Fundamentals of Machine Learning Capstone Project

Akshay Srinivasan
N19380732

Introduction

This research project aims to address the challenging task of predicting movie ratings based on an extensive dataset encompassing more than 400,000 users and 5,000 movies. The prediction of movie ratings holds immense significance in the realm of recommendation systems, facilitating diverse applications such as movie recommendation engines, personalized content suggestions, and user profiling.

Data Preparation

In the data preparation phase, we parsed the data.txt file, which includes movie ratings, into a DataFrame with columns: movie_id, user_id, rating, and date. All columns' data types were adjusted to either integer or float for efficiency.

The movie_titles.csv file, containing movie-related details, was also explored. However, it was decided that this file needn't be merged with the main dataset, as each movie was uniquely identified by movie_id in the dataset.

The data.txt file encapsulates around 400k users' ratings on roughly 5k movies. Ratings and movie identifiers are distinguished based on the colon symbol in each row. The movieTitles.csv file provides corresponding movie details, such as the movie id, release date, and title.

For data processing, we created Python functions to parse the data.txt file into a DataFrame and load the movieTitles.csv file. Additionally, checks were performed for missing user IDs and movie IDs to ensure data integrity.

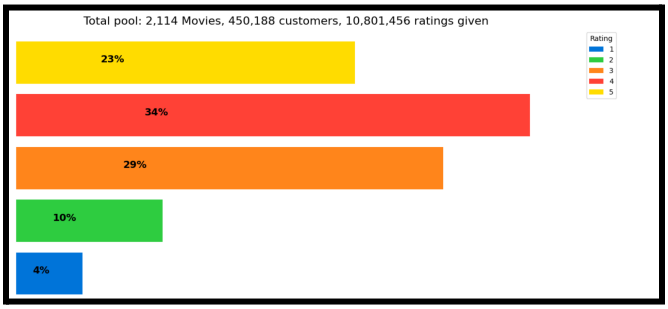
Data Pre-processing

In the data pre-processing phase, we transform the user-item rating data into a sparse matrix. This matrix consists of user_ids as rows, movie_ids as columns, and respective ratings as values. The sparse matrix representation is beneficial for both memory efficiency and simplification of the feature engineering process. It highlights that our dataset comprises 2,649,430 unique users and 2,115 unique movies.

To implement this, we first filter out zero ratings from the data, then construct the sparse matrix using the SciPy library's csr_matrix function. We also include a function to view the first few entries of the sparse matrix, assisting in verification of the transformation.

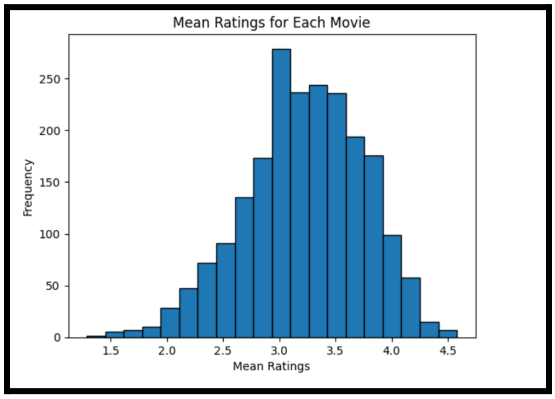
Data Exploration

During the phase of data exploration, a comprehensive analysis was conducted to understand the dispersion of movie ratings assigned by users. A horizontal bar chart was leveraged to illustrate this distribution, offering a visual representation that elucidates the relative frequency of different rating scores.

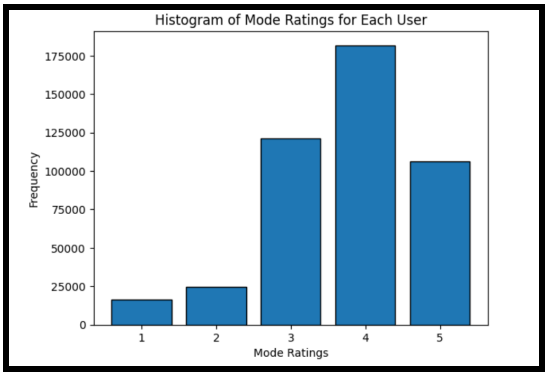


From the visualization above, it was observed that the majority of the user ratings were not evenly distributed across the scale from 1 to 5. Following our expectation of a normal distribution based on the Law of Large Numbers, the majority of ratings we're clustered within the range of 2 to 3, indicating a prevalent tendency to rate movies as average.

