MAS 651 - Machine Learning II

Exploring Movie Recommendations: A Comparative Study of User-Based KNN, Item-Based KNN, SVD, and GridSearch CV Algorithms

Andreina Diaz & Andrew Wolfe

Data Source: https://www.kaggle.com/datasets/eswarchandt/amazon-movie-ratings/data

GitHub: https://github.com/awolfe4/ML-Project

Abstract

This report explores the application of algorithms like user-based KNN, item-based KNN, Singular Value Decomposition (SVD), and GridSearch Cross-Validation (GridSearch CV) to construct a sophisticated recommendation system. Using the Amazon dataset obtained from the Kaggle platform, we analyze 4,848 unique user IDs data alongside ratings for 206 distinct movie titles. In this dataset, each row represents a distinct user, while the columns include ratings provided for different movies, with NA values indicating instances where users did not rate specific films.

Our investigation focuses on training, evaluating, and comparing four different models: User-based KNN, Item-based KNN, SVD, and GridSearchCV for hyperparameter tuning. The primary objective of this report is to identify the most effective model based on the Root Mean Square Error (RMSE) metric. By analyzing the RMSE values, our goal is to identify the most accurate and reliable Amazon movie recommendations, ultimately enhancing user satisfaction and engagement.

Introduction

The rise of streaming platforms has transformed user movie consumption, leading us in a new era of convenience and movie accessibility. With the incorporation of recommendation systems, older and lesser-known movies now have the opportunity to reach a bigger audience. Streaming platforms have addressed the challenge that traditional movie stores like Blockbuster faced in promoting older films, this challenge often led to lost sales and profits. Big streaming platforms such as Netflix, Amazon, Peacock, and Max have revolutionized the movie-watching experience by offering individualized movie recommendations tailored to personalized tastes. These recommendation systems not only increase revenue for streaming companies but also provide insights for both users and content creators. By analyzing user data and preferences, companies gain valuable insights into the performance of their content catalog, enabling them to tailor movie offerings to better meet user demand.

Recognizing the importance of recommendation systems, this report analyzes four different Machine Learning algorithms aimed at creating the most effective recommendation system based on Root Mean Square Error (RMSE). The dataset utilized in this study includes user IDs, movie ratings by users, and movie titles. Our primary objective is to develop and evaluate four distinct Machine Learning algorithms to determine the optimal model for recommending Amazon movies.

Through comprehensive analysis, we aim to uncover patterns within the Amazon dataset, showcasing the complexities in movie ratings. Descriptive analytics techniques such as barplots and correlation plots are employed to visualize the relationships between user ratings and movies. The evaluation process consists of assessment based on RMSE and interpretability, providing important insights into the effectiveness and reliability of each recommendation model.

Ultimately, this study aims to create a sophisticated recommendation system capable of delivering personalized movie recommendations that enhance user satisfaction and engagement in this new era of movie consumption.

Descriptive Analysis

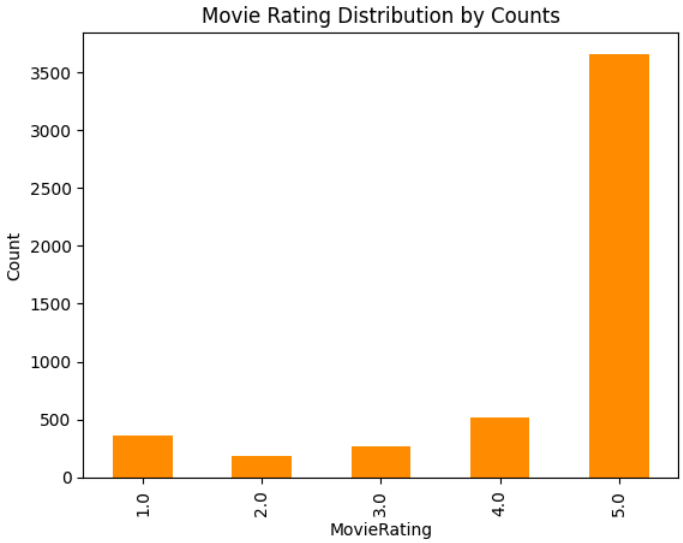

Figure 1. *Showcases the distribution of Movie Ratings by count, revealing that most movies in this dataset are rated at the higher end of the scale.*

