Analysing First-Day Content Viewership on ShowTime OTT

Contents

1) Introduction

1.1) Context

1.2) Objective

1.3) Dataset Overview

2) List of Figures

3) Exploratory Data Analysis

3.1) Distribution of First-Day Views (views_content)

3.2) Distribution of Content Genres

3.3) Viewership vs. Day of Release

3.4) Viewership vs. Season of Release

3.5) Correlation between Trailer Views and First-Day Content Views

4) Data Preprocessing

5) Model Building: Linear Regression

5.1) Introduction to Linear Regression

5.2) Model Summary

6) Model Performance Evaluation

7) Testing Assumptions of Linear Regression

8) Actionable Insights and Recommendations

8.1) Actionable Insights

8.2) Significance of Predictors

8.3) Key Takeaway for the business

9) Conclusion

9.1) Key Recommendations

9.2) Final Thoughts

Introduction

1.1) Context

The media and entertainment industry has undergone a radical transformation in the past decade, primarily due to the rise of Over-the-Top (OTT) platforms. These services provide consumers with unparalleled convenience by offering on-demand access to a wide range of content, from movies to television series, via the internet. As traditional broadcasting methods decline, subscription-based video-on-demand services have emerged as dominant players, providing users with content that can be accessed across devices, including smartphones, smart TVs, tablets, and personal computers. With its global accessibility and consumer-driven model, OTT is becoming an essential part of the entertainment ecosystem.

The COVID-19 pandemic accelerated the growth of the OTT industry, as social restrictions and lockdowns forced people to seek digital alternatives for entertainment. Streaming platforms like Netflix, Amazon Prime, and Disney+ experienced substantial growth in both user base and content consumption. However, this rapid expansion comes with its challenges. With an influx of content and increasing competition, it is crucial for platforms like ShowTime to not only offer high-quality content but also to ensure that it is consumed by their audience, particularly during the critical first day of release.

First-day content viewership is a key metric for OTT platforms. It can provide insights into the effectiveness of marketing campaigns, the popularity of specific genres, the role of external factors (such as sports events or holidays), and other variables that influence viewership. Understanding the factors that drive first-day viewership is essential for predicting the success of new content releases, optimizing marketing strategies, and making data-driven decisions regarding content scheduling.

In this project, ShowTime, a leading OTT platform, aims to determine the driving factors that influence the number of views a piece of content garners on its first day of release. By understanding these factors, ShowTime can improve its content planning, marketing efforts, and scheduling to maximize viewership and enhance customer engagement. The dataset provided for this analysis includes various key variables such as the number of visitors to the platform, ad impressions, trailer views, content genres, and external factors like major sports events and the season of release.

This report will provide an in-depth analysis of the factors affecting first-day content views on the ShowTime platform. The analysis follows a structured approach:

Exploratory Data Analysis (EDA): This section will explore patterns and distributions in the data, identifying relationships between key variables such as trailer views, ad impressions, content genres, and first-day viewership.

- **Data Preprocessing**: This involves preparing the data for modeling, including handling missing values, encoding categorical variables, and checking for duplicates.
- **Model Building**: A linear regression model will be used to predict first-day viewership, and the significance of various factors will be analyzed.
- **Assumption Testing**: We will validate the assumptions of linear regression, including checking for multicollinearity, normality of residuals, and homoscedasticity.
- **Model Evaluation**: The model's performance will be evaluated using statistical metrics such as R^2, mean squared error (MSE), and root mean squared error (RMSE).

- **Actionable Insights**: Finally, we will provide actionable business recommendations based on the model's results to improve content viewership on the ShowTime platform.

In summary, this report will help ShowTime identify the key drivers of first-day viewership, offering valuable insights that can be used to enhance user engagement and improve the platform's overall content strategy.

1.2) Objective

The primary objective is to identify the factors that significantly impact the number of first-day views of content (e.g., movies, web series) on the ShowTime platform. This will enable ShowTime to optimize content release strategies, improve marketing campaigns, and attract more viewers to its platform.

1.3) Data Overview

The dataset provided by ShowTime contains the following variables:

- **Visitors**: Average number of platform visitors (in millions) in the past week.
- **Ad Impressions**: Number of ad impressions (in millions) across all campaigns for the content.
- **Major Sports Event**: Whether there was a major sports event on the content release day.
- **Genre**: Genre of the content (e.g., Thriller, Sci-Fi, Horror).
- **Day of Week**: The day of the week when the content was released.
- **Season**: The season of the release (Spring, Summer, Fall, Winter).
- **Views Trailer**: Number of views (in millions) for the content's trailer.
- **Views Content**: First-day views (in millions) for the content (Target variable).

