PREDICTIVE MODELING PROJECT – BUSINESS REPORT

BY,

SHILPASHREE C M

Contents

1.1 Business Content 2
1.2 Objective 2
1.3 Data Description 2
2. Data Overview 3
3. Exploratory Data Analysis 4
4. Data Preparation and Modelling 5
5. Model Building – Linear Regression 5
6. Checking the linear Regression Assumption 6
6.1 Test for Multicollinearity 6
Dealing with high p-value variables 7
6.2 Test for Linearity and Independence 7
6.3 Test for Normality 8
6.4 Test for Homoscedasticity 10
6.5 Predictions on test data 10
7. Final Model 10
8. Conclusions and Recommendations 11

Figure 1: Correlation heatmap 4
Figure 2: viewership vary with the season of release 4
Figure 3: fitted vs. residual curve 8
Figure 4: Normality of residuals 9
Figure 5: probability plot 9

1. PROBLEM STATEMENT

1.1 Business Content

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.

1.2 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 spend, 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 analyze the data and come up with a linear regression model to determine the driving factors for first-day viewership.

1.3 Data Description

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

- 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

2. Data Overview

- The give data contains 1000 rows and 8 columns.
- There is no null values in the data.
- The datatype of the data contains 4 float columns, 3 object columns and 1 integer column.
- There is no duplicate values in the data.
- The statistical summary of the data is given below.
    - The average number of visitors is around 1.7 million, with a fairly narrow spread given the small standard deviation (0.232). The interquartile range (IQR) is from 1.55 to 1.83 million.
    - There is a wider spread in the number of ad impressions, with a mean of approximately 1434.712 million and a standard deviation of about289.535 million. The IQR ranges from 1210.330 to 1623.670 million.
    - The binary nature of this variable shows that 40% of the records had a major sports event, with a relatively high standard deviation due to its binary nature.
    - The average number of trailer views is around 66.916 million, with a notable spread as indicated by the standard deviation (35.001).
    - The first day views of content average at 0.473 million, with a narrow spread (standard deviation of 0.106).

3. Exploratory Data Analysis

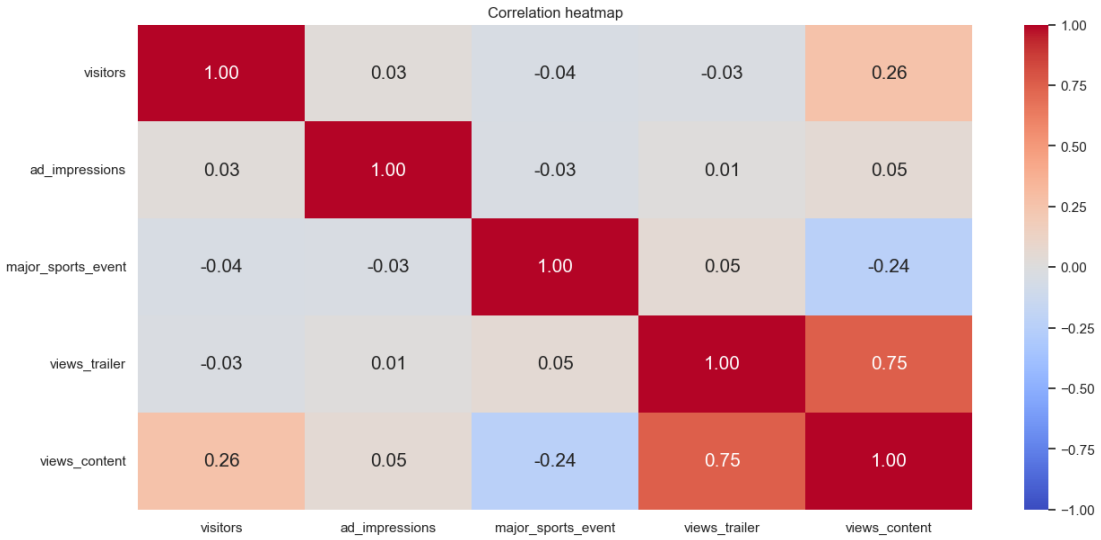

FIGURE 1: CORRELATION HEATMAP

- The correlation between trailer view and content view is approximately 0.75. This suggest a positive correlation.
- As the number of trailer views increases, the number of content views tend to increase as well.

