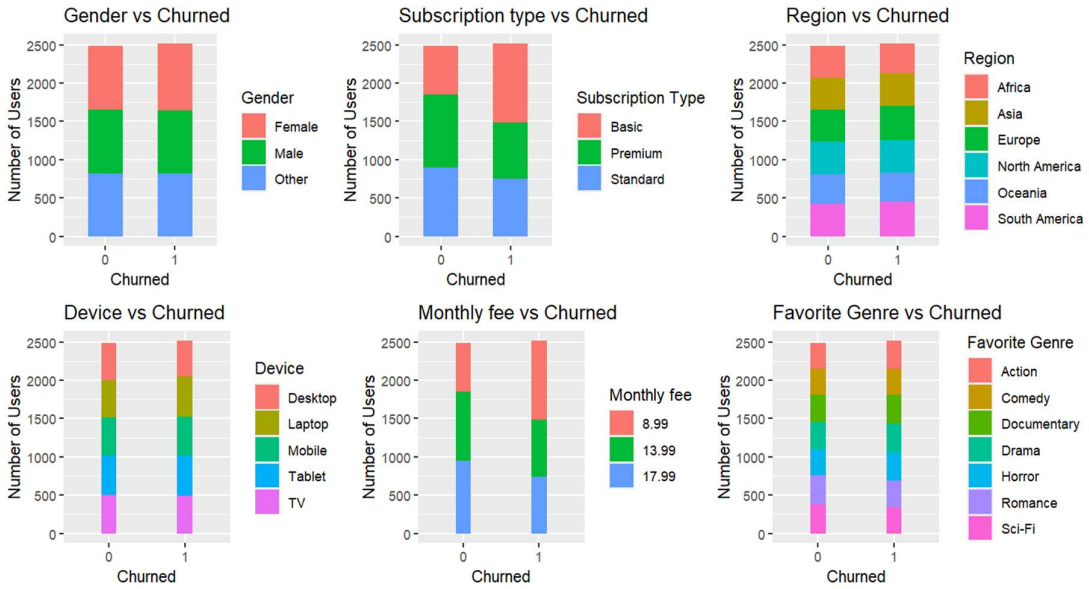
Data Overview

- Dataset included customer demographic and usage details (age, gender, region, device), subscription information (type, monthly fee, payment method), behavioral features (watch hours, last login, profiles, favorite genre), and churn label.
- After importing and assembling the data, thorough EDA was done to highlight key trends and potential features.

