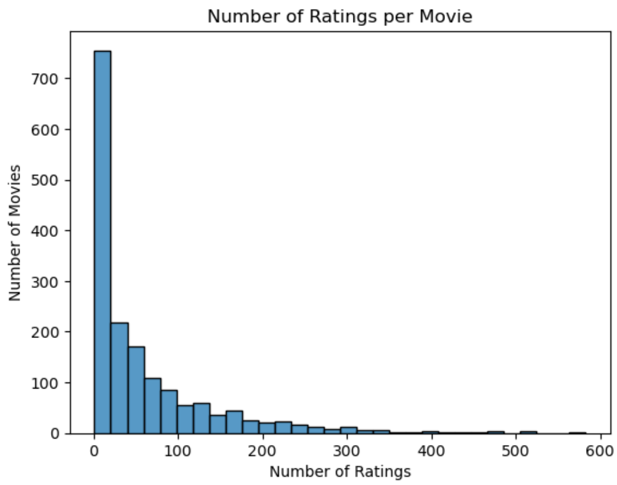
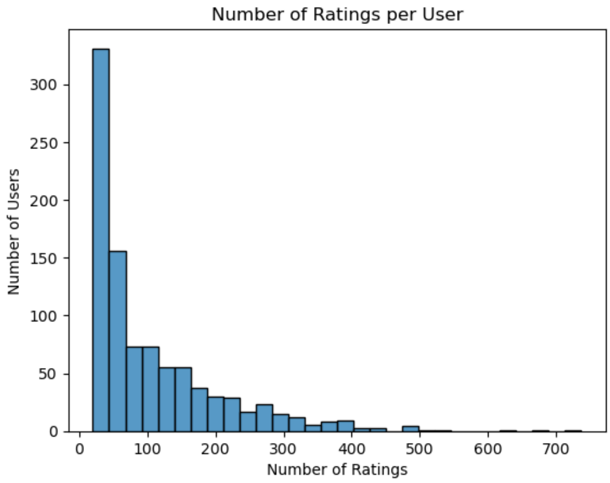
Movie Recommender System Final Report

Introduction:

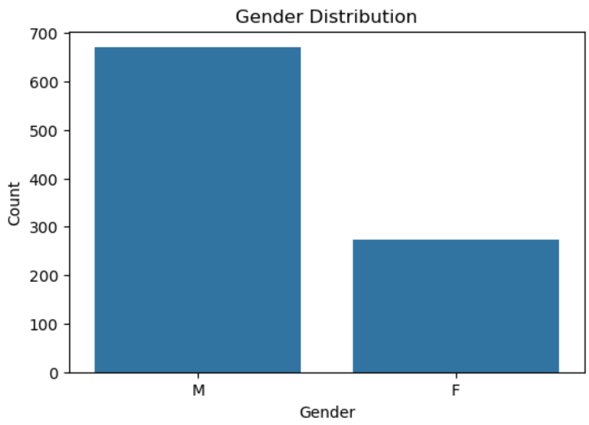
In this report, we present the development and evaluation of a movie recommendation system using collaborative filtering and deep learning techniques. We leveraged the MovieLens 100K dataset to build a personalized movie recommendation system, optimizing hyperparameters with Optuna and evaluating performance using metrics such as RMSE and DCG. This report summarizes our journey from data preprocessing to model training and presents key findings and insights.

Data analysis:



