Introduction:

1. Motivation

The motivation for working on this project came from our experience with streaming services like Netflix and their amazing recommendation system. With this project we seek to learn and implement recommendations system, it's evaluation metrics.

1. How?

For the project, we used the MovieLens data set (\*Insert link) and segmented the users into two groups, namely short term users and long term users. We defined short term users as those who have watched less than three movies while long term users are classified as those who have watched more than three movies.

To predict the ratings of the movies by the users, we used two approaches, a) Baseline Model and b) Collaborative Filtering.

For the project we constructed two models

1. Baseline
2. Collaborative filtering

We used the Movie-lense dataset which contains userID, movies they have watched and what rating they give for that particular movie. Using the data we first constructed a user similarly matrix and used euclidean and cosine distance for finding the similar users.

We used 2 ways to find similar users:

1. Euclidean Distance

This measure the straight line distance between any two points on the graph and the shorter distances implies greater similarity.

1. Cosine Distance

This measures the cosine of the angle between any two points and small angles implies greater similarity. (similarity = 1- cosine(angle))

We also use something called the centered cosine distance to remove any bias from the data and gauge similarity more accurately.

1. Collab Filtering:

Collaborative filtering algorithm finds similar users to recommend items i.e. in our case movies. The idea of collaborative filtering is that if A and B users have similar views/opinions about a particular topic then for another topic B should have similar opinions as A.

In our case if A and B have watched and like same kind of movies then B is likely to enjoy movies that are not seen by him/her but are seen by A

Why is Collaborative filtering good?

1. We don’t need to know anything apart from liked and disliked items. All other things are automatically learned
2. This algorithm helps users explore new interests.

Initial Approach - Cosine Similarity with User Genre Matrices

1. Pros and Cons

Second Approach - Euclidean with User Genre Matrices -

1. Pros and Cons

Data Processing :

1. Training and testing set. ( Two approaches)
   1. 70-30%
   2. Only 3 movies in testing rest in train

Baseline model - Used as the control set for the experiment

1. Approach 1 - Average based on historical data - ( for users without genre assigned rating of 0)
2. Approach 2 - Weighted average and users without genre assigned average rating.

Results /Graphs:

Conclusion: