Lenny Remache

The movieTitles.csv dataset required some initial cleaning up to do because when reading in using the pandas read_csv method, the data frame had two extra columns with no useful information at all. There were three total columns that were kept in the dataframe. The columns contained movie ID numbers, the release year, and the title of the movie for that specific movie ID. Some exploratory data analysis was done on this dataset in order to get a better understanding of the size of the dataset. There are a total of 5000 movies with two movies containing no release year information.

The data.txt file required parsing in order to extract specific ratings given by each user id on a certain date for a specific movie. This was done by storing the information in a specific format mentioned in the python file with the help of a dictionary. The main dictionary that contains movie id keys maps to another dictionary that contains user ID keys that each map to the rating they gave to the movie. Once the dictionary of dictionaries is properly storing all the necessary information that can be retrieved from the text file, it is converted into a pandas data frame. The data frame allows us to see that there are a total of 472542 users that provided ratings. However, most of the ratings are missing since most users are not providing a rating for all 5000 movies that are available.

In order to explore the missing data some more I used bar graphs along with the sum of missing data per movie. Each of the 5000 movies had over 278601 ratings missing. For a total of over 2 billion ratings that are not given by users. When looking at missing ratings totals and bar graphs there is not much detail you can extract from it since all you can tell is that it is true that most of the data is missing from the dataset. The bar graph for the first 50 movies missing ratings total showed exactly that. This exploratory data analysis also allowed me to conclude that movie id 4807 with the title ‘The Many Faces of Zorro’ had the least amount of ratings given with 472529 missing ratings.

So instead of looking at each movie's missing ratings total, I looked at each movie's total number of ratings given. This would allow me to see which movies had more users rating it and which movie overall was the most rated. Each movie had over 13 ratings given at minimum with the most ratings given being 193941. This could be used as a metric for movie popularity. In other words, it could show us a reason as to why some movies do not have a lot of ratings. The movie could simply just not be popular or good enough for more users to want to give it a rating. I provided a bar graph that shows that there are some movies that do in fact have a significant amount of ratings given to it. The movie with the most amount of ratings is movie id 1906 with the title ‘The Knights Templar.’ It had 193941 total ratings given to it. Which to me has more meaning since regardless of the significant amount of missing ratings this particular movie has, we can now detect something that differs each movie from each other. It is pretty difficult to differentiate between movies based on missing ratings through the bar graph.

So initially, the plan to reduce the dataset’s missing values was by excluding columns of movie ids that contained less than 5402 ratings. 5402 because the average number of ratings that

a movie gets is 5402 ratings. However, this would end up removing movie ids which are supposed to be a part of the target data. We are trying to predict ratings for all 5000 movies available that a user would give. This way of removing data would completely defeat the purpose of the project.

Thus, more exploratory analysis was done but this time based on the number of ratings that a user gives. I wanted to see if there was a relationship between user ids and the number of ratings they give. Meaning which users are more active at giving ratings and on average how many ratings do these active users give. Based on the bar graph there were definitely users that are way more active than others. The most active user had user id 305344 with a total of 4963 ratings given. That's almost close to having given ratings to all possible movies in the movieTitles.csv dataset. The average number of ratings that a user gives is about 57 ratings total. So for my case, I considered users with more than 57 ratings given to be an active user.

Only active users would be included in the linear regression model that is built later on. This reduced the dataset to having about 143502 rows of data left compared to the 472542 rows of user ratings at the start. This did help remove a lot of the missing data present in the dataset, but it does not remove all the missing ratings. Now we are able to look at the most important users that give ratings, meaning the users that have more of an impact on the number of ratings that a movie received.

Thus, some imputation was done to the new dataframe with fewer rows of user ids. The imputation involved the usage of the minimum rating given to a movie. Since the users left are considered to be on average an active user, we can assume that their participation in rating movies should have some form of accuracy on what most people would rate a movie. Also taking into account the fact that there are missing ratings because a user might have made the choice to not see the movie because they knew they were not going to enjoy it. So using the minimum rating of each movie id column makes sense to replace all the missing ratings.

The test set was set to user id 180687, while the training set was the rest of the data. A linear regression model was created and trained on the training set. Predictions on the ratings given to movies were made with both the training set and the test set. The RMSE value when predicting using the test set was ~0.4182 whereas with predictions when using the training set the RMSE value was ~0.2669. There is a clear difference in the rsme scores between the training and test set so I would definitely seek to reduce this difference to a point where values are fairly similar and represent a RMSE value that signals that the model is a good representation of the ratings that movies could get.