Replication Study: Developing and Evaluating a University Recommender System

Max Enderlein and Choudhry Rafay

May 4, 2026

1 Abstract

This project replicates and extends the offline evaluation from the paper "Developing and Evaluating a University Recommender System" by Elahi et al. The original paper compared collaborative filtering models using user ratings of universities. Since the original dataset was not publicly available, we created our own dataset using real university ranking data and simulated user ratings.

We replicated the evaluation using SVD, KNN Basic, and KNN with baselines, using 5 fold cross validation and RMSE. We also extended the study by adding MAE and R squared, implementing a neural network model, performing hyperparameter testing, and evaluating on a second dataset.

Our results differ from the original paper. The paper found SVD to be best, while we found KNN with baselines to perform best on our dataset. We explain this difference based on dataset structure and show that model performance depends heavily on the data.

2 Introduction

Choosing a university is a difficult decision because students must compare many options across many factors. General university rankings are not personalized and do not reflect individual preferences.

The original paper proposes a recommender system that uses user ratings to generate personalized university rankings. Instead of using a one size fits all ranking, the system learns what each user prefers.

In this project, we replicate the offline evaluation from the paper. Since the original dataset was not available, we created our own dataset using CWUR university rankings and simulated ratings. We also extend the work by adding additional experiments and analysis.

3 Original Paper Summary

The original paper collected ratings from 80 users across 551 universities. Ratings were given on a scale from 0 to 100. The authors evaluated several recommender models using 5 fold cross validation and RMSE.

They tested multiple algorithms, but the main ones were: SVD, KNN Basic, and KNN with baselines.

Their results showed that SVD had the lowest RMSE and performed best. KNN with baselines was second, and KNN Basic performed worse.

4 Methodology

4.1 Dataset

Since the original dataset was not available, we created a new dataset using the CWUR university rankings dataset. We selected the top 100 universities from the most recent year.

We simulated 80 users. Each user rated 20 universities, giving a total of 1600 ratings. This is similar in size to the original dataset.

4.2 Rating Simulation

Ratings were generated based on university rank and score, along with user specific preferences and randomness.

Each user had: a different weight for prestige, a different bias in ratings, and different randomness.

The general idea was:

rating = weighted combination of rank and score + user bias + random noise

Ratings were limited to the range 0 to 100.

4.3 Models

SVD: The SVD model uses matrix factorization. It predicts ratings using: global average + user bias + item bias + interaction between user and item factors.

KNN Basic: KNN Basic finds similar users using cosine similarity and predicts ratings using the average of neighbors.

KNN with baselines: This model adds baseline estimates using global average, user average, and item average, then combines this with neighbor predictions.

Neural Network: We added a neural network model using user and item embeddings passed through fully connected layers.

4.4 Evaluation

We used 5 fold cross validation and RMSE.

RMSE measures prediction error as: square root of the average squared difference between predicted and actual ratings.

We also used: MAE, which measures average absolute error, and R squared, which measures how well predictions explain variance.

5 Replication Results

Algorithm	Mean RMSE	Std RMSE	Min RMSE	Max RMSE
KNN Baseline	20.80	0.66	19.82	21.83
KNN Basic	21.96	0.76	20.80	23.09
SVD	24.59	1.15	22.66	26.28

Table 1: Replication results on simulated university dataset

KNN with baselines performed best, followed by KNN Basic, and then SVD.

6 Why Results Differ From the Original Paper

The original paper found SVD to perform best, but our results show KNN with baselines performs best.

There are several reasons for this.

First, our dataset is simulated. Ratings are based on rank and score, so they are more structured.

Second, users in our dataset are more similar. This makes it easier for KNN to find similar users.

Third, KNN performs well when user behavior is consistent.

Fourth, SVD performs better when there are complex hidden patterns in data. Real users likely have more complex preferences than our simulation captures.

Finally, the original implementation details are not fully available, so our SVD implementation may differ slightly.

This shows that model performance depends strongly on the dataset.

7 Extension 1: Additional Metrics

Algorithm	RMSE	MAE	R squared
KNN Baseline	20.80	16.22	0.549
KNN Basic	21.96	17.34	0.497
Neural Network	22.14	17.20	0.489
SVD	24.59	19.53	0.369

Table 2: Extended metrics results

The ranking remains the same across all metrics. KNN Baseline performs best, and SVD performs worst.

8 Extension 2: Neural Network

The neural network performed between KNN Basic and SVD. It did not outperform KNN Baseline.

This suggests that more complex models do not necessarily improve performance, especially with smaller datasets.

9 Extension 3: Hyperparameter Testing

We tested different values of k for KNN and different factor sizes for SVD.

KNN Baseline performed best at k = 60, but results were stable across values.

KNN Basic improved with larger k.

SVD improved with more factors, but still did not outperform KNN.

The neural network performed best with smaller architectures.

10 Extension 4: GameQueue Dataset

We tested all models on a second dataset from GameQueue. The dataset contained users and their ratings for video games.

Algorithm	RMSE	MAE	R squared
SVD	15.63	12.57	-0.005
Neural Network	15.70	12.46	-0.015
KNN Baseline	16.82	12.71	-0.174
KNN Basic	17.53	13.40	-0.273

Table 3: GameQueue dataset results

On this dataset, SVD performed best. This is different from the university dataset, we actually had real users ratings.

R squared values were negative, meaning all models performed worse than predicting the average.

This shows that dataset size and structure affect performance. This dataset is from an early access version with very limited users and less accurate ratings. So the results were not too surprising.

11 Conclusion

We replicated the offline evaluation from the original paper and extended it with additional experiments.

We did not reproduce the exact result from the paper. Instead of SVD, KNN with baselines performed best on our dataset.

However, our extensions show that this result depends on the dataset. On a different dataset, SVD performed best again.

The main conclusion is that recommender system performance depends heavily on the data, and results from one dataset may not generalize to others.

References

[1] Elahi et al. Developing and Evaluating a University Recommender System.

[2] CWUR Dataset.