Midterm project proposal

Zhihui Zhang

Personal Statement:

I would like to be a full-stack data scientist in tech firms after graduation. The full-stack means that as a data scientist, I will work in the whole data science life cycle - from generating business questions to data collection/ identification, pipeline, exploratory data analysis, creating features, then, building machine learning models, and finally deploying the models or draw final decisions. Therefore, I want to practice my ability from generating a business insight to making suggestions about how to improve the products using statistical methods.

Taking all considerations, Netflix data is the best fit for me. As one of the largest pay-television planforms in the world, Netflix not only provides trending movies but also original programming productions to their users. In my opinion, attracting new users and keeping the old customers are things they care about. And recommendation algorithms might be at the core of their products. The dataset contains 26 features of Netflix movies and TV series including genres, runtime, box office, availability as well as Rotten Tomatoes and IMDb scores, etc. It will allow me to discover the relationship between movies scores and characteristics of movies including genres, languages, number of awards received, etc. This might be helpful for further improvement of recommendation systems. If the users prefer to watch movies with high scores, what type of movies we should recommend? Besides, the owner of this data set calculates a column called hidden gem score. A hidden gem in Netflix means the movies with an underrated title but with strong performances and complex storylines. I also take a deeper look at the data set and find what kinds of movies are more likely to be hidden gems. If people have already watched most of the movies which ranked high, instead of putting movies with lower scores on the next play, recommendations of hidden gems might surprise them and make them more stick to the platforms.

Question:

What is the relationship between movies scores and features of movies including genres, runtime, language, director, writer, etc.?

Which types of movies/ TV series will be considered to have better quality (i.e., Higher average scores across different types of movie rating websites?)

Will movies/ TV series provided by Netflix have higher scores than other videos on Netflix?

Which types of movies/ TV series are more likely to be considered hidden gems?

The data source (s):

The data comes from Kaggle and the owner collected data using a variety of APIs across the different websites. It is well organized and will be helpful to address my questions.

https://www.kaggle.com/ashishgup/netflix-rotten-tomatoes-metacritic-imdb

Proposed Timeline of work:

- EDA: 11.4 – 11.10
- Data Processing: 11.10 – 11.14
- Modeling and Validation: 11.14 – 11.27
- Write up: 11.27 – 12.1