Assignment II: Movie Recommendation System

Robert Chen

December 2023

1 Introduction

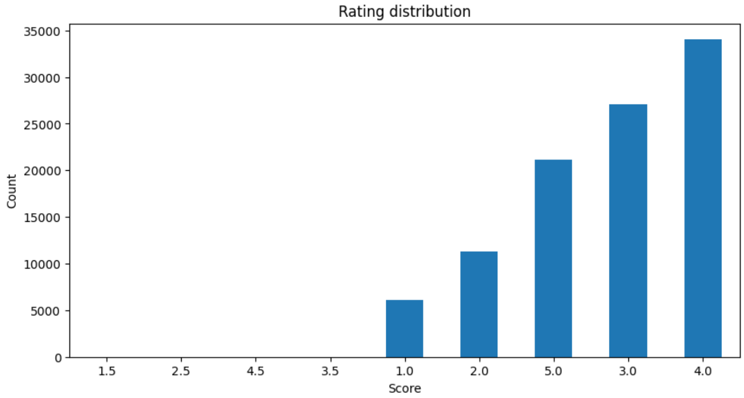
Despite being a very hot topic in the production-deployed AI today, most of the innovative approaches and methods are kept secret by companies. That is why there is a lack of public research on the recommendation algorithms as opposed to other areas like *computer vision* or *NLP*. Currently, the trends are a bit more favored towards the static algorithms, however, some of the machine learning approaches still can show superior performance.

2 Data Analysis

The MovieLens 100K dataset is a widely used version of the MovieLens dataset, particularly known for its manageable size, making it ideal for educational purposes and initial experimentation in recommender systems and machine learning. Short introduction to the dataset:

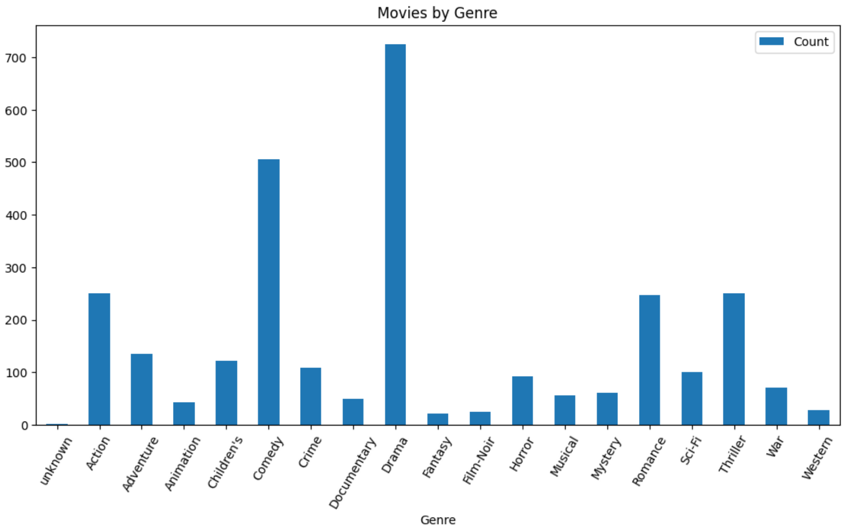
- Consists of 100,000 ratings from 943 users on 1682 movies
- Ratings are ranged from 1 to 5
- Each user has rated at least 20 movies
- Contains demographic info (age, gender, occupation, zip code)

First, we can analyze the rating distribution of the dataset:

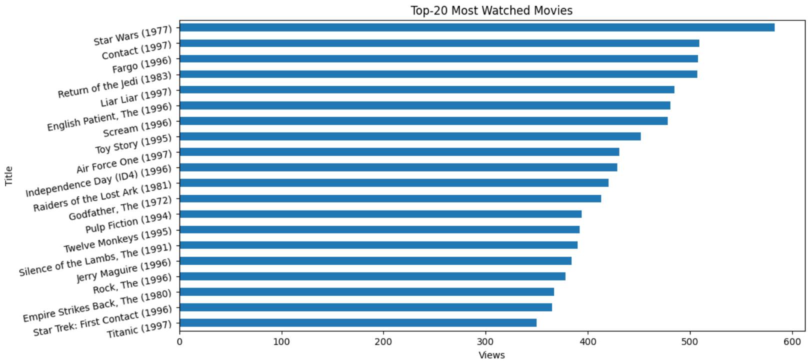


Note that there are additional non-integer ratings in the distribution. This happens because in very rare cases the user could give multiple different ratings for the same movie they watched. In order to resolve this, mean score is taken as a value.

Also, some insight on genre distribution also could help:



As we can observe, **Drama**, **Comedy** and **Action** are top-3 most popular genres across the viewers. And, finally, we can take a look at the most viewed movies in the dataset:



3 Model Implementation

Most current recommendation models rely on either collaborative filtering, such as the Simple Algorithm for Recommendation (SAR), or content-based filtering, like TF-IDF. However, the Wide & Deep model stands out as a hybrid approach. This model integrates the strengths of both *linear models* and *deep neural networks (DNNs)*. Linear models, with their broad feature sets, excel in memorizing feature interactions through co-occurrence. Meanwhile, DNNs specialize in generalizing feature patterns by transforming sparse features into dense, low-dimensional embeddings. The Wide & Deep model harnesses these dual capabilities through joint training, effectively blending memorization and generalization. This fusion results in a powerful tool for recommendation systems.