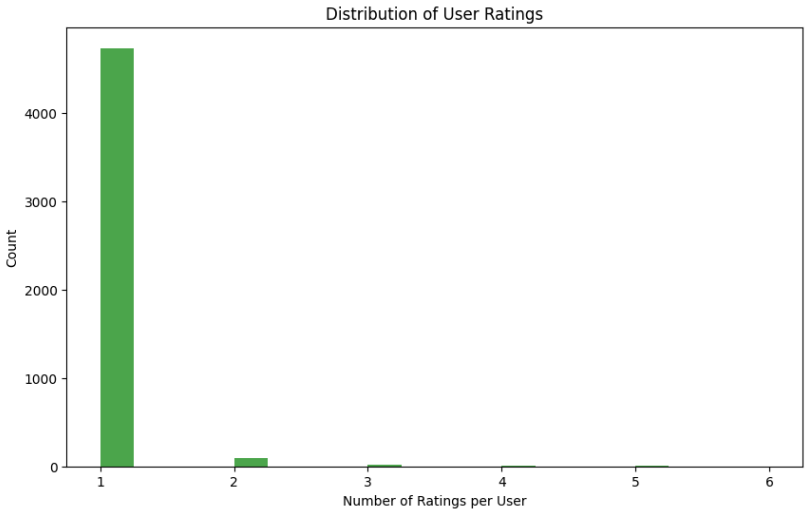

Figure 2. *Depicts the distribution of User Ratings, indicating that a significant portion of viewers in this Amazon dataset have only reviewed 1 movie.*

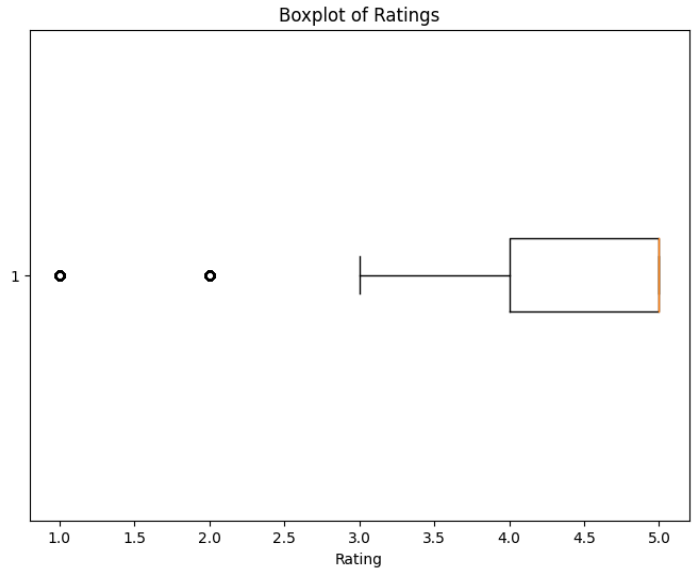

Figure 3. *Illustrates a Boxplot with Skewness Coefficient of -1.913, this reveals that the data is primarily composed of positive reviews, with outliers and lower ratings influencing the plot.*

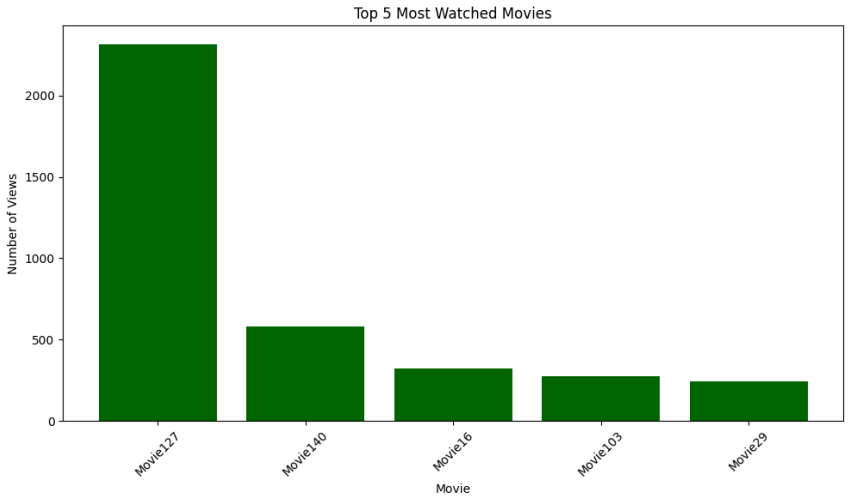

Figure 4. *Showcases the Distribution among the Top 5 Most Watched Movies (Movie127, Movie140, Movie16, Movie103, and Movie29).*