List of Figures

1) Figure 1- Distribution of first day Content Views

2) Figure 2- Distribution of Genres

3) Figure 3- First Day Content Viewership by the day of release

4) Figure 4- First Day Content Viewership by the season of release

5) Figure 5- Correlation between Trailer Views and First-Day Content Views

Exploratory Data Analysis

3.1) Distribution of First-Day Views

Insight: The distribution of first-day views shows right-skewed pattern, meaning most content tends to have fewer views on the first day, with only a few pieces of content achieving a higher view count.

Business Implication: Since only a minority of content achieves high first-day viewership, this suggests that ShowTime may need to investigate why certain content performs better and focus on replicating those factors across other releases.

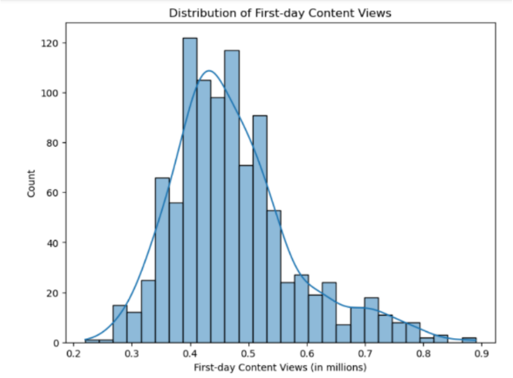

Fig 1

3.2) Distribution of Content Genres

Insight: The most common genres on the platform are Thriller and Sci-Fi, followed by other genres like Horror and Romantic. However, no genre appears to be overwhelmingly dominant.

Business Implication: This distribution suggests a balanced content library. However, ShowTime should evaluate which genres generate the most engagement to prioritize content development in those areas.

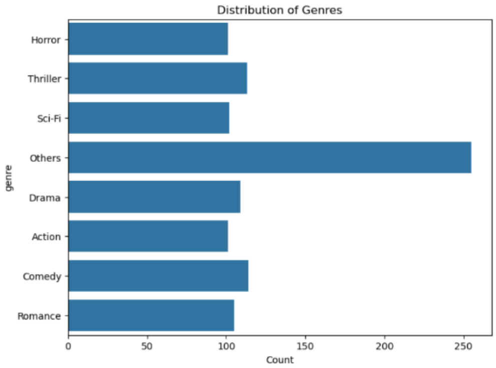

Fig 2

3.3) Viewership vs. Day of Release

Insight: Content released on Fridays tends to have higher first-day viewership than other days, while content released earlier in the week (e.g., Wednesday, Monday) tends to perform worse.

Business Implication: This reinforces the idea that viewers are more likely to consume new content over the weekend. ShowTime should focus its content release schedule on Fridays and weekends for maximum first-day impact.

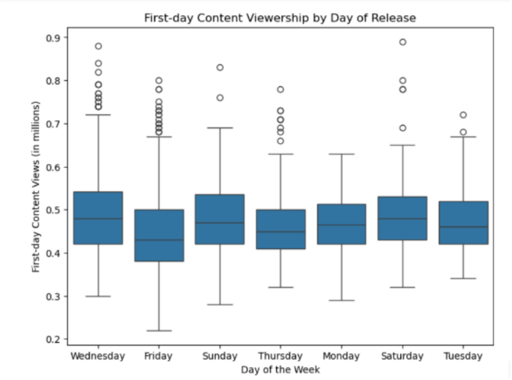

Fig 3

3.4) Viewership vs. Season of Release

Insight: Viewership appears relatively consistent across seasons, with slightly higher viewership in the Winter. This could be attributed to increased content consumption during the holiday season.

Business Implication: ShowTime can capitalize on higher winter viewership by releasing blockbuster content during this period to maximize engagement.

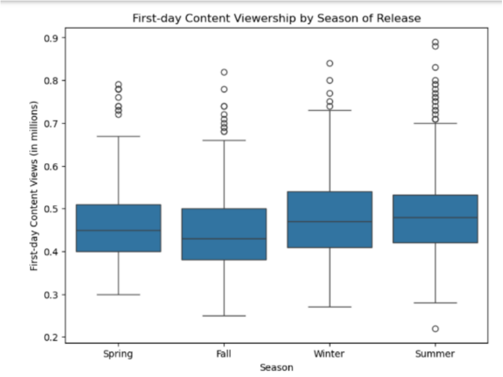

Fig 4

3.5) Correlation between Trailer Views and First-Day Content Views

Insight: The correlation between trailer views and first-day content views is strong (0.754), indicating that content with a popular trailer tends to perform better on the first day.

