Data Mining Exploration of Netflix Movies and TV Shows

Jingrui Gan: jingrui@umich.edu Yu Yan: kuminia@umich.edu

School of Information, University of Michigan

Introduction

In the dynamic landscape of digital streaming, this data mining project delves into Netflix's extensive Movies and TV shows dataset. Through a multifaceted approach involving time series analysis, network structures, and a recommendation system, we aim to decipher evolving viewership patterns and diverse user preferences within the Netflix ecosystem.

Dataset

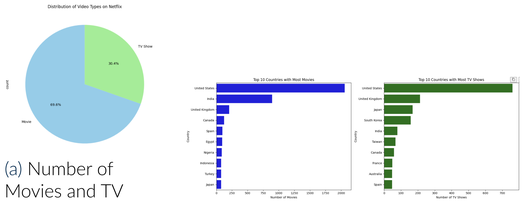
The data for our data mining project comes from the Kaggle website, the link is https://www.kaggle.com/datasets/shivamb/netflix-shows/data. This tabular dataset consists of listings of all the movies and tv shows available on Netflix , and contains various information related to these movies and TV series, including title, director, actors, country, release date, rating, duration, description and other features. The data contains a total of 12 features and 8807 movies and TV Shows. We will use these features to analyze and mine Netflix's movie and TV series data.

Method

In this comprehensive data mining project, we leverage various techniques such as data processing, exploratory data analysis (EDA), time series analysis, network analysis, and recommendation system development using CountVectorizer and Cosine Similarity. EDA, a foundational step, involves examining and visualizing dataset characteristics to identify patterns and anomalies. Our time series analysis employs data visualization, decomposition, stationary tests, and autocorrelation function (ACF) and partial autocorrelation function (PACF) plots, leading to the selection of a Seasonal AutoRegressive Integrated Moving Average (SARIMA) model for Netflix's monthly movie additions. Utilizing network analysis, we investigate the relevance and popularity of directors and actors in Netflix's US movies, calculating Degree Centrality, Betweenness Centrality, and Eigenvector Centrality to unveil different facets of the data's structure. Lastly, our recommendation system integrates CountVectorizer for text vectorization and Cosine Similarity for measuring semantic closeness between texts, facilitating personalized content suggestions. These methodologies collectively empower us to predict Netflix's movie trends, explore director-actor relationships, and enhance user engagement through tailored recommendations.

Exploratory Data Analysis (EDA)

In the EDA, we created a pie chart comparing movies and TV shows on Netflix, revealing a ratio of approximately 7:3. The bar chart identified the top movie-contributing countries as the United States, India, and the United Kingdom, while TV shows were primarily from the United States, the United Kingdom, and Japan. The U.S. dominates Netflix's content. Japan's high TV show count may be due to its exclusive anime category.



(a) Number of Movies and TV Shows in the Dataset

(b) Number of Different Ratings in the Dataset

Figure 1. Some Visulizations

Time Series Analysis

We added up the number of movies per month to calculate the number of movies added to Netflix each month, and then drew a time series model of the number of movies added.

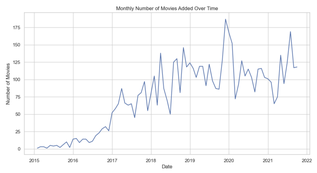


Figure 2. Time Series Plot of the Number of Movies Added

Then we started to dig deeper into the time series data. We first decomposed the data. We can clearly find that the data has obvious increasing trends and seasonal patterns. Before building the model, we need to analyze whether the time series is stationary. Using the Augmented Dickey-Fuller (ADF) test, we get the p-value is approximately 0.5419, which is significantly larger than 0.05, so we reject the null hypothesis and conclude that the data is not stationary. Also, from the ACF plot and PACF plot, the ACF decreases gradually and the PACF has a sharp drop, so we need a first-order differencing.

