PREDICTIVE MODELLING (PM)

Coded Project Report

Submitted to

by

Subhadeep Seal

in Partial Fulfillment of

PGP-DSBA

Table of Contents

Context

An over-the-top (OTT) media service is a media service offered directly to viewers via the internet. The term is most synonymous with subscription-based video-on-demand services that offer access to film and television content, including existing series acquired from other producers, as well as original content produced specifically for the service. They are typically accessed via websites on personal computers, apps on smartphones and tablets, or televisions with integrated Smart TV platforms.

Presently, OTT services are at a relatively nascent stage and are widely accepted as a trending technology across the globe. With the increasing change in customers' social behaviour, which is shifting from traditional subscriptions to broadcasting services and OTT on-demand video and music subscriptions every year, OTT streaming is expected to grow at a very fast pace. The global OTT market size was valued at $121.61 billion in 2019 and is projected to reach $1,039.03 billion by 2027, growing at a CAGR of 29.4% from 2020 to 2027. The shift from television to OTT services for entertainment is driven by benefits such as on-demand services, ease of access, and access to better networks and digital connectivity.

With the outbreak of COVID19, OTT services are striving to meet the growing entertainment appetite of viewers, with some platforms already experiencing a 46% increase in consumption and subscriber count as viewers seek fresh content. With innovations and advanced transformations, which will enable the customers to access everything they want in a single space, OTT platforms across the world are expected to increasingly attract subscribers on a concurrent basis.

Objective

ShowTime is an OTT service provider and offers a wide variety of content (movies, web shows, etc.) for its users. They want to determine the driver variables for first-day content viewership so that they can take necessary measures to improve the viewership of the content on their platform. Some of the reasons for the decline in viewership of content would be the decline in the number of people coming to the platform, decreased marketing spent, content timing clashes, weekends and holidays, etc. They have hired you as a Data Scientist, shared the data of the current content in their platform, and asked you to analyse the data and come up with a linear regression model to determine the driving factors for first-day viewership.

Data Description

The data contains the different factors to analyse for the content. The detailed data dictionary is given below.

Data Dictionary:

- **visitors:** Average number of visitors, in millions, to the platform in the past week
- **ad_impressions:** Number of ad impressions, in millions, across all ad campaigns for the content (running and completed)
- **major_sports_event:** Any major sports event on the day
- **genre:** Genre of the content
- **dayofweek**: Day of the release of the content
- **season:** Season of the release of the content
- **views_trailer:** Number of views, in millions, of the content trailer
- **views_content:** Number of first-day views, in millions, of the content

Understanding the structure of the data:

```
   visitors  ad_impressions major_sports_event     genre  dayofweek  season  \
0      1.67         1113.81                 0    Horror  Wednesday  Spring
1      1.46         1498.41                 1  Thriller     Friday    Fall
2      1.47         1079.19                 1  Thriller  Wednesday    Fall
3      1.85         1342.77                 1    Sci-Fi     Friday    Fall
4      1.46         1498.41                 0    Sci-Fi     Sunday  Winter

   views_trailer
0          56.70
1          52.69
2          48.74
3          49.81
4          55.83
0    0.51
1    0.32
2    0.39
3    0.44
4    0.46
Name: views_content, dtype: float64
```

The provided data structure is a dataset related to an OTT (Over-the-top) media service, which is a streaming service that offers online content to users. The dataset contains 5 rows, each representing a single observation or record, and 8 columns, each representing a feature or variable.

•The Dataset has been loaded properly.

•Dataset consists of several columns displaying the various attributes related to an OTT (Over-the-top) media service, which is a streaming service that offers online content to users.

•**visitors**: This column represents the number of visitors to the OTT platform.

•**major_sports_event:** This column is a binary indicator (0 or 1) representing whether if a major sports event occurred on the corresponding day. It is a relevant information for analyzing the impact of sports events on user engagement.

•**views_trailer:** This column represents the number of views for trailers on the platform.

•The day of the week and season might have an impact on user engagement.

•The Trailer views are relatively high as compared to the content views, which might indicate that users are more interested in previewing any content before watching it.

Showtime - Problem Statement:
Exploratory Data Analysis –

Problem definition, questions to be answered - Data background and contents - Univariate analysis - Bivariate analysis - Answers to the key questions provided - Insights based on EDA.

