Name – Bhavin Bhatt

Customer Segmentation and Churn Analysis for Streaming Service Platform.

Business Understanding: Problem Definition

In today's highly competitive streaming service industry, retaining customers is essential for profitability. High customer acquisition costs make it crucial to predict customer churn and reduce attrition. Additionally, understanding customer viewership patterns allows for better segmentation, helping companies personalize services and improve retention efforts.

Objective:

1. Customer Segmentation Based on Viewership: Classify customers into segments based on viewership behavior to deliver targeted marketing, personalized content recommendations, and customized subscription plans.

2. Predict Customer Churn: Identify customers likely to churn to proactively intervene and retain them.

Key Business Questions

1. **Churn Prediction:**

- Which customers are most likely to cancel their subscription?
- What features (e.g., MonthlyCharges, PaymentMethod, Viewership patterns) are highly predictive of churn?

2. Customer Segmentation:

- How do viewing habits differ among various customer segments?
- Which segments are high-risk for churn, and which ones are loyal or highly engaged?
- Can certain content types or subscription plans be optimized to improve retention?

I) Data Understanding:

1. There are total 2,43,787 customers and around 21 features to asses and evaluate the churn reasons amongst the customers.
2. There are no null values in the data.

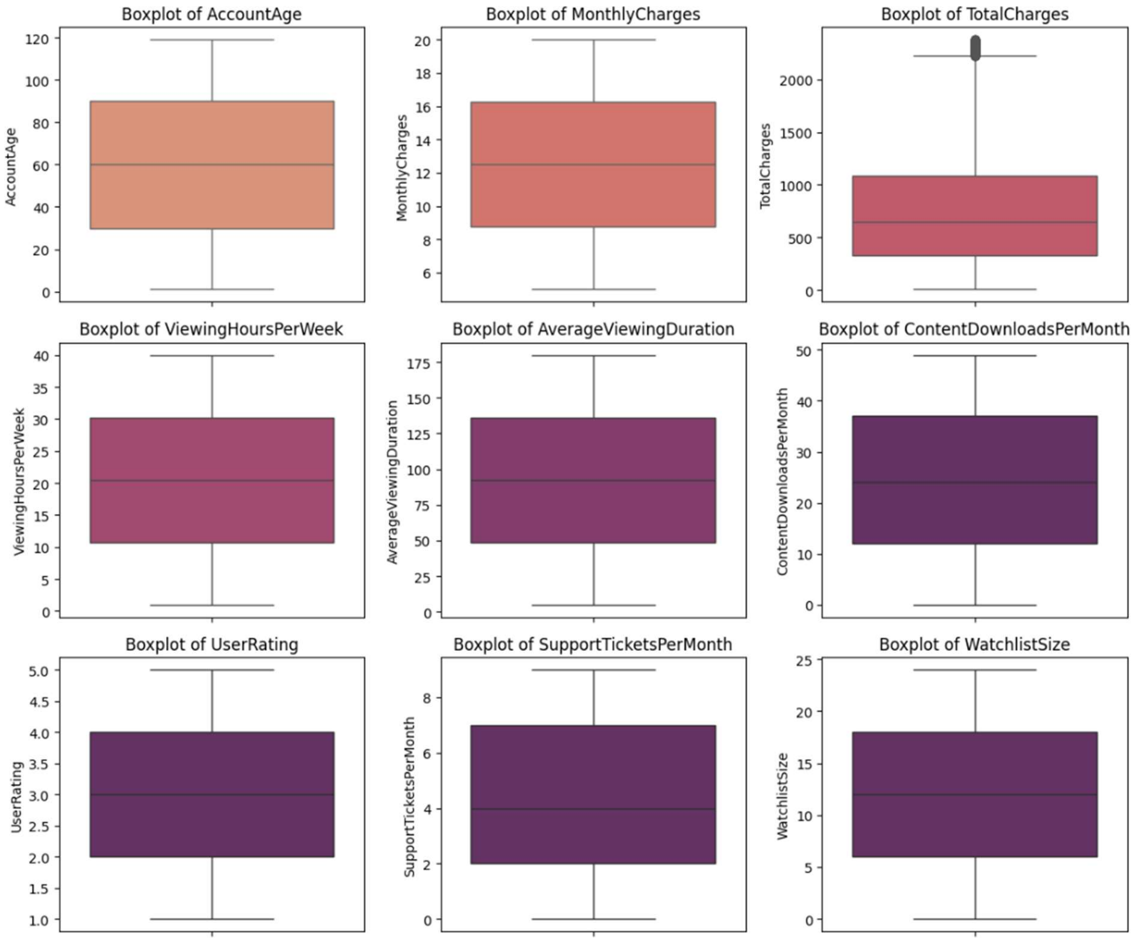
- **CustomerID:** Unique identifier for each customer
- **SubscriptionType:** Type of subscription plan chosen by the customer (e.g., Basic, Premium, Deluxe)
- **PaymentMethod:** Method used for payment (e.g., Credit Card, Electronic Check, PayPal)
- **PaperlessBilling:** Whether the customer uses paperless billing (Yes/No)
- **ContentType:** Type of content accessed by the customer (e.g., Movies, TV Shows, Documentaries)

- **MultiDeviceAccess:** Whether the customer has access on multiple devices (Yes/No)
- **DeviceRegistered:** Device registered by the customer (e.g., Smartphone, Smart TV, Laptop)
- **GenrePreference:** Genre preference of the customer (e.g., Action, Drama, Comedy)
- **Gender:** Gender of the customer (Male/Female)
- **ParentalControl:** Whether parental control is enabled (Yes/No)
- **SubtitlesEnabled:** Whether subtitles are enabled (Yes/No)
- **AccountAge:** Age of the customer's subscription account (in months)
- **MonthlyCharges:** Monthly subscription charges
- **TotalCharges:** Total charges incurred by the customer
- **ViewingHoursPerWeek:** Average number of viewing hours per week
- **SupportTicketsPerMonth:** Number of customer support tickets raised per month
- **AverageViewingDuration:** Average duration of each viewing session
- **ContentDownloadsPerMonth:** Number of content downloads per month
- **UserRating:** Customer satisfaction rating (1 to 5)
- **WatchlistSize:** Size of the customer's content watchlist