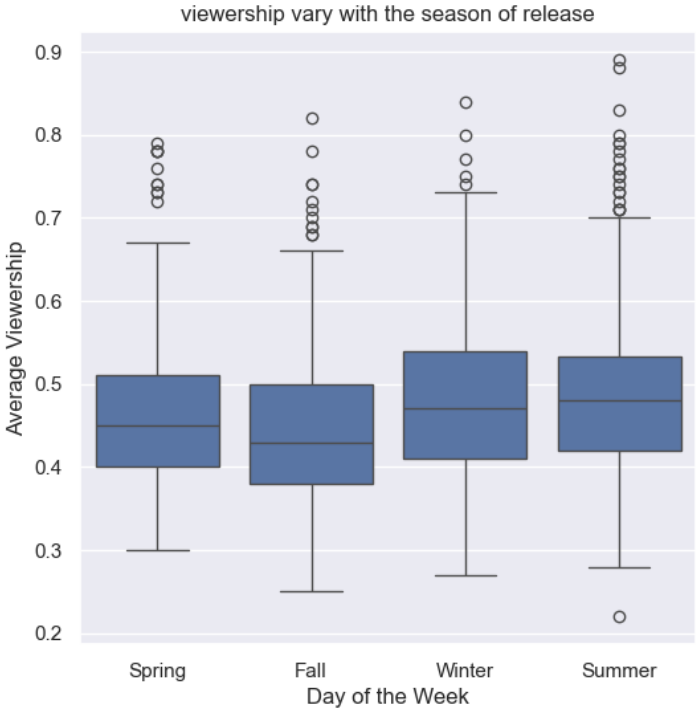

FIGURE 2: VIEWERSHIP VARY WITH THE SEASON OF RELEASE

- Winter seems to have the highest median viewership and many high outliers, suggesting it might be the best season for releasing content to achieve higher viewership.
- Spring and summer both have median viewership around 0.5 significant outliers, indicating potential for high viewership but also more variability.
- Fall shows a lower median viewership and fewer high outliers, it might be a less optimal season for releasing content compared to the other season.

4. Data Preparation and Modelling

- We want to predict the viewership of the content.
- Before we proceed to build a model, we'll have to encode the categorical feature.
- We will split the data into train and test to be able to evaluate the model that we build on the train data.
- We will build a Linear Regression model using the train data and then check its performance.
- Splitting the data in 70:30 ratio for train and test data. Number of rows in train data is 700 and number of rows in test data is 300.

5. Model Building – Linear Regression

- Adjusted R-squared: It reflects the fit of the model.
    - Adjusted R-squared values generally range from 0 to 1, where a higher value generally indicates a better fit, assuming certain conditions are met.
    - In our case, the value for adj.R-squared is 0.729, which is good.
- Const coefficient: It is the y-intercept.
    - It means that if all the predictor variable coefficients are zero, then the expected output (i.e., y) would be equal to the const coefficient.
    - In our case, the value for const coefficient is 0.0602.
- Coefficient to a predictor variable: It represents the change in the output y due to a change in the predictor variable (everything else held constant).
    - In our case, the coefficient of visitors is 0.1295.

Model performance.

- We will be using metric function defined in sklearn for RMSE, MAE, and r^2.
- We will define a function to calculate MAPE and adjusted R^2.
    - The mean absolute percentage error (MAPE) measures the accuracy of predictions as a percentage, and can be calculated as the average absolute percent error for each predicted value minus actual values divided by actual values. It works best if there are no extreme values in the data and none of the actual values are 0.

