Term Project – Intelligent Data Analytics 5103 – 995

Predicting Netflix Movie Rating and User-Based Movie Recommendation

Group 5

Amina Asghar (amina.asghar-1@ou.edu) | Austin Mould (austin.k.mould-1@ou.edu)

Chase Sturgill (chase.sturgill@ou.edu) | Ducha Hapsari (ducha.n.hapsari-1@ou.edu)

Executive Summary

The streaming industry has become a constant in the global market, with numerous companies competing for their audiences' attention. Hence, it is imperative to these companies that when you watch a movie or series on their platform, they want to recommend content to you that will keep you engaged. However, recommending more content is not an easy task. There are many variables involved that could potentially affect how audiences react. With so many possible factors at play, it is essential to understand which factors genuinely affect viewers' likeliness to keep watching content they like. Thus, this research aimed to predict user review ratings based on historical user review data.

One assumption of this data was that the users watched each film entirely. In application, however, this may not always be the case. There may be times when a user does not finish a film because they do not like how the film was presented up to that point, but upon further viewing may enjoy the rest of it. Another primary assumption is that not every viewing of a film gets reviewed by every user. There are many real-world instances where a user watches a movie and does not leave a review of it, regardless of if they enjoyed it or not.

While this project intended to predict user rating based on movie data, further investigation did not support this idea. There is little evidence suggesting that predictive modeling effectively predicts average movie rating as a function of historical data. However, more investigation into customer metadata, such as their age, location, and favorite genre may be required to provide more personalized recommendations rather than simply relying on their past rating. This would allow more accurate prediction based on their background similarity. Meanwhile, the information presented in this research was only focused on movie metadata. This presents the issue of subjectivity in movies. A film cannot be judged by its genre, runtime, or other descriptors. Movies and other media are forms of art, and the beauty of art depends on the eyes of the beholder. Hence it may be more beneficial to investigate the effects of customer descriptors on movie ratings in future research.

Introduction

Streaming has become an invaluable pastime for many people across the world. Whether they are casual users like families getting together for a movie night or seasoned experts of binge-watching an entire season of “The Office” or “Friends,” there is no denying that the streaming service industry has left a significant impact on the world. The streaming service market has become a juggernaut of industry, with many companies competing for the attention of their users. In recent years, most entertainment conglomerates have shifted their focus to streaming. Some of the largest streaming services, such as Netflix, Amazon, Disney Plus, and HBO, have more than six hundred million users. Streaming is now considered the future of entertainment. All these streaming services compete for users’ attention. Therefore, it is of paramount importance for these streaming services to predict what content their users will like and recommend that content to their users. If a user is recommended content that is highly engaging and to their liking, they are more likely to use/continue using these services. However, one issue of streaming applications is to accurately predict how well a film or series will perform before release.

One way of doing this is to look at earlier releases of films and media on streaming platforms to see what factors affect user ratings. In 2020, Frank M. Schneider, Emese Domahidi, and Felix Dietrich studied what factors affect user descriptions and comments on a film. Their analysis found that users more often commented on a film’s actor performance, hedonism (comedy), and narrative. This analysis makes sense since most people watch films for entertainment, and these are all qualities that tend to aid in the immersion and suspense of disbelief in a movie. People want to be engaged with the film they are watching, and that engagement leads to more people sharing their thoughts on that movie and thus recommending it to their friends. The issue with these metrics is that they are qualitative and thus subject to interpretation, much like their movies.

Additionally, this approach only analyzes what users are commenting on a movie rather than relating these responses toa quantifiable quality measurement, such as user reviews. Furthermore, if a user review score can be predicted on a movie, this may help the streaming platforms recommend movies and series like what a user has already rated highly. Hence, it is imperative to find what factors affect a user score from sites like IMDb, qualitative or quantitative.

