**CSE523 - Machine Learning**

**Movie Recommendation System using Machine Learning**

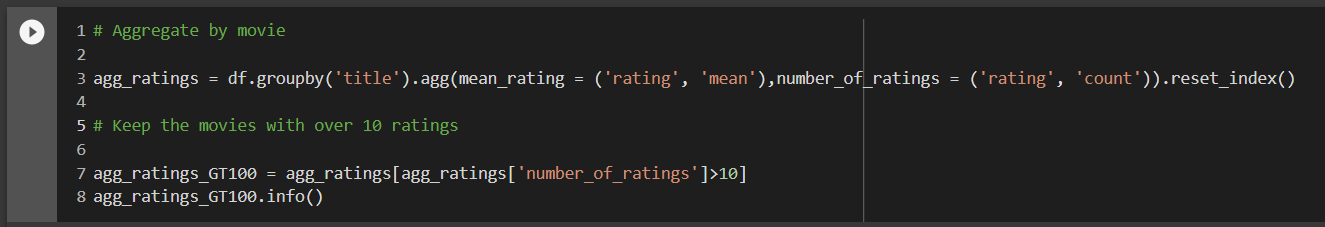
**Faculty - Prof. Mehul Raval**

**Weekly Report 6**

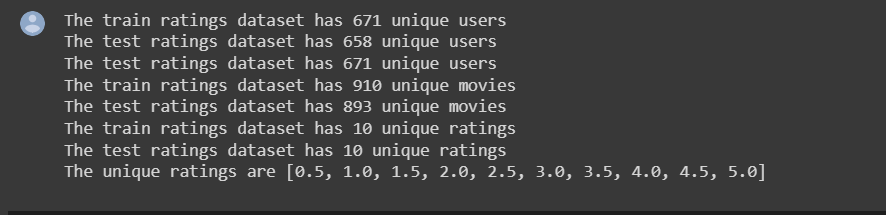
**Submission Date: 25-03-23**

**Group: Tech Titans**

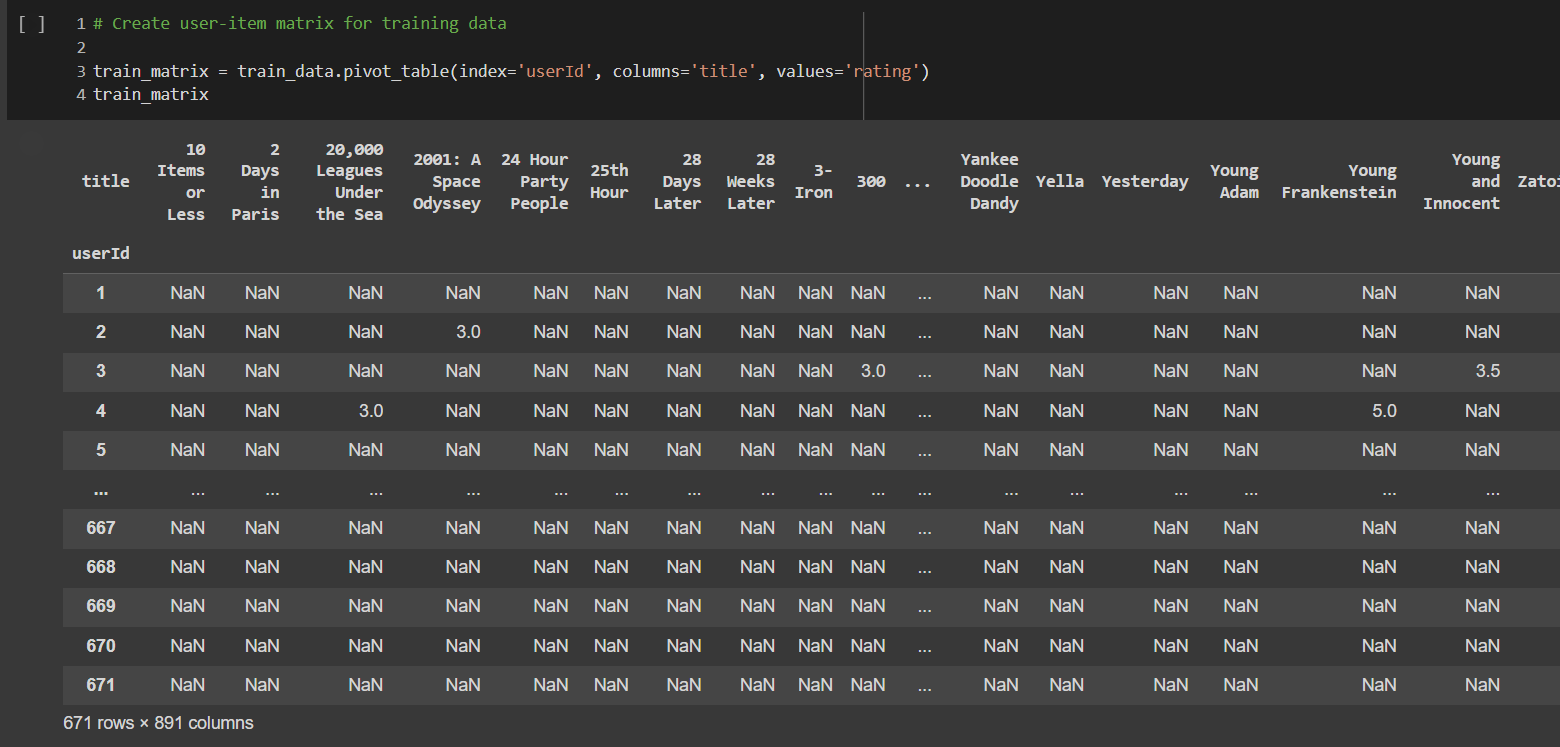
| **Name** | **Roll Number** | **Email** |
| --- | --- | --- |
| Jainam Shah | AU2040186 | jainam.s4@ahduni.edu.in |
| Yash Chotaliya | AU2040193 | yashkumar.c@ahduni.edu.in |
| Akshay Parmar | AU2040199 | akshay.p@ahduni.edu.in |
| Shubham Bhatt | AU2040206 | shubham.b1@ahduni.edu.in |