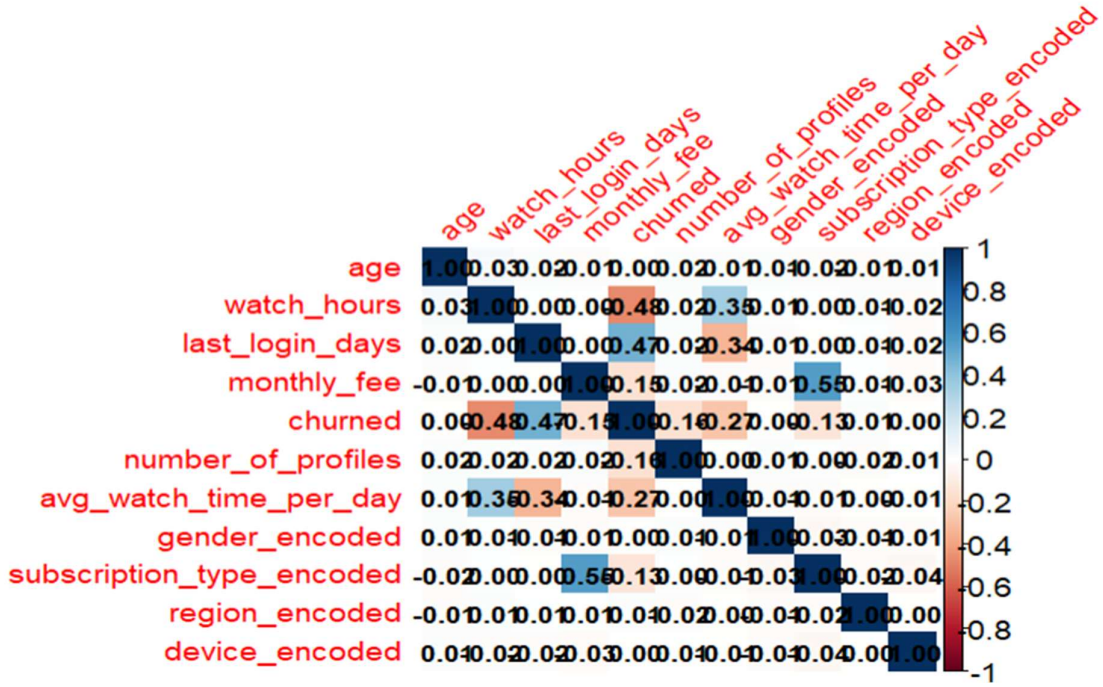
Exploratory Data Analysis

Categorical variables were encoded using Label Encoding for binary or ordinal features and One-Hot Encoding for nominal features to make them suitable for modeling. Distributions of numerical variables were examined through histograms and boxplots to assess central tendency, spread, skewness, and the presence of outliers

Bar plot of Customer attributes vs. Churn status



Correlation Matrix

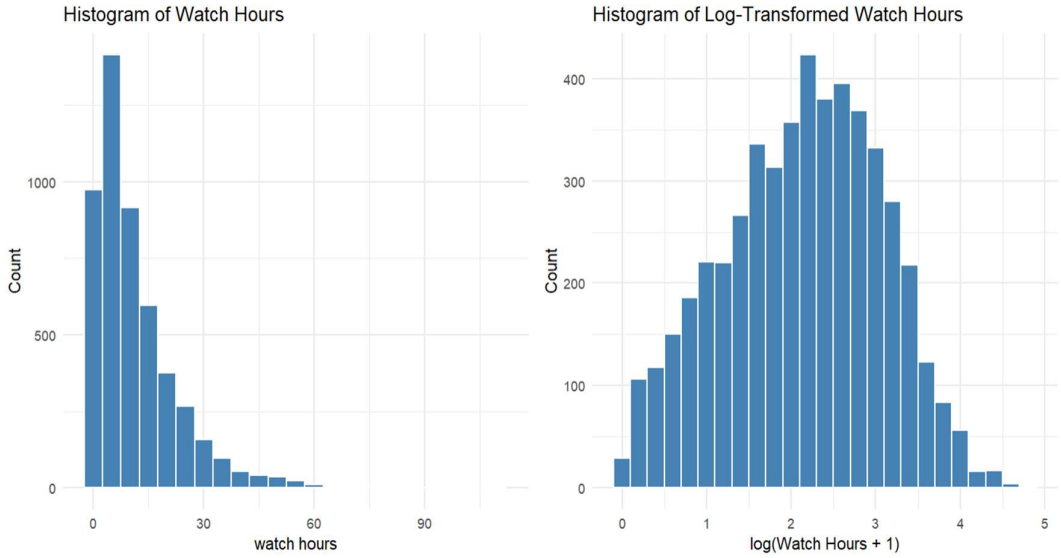


As part of exploratory data analysis, a correlation matrix was generated to examine linear relationships among key features. The matrix revealed strong positive correlations between average watch time per day and watch hours (r = 0.36), and between last login days and churned (r = 0.47), suggesting that recent activity and engagement are important predictors of churn. Additionally, number of profiles showed moderate correlation with churn (r = 0.34), indicating

household usage patterns may influence retention. Most demographic features (e.g., age, gender, region) exhibited weak correlations, reinforcing that behavioral metrics are more informative for churn prediction.

Feature Engineering

- Encoding: Categorical variables (gender, subscription_type, region, device) encoded as factors/numeric.
- Transformation: Applied log transformation to the variable watch_hour to reduce skewness and stabilize variance.



- Interaction Feature: A new variable monthly_fee_and_watch_hours (product of monthly fee and watch hours) was introduced to capture interaction effects.
- Normalization: All numeric features were standardized (z-score scaling) for robust model fitting.