Data Description and Exploratory Analysis

The dataset used in this project was based on an actual Netflix database detailing the user ID, rating, date of the rating, and movie ID. The movie title and release year catalogs are listed on a separate database for the corresponding movie ID. The two datasets are joined to obtain the correct reference of the movie the users have rated. The compounded dataset is further joined with IMDb catalogs that had more information about the movies, such as their genre, director, production company, main actors, duration, country, number of votes, and average votes on IMDb. Joining the Netflix user data to IMDb data helped cut lesser-known movies not listed on IMDb. The data cleaning process was further carried out to remove movies that receive few ratings, i.e., less than the average movie rating entries of 51. This process helped narrow down the size of the dataset to the following size:

Total rated movies	546 movies
Total users	143,453 users
Total given ratings	3,625,073 entries

Netflix's rating system scales from 1 to 5 as the lowest and highest rating, respectively. The ratings given by users for all the movies are shown in **Fig. 1**. Ratings of 3 and 4 dominate the user judgment and makeup 33.09% and 31.7% of the entire entries. It is followed by a rating of 5 with 18.37%, 2 with 11.78%, and 1 with just 5.06%.

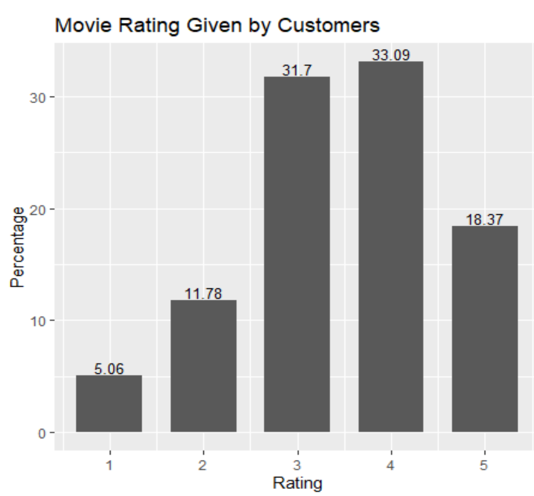
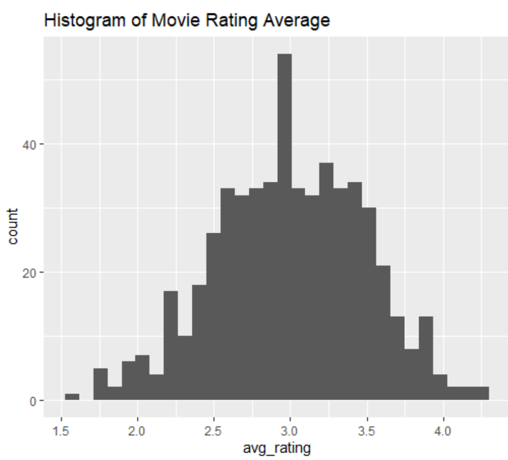


Figure 1 – User movie rating distribution

Out of those entries, it is found that the movie Harakiri received the highest average rating of 4.27, while the movie Midnight Mass received the lowest average rating of 1.58. The average rating of those 546 movies lies on 3.479, which indicates that, on average, the responses are positive. This average movie rating distribution is visualized in **Fig. 2** in a roughly symmetric distribution**.** With regards to users, the average movie rating given is 3.534. **Fig. 3** portrays its distribution in a bar chart which shows a left-skewed distribution. This figure shows that users tend to leave a positive rating on movies they like and are less likely to leave a negative rating on movies they dislike. Often, users leave the movies unfinished when they dislike them and do not bother to provide any ratings.



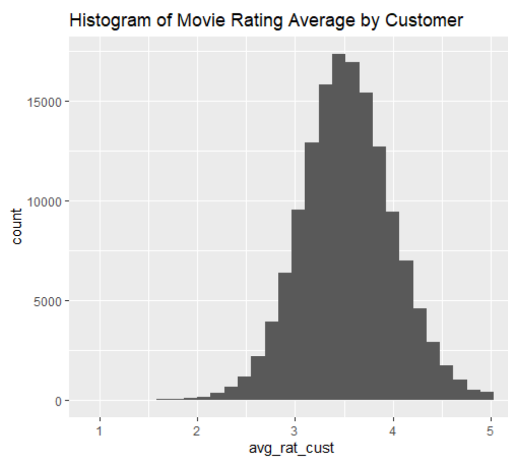


Figure 2 – Movie rating average

Figure 3 – Movie rating average by user

Throughout the exploratory data analysis, the log history for each user can be retrieved to learn more about their movie preferences. For example, the random user with an ID of *54333* gave high ratings on the following movies, suggesting their favorites.

