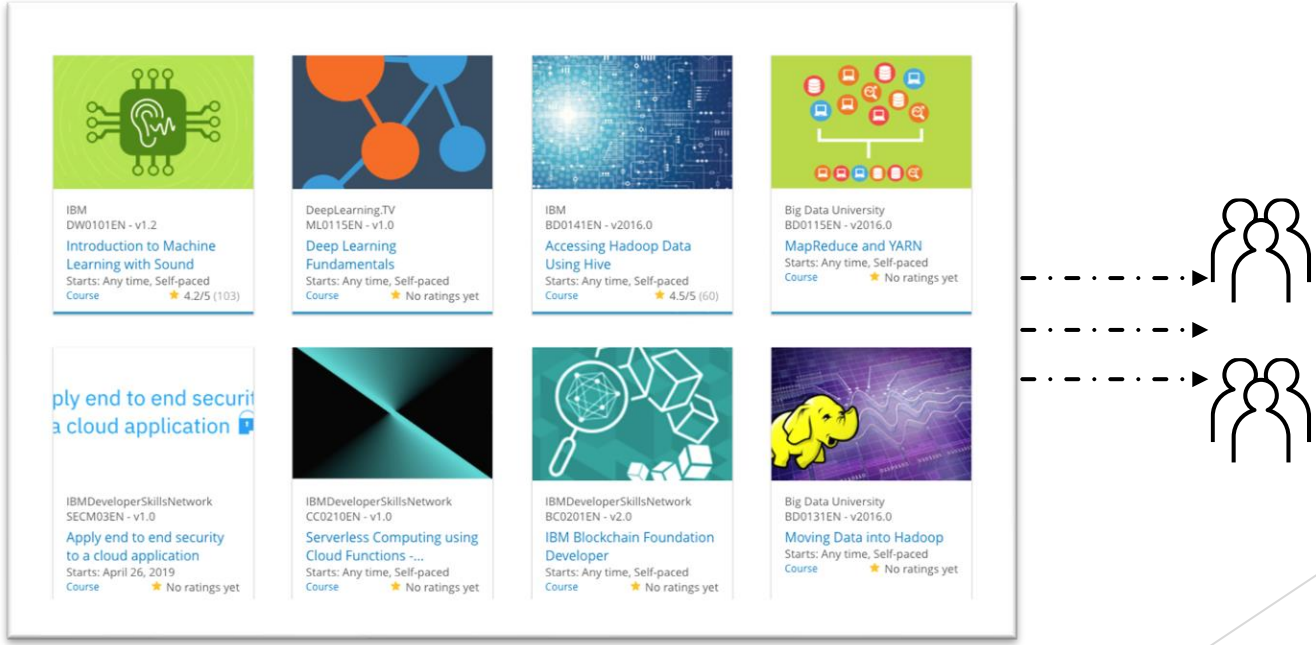
Build a Personalized Online Course Recommender System with Machine Learning

Young Rha
2023-12-18



Outline

- Introduction and Background
- Exploratory Data Analysis
- Content-based Recommender System using Unsupervised Learning
- Collaborative-filtering based Recommender System using Supervised learning
- Conclusion
- Appendix