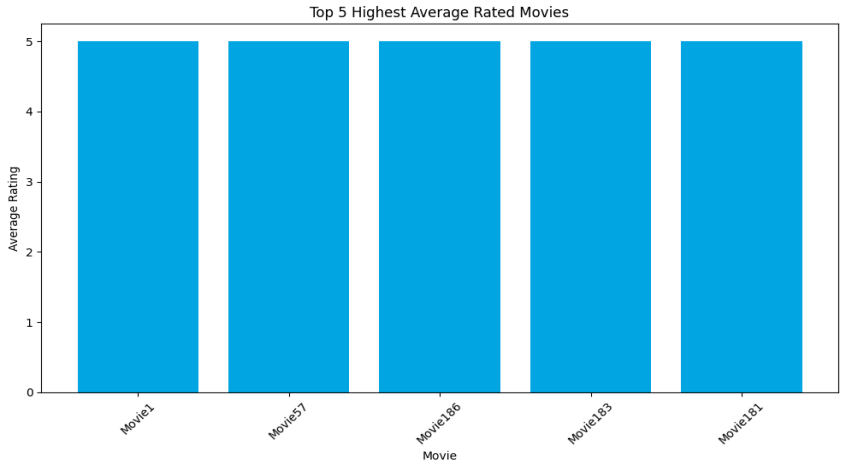

Figure 5. *Depicts the distribution of the top highest Average Rated Amazon movies (Movie1, Movie57, Movie186, Movie183, and Movie181)*

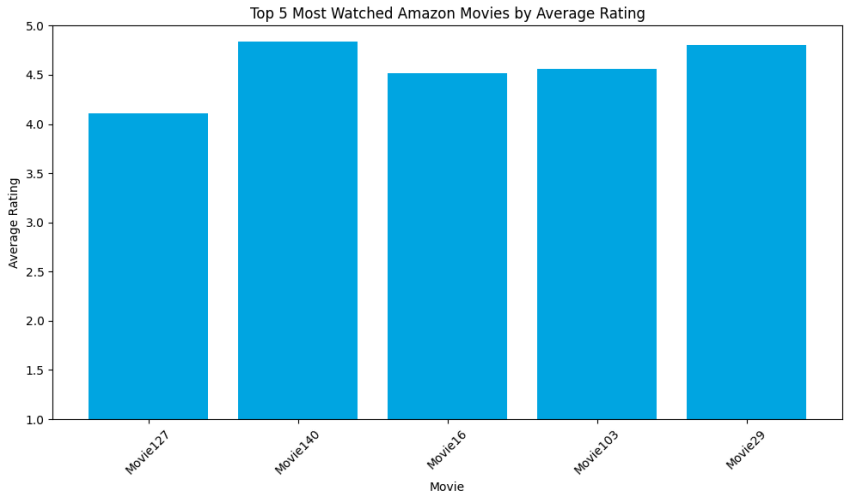

Figure 6. *Illustrates a bar plot showcasing the top 5 Most watched movies by the average rating distribution. These movies appear to*

have high ratings, with 'Movie140' having the highest average rating of 4.833 out of 5.

