# Netflix Customer Churn Prediction and Customer Segmentation

Machine Learning Individual Project

*By Kamya Sarda*

---

## Introduction

This project analyzes customer data from Netflix to solve two problems:

1. Predict which customers will cancel their subscriptions (churn)
2. Group similar customers into segments to understand different user types

Customer churn is a major problem for streaming services like Netflix. Every canceled subscription means loss in monthly revenue. If we can predict which customers are likely to leave, Netflix can take some action to keep them.

The dataset contains 5,000 Netflix customer records with information about their viewing habits, subscription details, and demographics. After removing data errors, the final dataset has 4,990 customers with 14 features.

I used Random Forest Classification to predict churn and K-Means Clustering to create customer groups.

---

## Data description

### Dataset overview

- Original size: 5,000 customers
- After cleaning: 4,990 customers
- Features: 14 variables
- Target variable: churned (0 = stayed, 1 = left)

### Data cleaning

I found some impossible values in the data. Specifically, some customers had avg_watch_time_per_day greater than 24 hours, which is physically impossible. These were likely data collection errors. I removed these 10 records to ensure the model learns from realistic behavior patterns.

I also removed the customer_id column because it's just a unique identifier, and we cannot use it to predict anything.

### Variables:

The dataset includes three types of information:

Engagement metrics:

- watch_hours - Total monthly viewing hours
- avg_watch_time_per_day - Average daily viewing time
- last_login_days - Days since last login

Account information:

- subscription_type - Basic, Standard, or Premium plan
- monthly_fee - Subscription cost in dollars
- payment_method - How they pay (credit card, PayPal, etc.)
- number_of_profiles - Number of profiles on the account
- device - Primary viewing device

Demographics:

- age - Customer age
- gender - Male/Female/Other
- region - Geographic location
- favorite_genre - Most-watched content type

All text columns (gender, region, payment method, etc.) were converted to numbers using Label Encoding to process them.

---

## Exploratory Data Analysis

### Churn distribution

The dataset is balanced:

- Stayed (0): 2,450 customers (49%)
- Churned (1): 2,540 customers (51%)

### Correlation analysis

I created a correlation heatmap to see which features are related to churn. The key findings were:

Strong relationships with churn:

- watch_hours and churn: -0.48 (negative correlation)
  → Customers who watch more are much less likely to leave

- last_login_days and churn: +0.47 (positive correlation)
  → Customers who haven't logged in recently are very likely to leave

- avg_watch_time_per_day and churn: -0.36 (negative correlation)
  → Daily viewing habits reduce churn risk

Weak relationships with churn:

- monthly_fee: -0.15 (slightly negative)
- age: ~0.00 (almost no relationship)
- region: ~0.01 (almost no relationship)

An important finding was that demographics like age, gender, and location barely correlate with churn. This suggests that the way people use Netflix matters much more than who they are. How matters more than who.

---

## Modelling

I chose Random Forest Classifier for churn prediction because it works well with both numbers and categories (my dataset has both), it tells us which features are most important for predictions, and doesn't overfit easily.

- Number of trees: 100
- Random state: 42

### Train-Test Split

I split the data into:

- Training set: 80% (3,992 customers)
- Test set: 20% (998 customers)

### Customer segmentation (K-Means clustering)

Besides predicting churn, I also used K-Means clustering to group customers into segments based on their behavior.

Features used for clustering:

- watch_hours - How much they watch
- monthly_fee - How much they pay
- age - What stage of life they're in

I chose these three because they represent engagement level, revenue value, and demographics: three important dimensions for understanding customer types.

Before clustering, I standardized the features using StandardScaler.

Number of clusters: I chose K=3 to create three meaningful customer groups that are practical for developing some sort of business strategy.

---

## Results

### Customer segmentation (K-Means clustering)

The algorithm found three distinct customer groups:

| Cluster | Watch Hours | Monthly Fee | Age | Group Name | Interpretation |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| 0 | 11.46 | $8.99 | 44 | Basic plan users | Older customers on basic plans with moderate viewing. These are stable users who could potentially upgrade to premium plans. |
| 1 | 10.99 | $15.99 | 30 | At-risk users | Younger customers paying standard fees but not watching much. High churn risk. |
| 2 | 12.33 | $16.04 | 57 | Premium users | Older customers on premium plans with good engagement. These are valuable customers to protect. |

Key insight: Cluster 1 is the most concerning. These customers are paying decent money but barely using the service, which means they might be getting poor value and will be likely to cancel soon.