II) Data preparation and exploratory data analysis

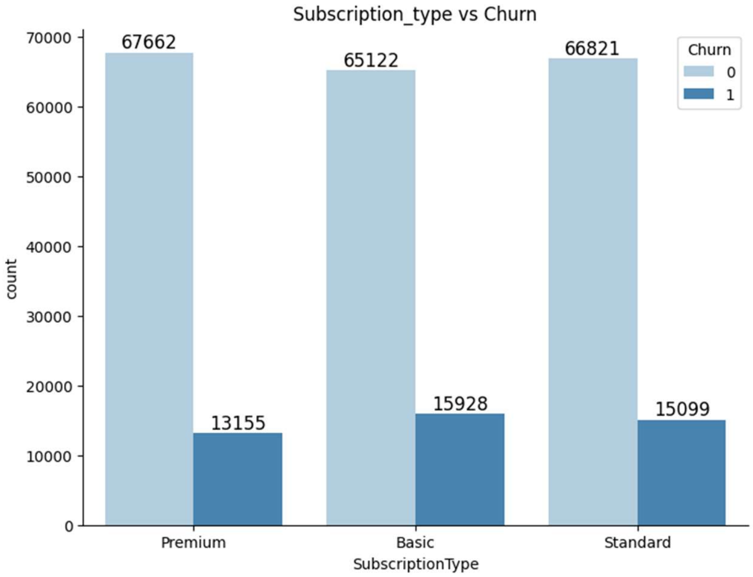
1. **Outlier detection:**

- AccountAge, ViewingHoursPerWeek, and AverageViewingDuration have a wide range with moderate spread, indicating diversity in user engagement and account age.
- MonthlyCharges and TotalCharges have some variation, with TotalCharges displaying outliers on the higher end, suggesting a few users with notably high cumulative charges.
- ContentDownloadsPerMonth, UserRating, SupportTicketsPerMonth, and WatchlistSize show relatively tight distributions with minimal outliers, indicating consistent usage patterns across these features.



2. Feature variations with Churn

- Subscription_type vs Churn

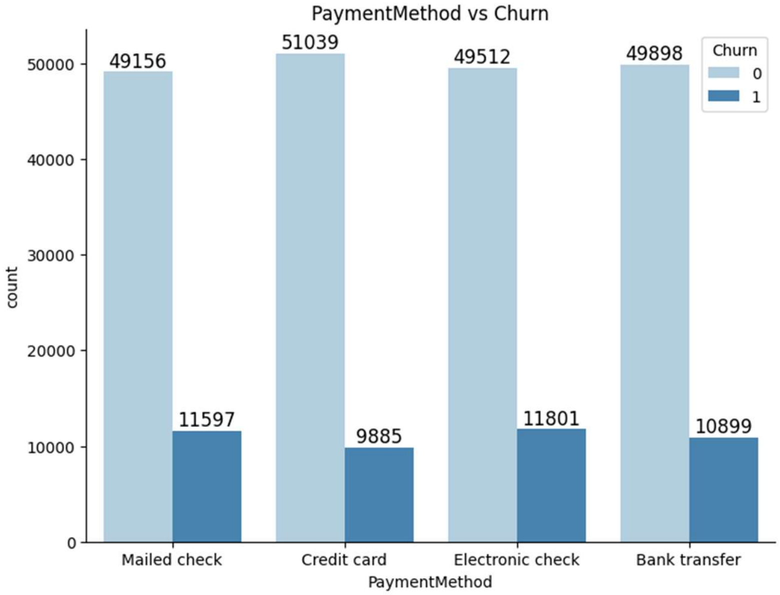


- Churn Rates Across Plans:
    1. The Basic and Standard plans have slightly higher churn compared to the Premium plan, with around 15,000 customers leaving each.
    2. Premium plan has the lowest churn at 13,155.
- Retention Trend:
    1. Retention is fairly consistent across all three subscription types, with each plan retaining between 65,000 to 67,000 customers.

Insights:

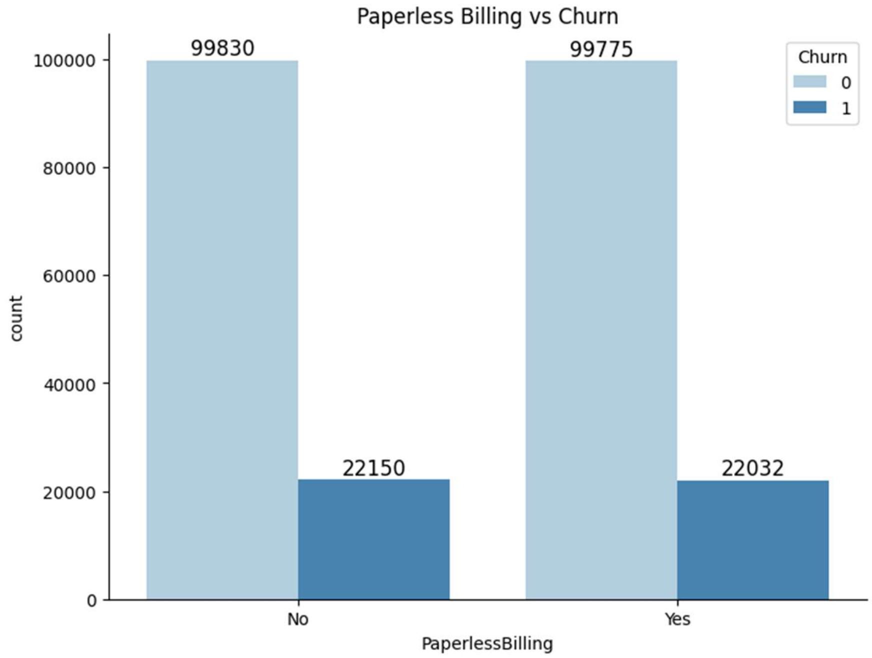
- Customers on Basic or Standard plans may be more likely to churn, possibly due to fewer features or lower satisfaction compared to the Premium plan.
- This suggests Netflix could potentially focus on upselling users from Basic/Standard to Premium to improve retention.

B) PaymentMethod vs Churn



- Churn Rates by Payment Method:
    1. Highest churn is seen with Electronic check users (11,801), followed by Mailed check (11,597) and Bank transfer (10,899).
    2. Credit card users have the lowest churn, with 9,885 customers leaving.
- Retention Overview:
    1. Retention is fairly consistent across all payment methods, with each method retaining between 49,156 to 51,039 customers.
- **Insights:** Customers using electronic checks show higher churn, possibly due to payment-related frustrations. Focusing on promoting automatic payment methods like credit cards or offering incentives for switching to more seamless payment options could help reduce churn.

C) PaperlessBilling vs Churn

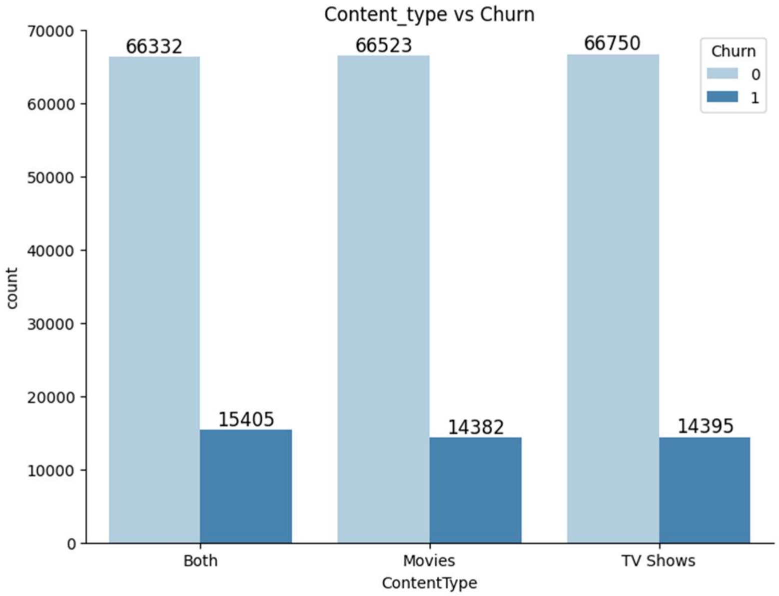


- **Churn Rates by Paperless Billing:**
  1. Customers with paperless billing have a churn count of 22,032. Those without paperless billing show a similar churn count of 22,150.
- **Retention Overview:**
  1. Retention is nearly identical across both groups, with around 99,775 to 99,830 customers retained.