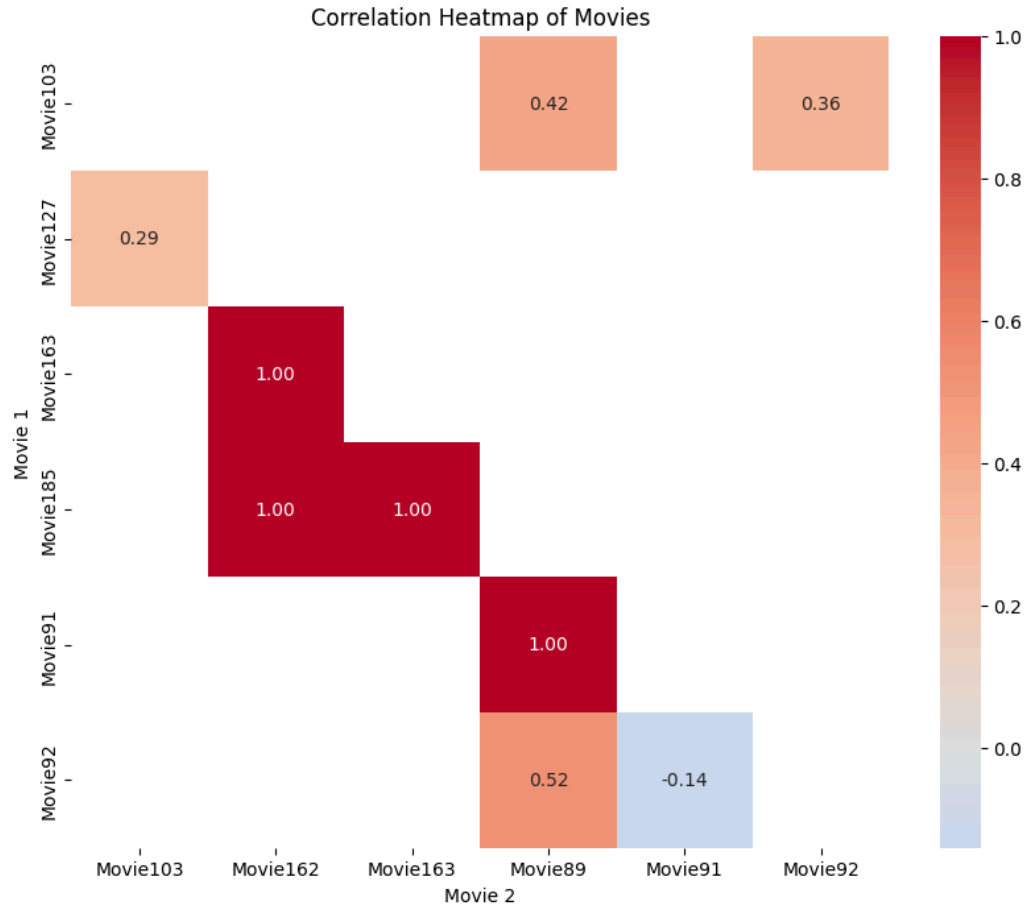

Figure 7. *Illustrates the Correlation Plot, which only includes movies that are correlated. Highlighting a few strong positive correlations, including between Movie91 and Movie89 (1.00), indicating that viewers who enjoy Movie91 tend to also enjoy Movie89. And a Weak negative correlation between Movie92 and Movie91 (-0.14), implying that viewers who like Movie92 may not necessarily enjoy Movie91, and vice versa.*

Preprocessing

Data Cleaning

The data was transformed from a "wider" format to a "longer" format, facilitating better analysis. During this process, a new column named 'rating' was introduced to accommodate the ratings. To handle missing values, which occurred for users who did not rate certain movies, these were filled with zeros. Additionally, to enhance the quality of the dataset, movies rated by fewer than 5 users were filtered out, ensuring a more robust dataset for analysis.

Train Test Split

For all algorithms the dataset was split into training and testing sets using a randomized approach. Specifically, 20% of the data was allocated to the test set while maintaining 80% for training. To ensure reproducibility, a fixed random seed value of 40 was utilized for the randomization process.

Modeling

The recommendation system was developed by selecting and implementing various algorithms, including user-based k-nearest neighbors (KNN), item-based KNN, Singular Value Decomposition (SVD), and GridSearchCV for hyperparameter tuning. These algorithms were chosen based on their effectiveness in collaborative filtering tasks and their ability to provide personalized recommendations.

Evaluation Metrics

RMSE (Root Mean Square Error) was employed as the primary evaluation metric to assess the performance of each recommendation algorithm. This metric was utilized to determine which algorithm yielded the most accurate and reliable recommendation system.