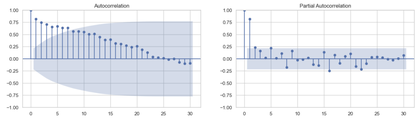


Figure 3. ACF Plot and PACF Plot of the Time Series

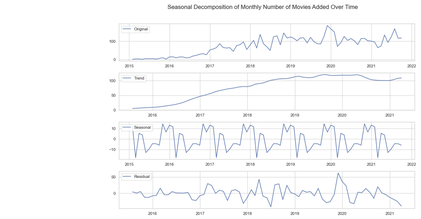


Figure 4. Decomposition Graph of the Time Series

According to the ACF plot and PACF plot, we can also get the values of q, d, and p in the SARIMA model. Through the SARIMA model, we obtained a prediction of movie growth in the next 12 months and the 95% confidence region of our prediction. Due to various environmental and resource uncertainties, although our forecast may not be perfectly accurate, the forecast results can still help Netflix conduct more effective resource planning. Understanding future content growth trends can help optimize server and bandwidth resources to meet user demand for new content.

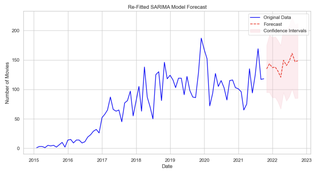


Figure 5. Prediction of Netflix Movie Added in the Next 12 Month

Network Analysis

In the Network section, we are going to study the partnership between directors and actors in American films on Netflix. We extract all the actors and directors in the data set and the number of times they appear, and select the top 50 (the number can be adjusted by modifying the code parameters) to become the nodes of our Network Graph and use red to represent directors and yellow to represent actors. If there has been a collaborative relationship between the director and the actor, they will be connected by edge. In addition, we also used degree centrality to find the node with the largest number of edges in the Network graph. The result was director Steve Brill, who has collaborated with 11 actors who appear in the top 50.

(a) degree centrality

(b) betweenness centrality

(c) eigenvector centrality

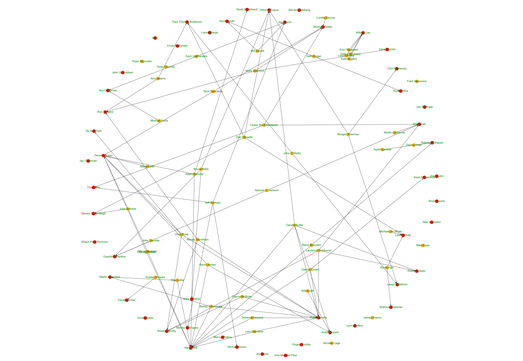


Figure 7. Network Diagram

We used degree centrality, betweenness centrality, and eigenvector centrality to analyze the popularity of directors and actors in US movies. We can conlcude that Samuel L. Jackson and Steve Brill are the actors and directors who have collaborated the most. They also serve as bridges or intermediaries between different groups of actors or directors in the collaboration network with the highest betweenness centrality. Eigenvector centrality signifies well-connected collaborations, exemplified by actors like Terrence Howard and directors like Elizabeth Banks.

Recommendation System

Our recommendation system employs content-based filtering, utilizing features like title, category, director, actor, country, rating, theme, and introduction. Through data processing and the application of CountVectorizer and Cosine Similarity, we transform and analyze the whole text feature to calculate similarity scores, facilitating the generation of personalized recommendations based on user preferences.

```
recommendations('Jaws')
✓ 0.0s

42                                                  Jaws 2
169                                               Poseidon
44                                          Jaws: The Revenge
7073          Indiana Jones and the Temple of Doom
7530                                  Mutiny on the Bounty
7879                                                 Rocky
7072    Indiana Jones and the Raiders of the Lost Ark
6496                                       Cleopatra Jones
7071          Indiana Jones and the Last Crusade
2059                                               Adrift
Name: title, dtype: object
```

Figure 8. Recommendation Example