Group Project report

## Group member:

Group Name: LLP

Haocheng Lu A0276434Y E1132167

Paul David Christopher Golling A0294203L E1345974

LUO Ming A0244396W E0860682

## Data processing

The data given is in a set of CSV files. The *u.data* captures the relationship between users and movies. The *u.item* contains information about a movie. It is also associated with *u.genre,* which contains genre information and uses item ID as a foreign key. The *u.user* contains information about users. The *u.train* and *u.test* are split data from *u.data*, which will be later used in training and evaluation processes for different approaches.

First, the data file is read from CSV into a Pandas DataFrame. Each file is read into a different DataFrame. Then, the DataFrame containing u.data will be used as the base to merge other data frames in.

The line number is checked using info() to ensure no lines are missing after the merge.

u.user data and u.item data are merged to the u.data based on user\_id and item\_id (Pandas, n.d.). Based on the key user\_id and item\_id, the method for merging is an inner merge, which means it will only keep the intersection part between data from u.user and u.item. This is because of the nature of the table where user\_id and item\_id are foreign keys in the u.data. After merging, the DataFrame info is checked. It can be confirmed that all columns in u.user and u.item have been merged into the DataFrame with no loss of rows.

Due to the final three approaches, there will be no temporal analysis applied as the timestamp is assumed to only bring noise. Therefore, this column is dropped. The columns “movie title,” “IMDb URL,” and “video release date” are also dropped from the DataFrame. The “movie title” serves a purpose similar to item\_id, which is only an identifier for the movie and does not provide much information for further analysis. The IMDb URL provides no further information that needs to be retrieved, so it is also considered non-contributory to the analysis. The “video release date” column is found to be empty, and the actual release information is contained in another column called “release date”; hence, this column is also dropped.

To unify the release date and make the data input more generalized, only the release year is kept, and the specific month and date are dropped. Keeping the date too specific provides very little benefit but introduces unnecessary noise to the model.

Also, the genre was one-hot encoded in the original data. At this point, it is not clear how each approach will use this feature, so it is reduced to one column.

Finally, the information in “zipcode” is not useful as it is only a code. We used an API to convert the zipcode into a state name. It is found that some of the zipcodes may not be U.S. zipcodes, in which case they will be marked as “Outside of USA.” Due to the process of using the API (zippopotam.us, n.d.) to convert all zip code data being time-consuming, a multi-threaded process is used to accelerate the whole process. Finally, due to the requirement to train our models using u.train and perform evaluation on them with u.test, they are also read into a Pandas DataFrame. A left merge is performed with the u.data DataFrame to get user information and movie information into the data for the training and test set.

## Approach 1 Neural Networks

# References:

Pandas. (n.d.). *pandas*. Retrieved 10 2024, from pandas: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.merge.html

zippopotam.us. (n.d.). *zippopotam*. Retrieved 10 2024, from zippopotam: http://api.zippopotam.us