CS1951V Capstone: Movie Recommendation System

Noah Atanda, Ezra Rocha, Jorge Sanchez

Goal

Watching movies is a popular pastime for millions of people around the United States. Everyone is familiar with popular streaming services such as Netflix, HULU, max, etc. For this project we wanted to gain insight into how recommendation algorithms work, and potentially improve upon ones that already exist. Given this background, we were most interested in the following goal: given movies that a user likes, find which movie from our dataset they are most likely to enjoy.

Data

There are several websites with movie data designed to train and test recommendation algorithms. We got our data from Kaggle, an online community for data science and machine learning and IMDb's Non Commercial Dataset. We created our dataset by joining tables from IMDb that included ratings on movies and basic identifying information. This was joined onto a data set from Kaggle; TMDb Top 10,000 Popular Movies Dataset. We joined these datasets based on the movie's title, and combined attributes such as the user id, their rating for the movie, the movie's budget and revenue, etc. We had to process the data in order to achieve similar formats for the titles. For example, we turned them all to lowercase, and removed roman numerals and extra white spaces. We removed rows with empty entries. Our final data table has the following fields: userId, movieId, rating, title, genres, release_date, vote_average, vote_count, popularity, budget,revenue.

Model + Interactive Component

We used an unsupervised K Nearest Neighbors with a custom distance function for our model. Within this custom distance function, we assigned the following weights: Vote Average (25%), Popularity (25%), Revenue (10%), Genre (40%). These metrics were the best way to determine movies that were similar to one another. For our interactive component, we created a web page where a user can input their top 3 movies and our model will recommend up to 6 movies they will like. We expect these recommendations to be of similar genres and popularity as we implemented a KNN algorithm, but there are instances of movie recommendations that are not necessarily popular. We used the dataset described above to generate these recommendations.

Findings

Claim #1: The movie recommendation system provides diverse recommendations, encompassing a range of genres, popularity levels, and revenue brackets.

Support for claim #1: The movies that were recommended by our model were evaluated based on various attributes such as genre, popularity, and revenue. This indicates that the model considers user satisfaction, by supplying them with a diverse set of movies based on different

attributes and giving them a broad spectrum of choices. This is a table showing the standardized attributes of movies recommended based on the input: ‘cars’

title	genres	popularity	Revenue
cars	Animation	1.724927	1.647094
despicable me	Animation	1.784896	2.747299
wreck-it ralph	Animation	2.090330	2.355322
bambi	Animation	1.316311	1.076428
charlie and the chocolate factory	Adventure	2.245133	2.378831

Claim #2: Movies with higher revenue tend to have higher popularity.

Support for claim #2: We ran statistical tests on our data table in order to find the correlation between the different variables we had in our dataset. We have a variable measuring the movie’s popularity, and another variable representing the movie’s revenue. We used a correlation matrix to show the correlation score between all of our variables. The correlations go from -1 to 1. -1 being perfectly negatively correlated, and 1 being perfectly positively correlated. The variables for revenue and popularity had a correlation of 0.61 indicating a strong positive correlation. The table below illustrates the correlation matrix:

	rating	revenue	popularity	budget
rating	1.000000	0.058536	0.058673	-0.004039
revenue	0.058536	1.000000	0.611002	0.715214
popularity	0.058673	0.611002	1.000000	0.462620
budget	-0.004039	0.715214	0.462620	1.000000

Socio-historical Context

The MovieLens database was created in 1997 by a group of researchers at the University of Minnesota. Originally a tool used by a limited number of users, it expanded in popularity after gaining traction in media, including a segment in ABC's *Nightline* and a favorable mention by movie critic Roger Ebert (Harper). With thousands of new users joining throughout the 2000s, the MovieLens database has expanded to include a wide range of users with ratings across all manner of movie genres. For this reason, we were confident in using MovieLens as a primary source of information for ratings to process in our KNN model.

Our project is primarily relevant for data scientists who are interested in improving the efficiency of recommendation systems. These systems may include other forms of media (music, books, games) or systems for more delicate subjects such as health. Movie enthusiasts may also be interested in our findings as the recommender can allow them to find new movies to enjoy, perhaps some they may have never heard of before. There is no group that can be harmed by our findings, not even on the level of market competition as our project is much simpler in scope. Many researchers have conducted their own studies using their own samples of the MovieLens database. Their studies have dived into statistical and personality-based aspects, including one that found that "users with different levels of personality trait have different preferences for how diverse, popular, and serendipitous" (Nguyen) their movie recommendations should be.

Ethical Considerations

The movie ratings used in our dataset are limited to ratings made by MovieLens users between 1995 and 2015, meaning that our movie selection only goes up to roughly March 2015 releases. This unfortunately means we were unable to study the major data trends that occurred within the movie industry as a result of the COVID-19 pandemic, which led to an extended period of inactivity for major studios and movie theaters alike. Although our model and hypothesis would greatly benefit from this data, we compensated by including as many reviews as possible across a variety of movie genres and release dates. Regarding bias, one study points out that DataLens can be influenced by "popularity bias… toward highly reachable products because these items are more likely to be treated as positive training instances [and] exposure bias: interactions collected from an online platform are influenced by its deployed recommender system" (Fan). From the limited selection of films we worked with, we opted to randomly select reviews to limit the amount of bias towards any one movie and randomly selected users as well.

The MovieLens site is intended to give users complete privacy regarding any identifying information, such as names, emails, IP addresses, etc. For research purposes, the database may include the age or city users are from, however these terms are clearly outlined for all users who wish to register an account. Therefore, users are aware that their data is primarily used in research circles to develop and study new methods of recommendation systems. If they choose to, they could opt out of the database (and by extension any related studies) simply by deleting their account. Ultimately, measuring the success rate of our recommender is not a reliable process given the subjective nature of user movie preferences. It's also important to note that this dataset is not up to date, as there are thousands of movie selections missing from both the time period these ratings were made and movies that have been released since.

Works Cited

Fan, Yu-Chen, et al. “Our Model Achieves Excellent Performance on MovieLens: What Does It Mean?”, *ACM Transactions on Interactive Intelligent Systems*, vol. 1, no. 1, Mar. 2024. https://doi.org/10.48550/arXiv.2307.09985

Harper, F. Maxwell, and Joseph A. Konstan. “The MovieLens Datasets: History and Context.” *ACM Transactions on Interactive Intelligent Systems*, vol. 5, no. 4, 1 Jan. 2016. https://doi.org/10.1145/2827872

Nguyen, Tien T., et al. “User Personality and User Satisfaction with Recommender Systems.” *Information Systems Frontiers*, vol. 20, no. 6, 3 Sept. 2017, pp. 1173–1189, https://doi.org/10.1007/s10796-017-9782-y