The below EDA section of the report will provide an overview of the data distribution and relationships between variables. The plots and statistics generated in this section will help to identify the patterns, outliers, and correlations in the data.

Important Note:
The following questions need to be answered as a part of the EDA section of the project:
1.What does the distribution of content views look like?

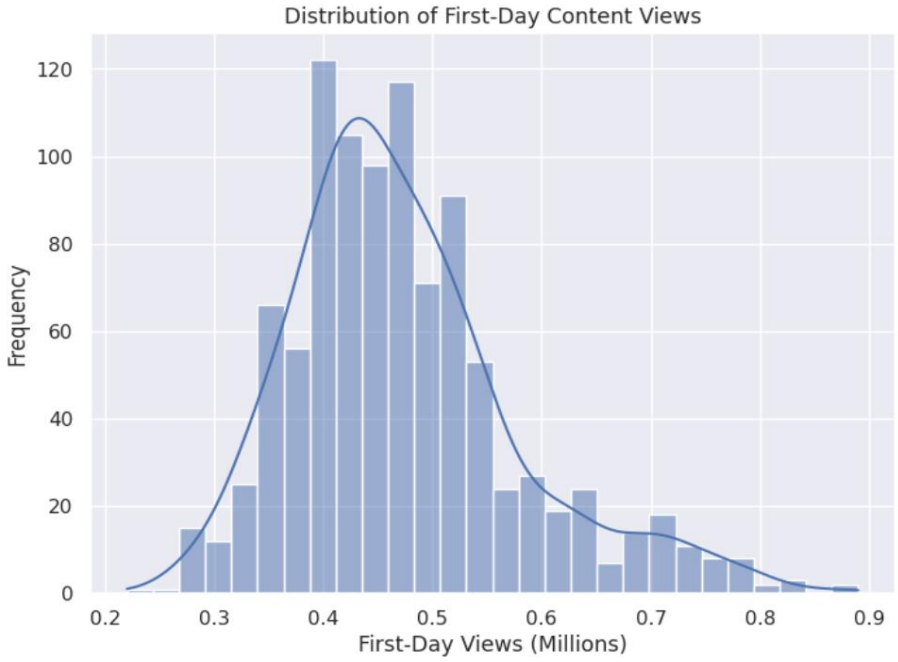


Fig: Distribution of Contents

The histogram shows a skewed distribution of content views, with most views concentrated around 0-1 million.

2.What does the distribution of genres look like?

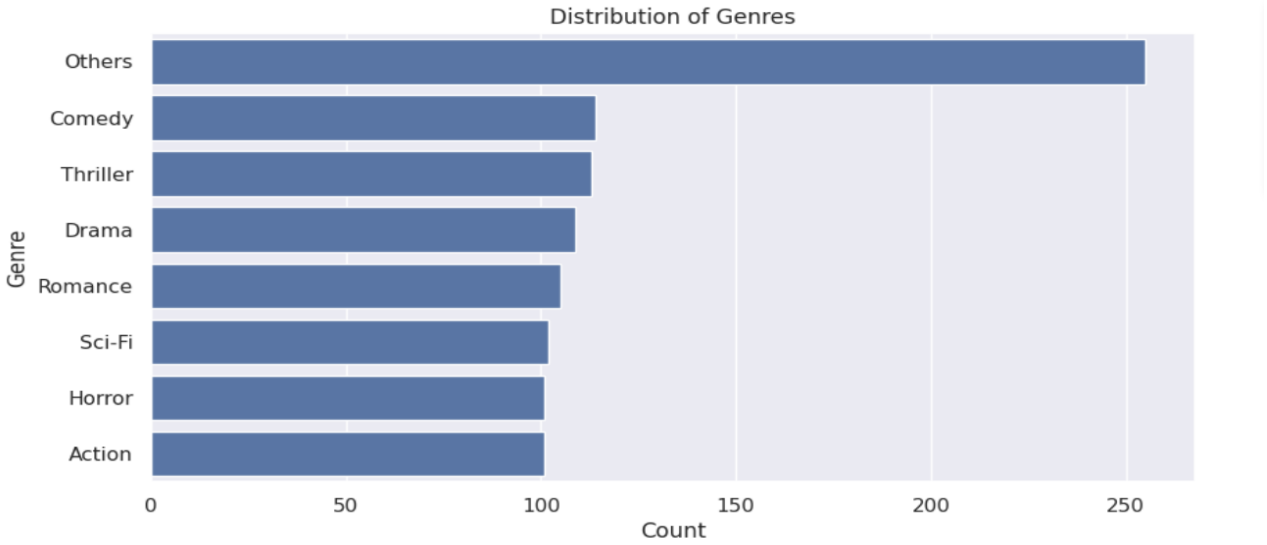


Fig: Distribution of Genres

The count plot reveals that Comedy is the most popular genre, followed by Thriller and drama.

