Midterm Project Proposal

Donghao Xue

Personal Statement

For my career goal, I plan to pursue the career in the direction of data scientist or research scientist so that I can work in a tech or media company. In this project, I choose a dataset that contains information of about five thousand movies from IMDB which is a popular website providing source for movies. This dataset is downloaded from the Kaggle website and it has 28 variables such as the genre of movies, the rating score of movies, the number of users for reviews, the number of Facebook fans of actors and so on. I will mainly focus on finding the relationship among these variables. I think dealing with this dataset can help me build a solid foundation for data analysis and get myself prepared for my future job.

Question

What are the factors that will have the impact on the rating scores of movies?

What are the factors that will influence the budget of the movie?

Can I make prediction of the rating score and budget of the movie given some basic information of the movie?

Data Source

https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset?select=movie_metadata.csv

Proposed Timeline of work

EDA: Week 11 (before 11/14)

Data Processing: Week 12 (before 11/21)

Modeling and Validation: Week 13 (before 11/28)

Write up: Week 14 (before 12/5)