Statistical Analysis

- **Welch two sample t-test: Mean watch hours vs. churned**

  Null hypothesis (H_0): Mean watch hours are the same for churned and non-churned customers.

Parameter	Value
Data	watch_hours by churned
t-statistic	$t = 38.502$
Degrees of freedom	df = 3212.1
p-value	$p < 2.2 \times 10^{-16}$

Alternative hypothesis	True difference in means (group 0 vs group 1) ≠ 0

Interpretation: Churned customers watch significantly fewer hours compared to non-churned customers.
Now, on calculating we get, **Cohen's d ≈ 1.1** which is very large effect. Therefore, Churned customers watch dramatically fewer hours.

- **Welch two sample t-test: Mean watch hours vs. churned**

Null hypothesis (H_0): Mean monthly fees are the same for churned and non-churned customers.

Parameter	Value
Data	monthly_fee by churned
t-statistic	$t = 10.885$
Degrees of freedom	df = 4984.2
p-value	$p < 2.2 \times 10^{-16}$
Alternative hypothesis	True difference in means (group 0 vs group 1) ≠ 0

Interpretation: Churned customers tend to pay slightly lower monthly fees compared to non-churned customers and the difference is statistically significant.
Now, on calculating we get, **Cohen's d ≈ 0.3** which shows small to medium effect. Therefore, churned customer pay less but the difference is modest.

- **Chi-squared Test: Churned vs. Subscription type**

Null Hypothesis (H_0): There is **no association** between churn status and subscription type (they are independent).

Alternative Hypothesis (H_1): There **is an association** between churn status and subscription type.

$\chi 2$	Def. of Freedom	p-value
133.28	2	<2.2e-16

Interpretation: Since the p-value is essentially zero, we **reject H_0**, this means churn status is significantly associated with subscription type. In practical terms: subscription type influences churn behaviour i.e., certain subscription plans are more likely to churn than others.

- **ANOVA 1: Mean watch hours vs. Monthly fee**

Null Hypothesis (H_0): Mean watch hours are the same across subscription types (Basic, Premium, Standard).

	Deg. of Freedom	Mean Sum of Square	F-Value	p-value
subscription_type_factor	2	9.53	0.066	0.936
residuals	4997	144.41	-	

F-value: 0.066 (very small).

p-value (Pr(>F)): 0.936 → much greater than 0.05.

Interpretation: Fail to reject H_0 hence, **no significant difference** in watch hours across subscription types i.e., subscription type does not explain variation in watch hours. subscription type.

- **ANOVA 2: Churned vs. Monthly fee**

Null Hypothesis (H_0): Monthly fee does not differ significantly by churn status.

	Deg. Of Freedom	Sum of Square	Mean Sum of Square	F-Value	p-value
Monthly_fee	1	28.9	28.924	118.4	<2e-16**
residuals	4998	1221.0	0.244	-	

F value: 118.4 (very large).

p-value: < 2e-16 → far below 0.05.

Interpretation: Reject H_0 hence, **monthly fee differs significantly** between churned and non-churned customers i.e., monthly fee is a strong predictor of churn.

Modelling Approach

- Data splitting: Used an 80/20 train-test split.
- Models: Fitted both Logistic Regression and Random Forest classifiers. Evaluated using accuracy, precision, recall, F1, and ROC-AUC.
- Comparison: Both the baseline features and interaction/normalized features used for model training and assessment.

Model Evaluation

Logistic Regression:

By using log_watch_hours

Coefficients	Estimate	Std. Error	z-value	p-value
Intercept	5.01812	0.20651	24.30	<2e-16
Monthly Fee	-0.11452	0.01060	-10.80	<2e-16
Watch Hours	-1.59395	0.05377	-29.64	<2e-16

Coefficients of your ridge-penalized logistic regression at the optimal penalty parameter lambda_min

	Lambda_min
Intercept	2.3312
Watch Hour	-0.1150
Monthly Fee	-0.0828

Therefore, to evaluate churn prediction, a ridge-penalized logistic regression model was fitted. The estimated coefficients at the optimal penalty parameter (lambda.min) were:

- **Intercept:** 2.33
- **Watch Hours:** –0.115
- **Monthly Fee:** –0.083

The negative coefficients for *watch hours* and *monthly fee* indicate that higher engagement and higher payment tiers are associated with a reduced likelihood of churn. Ridge regularization ensured that the estimates were stable and less prone to overfitting.