3.The day of the week on which content is released generally plays a key role in the viewership. How does the viewership vary with the day of release?

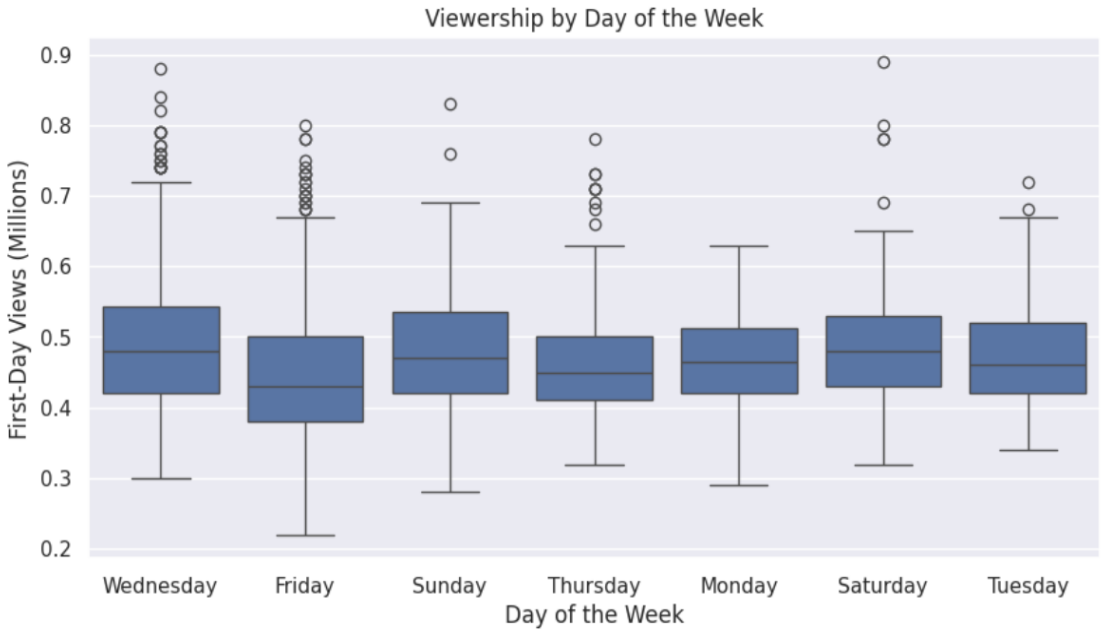


Fig: Viewership by Day of Release

The box plot shows that viewership is highest on Wednesdays and lowest on Thursdays.

4.How does the viewership vary with the season of release?