Insights:

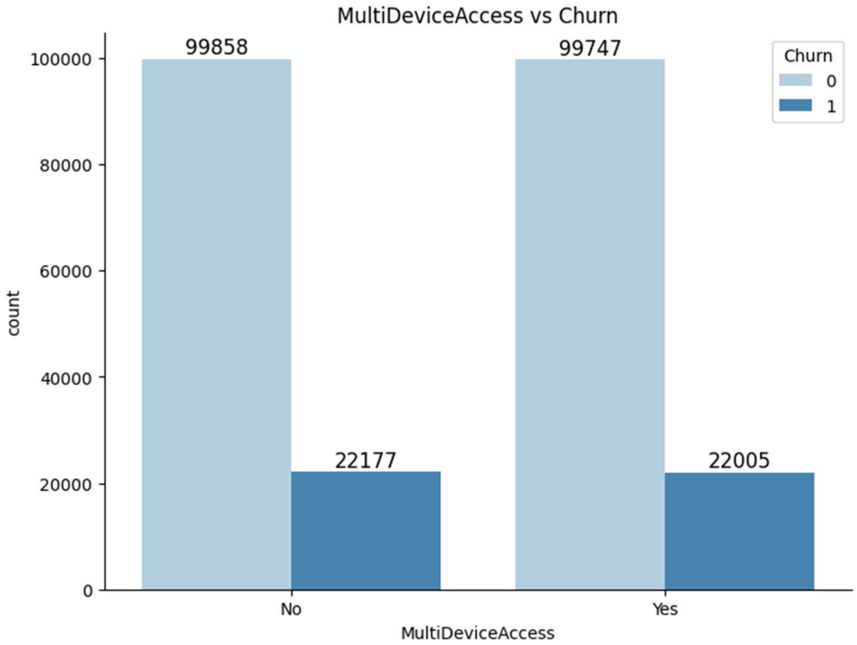
- Paperless billing does not appear to significantly impact churn rates. Other factors, such as payment method or subscription experience, might have a stronger influence on churn.

D) ContentType vs Churn



- The bar chart below depicts the relationship between Content Type and Churn. It shows the count of customers who accessed different types of content (Both, Movies, TV Shows) and their churn status (0 = Not Churned, 1 = Churned).
- The churn rates appear similar across all content types. However, customers accessing both types of content show a slightly higher churn count than those consuming only Movies or TV Shows, suggesting that broader content consumption alone may not be enough to prevent churn.

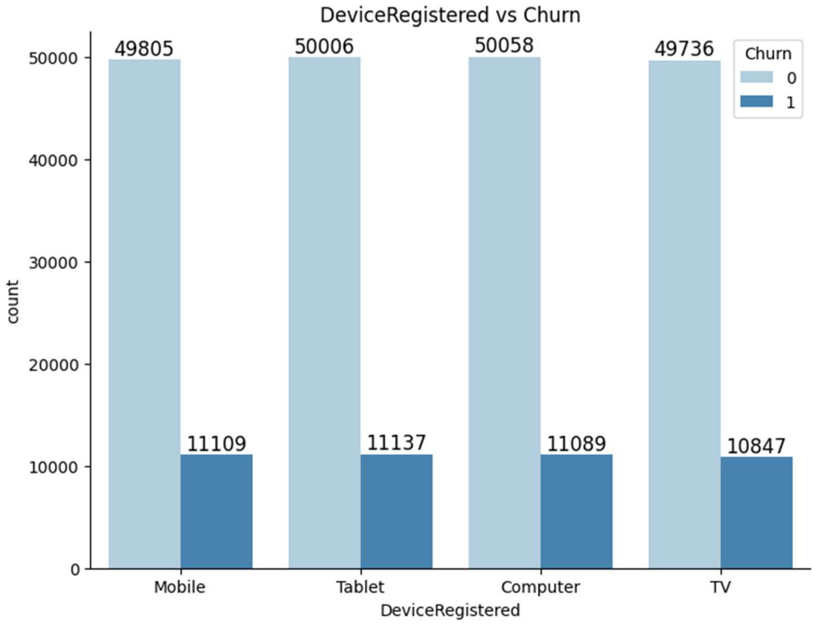
E) MultiDeviceAccess vs Churn



- **Churn Rates by Multi-Device Access:**
  1. Customers without multi-device access have a churn count of 22,177.
  2. Customers with multi-device access have a slightly lower churn count of 22,005.
- **Retention Overview:**
  1. Retention is similar across both groups, with around 99,747 to 99,858 customers retained.

Insights: Having multi-device access does not significantly reduce churn. Other factors, such as personalized features or better engagement strategies, might be required to improve retention.