Cust_Id	Rating	name	genre	release_year	country	duration	avg_vote	director	actor1
54333	5	Shrek 2	Animation	2004	USA	93	7.2	Andrew Adamson	Mike Myers
54333	5	In Good Company	Comedy	2004	USA	109	6.4	Paul Weitz	Dennis Quaid
54333	5	Coach Carter	Biography	2005	USA	136	7.3	Thomas Carter	Samuel L. Jackson
54333	5	Bad Boys	Action	1995	USA	119	6.9	Michael Bay	Lisa Boyle

Table 1 – Favorite movie user *54333*

The following table details the movies, the user *54333* moderately rated as 3, spanning various genres and years released.

Cust_Id	Rating	name	genre	release_year	country	duration	avg_vote	director	actor1
54333	3	Wonder Boys	Comedy	2000	USA	107	7.2	Curtis Hanson	Michael Douglas
54333	3	Titan A.E.	Animation	2000	USA	94	6.6	Don Bluth	Drew Barrymore
54333	3	The Rocky Horror Picture Show	Comedy	1975	UK	100	7.4	Jim Sharman	Tim Curry
54333	3	The Final Cut	Drama	2004	USA	95	6.2	Omar Naim	Robin Williams
54333	3	The Final Countdown	Action	1980	USA	103	6.7	Don Taylor	Kirk Douglas
54333	3	Starman	Romance	1984	USA	115	7.0	John Carpenter	Jeff Bridges
54333	3	Stargate	Action	1994	USA	116	7.1	Roland Emmerich	Kurt Russell
54333	3	Signs	Drama	2002	USA	106	6.7	M. Night Shyamalan	Mel Gibson
54333	3	Rocky V	Drama	1990	USA	104	5.3	John G. Avildsen	Sylvester Stallone
54333	3	Predator 2	Action	1990	USA	108	6.3	Stephen Hopkins	Kevin Peter Hall
54333	3	Mean Girls	Comedy	2004	USA	97	7.0	Mark Waters	Lindsay Lohan
54333	3	Hercules	Animation	1997	USA	93	7.3	Ron Clements	Tate Donovan
54333	3	Evolution	Comedy	2001	USA	101	6.1	Ivan Reitman	David Duchovny
54333	3	Don't Say a Word	Drama	2001	USA	113	6.3	Gary Fleder	Michael Douglas
54333	3	Dark City	Mystery	1998	Australia	100	7.6	Alex Proyas	Rufus Sewell
54333	3	Commando	Action	1985	USA	90	6.7	Mark L. Lester	Arnold Schwarzenegger
54333	3	Black Dog	Action	1998	USA	89	5.5	Kevin Hooks	Patrick Swayze
54333	3	American Psycho	Comedy	2000	USA	101	7.6	Mary Harron	Christian Bale
54333	3	2 Fast 2 Furious	Action	2003	USA	107	5.9	John Singleton	Paul Walker
54333	3	13 Ghosts	Horror	1960	USA	85	6.1	William Castle	Charles Herbert

Table 2 – Moderately rated movies by user *54333*

User *54333* only has one disliked movie with a rating of 1.

Cust_Id	Rating	name	genre	release_year	country	duration	avg_vote	director	actor1
54333	1	Napoleon Dynamite	Comedy	2004	USA	96	6.9	Jared Hess	Jon Heder

Table 3 – Movie disliked by user *54333*

Description of Modeling Approach

The joint Netflix and IMDb dataset was split into a testing dataset and a training dataset during the primary preparation of the data. Once split, a random sample of 50,000 observations was taken from each dataset to reduce the data's complexity and runtime. This random sample seemed to be the best sample size to yield more reliable results while keeping a reduced runtime. These random samples were investigated to see if the metrics within these observations seemed random. Upon further investigation, these observations did appear to have varied user ratings and movie titles to satisfy the appearance of randomness.

