Reproducing Experimental Results from “Everyone’s a Winner! On Hyperparameter Tuning of Recommendation Models”

TU Wien, Experiment Design for Data Science, WS 2024/25

Maximilian Laurent Heß, 12347554

Agnes Sinka, 12412694

Arjol Panci, 12347494

Hasan Berke Bankoglu, 12432802

Muhammad Azeem Shahzad, 12346021

Disclaimer: ChatGPT was used to assist in improving our code and the wording of this report.

1 Introduction

In this report, we document our efforts to reproduce the experimental results from the paper "Everyone’s a Winner! On Hyperparameter Tuning of Recommendation Models" by Faisal Shehzad and Dietmar Jannach, published at RecSys '23. The paper investigates the impact of hyperparameter tuning on the performance of recommendation models and highlights the importance of proper tuning when comparing algorithms. Our goal was to reproduce the experiments described in the paper, verify the reported results, and assess the reproducibility of the study.

2 Experimental Setup

2.1 Paper Overview

The paper uses the Elliot framework to evaluate seven deep learning-based recommendation models and a non-personalized baseline (MostPop) on three datasets: MovieLens-1M (ML-1M), Amazon Digital Music (AMZm), and Epinions. The performance of tuned and non-tuned models is compared with the Normalized Discounted Cumulative Gain (nDCG@10) metric. The authors emphasize the importance of hyperparameter tuning and demonstrate that tuned models perform significantly better than any model with non-tuned hyperparameters. The key takeaway is that improper tuning of baseline models can lead to misleading conclusions about model performances.

2.2 Reproduction Strategy

To reproduce the experiments, we followed these steps:

1. **Setup the Elliot Framework:** We installed the Elliot framework on both personal systems and cloud platforms. The framework already provides implementations of some models (others were provided by the authors) and configurable experiments targeted at evaluating those models both with tuning and without tuning.
2. **Dataset Preparation:** We used the same datasets (ML-1M, AMZm, Epinions) as the ones described and provided by the paper. In later stages of our work, these data sets were reduced due to computational constraints.
3. **Ran the evaluation experiments:** Using the Elliot configuration files provided by the paper’s authors, we attempted to run the same evaluation experiments to test the validity of their results. However, due to computational challenges (discussed in the next section), we had to restrict our experiments.

4. **Statistical Testing:** We performed significance testing on the untuned results to verify the paper's results. This involved running the (untuned) experiments multiple times and performing one-sample t-tests based on the results.

2.3 Challenges Faced

During the reproducibility process, we encountered two main challenges:

- **Runtime and Hardware Constraints:** Due to the computationally intensive nature of training models (both for hyperparameter tuning and untuned configurations) on large datasets, running the experiments on personal computers was highly time-consuming. Utilizing only CPUs, it took several days to complete the untuned training configurations for all seven models, and tuning a single model also required several days, even with a GPU. Additionally, we faced storage issues due to the large volume of files generated (amounting to several gigabytes).
- **Compatibility Issues:** The Elliot Framework has not been updated in recent years, which has led to compatibility issues with modern CUDA drivers. Without clear documentation on compatibility by the authors, we had to spend hours of troubleshooting to downgrade the drivers on both the cloud and one of our laptops. The usage of GPUs significantly shortened our runtimes, however, even with this adjustment, training full configurations still required an excessive amount of time.

3 Experiments

3.1 Untuned Models

To validate the results for the untuned models, we performed multiple runs for each model on each dataset to allow for statistical significance testing. To reduce running time, each dataset was downsampled to contain around 20,000 to 32,000 samples in each training split. To achieve this, the original datasets were downsampled before filtering was applied by the Elliot framework. Since the filtering removed a different number of samples from each dataset, different proportions of the dataset were required. Specifically, 65% of the dataset was used for the Amazon dataset, while 9% was used for MovieLens and 12% for the Epinions dataset.

For the Epinions and MovieLens datasets, each model was trained and tested 30 times, while for the Amazon dataset, only 13 runs were possible due to multiple internal Elliot crashes. Before each run, a new subsample of the original dataset was drawn so that the training and testing data varied for every run. Table 1 shows the results as the mean of all runs for each model-dataset combination.

Table 1: Untuned models' results

Epinions		ML-1M		AMZm	
Model	Mean nDCG@10 (sd)	Model	Mean nDCG@10 (sd)	Model	Mean nDCG@10 (sd)
Mult-DAE	0.01253 (0.00195)	Mult-DAE	0.01726 (0.00260)	Mult-DAE	0.00610 (0.00178)
ONCF (ConvNeuMF)	0.00320 (0.00062)	ONCF (ConvNeuMF)	0.01103 (0.00260)	GMF	0.00438 (0.00050)
GMF	0.00233 (0.00036)	NeuMF	0.01098 (0.00099)	Mult-VAE	0.00402 (0.00092)
NGCF	0.00187 (0.00016)	ConvMF	0.01098 (0.00099)	ONCF (ConvNeuMF)	0.00239 (0.00052)
NeuMF	0.00155 (0.00047)	Mult-VAE	0.01085 (0.00100)	NGCF	0.00180 (0.00026)
Mult-VAE	0.00155 (0.00047)	GMF	0.01033 (0.00077)	NeuMF	0.00105 (0.00013)
ConvMF	0.00155 (0.00047)	NGCF	0.00910 (0.00047)	ConvMF	0.00105 (0.00013)

Although the ranking for the untuned models turned out slightly different for us, the general performance trends are quite similar, with only minor differences. The actual metric values are also quite close, so the slight differences in ranking do not really reflect a big difference in actual performance.