F) MultiDeviceAccess vs Churn

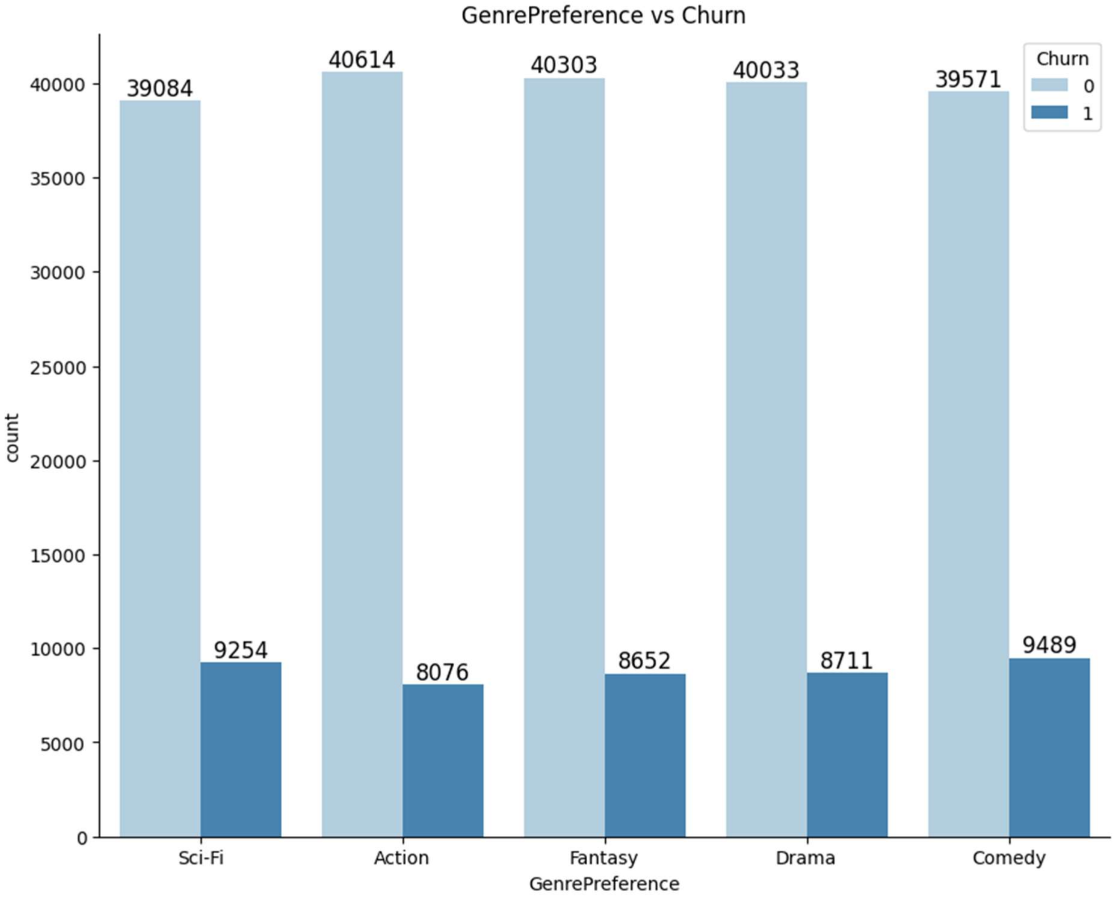


- Churn rate by Device type:
  1. Mobile: 11,109 churned, 49,805 retained
  2. Tablet: 11,137 churned, 50,006 retained
  3. Computer: 11,089 churned, 50,058 retained
  4. TV: 10,847 churned, 49,736 retained
- Retention is similar across all device types, with each device category retaining around 49,700 to 50,000 customers.

Insights:

The type of device registered does not significantly impact churn. Other factors, such as personalized content or improved engagement strategies, may be necessary to increase retention.

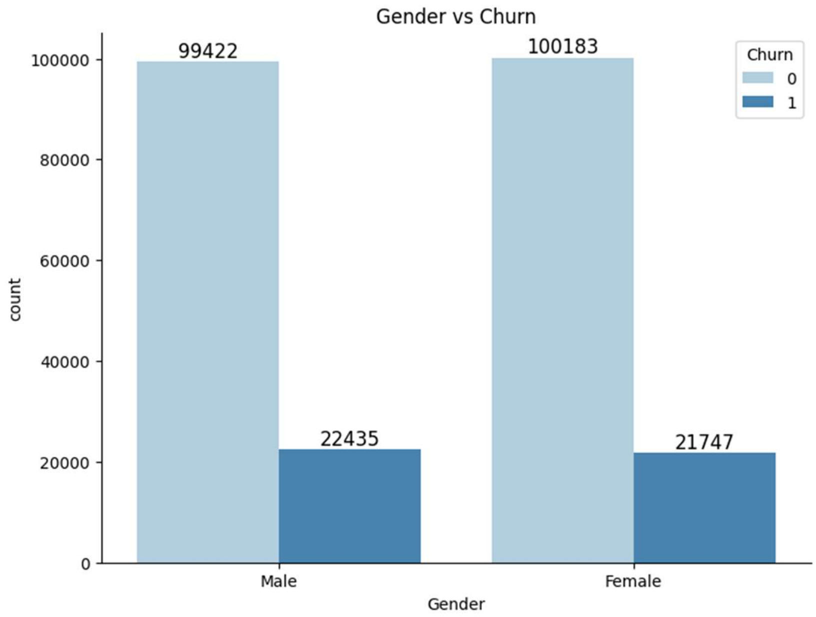
G) GenrePreference vs Churn



- Churn Rates by Registered Device:
    1. Tablet users show the highest churn (11,137), followed by Mobile (11,109), Computer (11,089), and TV (10,847)
- Retention Overview:
    1. Retention is consistent across all devices, with around 49,000 to 50,058 customers retained for each device type.

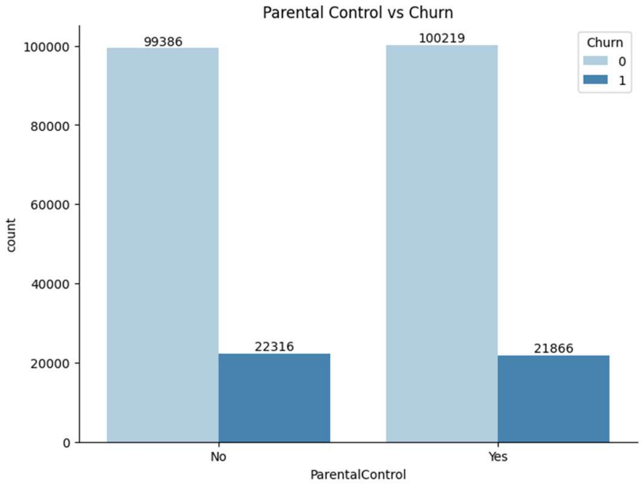
- **Insights:**
  The type of registered device has minimal impact on churn. Efforts to improve retention may need to focus on other areas like content engagement or subscription benefits rather than device-specific strategies.

H) Gender vs Churn



- Retention rates are nearly identical across genders, with slightly more males being retained.
- Churn rates are slightly higher for males (22,435) compared to females (21,747), though the difference is minor.
- Gender does not appear to have a significant impact on churn behavior, indicating that other factors may play a more prominent role.

I) ParentalControl vs Churn



- **Churn Rates by Parental Control:**
  1. Customers without parental control have a churn count of 22,316.

2. Customers with parental control enabled show a slightly lower churn of 21,866.

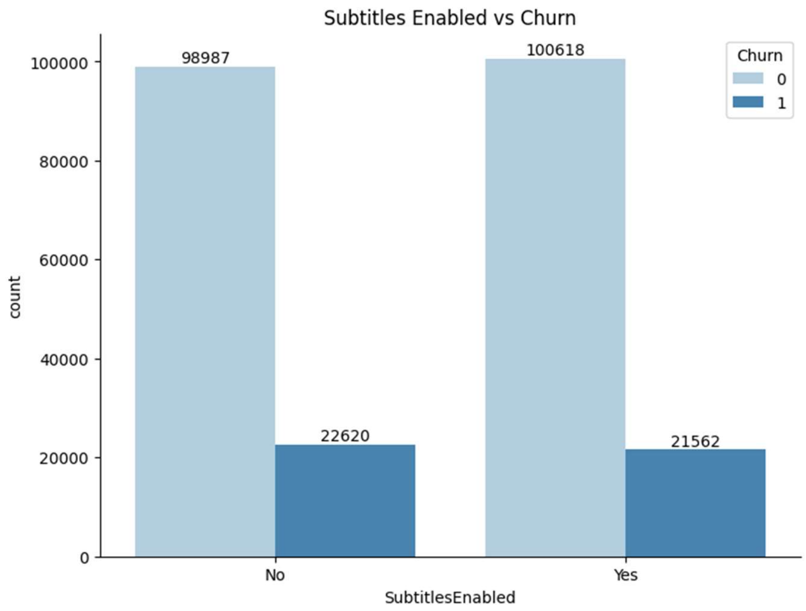
- **Retention Overview:**

3. Retention is consistent across both groups, with around 99,386 to 100,219 customers retained.

Insights:

- Enabling parental control has a minimal impact on churn reduction. Other engagement strategies might be needed to improve customer retention.