OLS (Ordinary Least Squares) regression was used to get a baseline predictive model of the data. This model was chosen due to the enormous size and complexity of the data. Since OLS is one of the simplest models to build and run computationally, it would make sense to use it as a baseline that can be improved later. With such a large dataset, the selected model needed to be simple enough to run in a reasonable amount of time. Furthermore, with such a large dataset, OLS was less likely to be affected by outliers in the data. Investigation into a more powerful model, SVM (Support Vector Mechanics), was also performed to achieve potentially more accurate results.

Attribute Selection

The following variables were chosen to construct an OLS regression model for movie rating prediction: user ID, genre, country, release year, leading actor, and production company. This information is selected as they contain each movie's pertinent information, which could help achieve accurate prediction. One issue encountered was handling a large number of factors for the non-numeric variables. For example, there are 504 unique values for directors, 494 unique main actors, and 405 unique production companies in the training set. A large number of these unique values pose a challenge in reducing the number of factors to the top few classes as there are too many variations. One approach that can be explored is to group the release year by decade instead of their yearly entry.

Results

Once training was conducted using OLS, a customer movie rating prediction was created for the test dataset. The preliminary OLS model was able to predict the rating up to 3 decimal points, and the predicted vs. actual rating results are visualized in **Fig. 4** below.

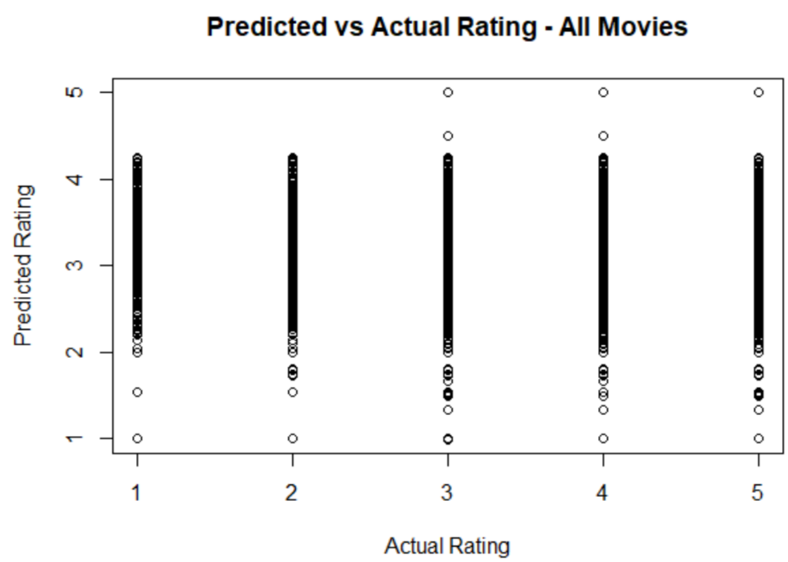


Figure 4 – Plot of predicted vs. actual rating for all movies

These preliminary results in **Fig. 4** show a relatively wide dispersion of predicted ratings (ranging from 1 to 4). This is partially due to the actual rating given as an integer, while the predicted outcome is given in up to 3 decimal points. Ideally, an accurate prediction would have a positive correlation between the actual and the predicted values and show a linear trend. Similar results were obtained from the SVM model but with an even wider dispersion. The following plots on **Fig. 5** display the behavior for an individual movie, with examples of Freddy vs. Jason and Mean Girls. Similar occurrences are observed.