- We will create a function which will print out all the above metrics in one go.
- The training R^2 is 0.79, so the model is not underfitting.
- The train and test RMSE and MAE are comparable, so the model is not overfitting either.
- MAE suggests that the model can predict anime ratings within a mean error of 0.04 on the test data.
- MAPE of 9.03 on the test data means that we are able to predict within 9.03% of the viewership of the content

6. Checking the linear Regression Assumption

We will be checking the following Linear Regression assumptions:

1. No Multicollinearity
2. Linearity of variable
3. Independence of error terms
4. No Heteroscedasticity

6.1 Test for Multicollinearity

- Multicollinearity occurs when predictor variables in a regression model are correlated. This correlation is a problem because predictor variables should be independent. If the correlation between variables is high, it can cause problems when we fit the model and interpret the results. When we have multicollinearity in the linear model, the coefficients that the model suggests are unreliable.

- There are different ways of detecting (or testing) multicollinearity. One such way is by using the Variance Inflation Factor, or VIF.
- Variance Inflation Factor (VIF): Variance inflation factors measure the inflation in the variances of the regression parameter estimates due to collinearities that exist among the predictors. It is a measure of how much the variance of the estimated regression coefficient βk is "inflated" by the existence of correlation among the predictor variables in the model.
    - If VIF is 1, then there is no correlation among the kth predictor and the remaining predictor variables, and hence, the variance of βk is not inflated at all.
- General Rule of thumb:
    - If VIF is between 1 and 5, then there is low Multicollinearity.
    - If VIF is between 5 and 10, we say there is moderate multicollinearity.
    - If VIF is exceeding 10, it shows signs of high multicollinearity.

- There is no columns with very high VIF values other than constant (intercept) indicating absence of strong multicollinearity.
- We will systematically drop numerical columns with VIF > 5.
- We will ignore the VIF values for dummy variables and the constants (intercept).
- Let's rebuild the model using the set of predictors variables

- std err: It reflects the level of accuracy of the coefficients.
    - The lower it is, the higher the lever of accuracy.
- P>|t|: It is p-value.
    - For each independent feature, there is a null hypothesis and an alternative hypothesis. Here is the coefficient of the βith independent variable.
    - H0 : Independent feature is not significant (ßi = 0)
    - Ha : Independent feature is that it is significant (ßi ≠ 0)
    - (P>|t|) gives the p-value for each independent feature to check that null hypothesis. We are considering 0.05 (5%) as significance level.
    - A p-value of less than 0.05 is considered to be statistically significant.
- Confidence Interval: It represents the range in which our coefficients are likely to fall (with a likelihood of 95%).
- There is no multicollinearity, we can look at the p-values of predictor variables to check their significance.

Dealing with high p-value variables

- Some of the dummy variables in the data have p-values > 0.05. So, they are not significant and we'll drop them.
- But sometimes p-values change after dropping a variable. So, we'll not drop all variables at once instead, we will do the following:
    - Build a model, check the p-values of the variables, and drop the column with the highest p-value.
    - Create a new model without the dropped feature, check the p-values of the variables, and drop the column with the highest p-value.
    - Repeat the above two steps till there are no columns with p-value > 0.05.
- Now no feature has p-value greater than 0.05, so we'll consider the features in x_train1 as the final set of predictor variables and olsmod1 as the final model to move forward with.
- Now adjusted R-squared is 0.786, i.e., our model is able to explain ≈79% of the variance.
- RMSE and MAE values are comparable for train and test sets, indicating that the model is not overfitting.

6.2 Test for Linearity and Independence

- Linearity describes a straight-line relationship between two variables, predictor variables must have a linear relation with the dependent variable.

- The independence of the error terms (or residuals) is important. If the residuals are not independent, then the confidence intervals of the coefficient estimates will be narrower and make us incorrectly conclude a parameter to be statistically significant.

To check the linearity and independence

- Make a plot of fitted values vs. residuals.
- If they don't follow any pattern, then we say the model is linear and residuals are independent.
- Otherwise, the model is showing signs of non-linearity and residuals are not independent.
- To fix this assumption, if this is not followed, we can try to transform the variables and make the relationships linear.