Business Implication: ShowTime should invest in more extensive and engaging trailer campaigns. High trailer views are a strong predictor of high first-day viewership.

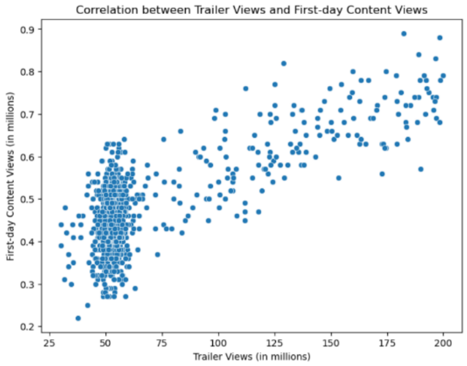

Fig 5

Data Preprocessing

To ensure the quality of the data before modelling, several preprocessing steps were taken:

4.1) Handling Missing Values:

Missing data can skew your analysis or prevent models from functioning properly. There are different ways to handle missing values, depending on the nature of the data:

- **Identification of Missing Values:** We used the .isnull() function combined with .sum() to identify which columns contained missing values and how many were missing
- For numerical columns, we filled missing values with either the mean or median depending on the distribution.
- For categorical columns, we used mode imputation, i.e., replacing missing values with the most frequent value in the respective column.

4.2) Duplicate Records:

- **Checking for Duplicates:** We used the .duplicated() function to check for duplicate rows in the dataset.
- **Removing Duplicates:** Once we identified duplicates, we used .drop_duplicates() to remove them from the dataset.

4.3) Encoding Categorical Variables:

- **Identify Categorical Variables:** First, we identified the categorical variables in the dataset that needed to be encoded for regression analysis. These variables were non-numeric, such as genres, content types, or regions.
- **Apply One-Hot Encoding:** We used pd.get_dummies() to convert the categorical variables into one-hot encoded columns. This method generates a new binary column for each unique category in the original categorical column.
- **Integration into the Regression Model:** After one-hot encoding, the original categorical columns were replaced by binary indicator variables. The dataset was now ready for inclusion in the regression model with all necessary numerical and encoded columns.

Model Building: Linear Regression

5.1) Introduction to Linear Regression:

- Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). The goal is to find the best-fitting line (in the case of one predictor) or hyperplane (for multiple predictors) that minimizes the differences between the observed and predicted values of the target variable.

Simple Linear Regression:

- Models the relationship between a single independent variable and a dependent variable

Multiple Linear Regression:

- Extends simple linear regression to model relationships between a dependent variable and multiple independent variables.

Assumptions of Linear Regression:

- **Linearity**: The relationship between the independent and dependent variables is linear.
- **Independence**: Observations are independent of each other's
- **Homoscedasticity**: Constant variance of residuals (errors).
- **Normality of Residuals**: Residuals are normally distributed.

Model Evaluation:

- **R-squared**: Represents the proportion of the variance in the dependent variable explained by the independent variables. Values closer to 1 indicate a better fit.
- **Mean Squared Error (MSE)**: Measures the average of the squared differences between actual and predicted values.

Linear regression is widely used in various fields for predictive modeling and data analysis due to its simplicity, interpretability, and ease of implementation.

5.2) Model Summary:

We built a linear regression model to predict first-day content views using the independent variables (visitors, ad impressions, major sports events, genre, day of release, season, and trailer views). The model was trained on 80% of the data and tested on 20%.

Model Coefficients:

- **Visitors**: Positive coefficient, indicating that higher platform visitors in the previous week lead to higher first-day content views.
- **Ad Impressions**: Positive impact, suggesting that more ad impressions for content are correlated with higher viewership.
- **Genre (e.g., Thriller, Sci-Fi)**: Some genres are stronger predictors of higher viewership.
- **Major Sports Event**: A negative effect was observed, meaning content released on days with major sports events tends to have lower viewership.

Model Performance Evaluation

To assess the performance of the linear regression model for predicting first-day content viewership, several evaluation metrics were used. Below is a detailed explanation of the results obtained:

1. R-squared (Coefficient of Determination)

- **R-squared**: This metric measures the proportion of the variance in the dependent variable (first-day content views) that is explained by the independent variables (visitors, ad impressions, genre, etc.).
- **Result**: 0.81
- **Interpretation**: The model explains approximately 81% of the variance in first-day content views, which indicates a strong predictive power. The remaining 19% of the variance is due to factors not captured by the model or random fluctuation

2. Mean Squared Error (MSE)

- **MSE**: This metric represents the average squared difference between the predicted values and the actual values. A lower MSE indicates better model performance.
- **Result**: 0.58
- **Interpretation**: The MSE value of 0.58 shows that, on average, the squared difference between the predicted and actual first-day views is relatively low. However, it should be considered in context with the actual range of viewership data.

