Business Movie Recommendation System Creation: Data Science Applications

Introduction:

Data science is the study of data in various forms to derive insights and in this context to use for business applications to achieve a predefined result. Businesses operate in many different ways and they value different ways of conducting business to achieve predefined results. Businesses within the movie domain, recommendation systems are built using a variety of techniques tailored to enhance user satisfaction and engagement (Papneja, 2022). Furthermore, businesses understand they can create a better user experience and customer engagement from using data to create recommendation systems to recommend movies to users that have high probabilities of the user liking and enjoying the recommendation. Recommendations are based on previous insight of user behavior and previous likes and reviews that's given. This data is analyzed and used to create a infrastructure to make recommendations and have a high statistical probability users will have a heightened experience therefore retaining customers long term, which increases life time value per user. Recommender systems have proven to generate measurable business value, especially in industries like online streaming, where user retention is critical (Jannach, 2019).

For example, one of the most popular recommendation systems is Netflix. When a business improves their recommendation system to an optimal level this in turn creates multiple effects, one dominant effect as stated before is user or customer retention. Retention is directly linked to higher revenues over the long term. All businesses desire to maximize profits and to grow continuously so by having high retention this equals larger revenue over longer periods of time. To achieve high retention and stabilize long term profits, in this context enhancing recommendation systems through data science techniques helps businesses continue to grow. Modern recommendation systems increasingly rely on deep learning techniques to model user preferences and improve predictive accuracy (Zhang, 2017)

As mentioned earlier, advancing recommendation systems are proven ways to help grow revenue but here I will explore the creation of a movie recommendation system using a movie dataset. Here, I will apply data science techniques to create a movie recommendation system. Once the movie recommendation system, also known as the algorithm, is created next I will train a machine learning algorithm using the inputs in one subset to predict movie ratings in the validation set. Once the algorithm reaches its final form, its final test is to see how accurate the algorithm is when it is to predict movie ratings. The algorithm is compared to the final holdout test set as if they were unknown. The Root Mean Squared Error (RMSE) was used to evaluate how close my predictions were to the true values in the final holdout test set. The business problem that's solved here is to create a movie recommendation system that helps increase user retention, which increases the lifetime value per customer and eventually leads to higher revenue per customer over a long period of time. If a movie recommendation system is created then this enhances the business value by retaining users longer so the result is more sales from paid subscriptions. Once the recommendation system is created the model that's developed will not only be evaluated using the root mean square error (RMSE) metric but also using a final hold out test set to test the actual model predictions for accuracy.