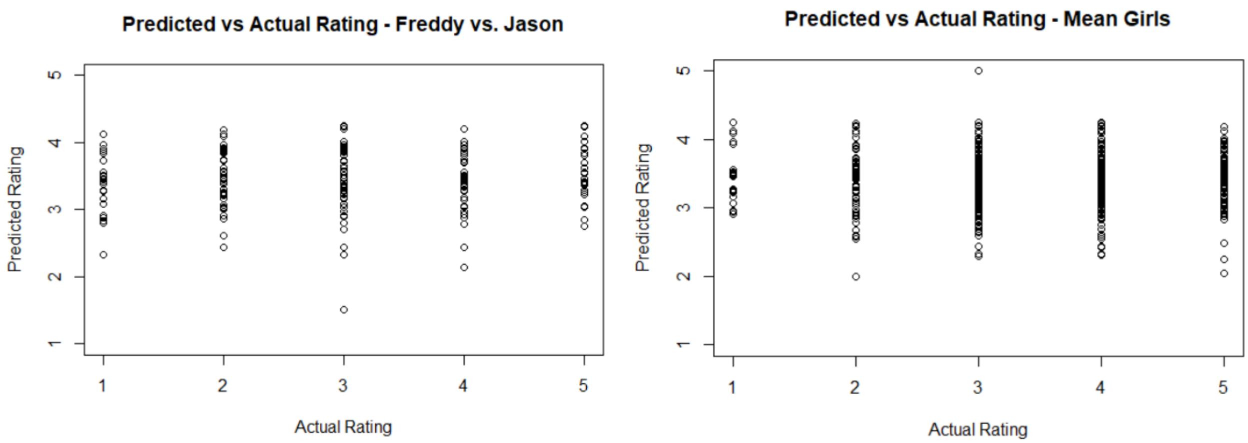


Figure 5 – Plot of predicted vs. actual rating for Freddy vs. Jason and Mean Girls

After obtaining results, a few individual users were analyzed to explore trends and further examine model accuracy. User *2439493* had the highest number of ratings in the sample dataset with nine different ratings. The table below highlights the predicted score along with all the associated details for all nine movies.

	predicted_rating	name	Cust_Id.y	Rating	genre	country	production_company	actor1
1	3.078	Manny & Lo	2439493	1	Comedy	USA	Pope Productions	Mary Kay Place
2	3.707	Species II	2439493	1	Action	USA	Metro-Goldwyn-Mayer (MGM)	Michael Madsen
3	3.371	Bad Boys II	2439493	1	Action	USA	Columbia Pictures	Martin Lawrence
4	3.713	Saajan	2439493	1	Drama	India	Divya Films International	Sanjay Dutt
5	2.438	Mr. Vampire	2439493	1	Action	Hong Kong	Bo Ho Film Company Ltd.	Ching-Ying Lam
6	3.924	Juice	2439493	1	Action	USA	Island World	Omar Epps
7	3.488	Rescue Heroes: The Movie	2439493	1	Animation	Canada	Nelvana	Norm Spencer
8	3.409	Jackpot	2439493	1	Drama	USA	Polish Brothers Construction	Jon Gries
9	4.122	Shalimar	2439493	1	Action	India	Judson Productions	Dharmendra

Table 4 – User *2439493* Table

The first thing to highlight with this user was that they were an anomaly; the user has given every movie a rating of one. The movies have a wide range of various attributes, making it difficult to determine the type of movie the user would enjoy. All movies also contained different directors, production companies, and actors. The model predicted that Shalimar would be rated the highest, although the user rated every movie differently in this case.

This type of user would be hard to predict since it cannot be known what they like. In cases like this, it would help to have more information about the user. In Netflix's case, they might know what the user watches regularly and doesn't rate, assuming what the user enjoys since their ratings are only negative.

User *684531* has the second-highest number of ratings at eight.

	predicted_rating	name	Cust_Id.y	Rating	genre	country	production_company	actor1
1	4.192	Ernest Goes to Jail	684531	1	Comedy	USA	Touchstone Pictures	Jim Varney
2	3.079	The Others	684531	4	Horror	Spain	Cruise/Wagner Productions	Nicole Kidman
3	4.127	American Beauty	684531	4	Drama	USA	DreamWorks	Kevin Spacey
4	3.361	Animal Crackers	684531	4	Comedy	USA	Paramount Pictures	The Marx Brothers
5	3.307	Evolution	684531	2	Comedy	USA	Columbia Pictures	David Duchovny
6	2.227	Life	684531	3	Comedy	USA	Imagine Entertainment	Eddie Murphy
7	3.341	Igby Goes Down	684531	3	Comedy	USA	United Artists	Kieran Culkin
8	3.392	Speed	684531	4	Action	USA	The Mark Gordon Company	Keanu Reeves