This notebook covers step-by-step training of the model, and here you can find an object-oriented implementation for the model API.

4 Model Advantages and Disadvantages

Advantages

+ Model is easy to implement

+ Very lightweight and fast in comparison to other ML models (50000 steps in 5 minutes)

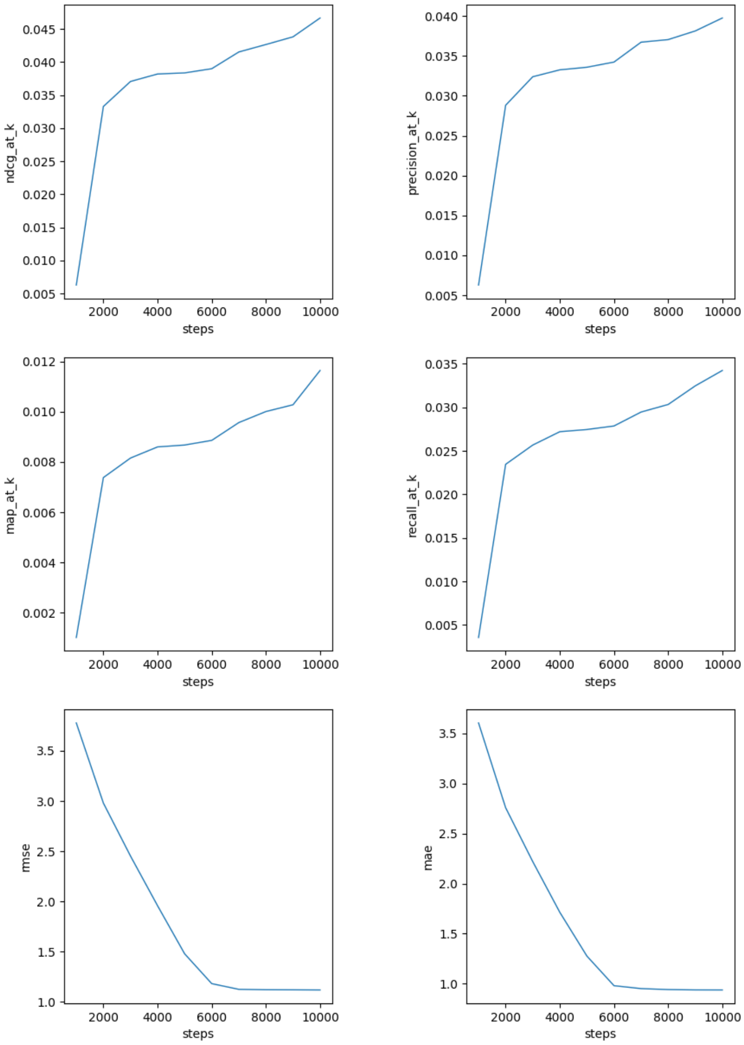
Disadvantages

- Static models can infer the result much faster

- Not the best model in terms of metrics either

5 Training Process

The model is trained on 10000 batches with the size of 32 samples—around 10 epochs in total. The train-test-split ratio of 90%/10% is used to achieve some semblance of convergence faster. The whole process takes around 5-10 minutes. Here is a loss graph for different metrics:



Hyperparameter setup can be found in this file.

6 Evaluation

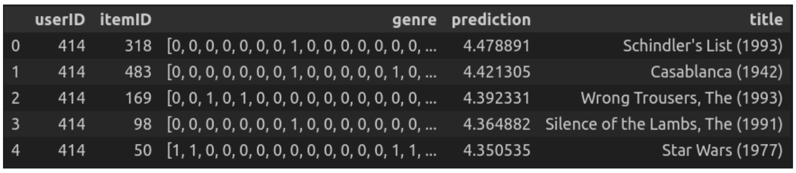
Here are some of the metrics compared to the other models:

Model	K	MAP@K	nDCG@K	RMSE	MAE
SAR	10	0.106959	0.379533	1.229246	1.033912
RBM	10	0.140828	0.41112	1.45538	1.14567
Wide & Deep	10	0.012126	0.048732	1.11828775	0.93635

As we can see, Wide & Deep produces underwhelming results in comparison to static Simple Algorithm for Recommendation (SAR) and Restricted Boltzmann Machine (RBM) in terms of precision, however, it is superior in terms of RMSE and MAE. On top of that, the model has an improvement potential with prolonged training process and thorough hyperparameter tuning.

7 Results

The model produces a `.csv` file with titles and predicted ratings by the model for a specified user:



	userID	itemID	genre	prediction	title
0	414	318	[0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ...	4.478891	Schindler's List (1993)
1	414	483	[0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, ...	4.421305	Casablanca (1942)
2	414	169	[0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	4.392331	Wrong Trousers, The (1993)
3	414	98	[0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ...	4.364882	Silence of the Lambs, The (1991)
4	414	50	[1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, ...	4.350535	Star Wars (1977)