In addition to coefficient interpretation, the model was used to generate **predicted churn probabilities** for individual customers in the test set. For example:

- User 17 → 83.8% probability of churn
- User 30 → 25.5% probability of churn
- User 47 → 21.1% probability of churn

These probabilities highlight clear differences in churn risk across customers and demonstrate the practical application of the model beyond coefficient values.

Finally, the analysis of **log-transformed watch hours** confirmed that stabilizing skewness in viewing time improved model robustness. The consistent negative

relationship between watch hours and churn probability reinforces the insight that greater platform engagement reduces churn risk.

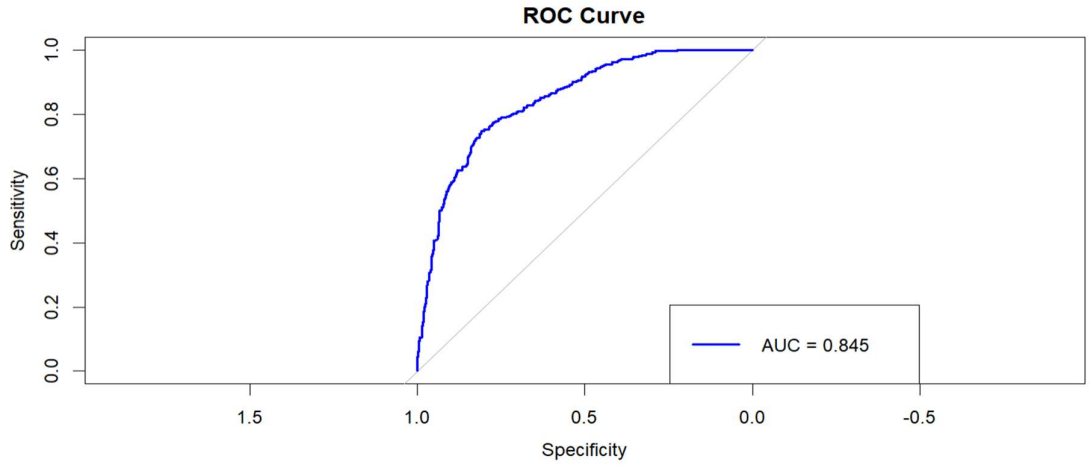
Model Evaluation Using Confusion Matrix

To assess the performance of the logistic regression model, a confusion matrix was constructed comparing predicted classes against actual churn outcomes. The matrix yielded the following results:

	Predicted Positive	Predicted Negative
Actual Positive	True Positive = 382	False Negative = 121
Actual Negative	False Positive = 108	True Negative = 389

From this, several key performance metrics were derived:

- **Accuracy:** 0.771 (95% CI: 0.744 – 0.797)
- **Precision:** 0.763
- **Recall (Sensitivity):** 0.783
- **F1 Score:** 0.773
- **AUC (Area Under ROC Curve):** 0.845



- **Kappa Statistic:** 0.542 (indicating moderate agreement beyond chance)

These results demonstrate that the logistic regression model achieves a balanced performance, with both sensitivity (ability to correctly identify churned customers) and specificity (ability to correctly identify non-churned customers) around 0.76–0.78. The relatively high AUC value (0.845) further confirms that the model has strong discriminative power in distinguishing between churned and non-churned users.