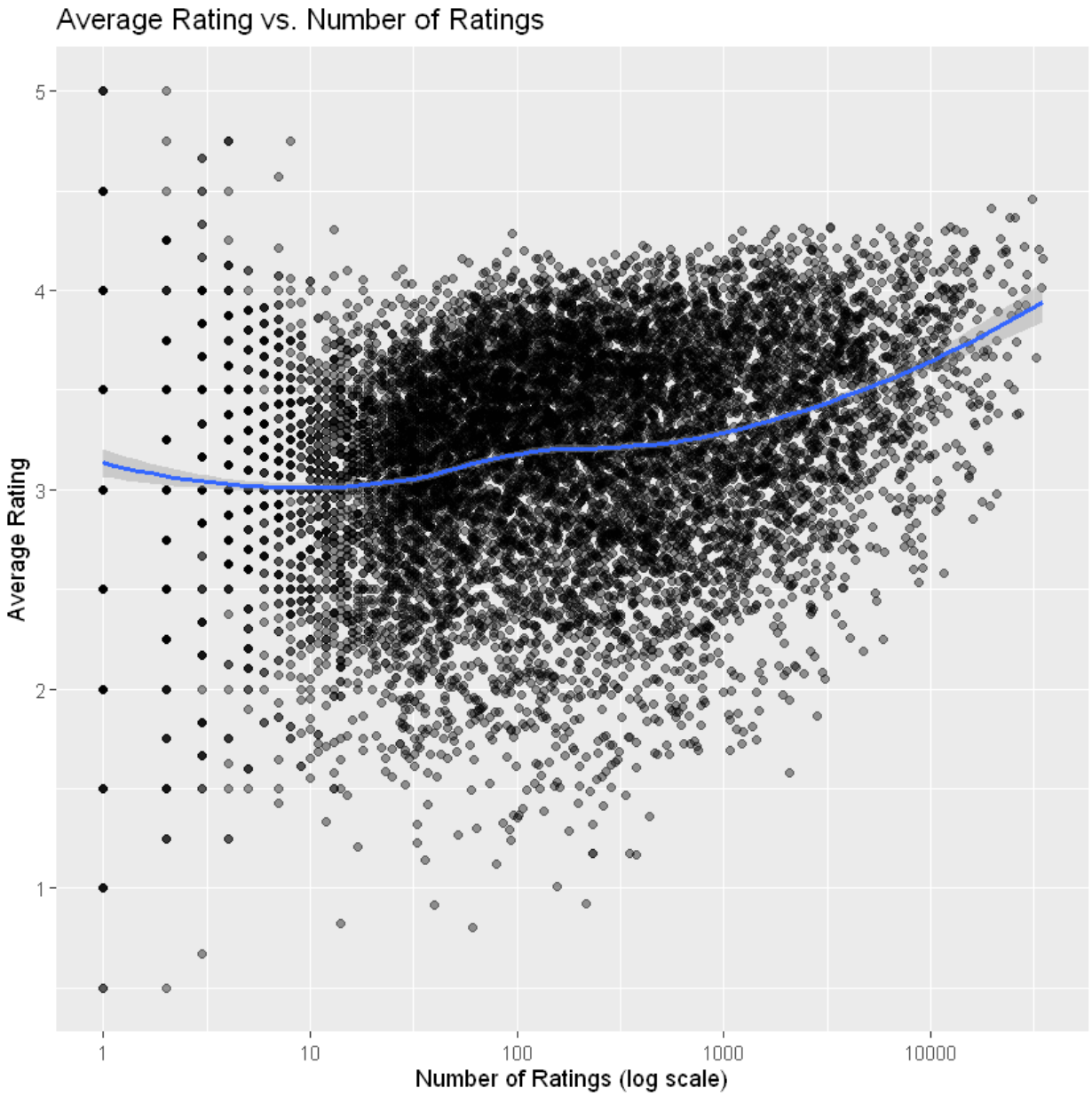
Data & Methods The dataset being used in this project is from a tiny subset of a substantially larger dataset that involves millions of ratings. The original mega data set is from the GroupLens organization. GroupLens is a research lab in the Department of Computer Science andEngineering at the University of Minnesota. The concentration is inrecommender systems, online communities, mobile and certain types of technologies, digital libraries, and local geographic information systems. A big part of solving problems using data is making sure data is from reputable sources and this group lens organization meets this criteria. It is important to not rush the process and to ensure the data being used is a good match for the problem that's

trying to be solved and that the quality of data is as close to superb as possible. As critical and important as the data is, the methods are also important. The methods that will be used in this project will be the Cross-Industry Standard Process for Data Mining, which is the CRISP-DM framework. Briefly, the CRISP-DM will follow the below phases: (1) confirming business understanding and problem to be solved as well as the desired business outcome or results. (2) Data understanding through observing overview of data. (3) Data preparation through data cleaning, constructing and organizing in preparation for data exploration to occur right after the preparation stage (4) Modeling stage is performed and designed is finalized (5) Evaluation is executed and next steps are determined along with future improvements noted (6) Deployment of algorithm as I monitor to improve and adjust accordingly relative to business objectives.

Data Explore

In this dataset, as I explored it there were 9000055 rows & 6 columns. All of the following is general information to help understand the context and body of the data.There are no zero star reviews and upon further review from other sources and society doesn't typically see zero star reviews. The four star review was selected the most compared to the others. There are 10,677 unique or distinct movies in total. The total user count is 69,878. I will not list the total ratings for each genre but here is a sample: Drama: 3,910,127 ; Comedy: 3,540,930 ; Thriller: 2,325,899 ; Romance: 1,712,100. The movie Pulp Fiction had the greatest number of ratings. The three most submitted ratings in order from the most to the least were the 4 star, 3 star and 5 star. Summary statistics were observed to gauge greater depth of dataset as it relates to the goal of predicting future movie ratings. The total average movie rating across the entire dataset was 3.5 stars. The overall goal is to predict future movie ratings so the variables selected for the predictors, influencers or factors are the userId, movieId and genres while the target variable is ratings. The model used will be a random forest. Key visualizations but not limited to these include (1) exploration of Average Rating vs Number of Ratings (2) Low Average Ratings vs Number of Ratings (3) High Ratings vs Number of Ratings (4) Average Movie Rating Over Time.