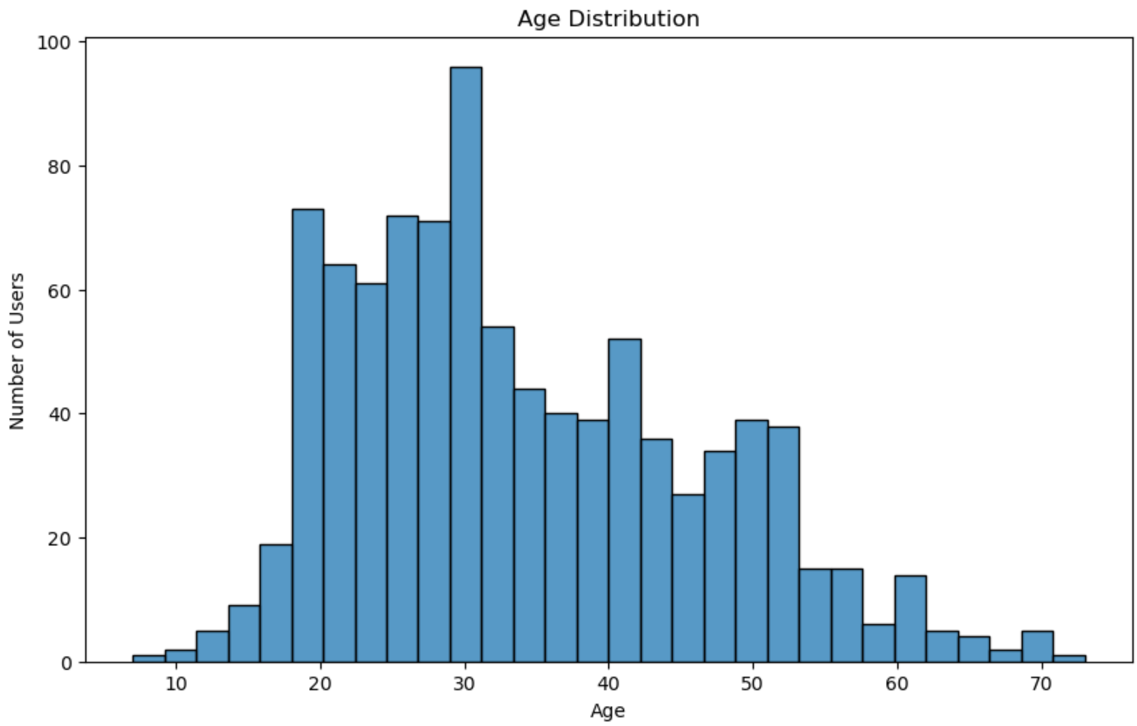


In our comprehensive exploration, we analyzed a movie ratings dataset encompassing 100,000 ratings, contributed by 943 users across 1682 distinct items. This rich dataset provided a detailed lens into the preferences and viewing patterns within a diverse movie-watching community. The analysis revealed a logarithmic distribution in both ratings per user and per movie, suggesting a common trend where few movies and users accumulate most of the ratings.

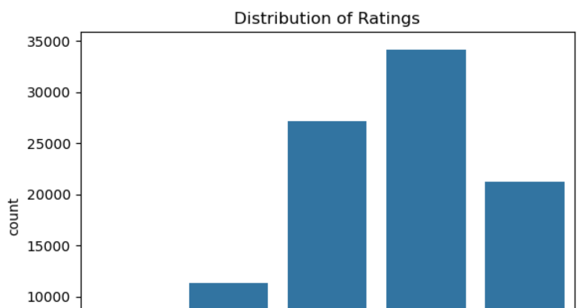


Gender distribution analysis highlights a disparity between the number of male and female users, with males being the predominant group.

The age distribution of users was examined, revealing a mean age of 34 with a standard deviation of 12, indicating moderate variability in the ages of the movie-watching audience. This demographic spread is depicted in the histogram

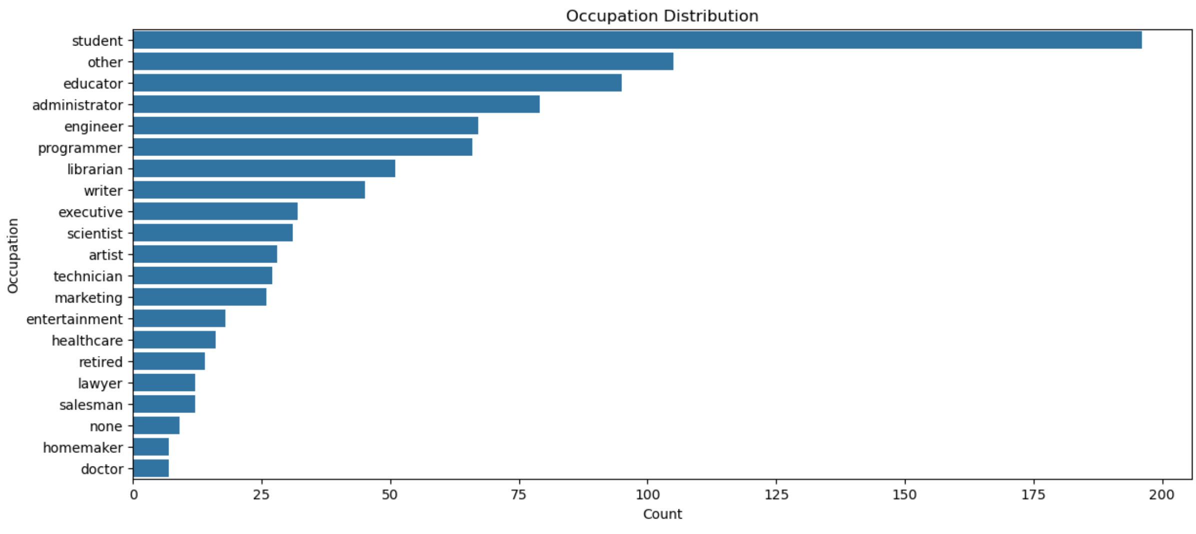


which shows a skew towards younger users, peaking in the 20-30 age bracket.