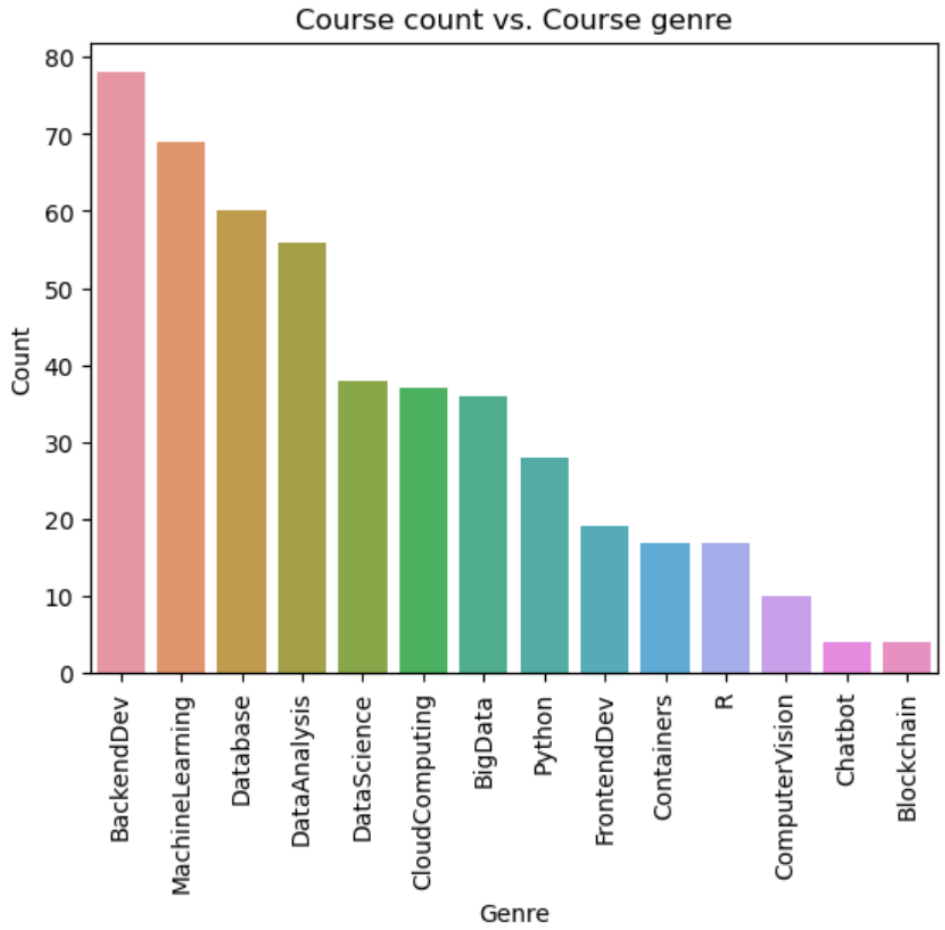
Introduction

- Goal
  - This project aims to explore recommendation system through analyzing online courses dataset. Both content based and collaborative filtering methods will be used to measure the performance between the various methods.

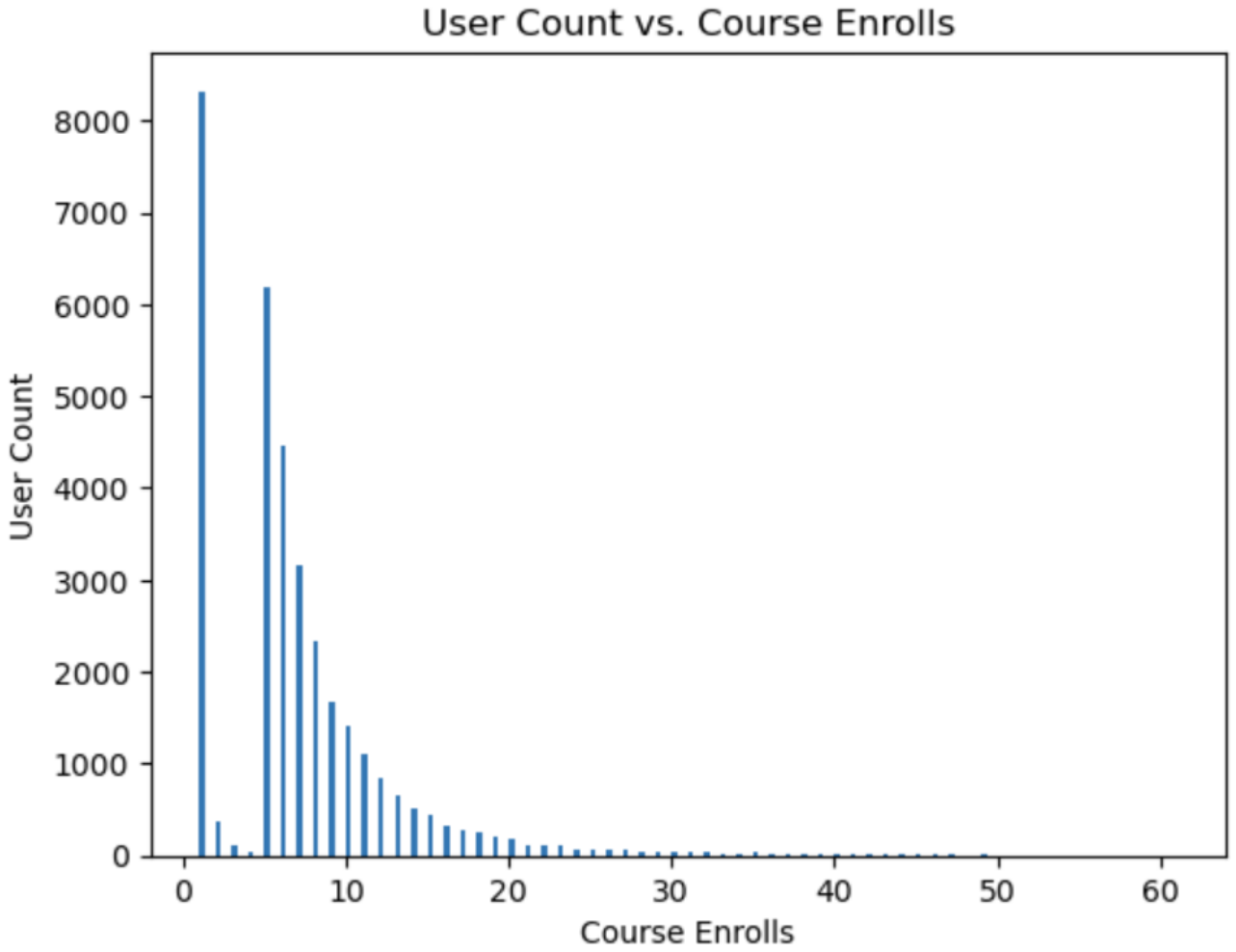
- Problem:
  - The online courses dataset consist of courses data and user profile data. The recommendation system aims to generate course recommendations to the user where the likelihood of accepted recommendation is optimized.
  - The recommendation system will assume several hypotheses such as:
    - A user will likely accept courses similar to courses they have taken in the past
    - Users with the similar interest are likely to accept courses taken by other users with the group