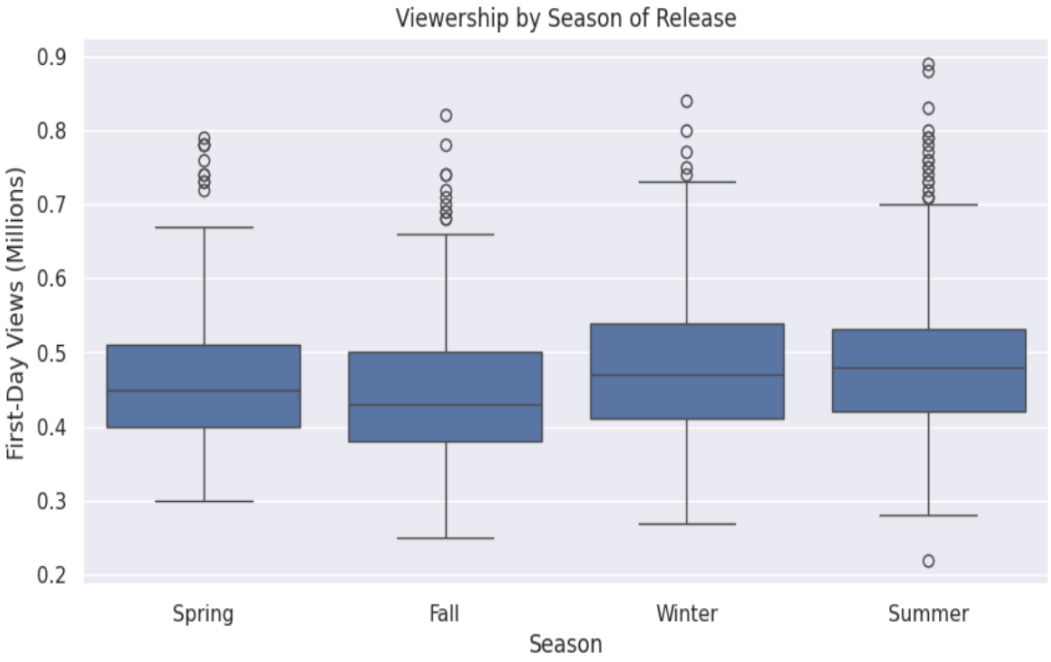


Fig: Viewership by Season of Release

The box plot indicates that viewership is highest in Winter and lowest in Fall.

5.What is the correlation between trailer views and content views?

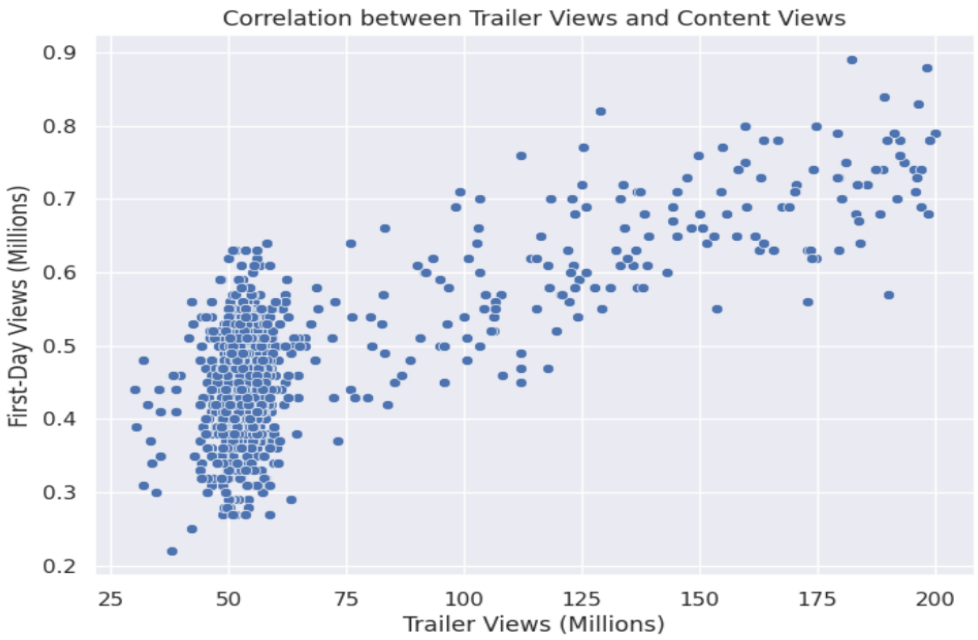


Fig: Correlation between Trailer Views & Content Views

The scatter plot and correlation coefficient (0.753) suggest a strong positive relationship between trailer views and content views.

Data preprocessing

Duplicate value check - Missing value treatment - Outlier treatment - Feature engineering - Data preparation for modelling.
Data Preprocessing includes the data for modelling by:
•Checking for duplicates.
•Checking for missing values.
•Detecting outliers in content views.
•Encoding categorical variables (genre, dayofweek, season) using one-hot encoding

a. Duplicate Value Check:

```
data.duplicated().sum()
```

```
0
```

No Duplicates are found in the given dataset.