To test whether our results differ significantly, we performed one-sample t-tests for every dataset-model combination. Our null hypothesis is that the sample mean from our runs is equal to the results, which are interpreted as the population mean, while our alternative hypothesis is that they differ.

For all combinations except ConvNeuMF on the Amazon dataset, under a significance level of 5%, our results differed significantly from those in the paper. Note that due to the smaller number of runs with the Amazon dataset, the degrees of freedom were lower (12) compared to the runs for the Epinions and MovieLens datasets (29).

Moreover, the significance test on the Epinions dataset for the models Mult-DAE, ConvNeuMF, and ConvMF was compromised by discrepancies in the parameters specified in the untuned run config file by the author and the results reported in the additional material. For these models, the number of epochs differed (10 epochs in the config file, 5 epochs in the reported results). Since the runs were computationally expensive, we were unable to repeat them with a modified config file.

In conclusion, our results, which are on the one hand statistically significant, suffer on the other hand from the limitation that we used a considerably smaller proportion of data to train and validate the models. Therefore, we are unable to disprove the untuned results from the original paper with confidence.

3.2 Tuned Models

Due to computational limitations, we were unable to run the tuning configuration files fully. As previously mentioned, we reduced the dataset sizes, however, even with smaller data fractions, the hyperparameter tuning process required training thousands of parameter combinations, resulting in extremely long runtimes.

To focus our efforts and verify the paper's key finding - “even the worst-performing tuned model outperforms the best model with non-tuned hyperparameters” - we limited our experiment to tuning only the worst-performing models identified in the original study.

The dataset reduction had a notable impact on our results: both the achieved nDCG@10 scores and the optimal hyperparameters differ from those reported in the paper. Interestingly, for the Amazon dataset (where we used 65% of the original data), our tuned models achieved better performance than those reported by the authors.

By comparing our tuned models to untuned model results, we confirm the paper’s core statement (assuming that the worst-performing models are representative of the broader set). In the case of the Amazon and Epinions datasets, our experiments even outperformed the untuned results from the full datasets, further emphasizing the importance of proper hyperparameter tuning.

Table 2: Tuned models’ results

Dataset	Worst Model Based on Paper	nDCG@10 (Paper)	nDCG@10 (Our Results)	Best Untuned (Paper)	Best Untuned (Ours)
ML-1M	NGCF	0.100	0.027	0.071	0.017
AMZm	ONCF (ConvNeuMF)	0.009	0.012	0.003	0.006
Epinions	NGCF	0.031	0.018	0.015	0.013

3.3 Shortcomings and Inconsistencies

The authors have provided the results for all datasets in a Excel file (*Results with multiple metrics.xlsx)* and the automatically generated result files by Elliot in TSV format, which can be found in the additional material folder of the GitHub repository referenced in the paper. However, we found that some of the values reported in the paper contradict the values in the Excel file.

For example, for the Amazon dataset, the value reported in the paper for the untuned ConvNeuMF (ONCF) is **0.0004**, but in the Excel file, it is **0.0026** (which is significantly higher). Similarly, for the untuned ConvMF, the value reported in the paper is **0.002**, but in the Excel file, it is **0.0004**. We put more trust in the Excel file values when conducting our experiments, as these match the values automatically generated by the Elliot framework.

Another issue in the results is that for certain models, the authors rounded values to four digits, but for others, they rounded to only three. Moreover, for ConvNeuFM, the paper reports **0.005**, but in the Excel file, it is **0.0059**, which should be rounded to **0.006**.

4 Considerations and Outlook

Two specific areas of concern regarding reproducibility are the required software versions and the use of alternative frameworks to obtain the results. A particular point of focus would be the CUDA version that is supported by the Elliot framework as this defines a limit on usable hardware. To make setting up the environment easier, it would also help to explicitly define the TensorFlow version in the requirements.txt file. Providing a more precise specification of dependencies, such as the CUDA version that aligns with the required TensorFlow version, could significantly enhance clarity and simplify the setup process. For example, when working with TensorFlow 2.13.0, CUDA 12.4 and the CUDA 11.8 are compatible.

Additionally, perhaps using a more up-to-date framework that has better support for modern hard- and software could boost both performance and the quality of the reported results. Frameworks like Cornac, Fidelity Mab2rec, FuxiCTR, and Recommenders are some good examples.

5 Conclusions

Although the paper provides sufficient information to reproduce the experiments, the computational demands and compatibility issues make it challenging to fully replicate the results. As a workaround for the hardware resources, we had to reduce the datasets, which may have introduced bias in our results. Addressing those issues through our suggested approaches or by some other means may assist in future reproducibility. Nevertheless, despite those limitations, our findings are generally aligned with the paper's conclusions, reinforcing the importance of hyperparameter tuning in recommendation systems and supporting the paper's thesis: "The best untuned model performs worse than the worst turned models for all the datasets."

References

Faisal Shehzad and Dietmar Jannach. 2023. Everyone's a Winner! On Hyperparameter Tuning of Recommendation Models. In Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3604915.3609488