## Churn prediction model performance

Random Forest:

Overall Accuracy: 97.60%

This means the model correctly predicted whether a customer would churn or stay in 97.6 out of 100 cases.

Detailed Performance Metrics:

| Metric | Class 0 (Stayed) | Class 1 (Churned) |
|---|---|---|
| Precision | 0.96 | 0.99 |
| Recall | 0.99 | 0.97 |
| F1-Score | 0.97 | 0.98 |

What these mean:

- Precision (0.99 for churn): When the model says someone will churn, it's correct 99% of the time. This means very few false alarms.

- Recall (0.97 for churn): The model catches 97% of people who actually churn. Only 3% go undetected.

- F1-Score (0.98): This combines precision and recall. A score of 0.98 is excellent and shows the model is balanced.

Why did it perform so well?
The behavioral features (watch time, login activity) are very strong predictors. When someone stops watching and stops logging in, it's a clear signal they're about to cancel.

## Feature Importance

Random Forest tells us which features matter most for predictions:

Top 10 Most Important Features:

| Rank | Feature | Importance | What It Means |
|---|---|---|---|
| 1 | avg_watch_time_per_day | 0.40 | Daily viewing habits are the strongest predictor by far |
| 2 | watch_hours | 0.22 | Total monthly watching matters a lot |
| 3 | last_login_days | 0.19 | How recently they logged in is crucial |
| 4 | number_of_profiles | 0.07 | Accounts with multiple profiles churn less |
| 5 | payment_method | 0.05 | How they pay affects churn slightly |
| 6 | monthly_fee | 0.04 | Price matters, but not as much as we would think |
| 7 | age | 0.04 | Age has minimal impact |
| 8 | subscription_type | 0.03 | Plan type doesn't matter much |
| 9 | favorite_genre | 0.02 | Genre preference barely matters |
| 10 | region | 0.01 | Location is almost irrelevant |

Key findings:

1. Behavior is important. The top 3 features (all behavioral) account for about 81% of the model's predictive power.

2. Demographics don't matter a lot. Age, gender, and region contribute almost nothing to predictions.

3. Multiple profiles help. Accounts with more profiles churn less, probably because family members share the account and it's harder for everyone to agree to cancel.

4. The payment method is important. Certain payment types (probably prepaid cards) might indicate less commitment than credit cards which auto-renew.

---

The results show that churn is a behavior problem, not a demographics problem.

Therefore, Netflix shouldn't focus on demographics ("Let's target 25-35 year old males"). The data proves that age, gender, and location barely predict churn.

The patterns that predict churn are:

Low watch time: Users watching less than 1 hour per day are at high risk

Inactive accounts: If someone hasn't logged in for 20+ days, they are likely gone

Single profiles: These accounts are more vulnerable than family accounts

## Business Recommendations

1. Monitor behaviour with time

Low risk: Watching regularly, logged in recently

Medium risk: Watch time declining or more than 10 days since login

High risk: Barely watching and inactive for more than 10 days

2. Creative campaigns

If someone's viewing drops 40%, send them personalized recommendations for their favorite genre

If 15+ days pass without login, send a "Your favorite show has a new season!" email

If someone has 1 profile and watches <5 hours/month, encourage them to add family profiles

3. Pause targeting based on demographics

Since age groups matter very less, as we've seen, focus on behavior instead.

4. Encourage multiple-profile accounts

Since accounts with more profiles churn less, promote the multi-profile feature. Many users don't know they can add family members for free.

---

# Conclusion

This project successfully built something that predicts Netflix customer churn with 97.6% accuracy. The Random Forest model identified that behavioral engagement (daily watch time, total hours, login activity) are the strongest churn predictors, while demographics (age, gender, region) barely matter.

K-Means clustering revealed three customer types (Please refer last image in output section):

Cluster 0 (1657 or approx. 33%): Basic plan users - stable, moderate engagement

Cluster 1 (1689 or approx. 34%): At-risk users - paying decent fees but not watching much