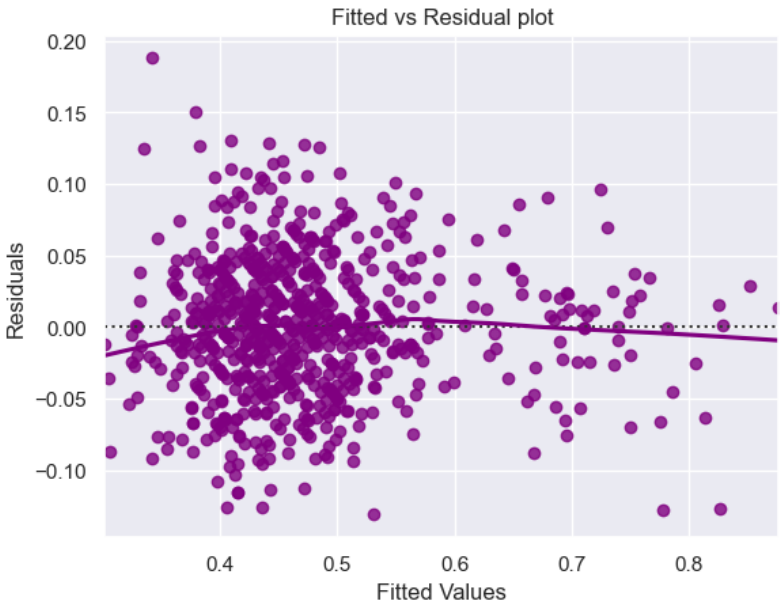

FIGURE 3: FITTED VS. RESIDUAL CURVE

- 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.

6.3 Test for Normality

- Error terms or residuals should be normally distributed. If the error terms are not normally distributed, confidence intervals of the coefficient estimates may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares. Non-normality suggests that there are a few unusual data points that must be studied closely to make a better model.
- To check normality, the shape of the histogram of residuals can give an initial idea about the normality.

- It can also be checked via a Q-Q plot of residuals. If the residuals follow a normal distribution, they will make a straight line plot, otherwise not.
- Other tests to check for normality includes the shapiro-wilk test.
    - Null hypothesis: Residuals are normally distributed.
    - Alternative hypothesis: Residuals are not normally distributed.
- To fix this assumption if it is not followed, we can apply transformations like log, exponential, arcsinh, etc. as per our data

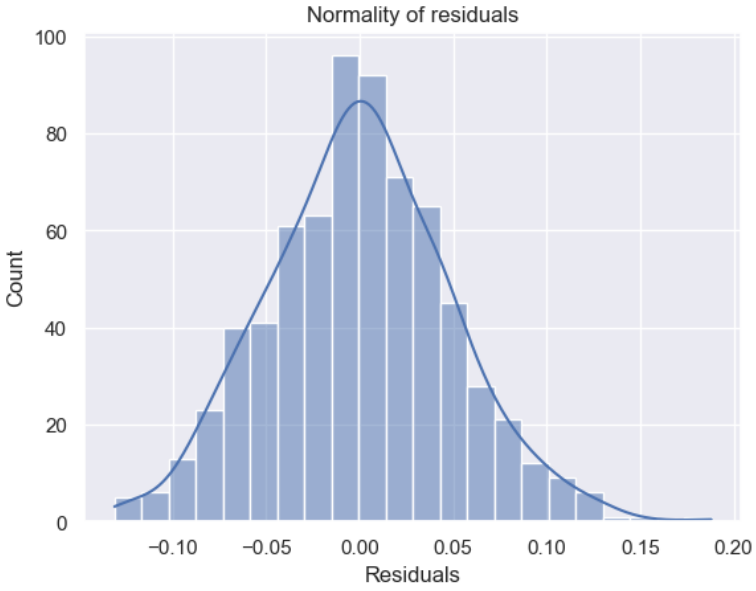

FIGURE 4: NORMALITY OF RESIDUALS

- The histogram of residuals have a bell shape.
- Let's check the Q-Q plot.

