Movies on streaming platform analysis

Walker

Good afternoon everyone, today we are going to present the movies on streaming platform analysis.

Problem statement: p.2

Nowadays, with the fast development of online movie streaming platforms, people can access more abundant resources for movies than without having to travel to the movie theatre. However, there are many streaming platforms available for selection, and often subscriptions for membership are required for consumers to access the content.

Also, some new movies produced by the streaming platform can only be viewed on their website and must be bought separately. Which streaming platforms should you choose so that you can access more content that you are interested in? Should you pay for that newly released movie that has few reviews? To help consumers make more informed choices before they purchase a membership or a new movie, we did a series of analyses with two datasets on four streaming platforms, which are Prime Video, Netflix, Hulu and Disney+. (翻頁）

p.3 1. Introduction of the dataset

Before diving into our analysis, we want to briefly introduce our two datasets.(翻頁）

p.4

1. The first set of data contains around 16 thousand movie records, and their features such as runtime and language, together with their IMDb rating. We used the first dataset in Part 1 and Part 3. And we used it in the case study.
2. The second dataset contains movie records, and their release date on the corresponding platform, which includes Netflix, Hulu and Disney+. (翻頁）

p.5 **Approach to the Problem**

Our analysis is structured in three parts:

1. Exploratory data analysis of our movie data to navigate platform-based characteristics of movies
2. Forecasting the number of new movies available on each platform(Hulu, Disney, Netflix)
3. Modelling the relationship between features of a movie and its IMDb rating

Thomas

1. Exploratory Data Analysis result

In case study 25, some analysis has been done to elucidate the characteristics of different platforms. Let us take a look.(Part 01)(翻頁）

1. *(First we examined) Platforms by the number of movies they have(First, we examine the quantities of movies across platforms)*

   The pie chart shows that movies from Prime Video take up 70% percent of all movies in the dataset. Assuming that the dataset is unbiased, we can conclude that Prime Video has the most abundant movie resources from the plot. While Netflix ranks second in quantity.

   Next, we look at the Bar chart of High quality movies on different platforms(翻頁）

2. *(Secondly, we look at) Platforms by the number and percentage of different qualities movies*

   Prime Video still possesses a dominant advantage when we limit our scope to movies with IMDb ratings larger than 8.5.

   As Prime Video always has an advantage in quantities, we transform our data to see the percentage of IMDb 8.5+ movies across platforms.(翻頁）From the bar chart, Prime Video surprisingly performs the best, with a percentage of 0.6% IMDb 8.5+ movies（翻頁）.

   Yet the four pie charts here also suggest that Prime Video has a higher percentage of bad movies(iIMDb rating 2-4) than other platforms, meaning that you have a higher chance of running into a bad movie if you randomly select a production to watch.

After that, we further explore platform-based characteristics on produced year, prevalent genre and contents exclusive to the platform.(翻頁）

3. *(Thirdly, we look at) Platforms by the number of movies available with different production years from 1915 to 2020.*

   Some audiences may be especially interested in movies produced during a specific period because they fancy the time's ideology and the films' unique textures that resulted from the filming techniques back then. E.g., movies in black and white. So we do a visualisation to see shares of movies on different platforms produced from 1915-2020. As a result, prime video has a larger share of movies across most of the periods. So if you like productions before 2000, Prime Video is a great choice.

4. (In addition, we examine) *Platforms by their top ten genres of movies*(翻頁）

   We generate four bar plots where the number of movies of the top ten genres in each platform is displayed. This can be a visual aid for the audience to select platforms with a perfect mix of their preferred genre. For example, if you like comedy, you can choose Netflix. For Drama, you may choose Prime video.

5. (After that, we further visualise different)*Platforms shares of top 50 movies genre focusing in documentaries, action, animation, and Adventure* (翻頁）

   We can see that while the prime video still has most of the best documentaries, other Platforms have fairer shares in action movies. When it comes to genres like animation and adventure, Disney becomes the optimal choice. Therefore, people may consider subscribing to Disney if they or their kids like animation.