Exploratory Data Analysis

Course counts per genre



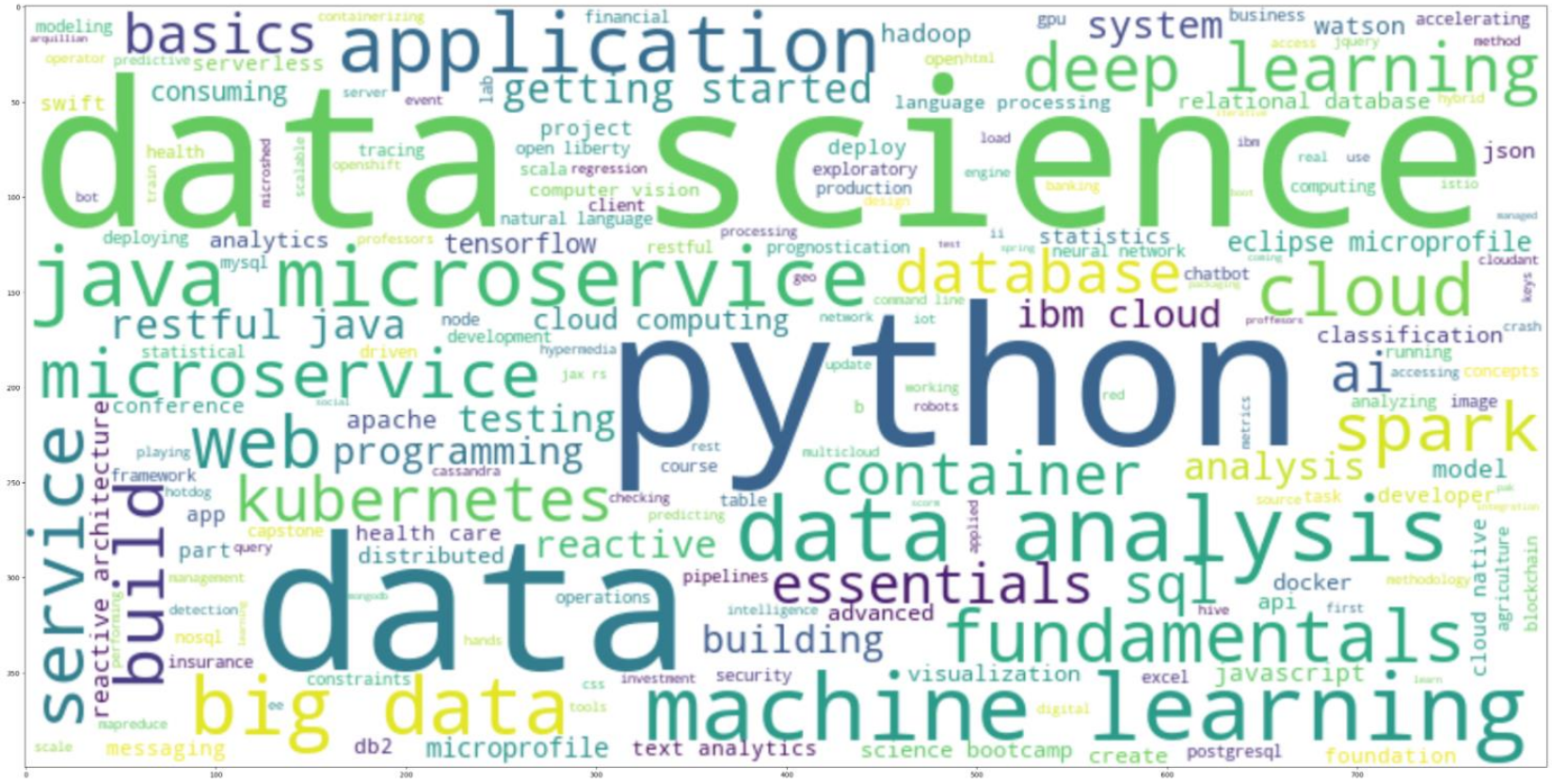
Course enrollment distribution



20 most popular courses

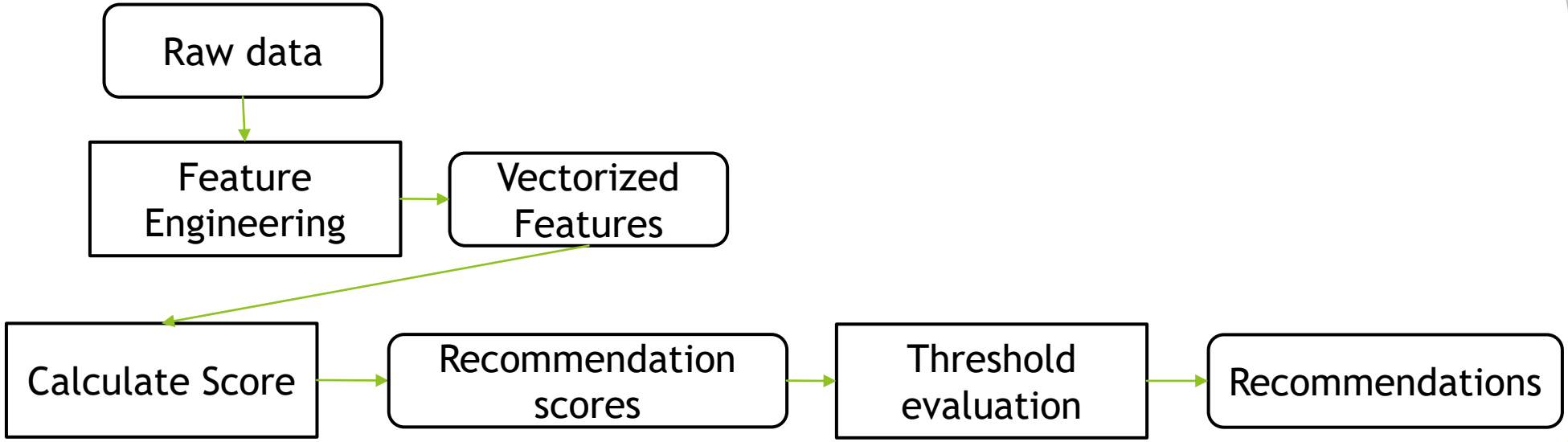
	TITLE	Enrolls
0	python for data science	14936
1	introduction to data science	14477
2	big data 101	13291
3	hadoop 101	10599
4	data analysis with python	8303
5	data science methodology	7719
6	machine learning with python	7644
7	spark fundamentals i	7551
8	data science hands on with open source tools	7199
9	blockchain essentials	6719
10	data visualization with python	6709
11	deep learning 101	6323
12	build your own chatbot	5512
13	r for data science	5237
14	statistics 101	5015
15	introduction to cloud	4983
16	docker essentials a developer introduction	4480
17	sql and relational databases 101	3697
18	mapreduce and yarn	3670
19	data privacy fundamentals	3624