J) SubtitlesEnabled vs Churn

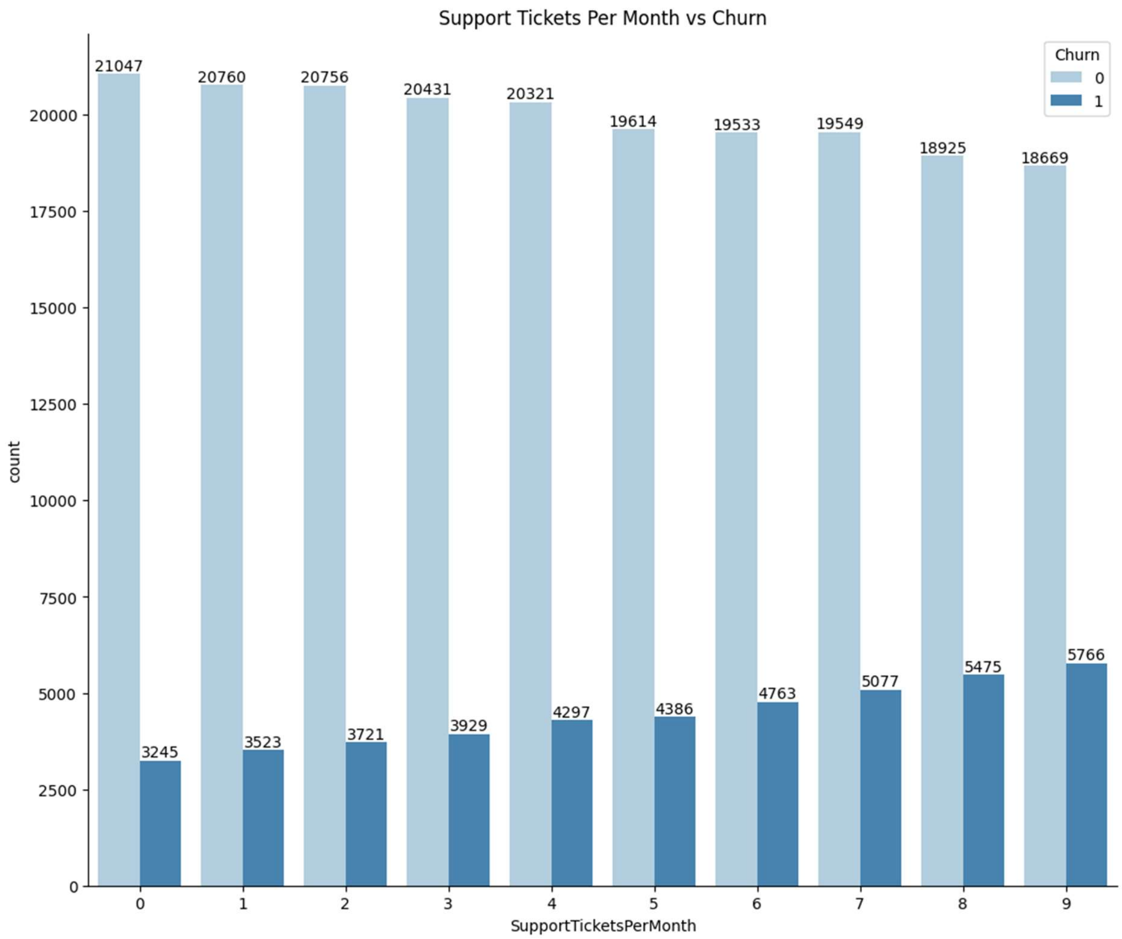


1. **Churn Rates by Subtitles Enabled:**
   - Customers without subtitles enabled have a churn count of 22,620.
   - Customers with subtitles enabled show slightly lower churn at 21,562.
2. **Retention Overview:**
   - Retention is consistent across both groups, with 98,987 to 100,618 customers retained.

Insights:

- Enabling subtitles has a minimal effect on churn reduction.
- Additional engagement or personalization strategies may be needed to improve retention.

K) SubtitlesEnabled vs Churn



1. **Churn Rates by Support Tickets per Month:**
   - Churn increases with the number of support tickets, with the highest churn at 5,766 for customers raising 9 tickets. Lower churn is observed among those with fewer tickets, e.g., 3,245 churn for customers with 0 tickets.
2. **Retention Overview:**
   - Retention decreases slightly as the number of support tickets rises, starting from 21,047 for 0 tickets to 18,669 for 9 tickets.

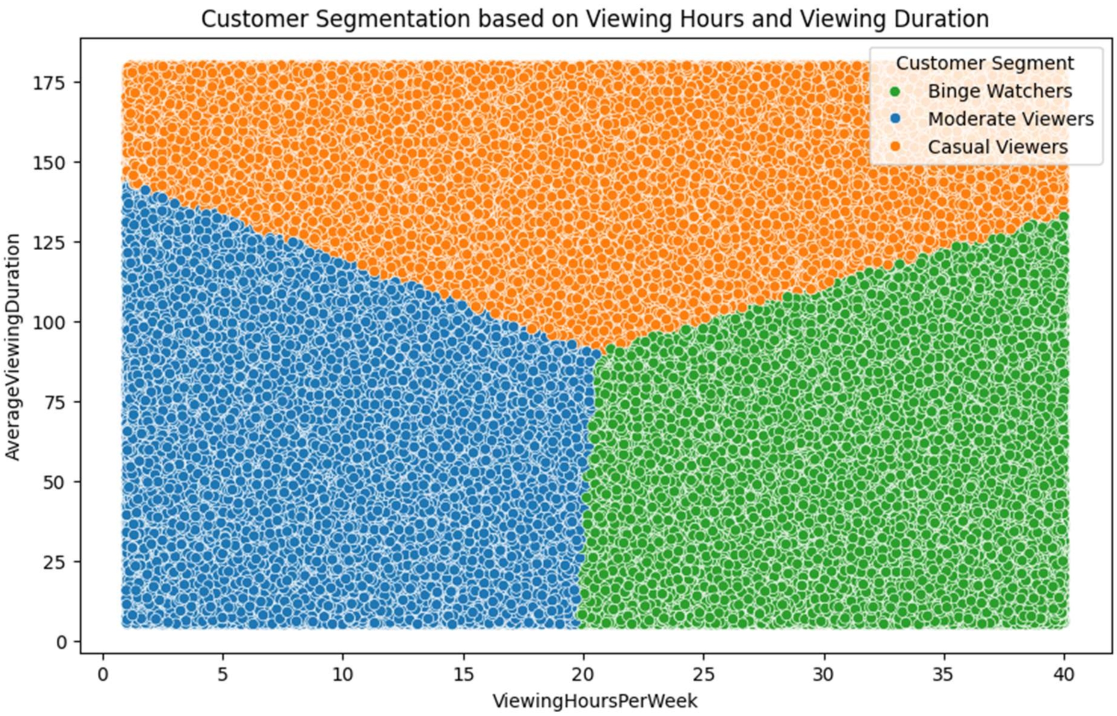
Insights:

- A higher number of support tickets correlates with increased churn, indicating possible customer dissatisfaction.
- Reducing ticket volume through proactive support and issue resolution could improve retention.

