Project: Predictive Modelling (OTT Media Service)

Prepared by: Manya Verma
Role: Data Scientist

Objective: To provide logical guidance for improving first-day content viewership on an OTT platform through exploratory data analysis and linear regression modelling.

1. Introduction

This business report presents a comprehensive study on predicting first-day content viewership for an OTT platform. The aim is to identify key influencing factors and derive actionable insights using data science tools and techniques. The growing trend of OTT content consumption makes this analysis valuable for improving business strategies.

2. Problem Context and Stakeholders

The OTT industry is evolving rapidly, with user behavior shifting from traditional broadcasting to on-demand video services. A core challenge faced by OTT platforms is ensuring high viewership on new content during its release. Stakeholders impacted include:

- Content creators (seeking wider reach)
- Marketing teams (allocating budgets efficiently)
- Business leaders (optimizing ROI)

The key business question: What factors influence first-day content viewership?

3. Dataset Description

The dataset comprises records of released content on the platform and includes features believed to affect first-day viewership.

Variables in the dataset:

- visitors: Average weekly visitors to the platform (in millions)
- ad_impressions: Number of ad impressions for the content (in millions)
- major_sports_event: 0/1 indicating presence of a major sports event on release day
- genre: Genre of the content (e.g., Action, Comedy)
- dayofweek: Day of release
- season: Season of the year
- views_trailer: Number of trailer views (in millions)
- views_content: Target variable (first-day views in millions)

Special Notes:

- major_sports_event is a binary external event variable which may impact viewer interest.
- genre, dayofweek, and season are categorical variables that need encoding for modelling.
- views_trailer is a strong candidate to influence actual content views based on preliminary assumptions.

4. Exploratory Data Analysis (EDA)

Key takeaways from EDA:

- Content viewership is right-skewed, with most content performing under 5M views.
- Genres like Thriller and Sci-Fi have higher median views compared to others.

- Fridays and Sundays show stronger viewership patterns.
- There’s a strong positive correlation between trailer views and actual content views.
- Major sports events can negatively impact viewership.
- Seasonal effects are subtle, with slight upticks in Winter.

These insights helped in selecting relevant variables for the regression model.

5. Modelling Approach

A linear regression model was selected as a baseline predictive method due to its interpretability and suitability for numeric prediction tasks. Before modelling, the following steps were performed:
- One-hot encoding of categorical variables (genre, dayofweek, season)
- Splitting data into training and testing sets
- Checking for multicollinearity using VIF
- Validating assumptions: linearity, normality, homoscedasticity, and autocorrelation

6. Model Evaluation

Model performance was evaluated using:
- R^2 Score: Measures variance explained by the model
- MAE and RMSE: Error-based metrics

Findings:
- The model demonstrated reasonable predictive power on unseen data
- RMSE was within an acceptable margin for business-level decisions
- Residuals were approximately normally distributed and homoscedastic

7. Actionable Insights and Recommendations

- ▯ **Trailer Views**: Strongest positive predictor. Boosting trailer views through social media & app engagement can increase content visibility.
- ▯ **Ad Impressions**: Positively impacts views. Optimize ad campaigns especially during peak viewing times.
- ▯ **Day of Release**: Friday and Sunday releases yield better performance. Schedule key content accordingly.
- ▯ **Major Sports Events**: Negative impact. Avoid releases during major sports events.
- ▯ **Visitors**: Higher recent traffic contributes positively. Promote general platform usage prior to content drops.

8. Conclusion

The project successfully identified significant drivers of first-day viewership. The regression model provides a solid foundation for forecasting performance and guiding business strategy. Further improvements could include testing nonlinear models and incorporating user demographic or session-level engagement data.