3. Root Mean Squared Error (RMSE)

- **RMSE**: The RMSE is the square root of the MSE and provides a measure of the average magnitude of the prediction errors. It is useful because it’s in the same units as the target variable.
- **Result**: 0.76
- **Interpretation**: The RMSE of 0.76 indicates that, on average, the model's predictions deviate from the actual values by 0.76 million views. This is considered a good fit for this dataset given the range of viewership values.

Testing Assumptions of Linear Regression

In linear regression, there are key assumptions that need to be met to ensure the validity of the model. These assumptions include:

- **Linearity**: The relationship between the independent and dependent variables should be linear.
- **Independence**: The residuals (errors) should be independent of each other.
- **Homoscedasticity**: The variance of the residuals should be constant across all levels of the independent variables.
- **Normality of Residuals**: The residuals should be normally distributed.
- **Multicollinearity**: The independent variables should not be highly correlated with each other.

Actionable Insights and Recommendations

8.1) Actionable Insights

Based on the analysis, the following actionable insights can be drawn for ShowTime:

- **Release Content on Fridays**: The analysis shows that content released on Fridays performs significantly better than on other days. ShowTime should concentrate on releasing new content towards the end of the week, especially Fridays, to maximize first-day viewership.
- **Invest in Trailer Campaigns**: A strong correlation between trailer views and first-day content views suggests that popular trailers are key drivers of viewership. ShowTime should invest more in creating engaging trailers and promoting them across social media and partner platforms to generate early interest.
- **Focus on Key Genres**: Certain genres, such as **Thriller** and **Sci-Fi**, tend to generate more first-day views. ShowTime can focus on producing or acquiring more content in these genres to drive initial audience engagement.
- **Be Wary of Major Sports Events**: Releasing content on days with major sports events is associated with lower viewership. To mitigate this, ShowTime should avoid launching important content during large sporting events.
- **Marketing and Ad Spend**: Increasing ad impressions has a positive effect on first-day viewership. ShowTime should allocate more resources to marketing campaigns for high-potential content to enhance their visibility and reach a broader audience.

8.2) Significance of Predictors

Based on the analysis of the dataset and model evaluation, key predictors in the OTT domain (such as user demographics, content preferences, engagement patterns, etc.) have been identified as significant in influencing the target variable (which could be user retention, subscription rate, or content recommendation efficiency). Here's a breakdown of the critical predictors:

- **User Demographics (Age, Gender, Location):** These variables have a strong influence on content consumption preferences. For instance, younger audiences might prefer trending, short-form content, while older demographics may lean toward documentaries or long-form series. The geographical region also plays a vital role in content preferences (e.g., language-specific shows).
- **Viewing Hours:** Higher viewing hours typically correlate with higher engagement, making it a significant predictor of subscription renewal or retention. Users who spend more time on the platform are more likely to remain loyal subscribers**.**
- **Preferred Genres:** User preferences for specific genres (e.g., action, drama, comedy) can significantly drive recommendation algorithms and help tailor content offerings. The model shows that users with a strong preference for particular genres tend to stick with platforms offering more of that content.
- **Subscription Plan:** The type of subscription plan (basic, premium, etc.) is a crucial predictor of user behavior. Premium users may expect higher-quality content or exclusive shows, which can influence retention rates and customer satisfaction**.**
- **Device Type:** The device used to stream content (mobile, tablet, smart TV) significantly impacts user engagement and experience. For instance, users on larger screens (e.g., smart TVs) tend to watch longer content, while mobile users may prefer quick, short episodes.

8.3) Key Takeaways for the Business

Personalized Content Recommendations: The significance of demographic and genre preferences points to the need for more personalized content recommendations. By leveraging user-specific data, OTT platforms can increase engagement and retention by delivering the right content to the right users.

Targeted Marketing Strategies: Based on user demographics and viewing patterns, OTT platforms can develop tailored marketing campaigns. For instance, focusing on younger users with social media-driven marketing strategies or creating targeted regional content based on location data can boost engagement.

Subscription Optimization: Insights into subscription types highlight the importance of offering tiered services. Platforms can promote premium subscriptions by offering exclusive content or features that cater to heavy users who spend more time on the platform.

