Mike Stacey

Capstone 3 Proposal

Background

Recommender systems appear in almost every online service we use these days. From recommending what to watch on netflix, to people we may know on social media, to other products we may like on online shopping services like amazon. As these services become more and more popular, it becomes more and more important to be able to make recommendations that consumers will like. The goal of this project will be to gain an understanding of how to build recommender systems.

As I will be using data from a database of movie ratings, an optimal client for this project could be any streaming service such as netflix, disney plus, or hulu, or it could be a company selling movies such as amazon. A good system for recommending movies will help provide a more satisfying user experience for their services by reducing the amount of time their customers spend browsing before choosing a film or television show.

Data and Approach

I will use a dataset from kaggle of movies and user ratings for amazon movies found here https://www.kaggle.com/eswarchandt/amazon-movie-ratings. This dataset contains ratings of 206 films from nearly 5000 users. For a model intended to be put into production, I would choose a dataset that has better documentation so I can be certain of the accuracy of the data, but as this is purely an academic exercise to understand how to create the system, this dataset will suffice.

This dataset does not appear to require much cleaning, but I will check to see if there are any strange values that need to be replaced by NaNs and make sure the ratings are in the same numeric format. I will then perform exploratory data analysis to see if there are any imbalances in the data that need to be accounted for when I build my model.