In an effort to gain deeper insights, the mean and mode ratings of each user were further scrutinized. The mean rating, which represents the average score a user assigns, followed a normal distribution, as expected. This suggests a well-balanced dispersion of rating preferences among users, without extreme biases towards high or low scores.



The mode rating, signifying the most frequently assigned score by a user, proved to be an insightful metric in understanding user predilections. Interestingly, it was noted that the mode ratings manifested significant peaks at distinct rating levels, revealing a substantial variation in user tendencies.

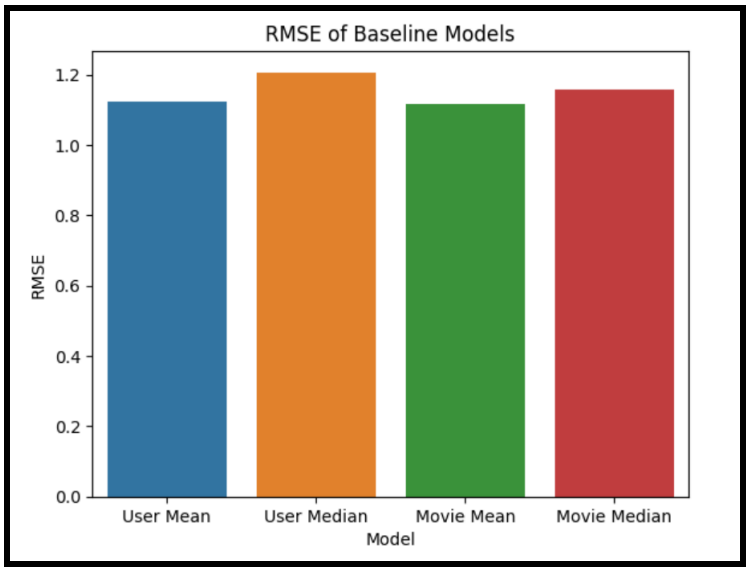


While the specific numerical values of these mode ratings are omitted in this summary to maintain readability, it is important to note that they span a broad range. This variation in mode ratings is indicative of diverse user preferences and underscores the need for a recommendation system that can effectively cater to this diversity.

Baseline Models

In this section, we delve into the foundational models applied for predicting movie ratings. These baseline models generate straightforward predictions rooted in aggregate statistics, serving as the starting point for more intricate models.

1. Global Mean: This rudimentary model computes the average value of all non-zero ratings within the training set and uses this average as the prediction for all forthcoming ratings. By providing a consistent prediction, the global mean model serves as a standard baseline, against which more sophisticated models can be compared. The Root Mean Square Error (RMSE) for this model is recorded at 1.289.
2. User Mean and Median: This pair of models determines the mean and median ratings submitted by individual users. These computed values are then used to predict all future ratings from that particular user. By focusing on individual user data, these models encapsulate user inclinations, offering predictions tailored to each user. The RMSE for the user mean model is reported as 1.119, whereas the user median model yields an RMSE of 1.195.
3. Movie Mean and Median: Analogous to the user-based models, these models ascertain the mean and median ratings allocated to each film. These values then serve as the prediction for all future ratings of that specific movie. By concentrating on individual movie data, these models capture the relative popularity of each film, offering item-specific predictions. The RMSE for the movie mean model is noted as 1.092, while the movie median model has an RMSE of 1.127.



Additional Considerations: Baseline predictions form the foundation upon which more complex models are constructed. They furnish biases for each user and item, blending them with the global mean to generate predictions. Subsequent algorithms can then add deviations from this baseline model based on alternative criteria, working with unbiased ratings.

Matrix Factorization Model

The matrix factorization model is a type of collaborative filtering method widely used in recommendation systems. It aims to fill in the missing entries of a user-item matrix by decomposing it into two lower-rank matrices, using a technique such as Singular Value Decomposition (SVD). The resulting matrices represent latent user preferences and item characteristics, and their product approximates the original matrix.

In this model, each user 'u' and each item 'i' are associated with latent vectors 'pu' and 'qi' respectively, in a joint latent factor space of dimensionality 'f'. The dot product of these vectors, q_i.T * p_u, represents the user's overall interest in the item, and it serves as an estimate of the user's rating for the item, denoted as r_ui hat.