Cluster 2 (1644 or approx. 33%): Premium users - high engagement, high value

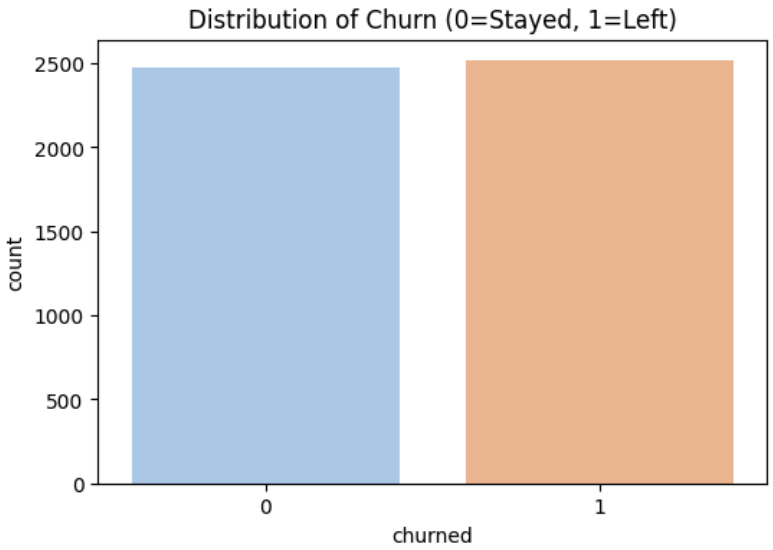
The main takeaway from this is that Netflix should focus retention efforts on monitoring user engagement patterns rather than targeting demographic groups.

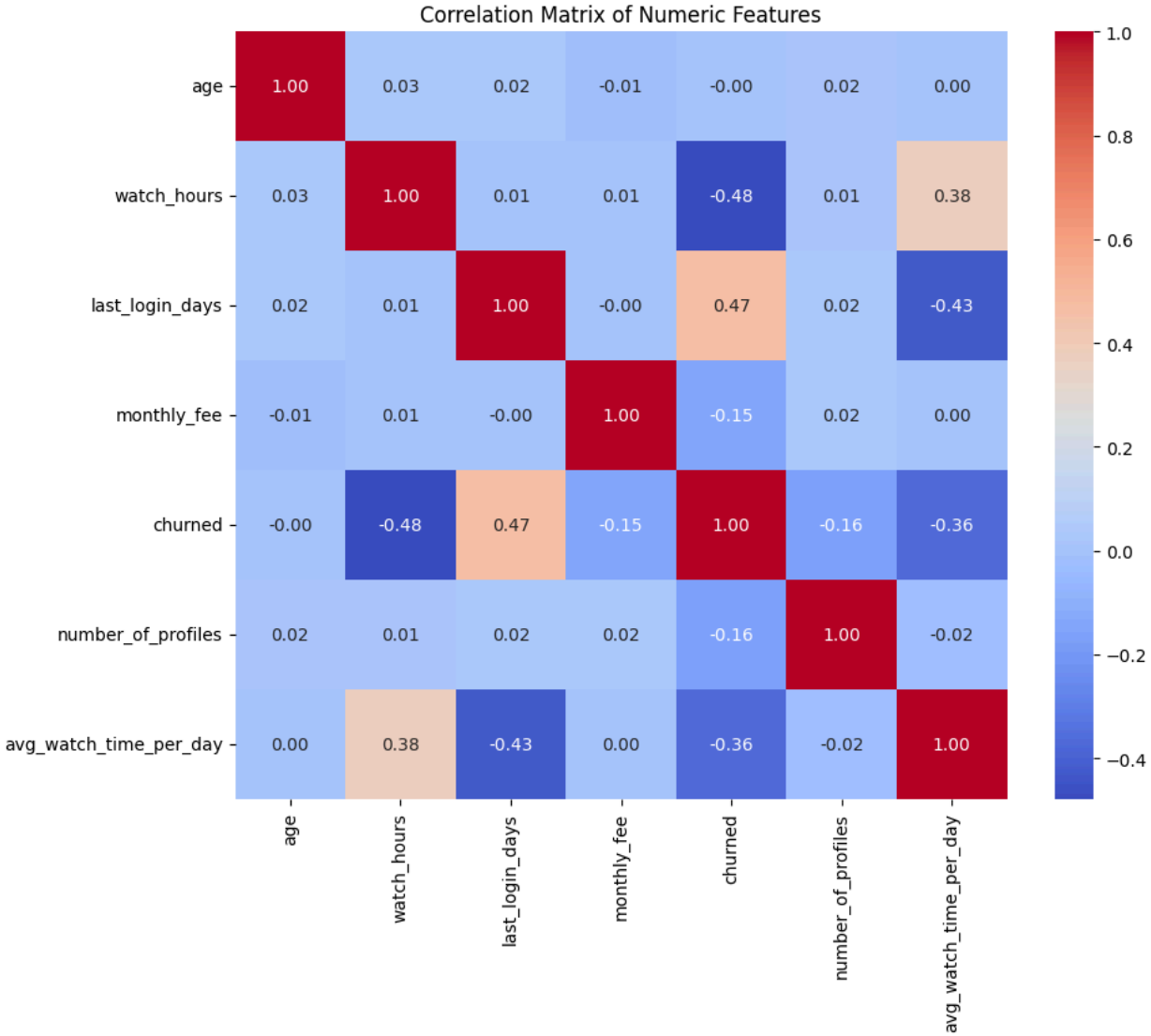
Customers who stop watching will cancel regardless of their age or location. The solution is to catch declining engagement early and fix it with personalized content recommendations.