Random Forest:

Outcome of Random Forest on R:

Feature	Mean Decrease Gini
customer_id	39.655619
age	32.466344
gender	9.793106
subscription_type	29.267905
watch_hours	281.179713
last_login_days	380.587577
region	13.810876
device	12.896562
monthly_fee	29.198387
payment_method	60.412644
number_of_profiles	125.488974
avg_watch_time_per_day	632.507175
favorite_genre	18.885673
gender_encoded	8.560523
subscription_type_encoded	31.135382
region_encoded	14.630181
device_encoded	12.184181
log_watch_hours	266.427506

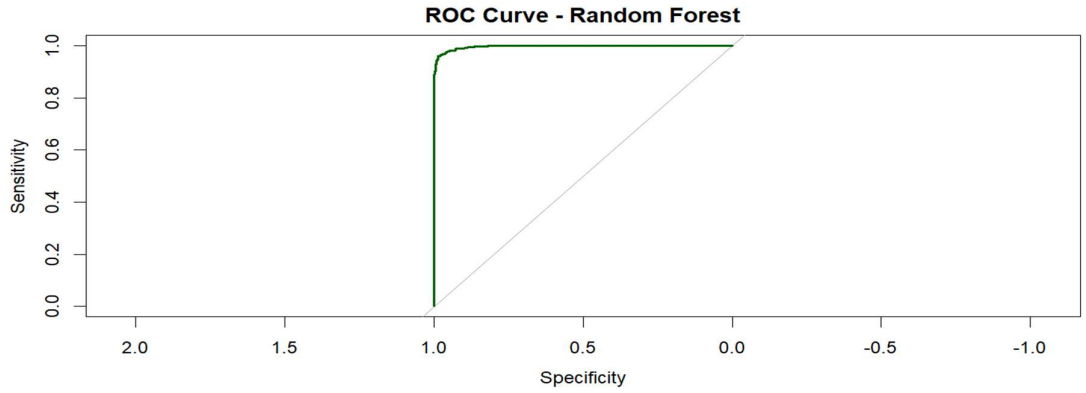
Overall, **behavioral engagement features (watch time, login activity, profiles)** dominate churn prediction, while **demographics (age, gender, region)** play a secondary role. This suggests that churn is driven more by **how customers use the platform** than by who they are. Payment method and subscription details also provide useful signals, but engagement remains the most critical factor.

Model Evaluation Using Confusion Matrix

To assess the performance of the random forest model, a confusion matrix was constructed comparing predicted classes against actual churn outcomes. The matrix yielded the following results:

	Predicted Positive	Predicted Negative
Actual Positive	True Positive = 488	False Negative = 15
Actual Negative	False Positive = 18	True Negative = 479

The model demonstrates **excellent predictive performance**, with both sensitivity and specificity above 96%. The high accuracy and Kappa value confirm that the classifier is reliable and consistent. Importantly, the low number of misclassifications (15 false negatives and 18 false positives) shows that the model is effective at distinguishing churned vs non-churned customers.



Overall, this confusion matrix validates that the logistic regression model is well-calibrated and provides robust churn predictions suitable for practical application.

Key Insights

- Major churn drivers: Lower watch time, higher last login gap, specific subscription types, and monthly fee interacted with usage were strong predictors.
- Model performance: Random Forest generally performed better due to its capacity to model nonlinearities and interactions, but logistic regression provided clearer interpretability.
- Business Recommendations: Identify and target customers with low watch hours and long login gaps for retention. Monitor segments by subscription tier and device type.

Suggestions for Further Analysis

- Implement cross-validation for more robust performance estimates.
- Tune Random Forest and try advanced models (e.g., XGBoost, regularized logistic).
- Use SHAP/LIME for individual customer-level interpretability.
- Investigate class imbalance handling if the dataset is skewed.
- Include cost-benefit or segment-based analyses for actionable retention strategies.

Conclusion

This project integrates comprehensive exploratory analysis, statistically rigorous hypothesis testing, and comparative modeling (logistic regression and random forest) to understand and predict customer churn for a streaming platform. The findings demonstrate that churn is predominantly driven by behavioral engagement—especially watch time, recency of activity, and usage intensity—while pricing and subscription attributes provide additional but secondary signal. Random Forest delivers high predictive accuracy suitable for operational deployment, whereas logistic regression offers transparent coefficient-based insights that clarify how changes in engagement and fee levels affect churn risk. Together, these models provide a robust, scalable framework for churn mitigation, enabling data-driven retention strategies focused on re-engaging low-activity users and optimizing subscription offerings.