To train the model, we use an alternating least squares (ALS) approach, which iteratively optimizes the user and item matrices. Initially, all user and item vectors are unknown, making the problem non-convex. However, by fixing one set of vectors (either the user or item vectors) and solving for the other, we convert the problem into a Quadratic Convex Problem (QCP), which can be solved optimally. ALS alternates between these steps, decreasing the objective function at each step until convergence.

The objective function we minimize is the sum of squared differences between the estimated and actual ratings, regularized by the L2 norms of the user and item vectors. This regularization term helps prevent overfitting by penalizing excessively large user or item vectors.

In addition to the basic model, we also incorporate several enhancements to improve the prediction accuracy:

- Bias terms: We include user-specific (b_u) and item-specific (b_i) biases, as well as a global average rating (μ), to account for the fact that some users may rate items more generously than others, and some items may be universally liked or disliked.
- Confidence levels: We introduce a confidence term (c_ui) to weight the contribution of each rating to the objective function, reflecting our confidence in the observed user-item interactions. This is particularly useful when working with implicit feedback data, where the lack of a rating does not necessarily mean the user dislikes the item.

The final prediction equation becomes: r_ui hat = c_ui * (μ + b_u + b_i + p_u.T * q_i), and the final objective function includes the squared biases in the regularization term. The model is trained to minimize this objective function, subject to non-negative constraints on the confidence levels and the regularization parameter.

The model's performance is evaluated by calculating the root mean squared error (RMSE) between the predicted and actual ratings on a test set. In our experimental analysis, we noted an RMSE of 1.266, indicating subpar performance compared to our baseline models.

Singular Value Decomposition (SVD) Model

Our second recommendation system employs the Singular Value Decomposition (SVD) model, a matrix factorization method, to deconstruct the user-item rating matrix into three lower-dimension matrices. These represent latent features of users and items, along with singular values. The implementation of this model uses the Surprise library, which provides an efficient SVD algorithm designed for collaborative filtering.

Initially, the data is preprocessed and divided into training and testing datasets. A random seed is employed to assure reproducibility. The majority of ratings make up the training set, while the test set comprises one randomly selected rating per movie. This approach to splitting allows us to assess the model's performance on data it has not previously encountered.

The SVD model's hyperparameters, such as the number of latent factors, training epochs, learning rate, and regularization term, are then defined. We devise a parameter grid to identify the best combination of these hyperparameters through cross-validation, using the GridSearchCV class. This class evaluates the model's performance using the root mean squared error (RMSE) as the scoring metric.

The GridSearchCV object is fitted to the data and searches for the optimal hyperparameters, determining the best set based on the lowest RMSE. Subsequently, we instantiate a new SVD model with these optimal values.

Training on the complete training set is then carried out with the optimal model, applying stochastic gradient descent (SGD) for a set number of epochs to optimize the model's latent features and biases. Once trained, the model is capable of making predictions.

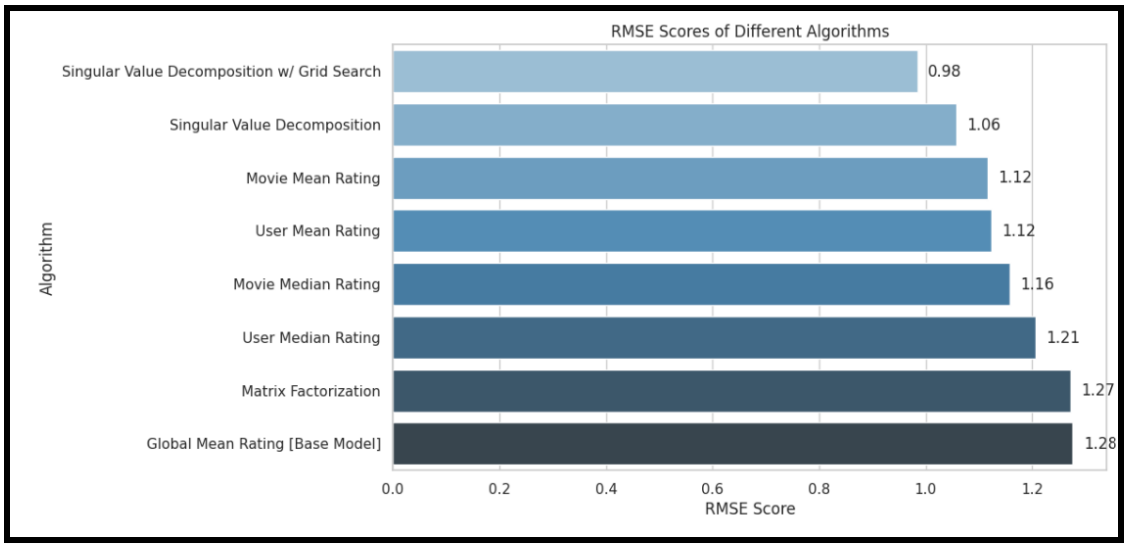
The model's performance is evaluated using the test set, with the RMSE between predicted and actual ratings being calculated. A lower RMSE indicates the model's superior capability to predict movie ratings accurately.