The evaluation of the ratings data showcased a trend where ratings of 3 and 4 are most common, suggesting a general positivity in movie reception. Contrary to what might be expected, the analysis demonstrated no significant difference in average ratings between genders, indicating a shared level of satisfaction with the movies watched. Furthermore, the preference for genres did not significantly differ by gender, which contradicts common stereotypes regarding gender-based preferences in movie genres.

The exploration of genre preferences across different occupations unveiled distinctive trends, reflecting the diversity of the audience's professional backgrounds. Certain genres resonated more strongly with specific occupational groups, suggesting that movie taste may be influenced by the cultural and intellectual milieus characteristic of one's profession. For instance, individuals in more analytical or technical fields might show a preference for genres that challenge the intellect, such as thrillers or science fiction, while those in creative industries might gravitate towards dramas and independent films that offer nuanced storytelling. These insights highlight the nuanced relationship between occupational identity and entertainment choices.



The "Occupation Distribution" chart reveals that students are the most represented demographic in the movie ratings dataset, followed by a diverse 'other' category and professionals like educators and engineers. This suggests a young audience with varied interests and implies potential for rich, occupation-specific insights into movie preferences.

Model Architecture:

Our Movie Recommender System is built on a neural network model designed to predict movie ratings for users. The model consists of:

- Embedding Layers: Transforming user and movie IDs into dense vectors to capture preferences and characteristics.
- Hidden Layers: Stacked layers for learning complex interactions.
- Activation Functions: ReLU activations introducing non-linearity.
- Output Layer: Predicting movie ratings, optimized using Mean Squared Error (MSE) loss.

Model Training:

Hyperparameters are optimized using Optuna. The specific model shapes and dimensions are determined during training for optimal performance. The model's architecture provides a foundation for making movie recommendations, with fine-tuning through hyperparameter optimization.

Model Advantages and Disadvantages:

Advantages:

- Personalized Recommendations:
  The model offers personalized movie recommendations based on user behavior, ensuring that users receive movie suggestions tailored to their tastes.

- Non-Linear Learning:
  With the inclusion of hidden layers and ReLU activations, the model can capture complex, non-linear patterns in user-movie interactions, improving recommendation quality.

- Embedding Layers:
  The use of embedding layers allows the model to represent users and movies as dense vectors, effectively encoding user preferences and movie characteristics.

Disadvantages:

- Limited to Observed Data:
  Collaborative filtering models, including this one, are constrained by the availability of observed user-movie interactions. Recommendations for niche or unpopular movies may be less accurate.

- Cold-Start Problem:
  The model faces challenges when making recommendations for new users or movies with limited data. It may require additional strategies to address this "cold-start" problem.

Training Process:

In the training process, our model's architecture is determined through hyperparameter optimization, which fine-tunes embedding sizes, hidden layer configurations, and learning rates. We employ the Adam optimizer for efficient parameter updates and minimize the Mean Squared Error (MSE) loss to enhance prediction accuracy. A learning rate scheduler is used to dynamically adjust learning rates during training for better convergence. These techniques collectively ensure the model's optimal performance in recommending movies.

Evaluation:

In this section, we evaluate the performance of our movie recommender system model. The evaluation is conducted using a validation dataset to assess how well the model generalizes to unseen data. We employ two key metrics for evaluation: Root Mean Squared Error (RMSE) and Discounted Cumulative Gain (DCG).

Metrics:
The following metrics are used to evaluate the model:

RMSE (Root Mean Squared Error): RMSE measures the average squared difference between the actual user ratings and the ratings predicted by the model. Lower RMSE values indicate better performance.

DCG (Discounted Cumulative Gain): DCG assesses the ranking quality of recommended movies. It considers both the relevance and the position of recommended items. Higher DCG values indicate better ranking quality.

Results

The evaluation results for our model are as follows:
RMSE: 0.998
DCG: 16.81
These metrics provide insights into the accuracy and ranking quality of our movie recommender system. A lower RMSE indicates that the predicted ratings are closer to the actual ratings, while a higher DCG signifies better recommendations.