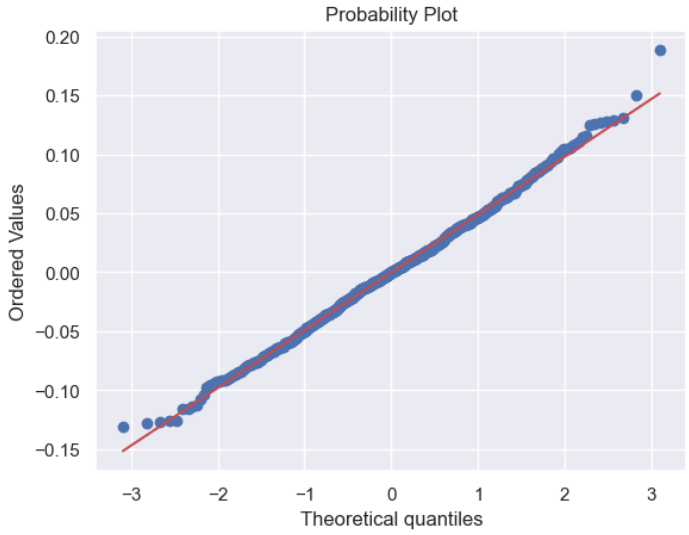

FIGURE 5: PROBABILITY PLOT

    - The residuals more or less follow a straight line.
    - Let's check the results of the Shapiro-Wilk test.

- Since p-value < 0.05, the residuals are not normal as per the Shapiro-Wilk test.
- Strictly speaking, the residuals are not normal.
- However, as an approximation, we can accept this distribution as close to being normal.
- **So, the assumption is satisfied.**

6.4 Test for Homoscedasticity

- **Homoscedasticity:** If the variance of the residuals is symmetrically distributed across the regression line, then the data is said to be homoscedastic.
- **Heteroscedasticity:** If the variance is unequal for the residuals across the regression lie, then the data is said to be heteroscedastic.
- The presence of non-constant variance in the error terms results in heteroscedasticity. Generally, non-constant variance areas in presence of outliers.
- To check for homoscedasticity, the residual vs. values plot can be looked at to check for homoscedasticity. In the case of heteroscedasticity, the residuals can form an arrow shape or any other non-symmetrical shape.
- The goldfeldquandt test can also be used. If we get a p-value > 0.05 we can say that the residuals are homoscedastic otherwise they are heteroscedastic.
    - Null hypothesis: Residuals are homoscedasticity.
    - Alternative hypothesis: Residuals have heteroscedasticity.
- To fix assumption if it is not followed, heteroscedasticity can be fixed by adding other important features or making transformations.

Since p-value > 0.05, we can say that the residuals are homoscedastic. So, this assumption is satisfied.

6.5 Predictions on test data

- Now that we have checked all the assumptions of linear regression and they are satisfied, let's go ahead with prediction.
- We can observe that our model has returned pretty good prediction results, and the actual and predicted values are comparable.

7. Final Model

- The model is able to explain ~75% of the variation of 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% of the viewers of content of the data.

- Hence, we can conclude the model olsmodel_final is good for prediction as well as inference purpose.

8. Conclusions and Recommendations

- The model is able to explain ~75% of the variation in the data and within 9.2% of the viewership of content on the test data, which is good. This indicates that the model is good and for prediction as well as inference purpose.
- If the visitors of an OTT platform increases by one unit then its viewers of content increases by 0.123 units. All other variables held constant.
- If the major of events on OTT platform increases by one unit then its rating decreases by 0.061 units. All other variables held constant.
- If the viewers of the trailer on OTT platform increases by on unit then its viewers of content increases by 0.0023 units. All other variables held constant.
- As the viewers of content on OTT platform increase with the increasing of the visitors to the OTT platform, the company need to improve its marketing activities to increase their viewers.
- As the viewers of the content increases with increase in its content, the company can look to add more content which is attracted by viewers to their portal.
- Company need to gather data about their users like age, gender, geographical location, occupation etc., to better understand the kind of genre and shows different users like.