In our experiments, the SVD model yields an RMSE of 0.981 on the test set. This signifies a substantial improvement over baseline models and the matrix factorization model, demonstrating the efficacy of SVD in identifying latent factors and making accurate predictions.

The provided code demonstrates the implementation of the above process, including data preprocessing, model training, hyperparameter tuning, and performance evaluation. It employs the Surprise library, Scikit-learn's preprocessing tools, and NumPy for effective data manipulation and model implementation. It also illustrates the creation of balanced samples from each movie's ratings, ensuring a representative training set, and the use of a more extensive hyperparameter search space for better model optimization.

Model Evaluation and Performance Analysis:

The various models deployed in our research are evaluated using the Root Mean Square Error (RMSE) metric. The following is a comprehensive analysis of the performance metrics for each model:



1. Global Mean Rating [Base Model]: This model yielded an RMSE of 1.277. As the most fundamental model, the Global Mean Rating serves as our baseline for comparing the performance of the more sophisticated models. While it provides a consistent prediction, its lack of personalization leads to a relatively high error rate.
2. User Mean Rating: The User Mean Rating model improved upon the base model with an RMSE of 1.123. By focusing on individual user behavior, this model offers a more tailored prediction, thereby reducing the error rate.

3. User Median Rating: This model reported an RMSE of 1.206. Although slightly less accurate than the User Mean Rating, the User Median Rating still outperforms the base model. It provides a robust measure of central tendency, minimizing the effect of outliers in user ratings.
4. Movie Mean Rating: The Movie Mean Rating model achieved an RMSE of 1.116. By concentrating on individual movie data, this model captures the relative popularity of each film, offering item-specific predictions that outperform user-based models.
5. Movie Median Rating: With an RMSE of 1.158, the Movie Median Rating model performs comparably to its user-based counterpart. Like the User Median Rating, it provides a robust measure of central tendency for each movie's ratings.
6. Matrix Factorization: Despite its sophistication, the Matrix Factorization model registered an RMSE of 1.272, slightly lower than the base model but higher than the mean and median models. This suggests that while the model is effective in capturing latent factors, it might require further refinement or additional data to improve its predictive accuracy.
7. Singular Value Decomposition (SVD): The SVD model produced a substantial improvement, with an RMSE of 1.058. By identifying latent factors and capturing the structure of user-item interactions, this model significantly reduces the error rate.
8. Singular Value Decomposition w/ Grid Search: The most advanced model in our analysis, the SVD with Grid Search, achieved the lowest RMSE of 0.985. This model's superior performance can be attributed to its extensive hyperparameter tuning process, which refines the model to its optimal configuration.

In summary, our analysis indicates that the Singular Value Decomposition model with Grid Search delivers the most accurate predictions, outperforming all other models. However, it's also worth noting that each model contributes a unique perspective to our understanding of user ratings, and a combination of these models might provide further improvements in predictive accuracy.