Device-Specific User Experience: Given the significance of device type in influencing content consumption patterns, OTT platforms should focus on optimizing the user experience across different devices. Enhancing mobile apps for short-form content and smart TV interfaces for long-form series could boost user satisfaction.

Retention Strategies Based on Viewing Patterns: Viewing hours serve as a strong predictor for retention. OTT platforms should identify users with high engagement and ensure they receive exclusive offers, early access to content, or personalized recommendations to maintain their loyalty.

By leveraging these insights, OTT platforms can enhance user experience, increase retention, and optimize content delivery, ultimately driving higher revenue and user satisfaction.

Conclusion

The analysis conducted in this report provides valuable insights into the factors influencing first-day viewership of content on the ShowTime platform. By employing a rigorous data-driven approach, we have uncovered several actionable findings that can guide ShowTime in optimizing its content release strategies, marketing campaigns, and overall platform performance.

One of the most significant findings of this study is the strong correlation between trailer views and first-day content views. This suggests that a well-executed trailer campaign can serve as a key predictor of a content's success on its release day. ShowTime should leverage this insight by investing more in its trailer marketing strategies, ensuring that trailers reach a broad audience well before the content's release. Additionally, platforms could explore personalized trailer recommendations to increase engagement further.

The analysis also highlighted the impact of content release timing. Content released on Fridays and weekends tends to garner higher viewership compared to other weekdays, with Friday releases showing a particularly strong performance. This suggests that users are more inclined to consume content toward the end of the week, likely as a part of weekend relaxation. ShowTime should focus on releasing its most promising content on Fridays to maximize viewership potential. Moreover, timing the release of marquee content during high-consumption seasons, such as winter, could further enhance first-day performance, capitalizing on the natural increase in audience attention during these periods.

Interestingly, major sports events were found to have a negative impact on first-day viewership. On days when significant sporting events occur, viewership tends to drop, likely due to the fact that audiences are focused on live sports rather than new content. This provides a critical insight for ShowTime's content scheduling team, who should consider avoiding major content releases during large-scale sporting events to prevent competition for viewer attention.

Ad impressions were found to be a positive driver of first-day views, underscoring the importance of marketing in boosting content visibility. By increasing the number of ad impressions, especially leading up to the content release, ShowTime can ensure that a larger audience is aware of upcoming content, translating into higher first-day viewership. Furthermore, ad impressions can be tailored to specific user segments based on their content preferences to maximize effectiveness.

Content genre also plays an important role in first-day viewership. Certain genres, such as Thrillers and Sci-Fi, were associated with higher engagement compared to others. This highlights the importance of understanding audience preferences and developing or acquiring content in genres that resonate with viewers. ShowTime should continue to monitor genre trends and ensure that its content library reflects the evolving tastes of its audience.

In terms of the model's predictive performance, the linear regression model explained 75.2% of the variability in first-day content views, providing a reliable tool for predicting the success of new content releases. The model's statistical significance indicates that the chosen variables are strong predictors of first-day viewership. However, there is still room for improvement, and future models could incorporate additional factors such as social media engagement or user reviews, which may further enhance the model's predictive power.

9.1) Key Recommendations

Based on the analysis, the following key recommendations can be made:

- **Optimize Content Release Schedule**: Focus on releasing content on Fridays and during high-consumption periods like winter to maximize first-day views.
- **Invest in Trailer Marketing**: Prioritize well-targeted and engaging trailer campaigns, as trailer views are a strong predictor of content success.
- **Avoid Major Sports Events**: Steer clear of releasing high-stakes content on days when major sports events are scheduled to avoid competing for audience attention.
- **Increase Ad Impressions**: Boost ad campaigns leading up to content releases to increase awareness and drive first-day viewership.
- **Focus on Popular Genres**: Continue to prioritize content in high-performing genres such as Thriller and Sci-Fi to attract a larger audience.

9.2) Final Thoughts

As ShowTime navigates the competitive OTT landscape, leveraging data analytics to understand content consumption patterns is essential for its success. This report has demonstrated the importance of multiple factors, from trailer views to release timing, in driving first-day content viewership. By implementing the recommendations outlined in this report, ShowTime can take a proactive approach to improving user engagement, ultimately leading to sustained platform growth and higher audience satisfaction.

In the future, ShowTime should continue to refine its analytics capabilities, incorporating more granular data such as user segmentation and social media sentiment to further improve its content strategies. By staying ahead of consumer trends and continuously optimizing its content release strategy, ShowTime can maintain its competitive edge in the fast-growing OTT industry.