Figure 1

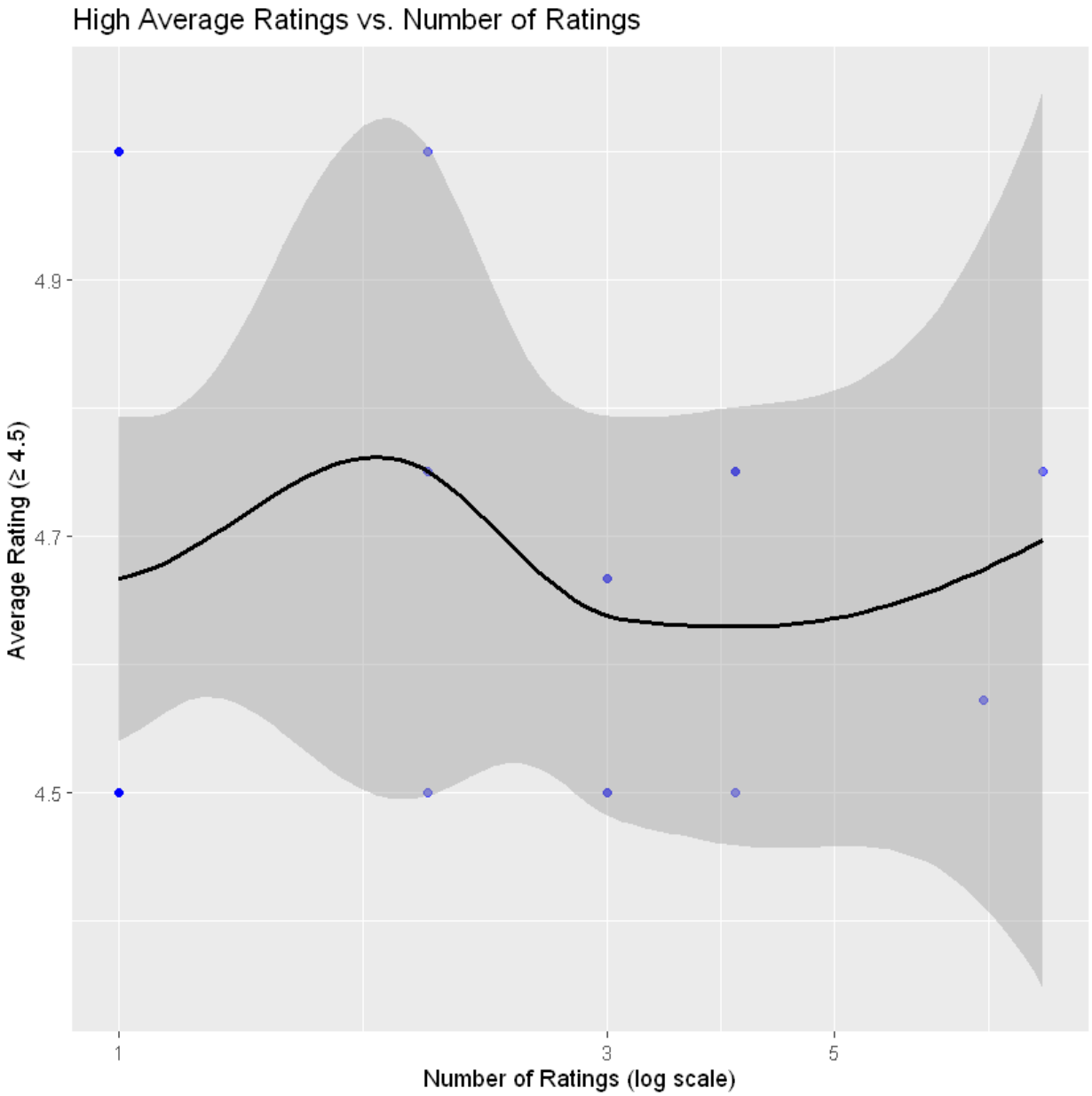


The average movie rating mentioned earlier in this project is 3.5 and this is confirmed from figure one above. Overall, in figure 1 there is a weak nonlinear expression from the data. This means that there is no strong correlation from the average rating vs the number of ratings. At the start of the average ratings the line begins not too far above from -3. As the number of ratings increase the average ratings decline to an all time low then rises again slowly to where it finishes just under -4. No real correlation here with these variables but still critical in pointing out to us the right direction even if it is away from these variables.