Initially we keep only those movies that received more than 100 numbers of ratings to reduce the size of the data. But to effectively divide the data into test dataset and training dataset we take movies that have more than 10 numbers of ratings. We make our dataset larger to train our models to learn the patterns and relationships present in the data. The more data we have, the better our models can capture the underlying patterns in the data. Using a large dataset to create training and test sets can help prevent overfitting and result in a more accurate and generalizable model.

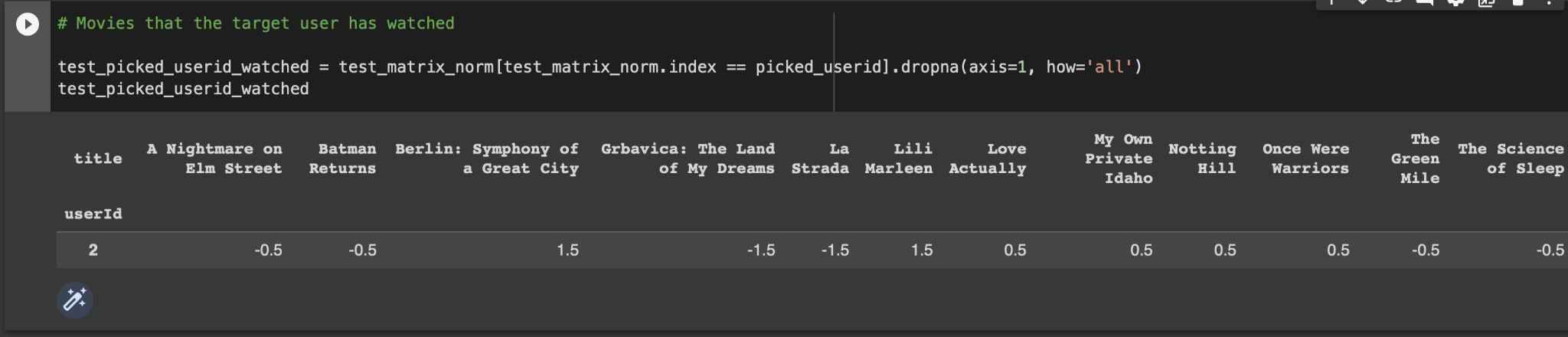


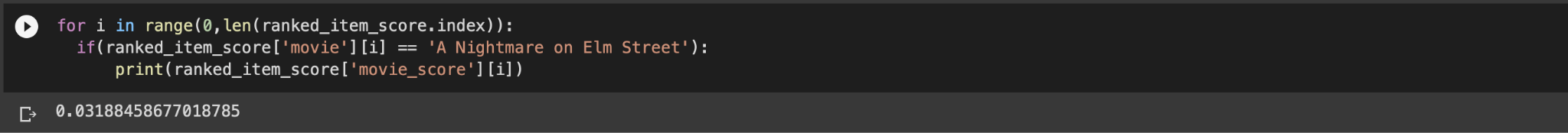
After that, we split data into 80% train and 20% test datasets. The reason for choosing 80% for training and 20% for testing is to strike a balance between having enough data to train the model effectively and having enough data to test the model's performance. This split is a commonly used heuristic, and it's not necessarily the best split for all datasets or problems, but it can work well in many cases. By using an 80/20 split, we can train the model on a larger portion of the data, which allows the model to learn more effectively from the data. At the same time, we reserve a portion of the data for testing, which allows us to evaluate how well the model generalizes to new, unseen data.



As we make our dataset larger we got a huge user-item matrix compared to before. Before it contains 597 rows and 134 columns but now it has 671 rows and 891 columns. Then we performed all operations which we performed earlier before for the smaller data. We again performed matrix normalization, Pearson correlation, and cosine similarity, picked one user to find its similar user and also found the movie score. This operation we already explained in the last report when we performed it for the small data.

Then, further, we printed the test dataset of the particular mentioned user and normalized it with the average of that user. Here we can also normalize the data using the mean of that user and divide it by the standard deviation of that particular user, so we can indicate the normalization we use in data science. After that, we predict the values of the normalized user as it, we have used a small data set to check it. We can see the predicted data of the model with the actual data, which is taken from the test dataset.





From the above image, we can see that after normalization of the test dataset for the movie “A Nightmare on Elm Street,” it is -0.5, and the model we prepared to get a prediction of the normalized value of 0.031. If we can say the model is 50 - 60 % correct for a particular mentioned movie as the value is not getting very high; further, we need to check it for all users and every movie a particular user has watched to see the actual performance of the model.