Customer Segmentation:

Segmentation was done based on the viewership by the user and we mainly formed clusters based on two features:

1. ViewingHoursPerWeek
2. AverageViewingDuration



1. **Binge-watchers (Green Segment):**

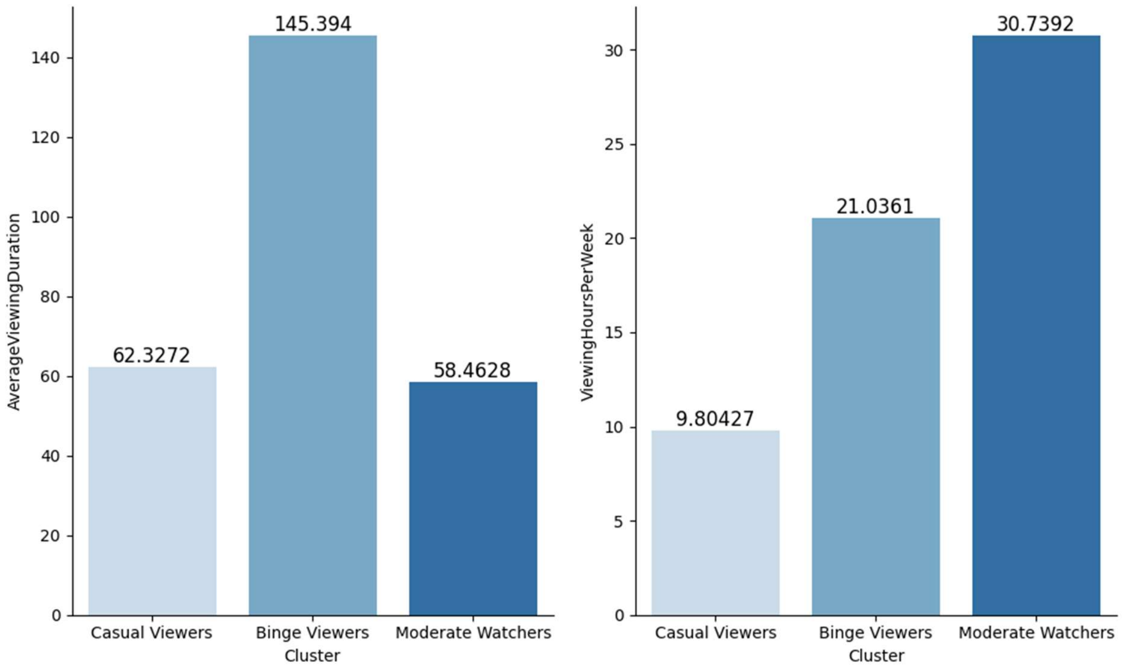
   - High Viewing Hours per Week and long session durations. These users are deeply engaged and likely to binge-watch content frequently. Strategy: Offer personalized recommendations for trending series, early access to new releases, and premium subscription upgrades.

2. **Casual Viewers (Blue Segment):**

   - Low Viewing Hours per Week and shorter session durations. These users engage infrequently, possibly only during free time. Strategy: Send engagement emails, suggest short-form content (e.g., documentaries, comedy specials), or offer discounts to retain them. Regular Viewers (Orange Segment):

3. **Moderate weekly hours and moderate session durations**

   - These customers show stable but not extreme engagement. Strategy: Promote family plans, introduce them to new genres, or offer mid-tier subscription options to encourage steady engagement.



Binge Viewers:

- Average Viewing Duration: 145.39 minutes
- Viewing Hours Per Week: 21.04 hours
- Engagement: High engagement with frequent binge-watching behavior.
- Strategy: Offer personalized recommendations, early access to new content, and premium subscription upgrades.

Casual Viewers:

- Average Viewing Duration: 62.33 minutes
- Viewing Hours Per Week: 9.80 hours
- Engagement: Low engagement, likely watching during free time.
- Strategy: Send engagement emails, recommend short-form content, and provide discounts to encourage retention.

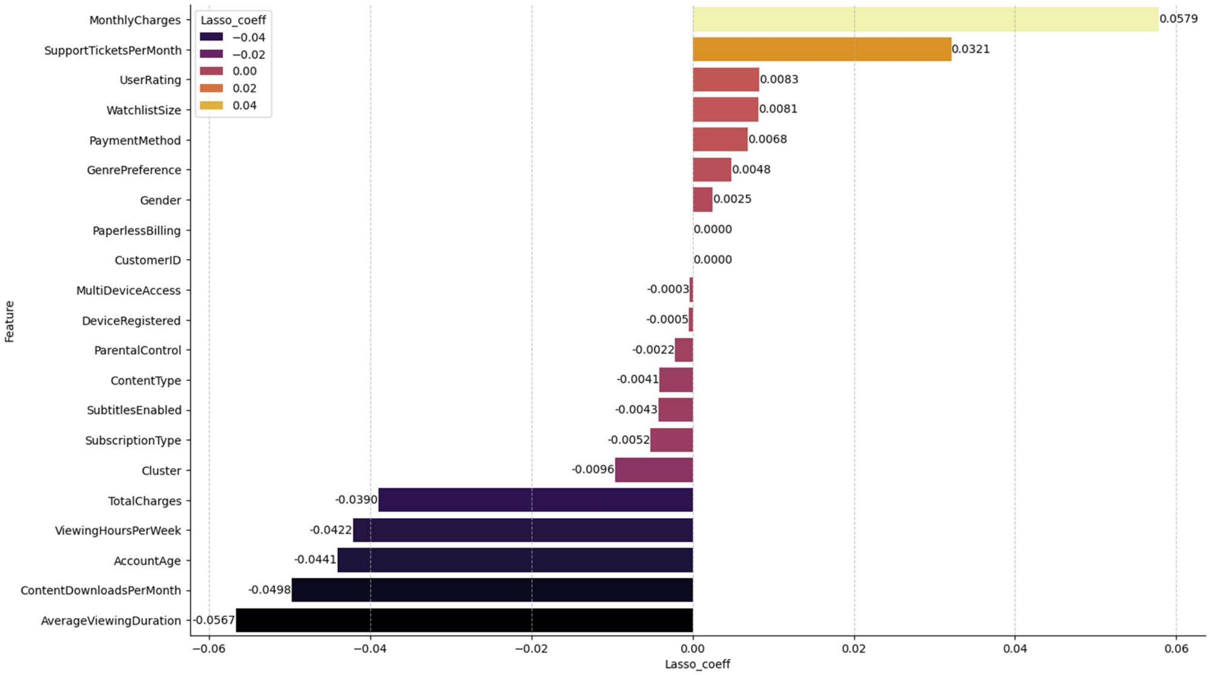
Moderate Watchers:

- Average Viewing Duration: 58.46 minutes
- Viewing Hours Per Week: 30.74 hours
- Engagement: Steady engagement with moderate viewing behavior.
- Strategy: Promote family plans, suggest new genres, or offer mid-tier subscription options to maintain steady engagement.