Figure 2

Altogether, figure 2 has a weak non-linear expression coming from low average ratings compared to the number of ratings. There is no strong correlation. At the start the low average rating is between 1.75 and 2. As the number of ratings rise the low average rating increases for a period then declines before rising again. The end point for the low average rating is in between 2.25 and 2.5. This gives us insight into understanding what variables to focus more on or less of when it comes to finding variables with strong correlations.

Figure 3



When examining the high average ratings compared to the number of ratings there is a weak correlation that is nonlinear. High average ratings start out at just under 4.7 then rises to just under 4.8 as the number of ratings go up. As the number of ratings keep going up the high average ratings drop to an all time lowest high average rating at slightly above 4.6. Lastly, as the number of ratings continuously climbs up the high average ratings end point is at or around 4.7. The variables here serve to push the data exploration toward or away from variables that have high correlations so adjustments can be made to uncover underlying trends.

Figure 4

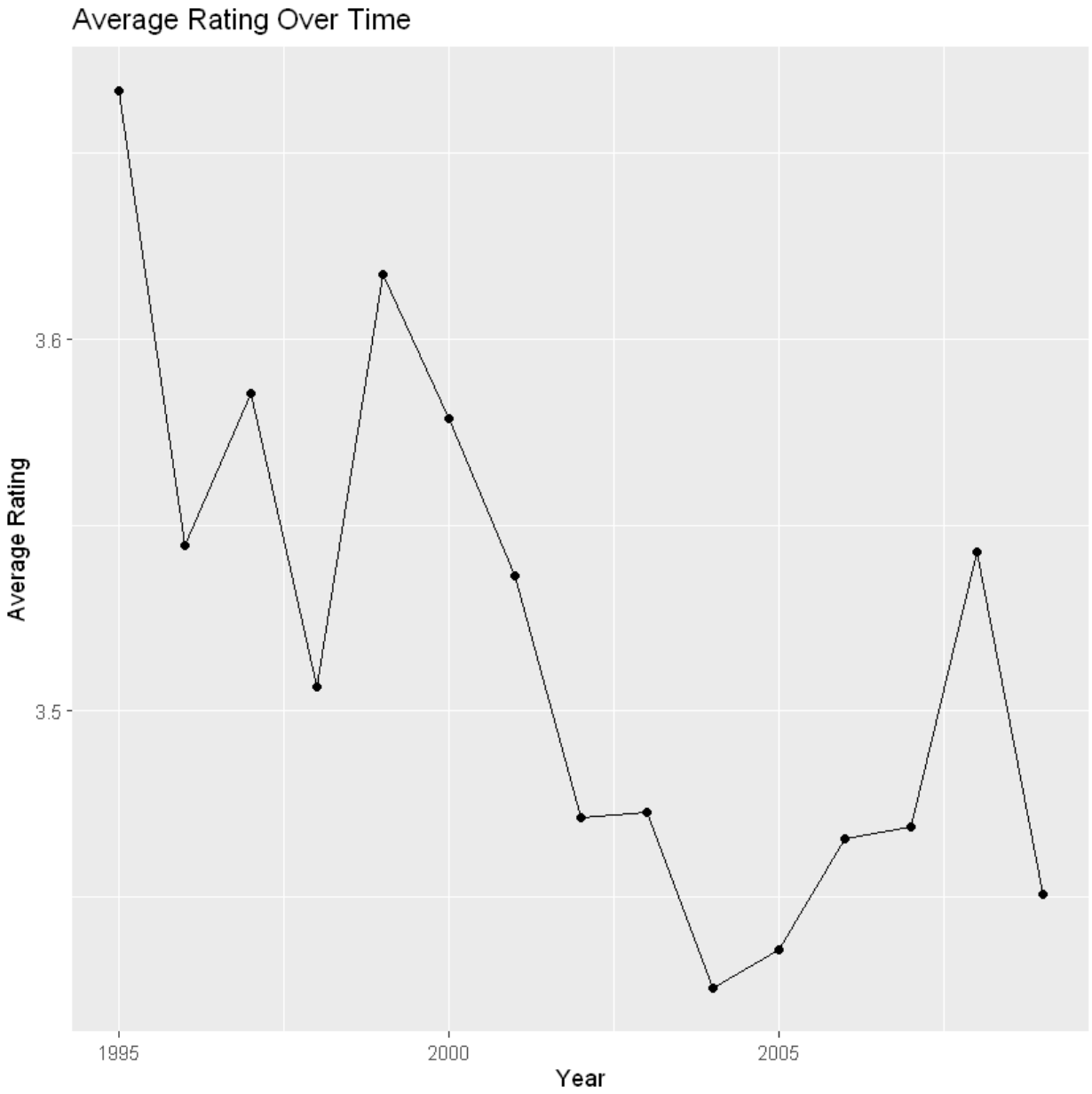
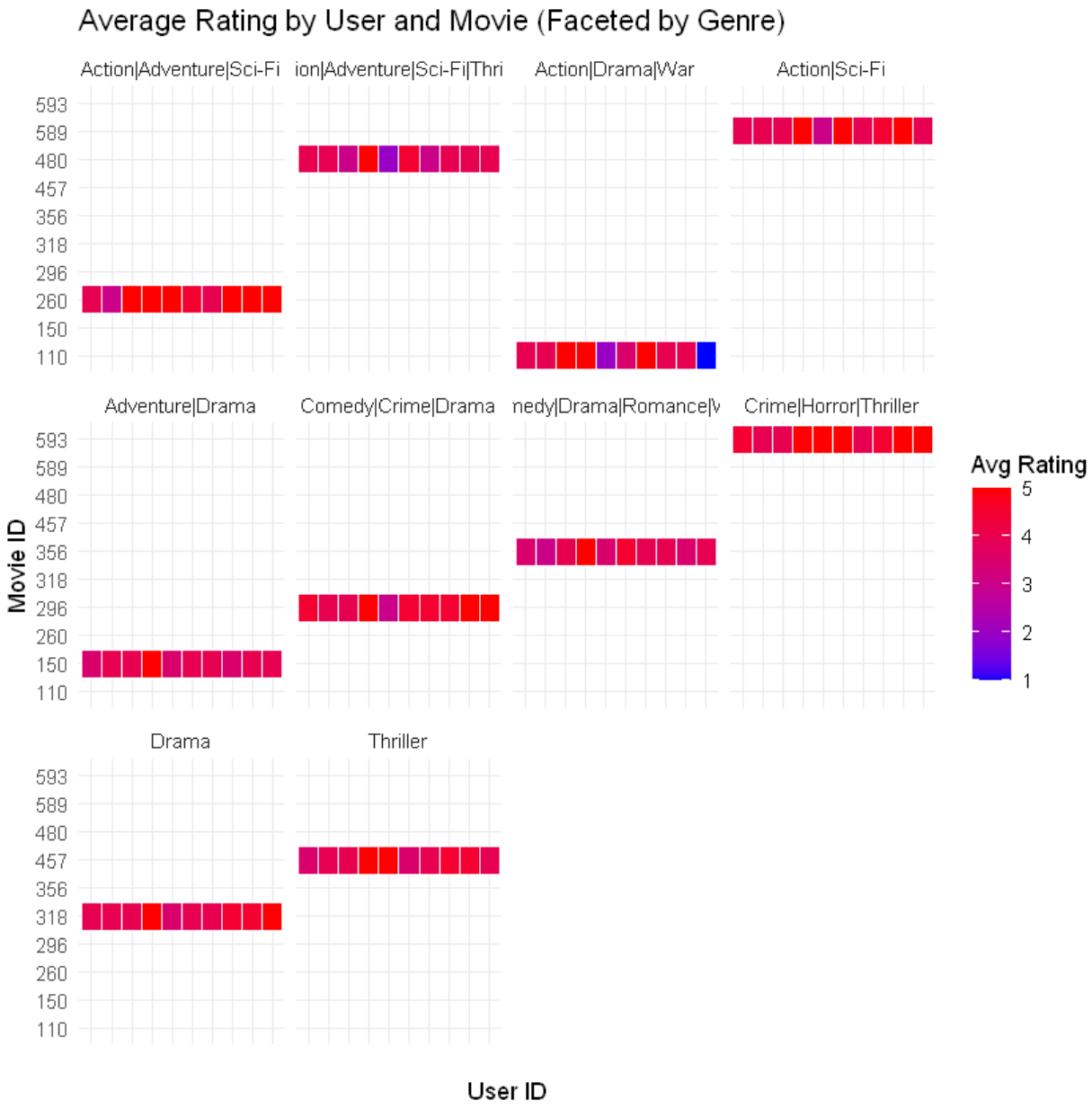