6. (Finally we look at)*Platforms by their percentage of exclusive contents*(翻頁）

   Exclusive contents are essential to platforms because people can only turn to one specific platform if they can't get what they want elsewhere. So we compute the percentage of exclusive content on each platform. And here are the results: Netflix: 90% Prime Video: 95% Disney+: 94% Hulu:71%. So you may think twice before subscribing to a Hulu membership because 30% of its content can be found on other platforms as well.

So that's all the content of expository analysis(翻頁). We have to note that the accuracy of our analysis depends on the completeness or representability of our dataset, which is difficult to verify since we don't know the procedure the owner used to obtain this set.

Sheng Chun Zhao:

（大标题，Part2第一页）For this part, we mainly focus on the total number of movies that will be released in 2023 and the number of movies that will be released on each streaming platform in 2023. Based on our prediction, the audience can make an informed choice in purchasing a yearly membership by selecting the platform with a larger or more stable production rate.

（翻页）Let's first look at the general trend in the movie market. As can be seen from the figure, the booming development of the film industry roughly began in 1993, and the film market rapidly developted around 2002. The film market reached its peak around 2017, with about 1,400 films being released on the market. But around 2020, the number of movies dropped off a cliff due to the COVID-19. But now that we've fully liberalized, I think the market will pick up, and through our predictions, the estimated number of films released in 2023 is around 800. This provides more choices for the movie-loving friends.

（翻页）Next, let's look at the number of movies played on streaming platforms. I found the data of three platforms from the KAGGLE website. However, the data set of amazon prime video mainly focuses on TV SHOWS, while there is very little data about movies. So, we only analyse and predict for Netflix, HULU and Disney.

First, we conducted data cleaning, removed the data unrelated to movies, and counted the number of movies released on the platform every month. Since the data in the earlier years was very small, we only took the data in recent years for analysis. Here is the plot of the number of movies across time. We can see that Netflix has the highest number of movies released each month. And from 2018 onwards, the number of movies released monthly is also relatively stable, maintaining an average of about 125 movies per month, while the number of movies released by Disney is the least. Therefore, the movie resources of the Netflix platform are more abundant.（翻页）Then we began to predict the number of movies based on the above data. First, we checked the data for seasonality. We found that Netflix and HULU have no significant seasonality, while Disney has fewer movies released in quarter one than in the remaining three quarters.

（翻页） Next, we use the last six months of the data as a test set, and then the rest is the training set.

（翻页）Here, we used AUTO-ARIMA, PROPHET, GLMNET, HYBRID machine learning methods to predict the data,

（翻页）So here's the result for the test data. We can see,

The Prophet model is more suitable for Netflix, the GLMNET is more suitable for Disney, and the PROPHET and the HYBRID model are more suitable for Hulu.

（翻页）Once the model is chosen, we make predictions about the data. Here is the result. The figure shows that the number of movies released by Netflix is relatively stable, but the fluctuation range is higher than before, which is basically around 140. As for the Disney platform, we can see that the number of movies will continue to increase in the future and will increase to about 100 movies by the end of 2023. For HULU, the PROPHET model has a smaller fluctuation with a slower growth rate, while the hybrid model gives us the opposite result.

（翻页）So, if you want to subscribe to one of these platforms, I prefer Netflix. The growth rate from Netflix is not that great as HULU and Disney, but the number of movies released monthly is more stable and larger than other platforms, giving you more variety choices.

（翻页）The weakness of our current predictions is that we did not add other variables into the model forecast, such as the profitability of these platform companies. Also, some uncontrollable factors, such as the another outbreak of COVID-19, will have a certain impact on the release of movies. Therefore, there may not be as many movies released as predicted.

Also, just like we said in part 1, our analysis is only based on a dataset obtained from Kaggle, meaning that the accuracy of our predictions heavily relies on the completeness or accuracy of the dataset, which is relatively hard to verify due to the lack of resources.