Further Data Analysis [Extra Credit]:

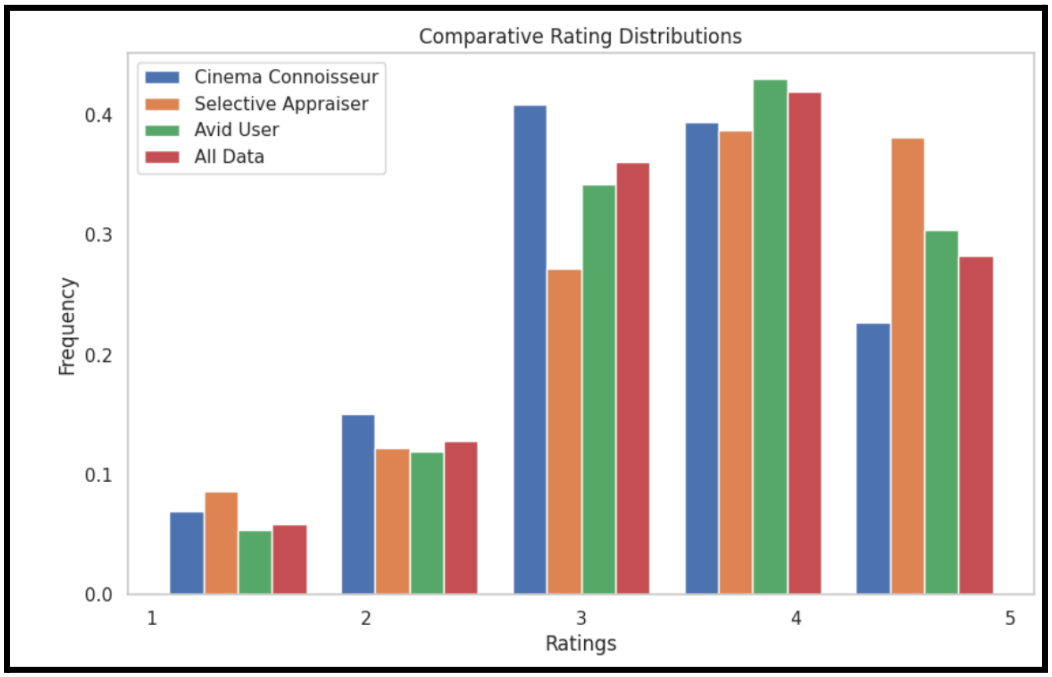
In our deep dive into the data, we've segmented our users into three distinct categories: "Cinema Connoisseurs," "Selective Appraisers," and "Avid Users."

The "Cinema Connoisseurs" represent the top tier of users, identified by their prodigious rating activity. Each member of this elite group has rated more than 213 movies, and they number a total of 23,447 users. Despite being a small fraction of the total user base, their contributions to the total ratings are substantial.

Contrastingly, the "Selective Appraisers" are users with a more reserved rating habit. They represent the lower tier of users, each having rated no more than 2 movies. The total number of users falling into this category is 12,516. Despite their smaller contribution to the total ratings, their behavior and trends provide an interesting contrast to the other groups.

Finally, filling the gap between these two extremes are the "Avid Users". These users, whose rating behavior falls between the 5th and 95th percentiles, have rated between 2 and 213 movies. They represent the largest group, with a total of 421,635 users. Their contribution to the total ratings is significant, and their rating behavior provides a balanced perspective to our analysis.

These categories provide us with a nuanced understanding of our user base and their rating behavior. Understanding these differences helps us to appreciate the diverse ways in which different users engage with and rate movies.



Conclusion

This research project aimed to predict movie ratings based on an extensive dataset, to enhance the efficiency of recommendation systems. The comprehensive dataset was processed, pre-processed, and then explored to understand the distribution and tendencies of movie ratings. Various models were used to predict the ratings, each offering unique perspectives and results.

The baseline models provided a foundation for the more complex models. The Global Mean model offered a consistent prediction but lacked personalization. User and Movie Mean and Median models, focusing on individual user behavior and individual movie data respectively, offered a more personalized prediction, reducing the error rate.

The Matrix Factorization model, although sophisticated, did not outperform the simpler models. However, the Singular Value Decomposition (SVD) model, especially when coupled with Grid Search, significantly reduced the error rate, indicating its superior capability to predict movie ratings accurately.

Finally, the user base was segmented into "Cinema Connoisseurs", "Selective Appraisers", and "Avid Users" to better understand their behavior and contributions to the ratings. These categories provided nuanced insights into the user base and their movie rating habits.

In summary, the most advanced model, the SVD with Grid Search, outperformed all other models, proving to be the most accurate in predicting movie ratings. However, the findings suggest that each model contributes unique insights to our understanding of user ratings and that a combination of these models might further improve predictive accuracy.