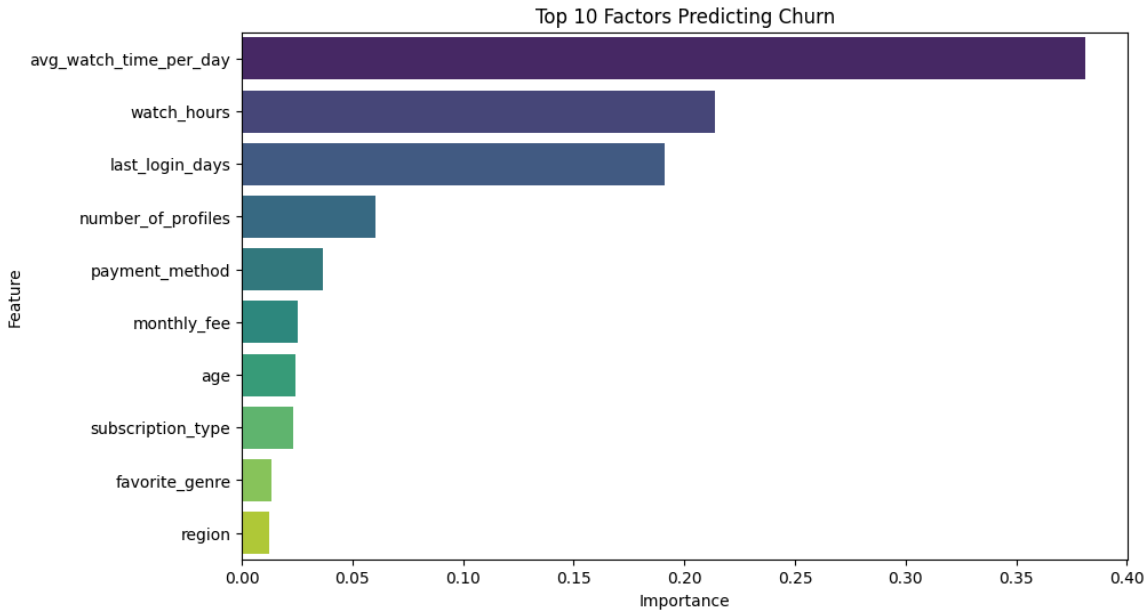
By implementing this, Netflix can potentially save a lot of money by shifting from reactive (responding after cancellation) to proactive (preventing cancellation).

Interestingly, the three clusters are roughly equal in size, suggesting Netflix's customer base is evenly distributed across engagement and subscription tiers rather than heavily concentrated in one segment.

VS Code Output & Visualisations:


Distribution of Churn (0=Stayed, 1=Left)
count
churned
0
1
2500
2000
1500
1000
500
0


Correlation Matrix of Numeric Features
age
watch_hours
last_login_days
monthly_fee
churned
number_of_profiles
avg_watch_time_per_day
1.00 0.03 0.02 -0.01 -0.00 0.02 0.00
0.03 1.00 0.01 0.01 -0.48 0.01 0.38
0.02 0.01 1.00 -0.00 0.47 0.02 -0.43
-0.01 0.01 -0.00 1.00 -0.15 0.02 0.00
-0.00 -0.48 0.47 -0.15 1.00 -0.16 -0.36
0.02 0.01 0.02 0.02 -0.16 1.00 -0.02
0.00 0.38 -0.43 0.00 -0.36 -0.02 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4




```
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.99      0.97       461
           1       0.99      0.97      0.98       537

    accuracy                           0.98       998
   macro avg       0.98      0.98      0.98       998
weighted avg       0.98      0.98      0.98       998
```

```
Cluster Profiles (Mean Values):
               watch_hours  monthly_fee        age
Cluster_Label
0                11.463060     8.990000  44.196138
1                10.994956    15.995921  30.208999
2                12.327354    16.041095  57.483577
```

```
Cluster_Label
1    1689
0    1657
2    1644
Name: count, dtype: int64
```

---

Dataset Source: Kaggle

"Netflix Customer Churn Dataset"

https://www.kaggle.com/datasets/abdulwadood11220/netflix-customer-churn-dataset