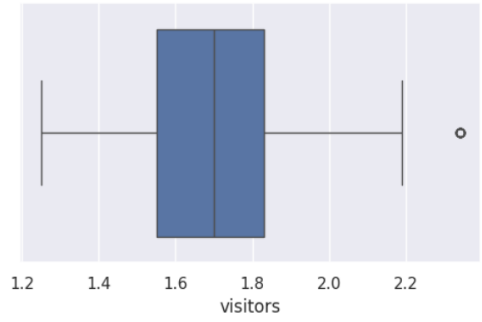
b. Missing Value Treatment:

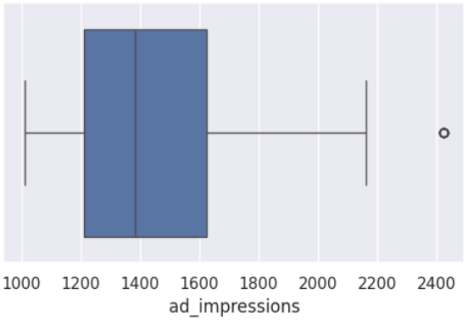
	0
visitors	0
ad_impressions	0
major_sports_event	0
genre	0
dayofweek	0
season	0
views_trailer	0
views_content	0

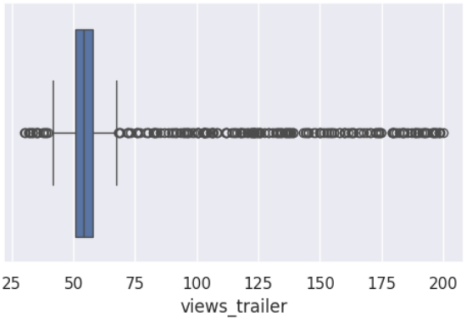
dtype: int64

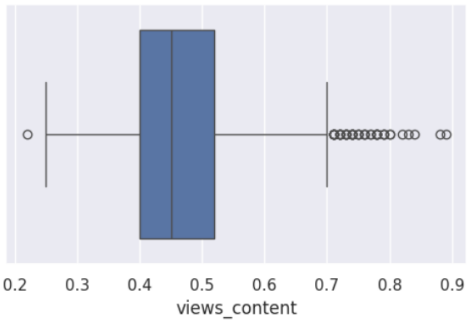
No Missing values are found in the given dataset.

c. Outlier Treatment:









- There are quite a few outliers in the data.
- However, we will not treat them as they are proper values.

d. Feature Engineering:
We have done the encoding for categorical variables (genre, dayofweek, season) using one-hot encoding.

Model building

Linear Regression - Build the model and comment on the model statistics - Display model coefficients with column names.
Building a linear regression model to predict content views based on the pre- processed data above and displaying the model's coefficients and statistics.
•Splitting the data into training and testing sets.
•Checking the shapes of the training and testing sets.

```
Number of rows in train data = 700
Number of rows in test data = 300
```

•Displaying the model coefficients with column names:

	Coefficient
const	0.000000
visitors	0.129451
ad_impressions	0.000004
views_trailer	0.002330
major_sports_event_1	-0.060326
genre_Comedy	0.009352
genre_Drama	0.012625
genre_Horror	0.009862
genre_Others	0.006325
genre_Romance	0.000551
genre_Sci-Fi	0.013143
genre_Thriller	0.008708
dayofweek_Monday	0.033662
dayofweek_Saturday	0.057887
dayofweek_Sunday	0.036321
dayofweek_Thursday	0.017289
dayofweek_Tuesday	0.022837
dayofweek_Wednesday	0.047376
season_Spring	0.022602
season_Summer	0.044203
season_Winter	0.027161

Testing the assumptions of linear regression model

- Perform tests for the assumptions of the linear regression - Comment on the findings from the tests

