MISM 6212: Project Proposal
Forecasting Flicks : Cracking the Code to Content Success

MISM 6212 - Group 11
Groom Dinkneh, Rachit Pandya, Rishabh Sharma, Apurv Garge, and Tanisha Babbar

—

Business Question

Predicting Future Content Success on Netflix

We seek to create a predictive model to assess the popularity of upcoming Netflix films and TV shows leveraging historical US Netflix data. We want to understand why some content performs more effectively than others by analyzing factors like genre, cast, rating, and duration. Analyzing this data will offer clues into trends that currently exist that might explain where future successful content may be produced. Our goal is to offer insights that will allow Netflix to tailor its library of movies to meet customer preferences, increasing subscriber satisfaction and loyalty. As competition heats up, retaining a diverse and compelling content library becomes increasingly important. Furthermore, by precisely forecasting content success, Netflix can reduce subscriber churn, especially in the face of price increases. Ultimately, this research will help to inform strategic decisions about content selection and payment structures.

We hope that our research will provide guidance on what investments Netflix should make in future programming. Because Netflix relies heavily on its algorithmic recommendation, we hope to provide insights on what programming is most desired and enjoyed on Netflix so we may have fresh material to recommend. Stale content will lead to discontent, getting bored watching reruns of shows. We'll provide customer segmentation information, define what genres have been successful and look at factors amongst movies and films with high runtimes and high ratings to run regressions and view if any characteristics are highly correlated.

Data Sources:

We will be utilizing data from publicly available websites **(such as IMDB, Kaggle)** that have gathered US Netflix viewing data for analysis.

The variables we will track to gain insight will include the movie/show title ID, the name of the movie/show, the type of show or movie, a description of the title, the release year, the age certification (suitable viewing age rating), the runtime (length of each episode for a show or movie), the list of genres, the countries that produce the title, and the number of seasons if it is a show. We may also use IMDB data that includes the title, audience score, number of votes, popularity, and overall score.

The following are links to various data sets that we will be using for our analyses:
● IMDB Movie Related data : https://developer.imdb.com/non-commercial-datasets/
● Netflix Dataset - https://www.kaggle.com/datasets/victorsoeiro/netflix-tv-shows-and-movies

Data Quality Concerns:

We are choosing to utilize data on Netflix aggregated by a 3rd party. Because we do not have the authority or time to request a research license from Netflix directly, we are utilizing other sources to assist in gathering this data. There are risks that the data may not be factually accurate.

Given that our data comes from multiple repositories, accurate merging is critical. We anticipate requiring some data cleansing to account for null values in the dataset. We may combine the two datasets we've selected, one with Netflix titles and another with Netflix Crew,which will need careful review and parse.

Aligning the data ensures smooth integration and improves our ability to come up with meaningful insights. By carrying out robust methods for resolving formatting discrepancies, we can maximize the value of our multi-source data for analysis and decision-making.

External Supporting Materials:

The links below are just a few outside sources that we have found that may be helpful in supporting our research:

- **Netflix Is Raising Prices By Killing Its Cheapest Ad-Free Plan**
  https://www.forbes.com/sites/erikkain/2024/01/25/netflix-is-raising-prices-in-the-most-cynical-way-and-its-only-going-to-get-worse/?sh=593777c115cf
- **Suits' Sets Another Streaming Record for 2023: Biggest Year Ever**
  https://www.hollywoodreporter.com/tv/tv-news/suits-year-end-streaming-record-2023-1235808594/amp/
- **Netflix says 80 percent of watched content is based on algorithmic recommendations**
  https://mobilesyrup.com/2017/08/22/80-percent-netflix-shows-discovered-recommendation/

We plan to use multiple regressions to analyze success on Netflix. We will train a model that will offer predictions on future success given the historical data available.

Proposed Questions

1. What are the most popular titles? Do they have the strongest ratings and viewings? Are there any observable similarities between the successful shows/movies and their popularity?
2. What characteristics are most relevant to maintaining high runtime? What is the average, median, and mode across each genre?
3. How do we compare the success of TV shows and movies based on the awards they have won?
4. In predicting future success, what variables are the best predictors? How can we create training and testing sets, and build a model to evaluate expected success?
5. How will we compare viewership ratings from external sources with our conclusions to validate our findings?
6. Which titles performed the best? What can we learn from them that may inform future title creation?

Challenges:

One challenge would be making inferences gathering insights on data from streaming that don't represent all streaming platforms since we are limited to only Netflix data

We also are limited in determining what factors are driving viewing, popularity, etc. External factors such as how much marketing was used to promote one show vs other shows may have a material
Impact on the final viewings

We also are unable to determine if issues specific to locale may impact and limit users from maximizing their usage of Netflix (web traffic being high making it difficult to stream in certain areas of the world)

Methods:

1. Customer Genre Segmentation and Behavior Analysis for Retail:
2. Utilize data mining techniques to segment customers based on their purchasing behavior, demographics, and engagement with a retail brand.
3. This project could involve analyzing transaction data, customer feedback, and social media engagement to identify distinct customer segments and predict future purchasing patterns.

TEAM CONTRIBUTION

Names	Contribution
Rishabh	Data formatting
Entire Team	Python regression modeling
Rachit and Apurv	Visualizations
Tanisha and Groom	Report writing and analysis
Tanisha and Groom	Preparing final presentation

Data Set Selection:

Found here: Netflix TV Shows And Movies