III) Modelling and Evaluation:

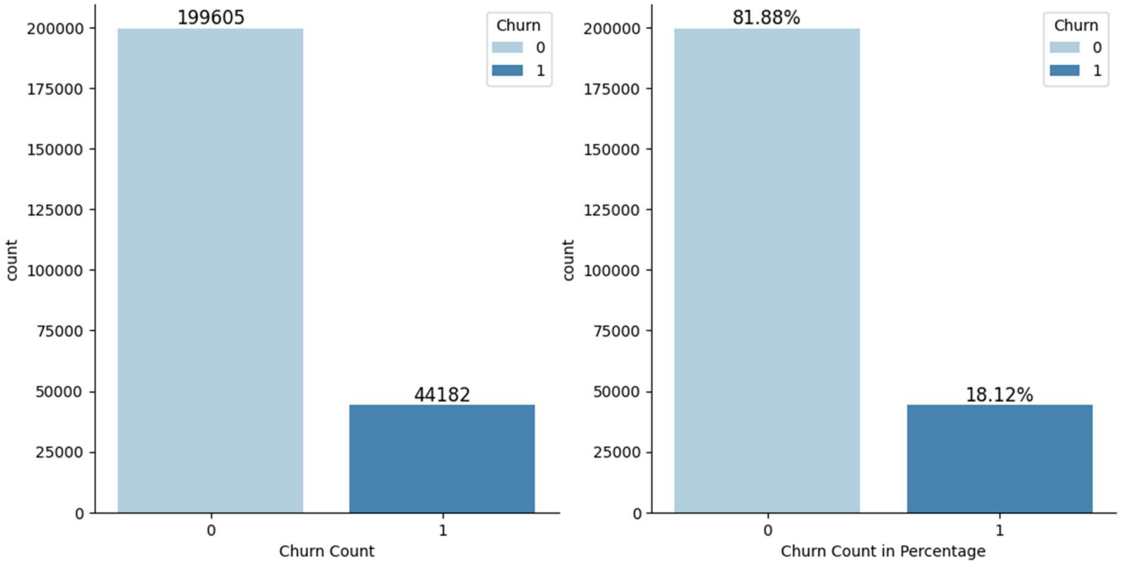
1. **Label Encoding:** Performed label encoding for the categorical columns below.
   SubscriptionType', 'PaymentMethod', 'PaperlessBilling', 'ContentType', 'MultiDeviceAccess', 'DeviceRegistered', 'GenrePreference', 'Gender', 'ParentalControl', 'SubtitlesEnabled', 'CustomerID'],dtype='object.

2. **Feature Selection:**
   There were many features in the dataset but we performed some feature selection method such as LassoCV to get the best features affecting our target variable.

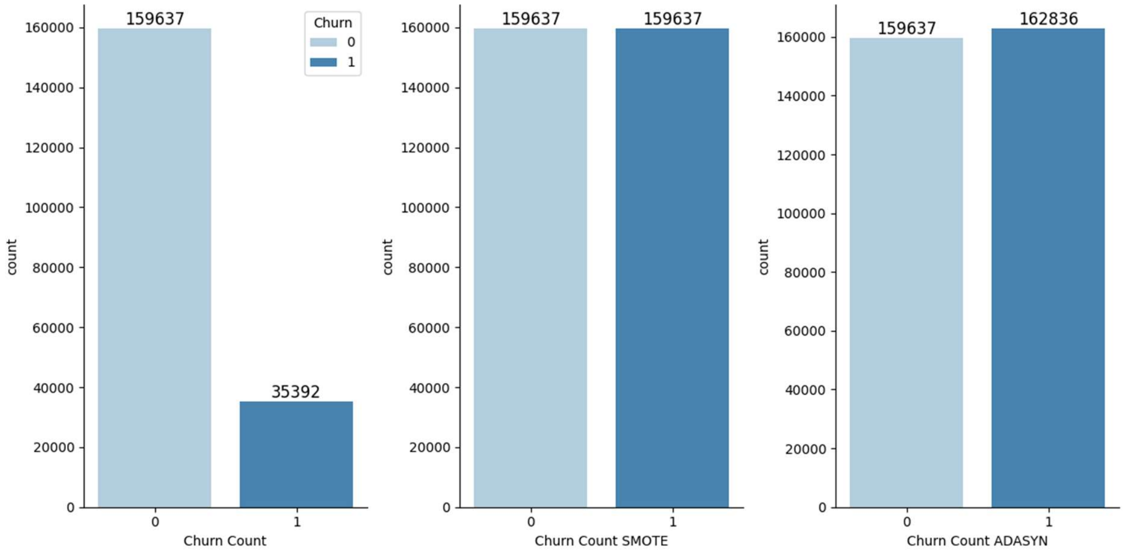


The features with co-efficient zero will be eliminated. Thus we eliminated Gender and Paperless Billing .

3. **Class Balance:**
    - Looking at the target class ('Churn') column we found that they are highly imbalanced which can affect our model training.



- It can be seen that the Churn class 1 just comprises of 18.12% whereas the Churn Class 0 comprises about 81.88% . This shows that our data is highly imbalanced.
- We can perform various class balance techniques to balance the target variable.
- We performed SMOTE (**Synthetic Minority Oversampling technique**) and ADASYN (**Adaptive Synthetic Sampling**) for class balance.
- The figure below shows the target class balance using SMOTE and ADASYN.



4. Model Accuracy and Classification Reports

- On the dataset after performing requisite preprocessing we mainly used Logistic Regression, Random Forest, KNN Classifier and XG Boost models in various combinations with SMOTE, ADASYN and class_weight=balanced which is a class balance feature in the ML algorithm itself.
- For KNN Classifiers instead of Class_weight balanced we use model without any class balance to check performance.
- In XGBoost, scale_weight is a parameter that adjusts for class imbalance by giving more importance to the minority class. It is calculated as the ratio of the number of non-churned to churned samples (Churn_no / Churn_yes).