Algorithms

User Based KNN

For the User-Based KNN model utilizing Pearson similarity, we used a KNNBasic algorithm with user-based similarity settings. This model was trained on the provided training data and tested on the test data. The RMSE was calculated and resulted in a value of 0.3198. This RMSE value indicates the average difference between predicted and actual ratings, with lower values suggesting better predictive accuracy. Therefore, the User-Based KNN model with Pearson similarity achieved a moderate level of accuracy in predicting movie ratings for the Amazon dataset.

Item Based KNN

We used an Item-Based KNN model using Pearson similarity by using a KNNBasic algorithm with item-based settings. Following training on the provided dataset, the model was tested on separate test data. The resulting RMSE for this model was 0.4966. A higher RMSE value suggests a greater average disparity between predicted and actual ratings, indicating relatively lower predictive accuracy. Hence, while the Item-Based KNN model with Pearson similarity offers a method for predicting movie ratings, it demonstrates only moderate performance on the Amazon dataset. This appears to be tied for our worst model performance score with GridSearch CV.

Singular Value Decomposition

We implemented a Singular Value Decomposition (SVD) Model. This model is great for matrix factorization and is able to predict missing values using recommendations. The RMSE is: 0.3157, which shows the average error between the actual ratings in the test dataset and the model's predictions. This score is our best model performance as it has the lowest RMSE score, which indicates high accuracy.

GridSearch CV

By using this method, we implemented GridSearchCV, this model was used for hyperparameter tuning and model selection. It uses the predefined list of hyperparameters below to produce the best model performance it can. The RMSE is: 0.4966, which shows the average error between the actual ratings in the test dataset and the model's predictions. This appears to be tied for our worst model performance score with Item Based KNN.

Algorithm Comparison

Algorithm Type	RMSE
User Based KNN	0.3198
Item Based KNN	0.4966
SVD	0.3157
GridSearchCV	0.4966

We evaluated four different recommendation algorithms: User-Based KNN, Item-Based KNN, SVD, and GridSearchCV (for hyperparameter tuning) based on their Root Mean Square Error (RMSE) values. The RMSE measures the average difference between predicted and actual ratings, with lower values indicating better predictive accuracy.

Among the algorithms tested, SVD was the most effective, with an RMSE of 0.3157. This suggests that the SVD model provides more accurate predictions of movie ratings compared to User-Based KNN (RMSE = 0.3198) and Item-Based KNN (RMSE = 0.4966). As mentioned previously, the SVD model is great for matrix factorization and is able to predict missing values using recommendations. This dataset contained a lot of missing values as many users did not rate every movie.

Therefore, for the Amazon movie dataset analyzed in this study, the Singular Value Decomposition (SVD) algorithm is recommended as the preferred choice for constructing a movie recommendation system due to its superior predictive performance.

Conclusion

In conclusion, this reports' goal was to construct a movie recommendation system using the following machine learning models: user-based KNN, item-based KNN, Singular Value Decomposition (SVD), and GridSearch Cross-Validation

(GridSearch CV) for hyperparameter tuning. By using training, testing, and model evaluation, we worked toward having a successful model based on the Root Mean Square Error (RMSE).

Our study recommends the Singular Value Decomposition Algorithm as our choice for constructing a movie recommendation system. The model revealed that the Singular Value Decomposition Model had the best RMSE score of 0.3157, followed by the User Based KNN RMSE score of 0.3198. The GridSearch CV and Item Based KNN had RMSE scores of 0.4966, revealing a much lower predictive accuracy.

Examining user satisfaction and engagement in contemporary streaming contexts is crucial as it directly contributes to business success. Satisfied users/viewers are more inclined to promote the platform to others. This is why building a successful and accurate Amazon movie recommendation is very important. Modern and older movies now have the opportunity to reach a bigger audience due to movie recommendation systems. Finally, recommendation systems, like the one discussed in this report, have significant importance in the field of machine learning. These systems not only enhance user experience but also widen the audience reach for both modern and classic movies. This showcases their pivotal role not only in shaping modern streaming platforms but also in machine learning applications.