1. Linearity:

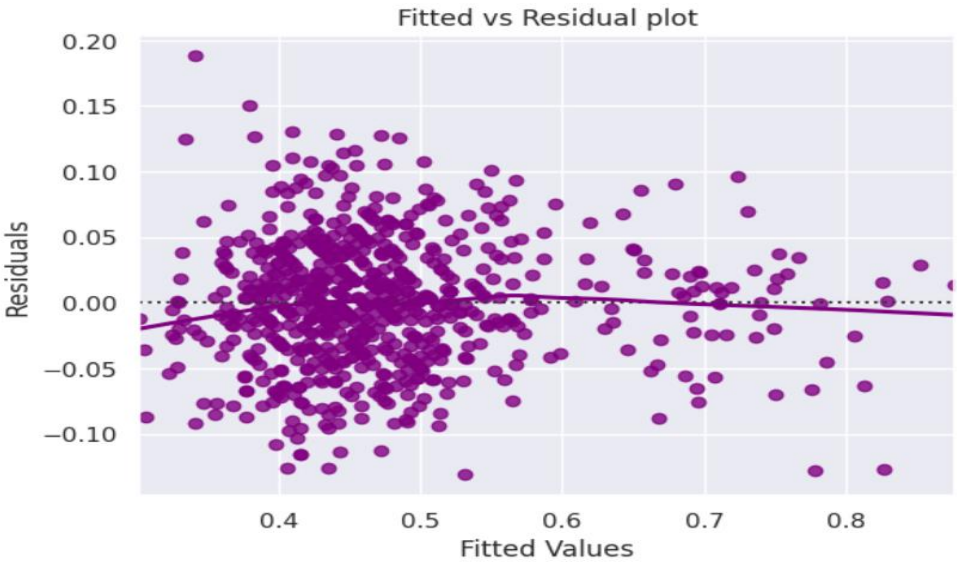


Fig: Residual Plot for Residuals vs Fitted values

- The scatter plot shows the distribution of residuals (errors) vs fitted values (predicted values).
- If there exist any pattern in this plot, we consider it as signs of non-linearity in the data and a pattern means that the model doesn't capture non-linear effects.
- **We see no pattern in the plot above. Hence, the assumptions of linearity and independence are satisfied.**

2. Normality of Residuals

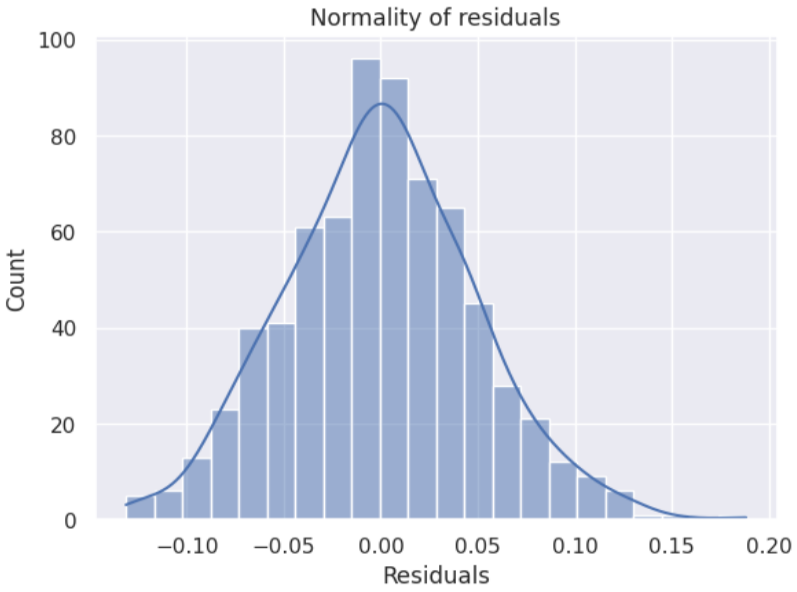


Fig: Histogram for Distribution of Residuals

- The histogram of residuals does have a bell shape.