1. Logistic Regression with SMOTE

- Test Set Accuracy: 72.08%
- Precision (class 0): 0.87 | Recall (class 0): 0.78| F1-score (class 0): 0.82
- Precision (class 1): 0.32 | Recall (class 1): 0.47 | F1-score (class 1): 0.38

2. Logistic Regression with Class Weight Balanced

- Test Set Accuracy: 67.68%
- Precision (class 0): 0.91| Recall (class 0): 0.67 | F1-score (class 0): 0.77
- Precision (class 1): 0.32 | Recall (class 1): 0.69 | F1-score (class 1): 0.44

3. Logistic Regression with ADASYN

- Test Set Accuracy: 71.64%
- Precision (class 0): 0.87 | Recall (class 1): 0.77 | F1-score (class 1): 0.82
- Precision (class 1): 0.31 | Recall (class 1): 0.47 | F1-score (class 1): 0.38

4. Random Forest with Class Weight Balanced

- Test Set Accuracy: 82.12%
- Precision (class 0): 0.83 | Recall (class 1): 0.99 | F1-score (class 1): 0.90
- Precision (class 1): 0.54 | Recall (class 1): 0.05 | F1-score (class 1): 0.10

5. Random Forest with SMOTE

- Test Set Accuracy: 77.45%
- Precision (class 0): 0.85 | Recall (class 1): 0.88 | F1-score (class 1): 0.86
- Precision (class 1): 0.35 | Recall (class 1): 0.29 | F1-score (class 1): 0.32

6. Random Forest with ADASYN

- Test Set Accuracy: 77.28%
- Precision (class 0): 0.85 | Recall (class 1): 0.88 | F1-score (class 1): 0.86
- Precision (class 1): 0.35 | Recall (class 1): 0.30 | F1-score (class 1): 0.32

7. KNN Classifier with SMOTE

- Test Set Accuracy: 66.38%
- Precision (class 0): 0.86 | Recall (class 0): 0.71 | F1-score (class 1): 0.78
- Precision (class 1): 0.26 | Recall (class 1): 0.46 | F1-score (class 1): 0.33

8. **KNN Classifier with ADASYN**

- Test Set Accuracy: 80%
- Precision (class 0): 0.86 | Recall (class 0): 0.70 | F1-score (class 1): 0.77
- Precision (class 1): 0.36 | Recall (class 1): 0.14 | F1-score (class 1): 0.20

9. KNNClassifier without class balance

- Test Set Accuracy: 75.12%
- Precision (class 0): 0.83 | Recall (class 0): 0.95 | F1-score (class 1): 0.89
- Precision (class 1): 0.33 | Recall (class 1): 0.38 | F1-score (class 1): 0.35

10. XGBoost with SMOTE

- Test Set Accuracy: 75.12%
- Precision (class 0): 0.86 | Recall (class 0): 0.83 | F1-score (class 1): 0.85
- Precision (class 1): 0.33 | Recall (class 1): 0.38 | F1-score (class 1): 0.35

11. **XGBoost with ADASYN**

- Test Set Accuracy: 74.83%
- Precision (class 0): 0.86 | Recall (class 0): 0.83 | F1-score (class 1): 0.84
- Precision (class 1): 0.33 | Recall (class 1): 0.38 | F1-score (class 1): 0.35

12. **XGBoost with Scale_weight for Class balance**

- Test Set Accuracy: 67.88%
- Precision (class 0): 0.91 | Recall (class 0): 0.68 | F1-score (class 0): 0.78
- Precision (class 1): 0.32 | Recall (class 1): 0.69 | F1-score (class 1): 0.44

Based on the model summary of all the above models, the best models that performed well on the overall accuracy alongwith a good recall score for Class 1 are as below.

- 2) Logistic Regression with Class_weight = balanced with
- 12)XGBoost technique with scale_weight.

5. Hyperparameter Tuning

- Performing Hyper parameter tuning on the above two best models using GridSearchCV to see if we can get a better accuracy and recall.
- After performing hyperparameter tuning on Logistic Regression Model with class_weight = balanced it was found that there was not considerable change in the accuracy, precision or recall for the model.
- Performing Hyperparameter tuning on XGBoost model gave us an increment of 3% in the accuracy although the recall for class 1 has reduced by 0.08%.

13. The specifics are as below

XGBoost with Scale_weight for Class balance(Hyperparameter Tuning)

- Test Set Accuracy: 70.58%
- Precision (class 0): 0.89 | Recall (class 0): 0.73 | F1-score (class 0): 0.80
- Precision (class 1): 0.33 | Recall (class 1): 0.61 | F1-score (class 1): 0.43

IV) Summary:

- In today's highly competitive streaming industry, reducing customer churn is essential for profitability, especially given the high costs associated with customer acquisition. This analysis focused on identifying the main factors driving customer churn and segmenting customers based on their engagement patterns to improve retention strategies.
- Using data on subscription plans, payment methods, device registrations, and content preferences, several machine learning models were evaluated for their predictive accuracy.
- The best performing models were Logistic Regression with balanced class weights (accuracy: 67.68%, Class 1 recall: 0.69) and XGBoost with scale weight after hyperparameter tuning (accuracy: 70.58%, Class 1 recall: 061) for class balance, although the latter experienced a minor decrease in recall following hyperparameter tuning.
- **Key insights highlighted** that churn rates were higher among customers on basic or standard subscription plans and those using electronic payment methods, likely due to fewer features or payment-related frustrations.
- Device type and access to multi-device usage showed minimal impact on churn, suggesting that other personalized engagement features might be more effective in retaining customers. Furthermore, customer segmentation identified three main groups—binge-watchers, casual viewers, and moderate watchers—each with unique engagement behaviors and needs. These insights indicate that targeted strategies based on both churn drivers and customer segments could be effective in improving overall retention.

V) Recommendation:

1. **Revenue Retention and CLV Enhancement**: Implement targeted retention strategies for high-risk customers using insights from churn prediction, such as offering discounts, personalized content, or loyalty rewards. This approach not only reduces revenue loss but also enhances customer lifetime value by encouraging longer-term engagement.
2. **Subscription Plan Optimization**: Address higher churn rates among Basic and Standard plan users by promoting upsells to Premium or introducing mid-tier plans with additional features that align with their preferences. This can be achieved through personalized campaigns that highlight the benefits of upgrading to improve satisfaction and reduce churn.

3. **Payment Method Optimization**: Given the increased churn among electronic check users, incentivize the adoption of automated, hassle-free payment methods like credit cards or bank transfers. Promotions or small discounts for auto-payment setups can help retain customers and reduce churn caused by payment-related issues.

4. **Segment-Specific Retention Strategies:**

- **Binge-watchers:** Keep these highly engaged customers satisfied by offering early access to new releases, premium recommendations, and subscription upgrades. Regularly updating content offerings with trending shows or exclusive access will strengthen their loyalty.
- **Casual Viewers:** Engage these infrequent users with targeted emails promoting short-form content (e.g., documentaries, specials) and offer incentives like time-limited discounts to increase their usage frequency.
- **Moderate Watchers:** Promote family plans, diverse genre recommendations, or mid-tier packages to cater to these users' steady engagement levels. Content exploration **incentives can increase satisfaction and retention in this segment.**

5. **Enhanced Customer Support:** Improve support experiences for users who frequently raise tickets by offering quicker response times or proactive outreach for recurring issues. Providing quality support can reduce dissatisfaction, lower churn risk, and build customer trust.
6. **Optimized Content and Marketing Strategy**: Leverage customer segmentation insights to deliver tailored content and marketing messages that match each segment's preferences, improving engagement and satisfaction. Personalized content recommendations and targeted campaigns will drive a more immersive, satisfying user experience.