3. Modelling

Li Yuji

Audiences all want to see high-quality and exciting films on streaming platforms. While the rating of a particular film often varies from person to person, the IMDb rating data of a film may serve as a reference for the quality of the film. With this aim, we want to investigate the impact of a film's features on its IMDb rating as a reference for viewers when streaming a film that doesn’t have an IMDb rating yet. This could help us to estimate the IMDb rating of a particular newly released film.

We next show how we fitted different models with data on the available features of movie records and their corresponding IMDb scores.

Before constructing the regression model, we pre-processed the data using five steps. Firstly, we removed records with missing data or filled the blank with column average. Due to the significant differences in scales between different features, such as the film's production year and films runtime, we standardised the numerical features to fit machine learning requirements. To include qualitative variables such as film genre, language and country in the regression analysis, we convert them into dummy variables. We found the column that specifies the movie's directors hard to utilize as we don’t have further information about them, so we drop the column. To validate the performance of our trained model later, we split our data into train and test sets.

We first fitted the existing data with the OLS model as a basis for further refinement. As can be seen, the OLS model did not work well, with severe overfitting problems. i.e. it fits too closely to the training dataset that it fails to study the general form and fails in making proper predictions for our test set.

To improve the linear regression model, we used the Ridge Regression model as a modified solution to OLS. It introduces bias into our model to mitigate overfitting. From the RMSE of the testing set, We can see that the Ridge Regression model performs much better than the OLS model. We also tried Lasso and SVR, but Ridge performs better than them in all testing set metrics.

Chen Taoyue

30 Most of the models above do the prediction linearly. Next, we tried the decision tree model, which builds regression models in tree structure.This method will partition films according to their features, such as genre, release year, and film duration. The classification results will be the predicted film scores. The important features of our dataset are shown in the figure. And the three most influential factors in the rating of a film are "Documentary", "Horror", and "Runtime ". with RMSE …

31 Since gaps exist between R^2 and RMSE of the train and test sets, we tried to improve the decision tree model. Here we used a random forest model. It can be understood as consisting of many unrelated decision trees. Like ridge regression, the random forest model also tackles with the problem of overfitting that tends to arise when training decision trees. The random forest results are as follows, which improve considerably upon the tree model. It can be seen that the three most influential factors for movie ratings are "Runtime", "Year", and "Documentary ". 32 We also tried boosting algorithms such as Adaboost and Gradient Boost, but they didn’t perform as well as a random forest after tuning.

In conclusion, the ridge regression and random forest methods improved the model's fit after dealing with the over-fitting issues. With our best model random forest, we obtain an R^2=0.437

and RMSE = 1.014. 33 This evaluating metric shows that the features employed can only partially account for the variance in IMDb ratings. Our weakness lies in need for other essential information about the movies. If we can find a way to incorporate information such as the award records of directors and the cast into our analysis, our model may perform better.

With the model we have just developed, we can predict the future IMDb rating results of a particular unreleased blockbuster, allowing those rating-focused moviegoers to decide better whether to see the film. Though many improvements are needed to boost our model's performance, it can still offer helpful information for the audience to decide when there are not enough reviews for a new movie online.
34
Finally, let’s quickly summarise the contents covered in our presentation:
Firstly, we did exploratory data analysis for each platform and found some informative platform-based characteristics for the potential customers of the streaming platform. For example, Prime Video provides the highest volume of movies and also high-quality movies.
Secondly, we used time series and machine learning approaches to forecast the number of new movies in 2023. And we recommend audiences choose Netflix over Hulu and Disney for a long-term membership if they expect rich new content.
Thirdly, we model the relationship between movie features and IMDb ratings with regression and tree-based models. We found that features "Runtime", "Year", and "Documentary " are the most important ones that affect the IMDb ratings. Though the available movie features in our dataset can only account for 44% of the variance in IMDb ratings, our model can still generate a helpful prediction based on simple features of a new movie that doesn’t have reviews yet.
35
And here comes the end of the presentation. Thank you(all) for listening.