3. Homoscedasticity

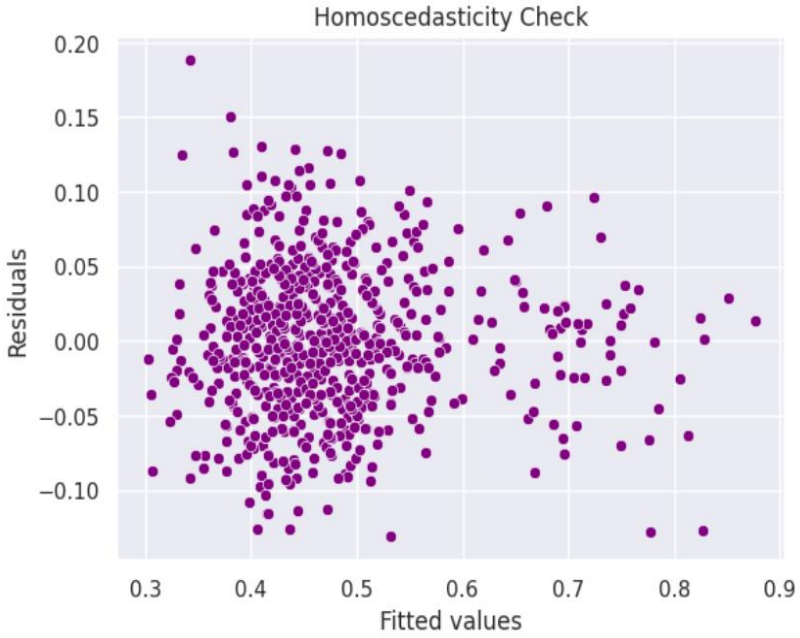


Fig: Scatter plot for Homoscedasticity

The scatter plot indicates constant variance of residuals across predicted views.

4. Independence

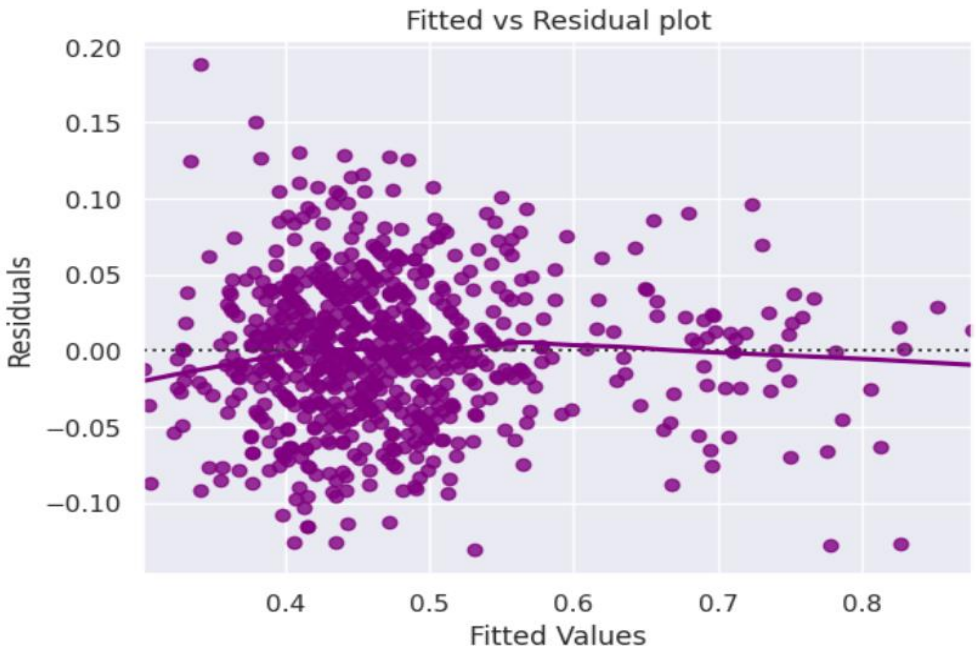


Fig: Scatter plot for Residuals vs Fitted values

The scatter plot shows the residuals are independent.

Model performance evaluation

Evaluate the model on different performance metrics

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.048841	0.038385	0.788937	0.785251	8.595246

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.051109	0.041299	0.761753	0.751792	9.177097

- **The model is able to explain ~79% of the variation in the data.**
- **The train and test RMSE and MAE are low and comparable. So, our model is not suffering from overfitting.**
- **The MAPE on the test set suggests we can predict within 9.2% on the first-day viewership.**
- **Hence, we can conclude the model is good for prediction as well as inference purposes.**

Actionable Insights & Recommendations

- Comments on significance of predictors - Key takeaways for the business

Below are the insights and recommendations based on the analysis:
• **Significance of Predictors:** The coefficients table shows that ad impressions, trailer views, and certain genres (e.g., Thriller) are significant predictors of content views.

- **Key Recommendations:**
  1. Increase marketing spend to improve ad impressions.
  2. Optimize content release schedules, avoiding clashes with major sports events.
  3. Promote trailers more to increase first-day viewership.

Overall, this report provides a comprehensive analysis of the OTT data, identifying key patterns, relationships, and predictors of content views. The insights and recommendations generated can inform business decisions to improve the OTT service's performance.