Figure 4 demonstrates that the average ratings for movies over time has fluctuated up and down from around 3.5 stars in the year 1995 to ending around 3.45 stars in 2010. The overall average movie rating from the data set is 3.5 stars but during the historical review beginning in 1995 and going until 2010 the average movie rating as mentioned previously stops at 3.45 stars.

Below is figure 5, which shows average rating from users and the movie plus it incorporates genre. The goal of this project is to create a movie recommendation system. During the data explore phase from previous visuals of many variables there were weak correlations related to predicting future movie ratings. However, as I explored further, I decided to select my predictor variables to be User ID, Movie ID and Genres. I believe these combined will produce accurate predictions of future movie ratings. Figure 5 expresses the user ID on the x-axis while the movie ID is on the y-axis. The average rating is marked by color and is anything from 1 star all the way to 5 stars. Figure 5



Modeling & Evaluation

Modeling in the field of data science,refers to the process of developing a mathematical or computational representation of a real-world system. This representation, known as a model, is constructed using historical or observed data to capture underlying patterns and relationships. These models are designed to either explain behaviors or predict future outcomes based on past trends.

When used specifically for forecasting, this process is called predictive modeling. Predictive modeling involves applying statistical or machine learning algorithms to historical datasets in order to make informed predictions about future or unseen data. These models are trained on known input-output pairs and then tested for their ability to generalize to new scenarios. There are many common techniques used in prediction. (Kuhn & Johnson, 2013).

The data was split 80% for training and 20% for testing because this balance gives a good amount of data for the model to learn trends while reserving an amount of unseen data to fairly evaluate its performance. This split helps make sure the model generalizes well without overfitting. As far as other ratios like 70/30 or

90/10 can be used depending on dataset size and goals, I decided on 80/20 because it is a common, effective standardized industry default for the overall project objectives.

For this project, I applied the random forest model because numerous publications have used this as a basis and is a trusted model in this sector for high accuracy. Model performance is evaluated using mathematical metrics. There are several and the one used in this project is the Root Mean Squared Error (RMSE). The application of predictive modeling for movies is movie rating predictions, which play a major role in recommendation systems. These systems predict a user's potential rating for a movie they have not yet watched, thereby enabling personalized movie recommendations. These predictions are a continuous rating value from 5 stars to 1 star ratings.

There are two modeling approaches but are used in combination, which are collaborative filtering and content-based filtering. Collaborative filtering focuses on patterns in user behavior. It assumes that if two users have historically rated movies similarly, then one user is likely to enjoy movies the other has liked. One predictor variable out of others used, userId, is present in the dataset so we know the collaborative approach is in motion. In contrast, content-based filtering relies on attributes of the movies themselves, such as genre, director, cast, or keywords. The system uses this information, along with user preferences, to predict ratings for movies with similar characteristics. The content-based modeling approach is present due to the movie genre selected as one of the predictors used to forecast movie ratings.

Both approaches are present in the system so this would be a hybrid modeling approach, which utilizes both collaborative and content-based approaches to advance prediction accuracy and makeup for the limitations of each individual method. These models benefit from the strengths of both strategies: the understanding of user behavior from collaborative filtering and the descriptive power of content-based features make movie prediction more accurate working together compared to working independently of each other. In total, the predictor variables I chose were User ID, Movie ID and genre while the target variable is movie ratings.

To develop these models, user-movie rating data is split into training and test sets. The training data is used to fit the model, and its performance is validated using the test set. Evaluation metrics such as RMSE determine how accurately the model can predict unseen ratings (Ricci, Rokach, & Shapira, 2015).