Table 5 – User *684531* Rating Predictions

This user gave four ratings to four out of the eight movies, all of which were different genres and directors, production companies, and actors. They rated the most comedies, with only one getting a four and the other receiving a three or lower ratings. Ernest Goes to Jail received the highest rating from our model, a comedy set in the US, and a similar year to the other comedies the user had rated, despite the user rating the movie a one. This is another case where additional details about the user could help to better understand movie ratings since none of the movies with ratings of fours from the user have any metadata in common.

These users highlight an underlying issue with trying to predict how a user will rate a movie. The movies the users rated did not have much in common based on the available metadata, making it challenging to uncover why they rated the movie with that specific score. User *2439493* brings up another potential issue with the data or potentially the sampling method, where the user rated every movie the same score, and in this case, it was the lowest score possible. Users like this would be difficult to predict as they deviate far from the expected user ratings.

More metadata about the movies would be beneficial in trying to find the common attribute in ratings. Another consideration is the user data, which was missing from the dataset. It is possible that it is better to predict movie ratings based on user commonalities instead of movie commonalities. With user 2439493, this could have helped highlight why they rated every movie they watched so low if people with similar backgrounds watched different types of movies. Additional user data would allow Netflix to show recommendations based on what other users with similar backgrounds (similar age range, location, etc.) enjoyed, which could lead the user to find movies they actually enjoy and encourage them to continue renewing their subscription.

Conclusion

In summary, the user-review data from both Netflix and IMDb were combined to try and predict user review scores. The data was preprocessed to reduce its size, runtime, and any confounding effects from outliers. First, movies with a total number of reviews less than the average number of reviews per movie were removed. The idea behind this was to remove movies that weren't as popular and could be highly affected by outlier review scores, thus inflating or deflating their average review scores. The size of the data was then further cut by sampling 50,000 reviews. Finally, only 6 factors that seemed to have the most impact on a user review score were used since they would likely be more useful in the data modeling. The data was modeled using OLS to achieve an initial baseline model to spearhead future research. However, this model would yield very little success in terms of accuracy. This might've been due to the simplicity of OLS and its properties. Specifically, OLS is highly affected by outliers and is linear in nature, which does not work well with data that isn't continuous like the user review scores.

Some issues faced in the analysis were the many outliers that could not be seen until further investigation into specific users that only left reviews that were skewed toward one extreme or another. An excellent example of this occurring in a real-world setting would be bot accounts designed to only give extremely low or high reviews for certain movies depending on the user preference of those bots. Another issue was the long runtime for the creation of the OLS model. While the data was sampled appropriately, runtimes were still extremely high. SVM was also used with a smaller sample size to attempt a more complex modeling technique, only to yield even less accurate predictions. It is possible that its smaller sample size impacted the SVM model's results. Most of these limitations can be attributed to limited hardware power and processing time. If given enough time or more powerful machines, an even larger sample size with potentially more accurate modeling techniques could be processed.

In conclusion, the data showed some potential with predictive modeling. Future research may provide better results, particularly with the use of more advanced modeling techniques. Doing this may require using a smaller sample size to reduce process time, but it could still lead to better results in a timely manner. Another consideration for future research could be the removal of skewed reviewers and their bot accounts. However, this would require more insight into the subject to avoid indiscriminate review deletions. While there was some potential, there is no certainty that predictive modeling effectively predicts user review scores as a function of user ID, genre, country, release year, leading actor, or production company. Instead, a film's performance and rating may be more closely related to other factors related to the user instead of the film.

References

Schneider, Frank M, et al. "What Is Important When We Evaluate Movies? Insights from Computational Analysis of Online Reviews." *Media and Communication (Lisboa)*, vol. 8, no. 3, 2020, pp. 153–163., https://doi.org/10.17645/mac.v8i3.3134.