Christian Rhodes
02-26-2023
Data Science 1

Project Proposal - Netflix

Motivation and Problem Statement

For my Data Science final project, I will be working with Netflix data and observing the kind of content that consumers most like to watch. In addition, I would like to evaluate the motivations behind the types of content people like to watch. Do they watch for escapism? Maybe they work in entertainment and watch content to study best practices or compare their style with that of others. Do viewers watch content because they want to learn about certain topics or locations? While all of these situations are true for some, my project will look at the intersection of popular media, its popularity and ratings, and demographics.

Working with Netflix data is a relevant area to study because of how popular it is and how much time is spent on the service. Netflix is a tech giant, and has large influence in the tech and entertainment industries, which both encapsulate billions of people worldwide. In 2022, Netflix generated $31.6 billion with 220 million subscribers worldwide. During the coronavirus lockdown in 2020, over 200 million collective hours were spent watching Netflix. I think this field of study is important simply because of how popular it is, and how much time is spent on the platform.

Related Work

Netflix, Who is Watching Now? By Cristóbal Fernández-Robin

This study looks at Netflix consumers using a UTAUT2 (Unified Theory of Acceptance and Use of Technology) model to evaluate their behaviors. This research ties into advertising and creating profiles for users to provide them with content they are most likely to watch and enjoy. Hedonic Motivation and Social Influence from the study are relevant to my study, but I will be aiming to make a macro analysis on these factors to determine for what reason people watch certain things, which differs from simply finding the type of content they watch.

The Netflix Effect: Technology and Entertainment in the 21st Century. By Kevin McDonald

This book looks at how Netflix took over video media and how and why people use it. This is relevant to my study because so many demographics use Netflix. This book looks at the economy and state of technology over the last two decades, and a lot can be learned from this about user motivations, and even types of content they enjoy.

Data Collection, Cleaning, and Exploration Plan

The dataset I will be working from comes from Kaggle. It features nearly 6,000 titles. The dataset includes features that one would expect from a video streaming service, including imdb ratings. The combination of content details with reception is a promising combination that will allow for robust and nuanced data exploration. Some features have null values and unconventional array formatting as strings. Data cleaning will be performed here and feature engineering may be performed to expand the use of the data at hand. One-hot encoding can be performed on categorical variables as well.

Modeling, Analysis, and Visualization Plan

From the start, a pair plot should be created to explore relationships across different features. The features (which already exist in the dataset) that are especially relevant to my problem statement are movie or show genres paired with IMDB statistics. The description data could also be of interest to explore common terms in relation to popularity and ratings. Different linear models will be created and tested to find and tune high performing models. Loss functions can be tested by predicting genres, keywords from the description column, and / or IMDB ratings with the IMDB popularity feature. These models will predict popularity given the features mentioned.

Needs Assessment and Contingency Plan

It is feasible to clean the data, explore it, and create a strong model to predict popularity. The assessing of ‘why’ of popularity and consumer motivations may be more complex and require taking larger gaps of predicting or inferring based on the raw data at hand. Aggregate data could be a useful addition, but may not be feasible given the scope.

Proposed Timeline

Week #	**Goal**
8	- Find partner for project - present project proposal - Exchange contact info with partner and establish communication method and general availability
9	- Verify validity of problem statement - specify or modify if necessary - Finalize stretch goals - Make meeting plan and agree on logistics - Traverse data to get familiar it
10	- Create GitHub repo and begin code - Create a pandas dataframe and begin data cleaning - Begin EDA - add / remove features, handle null values, etc.
11	- Explore relationships of genre / IMDB ratings and IMDB popularity - Explore description text data - bag-of-words, term frequency - Create linear model (if time permits)
12	- Create linear model (if wasn't accomplished last week) - Tune model; audition different models and inputs - Create visualizations to help assess model effectiveness and make readable conclusions
13	- Finalize model - test the test and validation datasets for accuracy - Make analyses - Begin report outline
14	- Make presentation and prepare to present to class. All visualizations should be prepared and the process of data cleaning, EDA, modeling, analysis, etc. that was documented should be gathered and presented succinctly.
15	- Write-up complete and present

References

https://www.kaggle.com/datasets/victorsoeiro/netflix-tv-shows-and-movies

https://www.businessofapps.com/data/netflix-statistics/

https://link.springer.com/chapter/10.1007/978-3-030-21902-4_15

https://books.google.com/books?hl=en&lr=&id=NpoyEAAAQBAJ&oi=fnd&pg=PP1&dq=netflix+demographics&ots=oqLAIIj8-z&sig=3nsBk5ZaZTpDVRQvyHYTp1I7Bxo#v=onepage&q=netflix%20demographics&f=false