Below is Figure 6 and it is the random forest regression model output information. Figure 6

```
Type:                             Regression
Number of trees:                  10
Sample size:                      8000045
Number of independent variables:  3
Mtry:                             1
Target node size:                 5
Variable importance mode:         impurity
Splitrule:                        variance
OOB prediction error (MSE):       0.9285253
R squared (OOB):                  0.1743014
```

Below here is Figure 7, which displays the RMSE score from the model's performance evaluation. Figure 7

```
rmse <- sqrt(mean((predictions - test_data$rating)^2))
cat("RMSE on test set:", rmse, "\n")
```

```
RMSE on test set: 0.9537687
```

The RMSE is too high at 0.9537 so an ensemble averaging approach is implemented to combine predictions of my random forest model and a baseline model with regularized movie plus user effects. In figure 8 below, the RMSE is 0.8622, which is an acceptable RMSE score. At first the RMSE score from the ensemble model was not acceptable. When the model was given different weighted averages so the weight of the predictions lean more toward the better model this gave the model an acceptable RMSE score.

Figure 8

```
w <- 0.7  # weight for baseline model
ensemble_preds <- w * baseline_preds + (1 - w) * rf_preds
final_rmse <- RMSE(final_holdout_test$rating, ensemble_preds)
cat("Weighted ensemble RMSE:", final_rmse, "\n")
```

```
Weighted ensemble RMSE: 0.8622617
```

Results, Discussion & Conclusion

The purpose of this project is to create a recommendation system to predict future movie ratings. Business problems can range but here let's assume a business wants higher user retention as they know the longer they keep a user watching movies the longer they pay subscription fees so it creates increased revenue and longer lifetime value per user. In the creation of this recommendation system, the business problem is solved. The results from this project show that using a hybrid collaborative and content-based modeling approach was the best option in constructing the random forest model for accurate future movie predictions. The result from the predictive ensemble model with the weighted averages, after evaluation has a RMSE score of 0.8622, which is appropriate for machine learning models. In conclusion, There are several actions that were not taken that could have led to possibly a more accurate predictive model such as adding more data or using different evaluation metrics. There is always room for improvement and future similar projects may consider other approaches and methodologies. From a global view, user behavior, technology and new inventions can totally change how we use data science to solve business problems. For businesses and professionals to keep their strong abilities to solve complex problems it's important to continue to learn and apply high performing methods to reach desired outcomes.

References:

"Hybrid Recommendation System for Movies Using Artificial Neural Network." Expert Systems with Applications, vol. 258, 15 Dec. 2024, p. 125194. ScienceDirect, doi:10.1016/j.eswa.2024.125194.

Jannach, Dietmar, and Michael Jugovac. "Measuring the Business Value of Recommender Systems." arXiv, 22 Aug. 2019. https://arxiv.org/abs/1908.08328.

Miao, Y. "Digital Movie Recommendation Algorithm Based on Big Data Platform." Mathematical Problems in Engineering, 2022, Article ID 4163426. Wiley Online Library, doi:10.1155/2022/4163426.

"Movie Recommendation through Multiple Bias Analysis." Applied Sciences, vol. 11, no. 6, 2021, p. 2817. MDPI, doi:10.3390/app11062817.

Mu, Yongheng, and Yun Wu. "Multimodal Movie Recommendation System Using Deep Learning." Mathematics, vol. 11, no. 4, 2023, p. 895. MDPI, doi:10.3390/math11040895.

General Use Purposes: OpenAi

Papneja, D., et al. "Movie Recommender Systems: Concepts, Methods, Challenges, and Future Directions." Sensors, vol. 22, no. 13, 2022, p. 4904. MDPI, doi:10.3390/s22134904.

Tripathi, S., et al. "A Hybrid Recommender System Based on Link Prediction for Movie Baskets Analysis." Journal of Big Data, vol. 8, no. 1, 2021, Article 59. SpringerOpen, doi:10.1186/s40537-021-00422-0.

Xia, Ziyuan, et al. "Contemporary Recommendation Systems on Big Data and Their Applications: A Survey." arXiv, 31 May 2022. https://arxiv.org/abs/2206.02631.

Yu, Fei, et al. "Network-Based Recommendation Algorithms: A Review." arXiv, 19 Nov. 2015. https://arxiv.org/abs/1511.06252.

Zhang, Shuai, et al. “Deep Learning Based Recommender System: A Survey and New Perspectives.” arXiv, 24 July 2017. https://arxiv.org/abs/1707.07435.