Word cloud of course titles

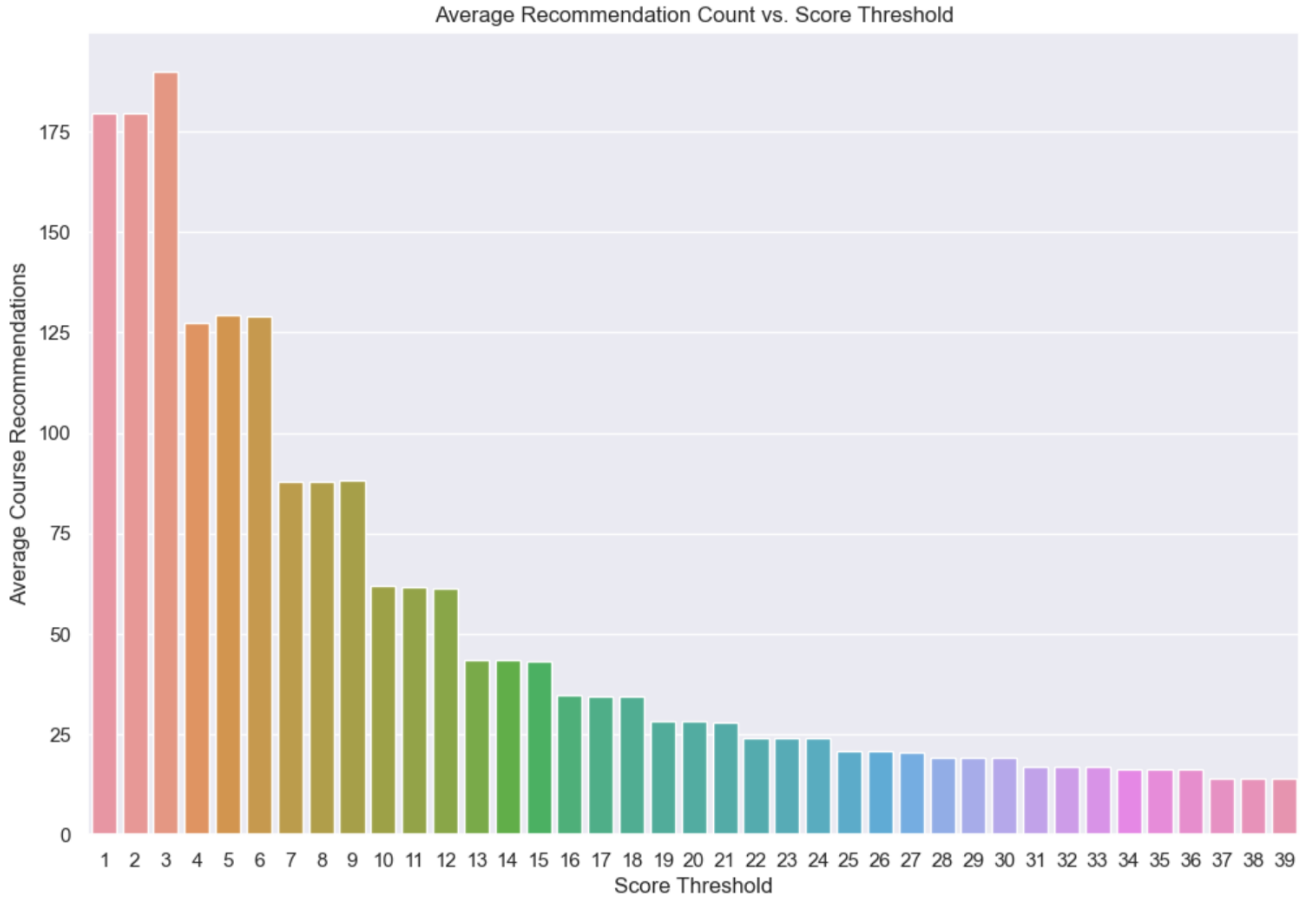


Content-based Recommender System using Unsupervised Learning

Flowchart of content-based recommender system using user profile and course genres



Evaluation results of user profile-based recommender system

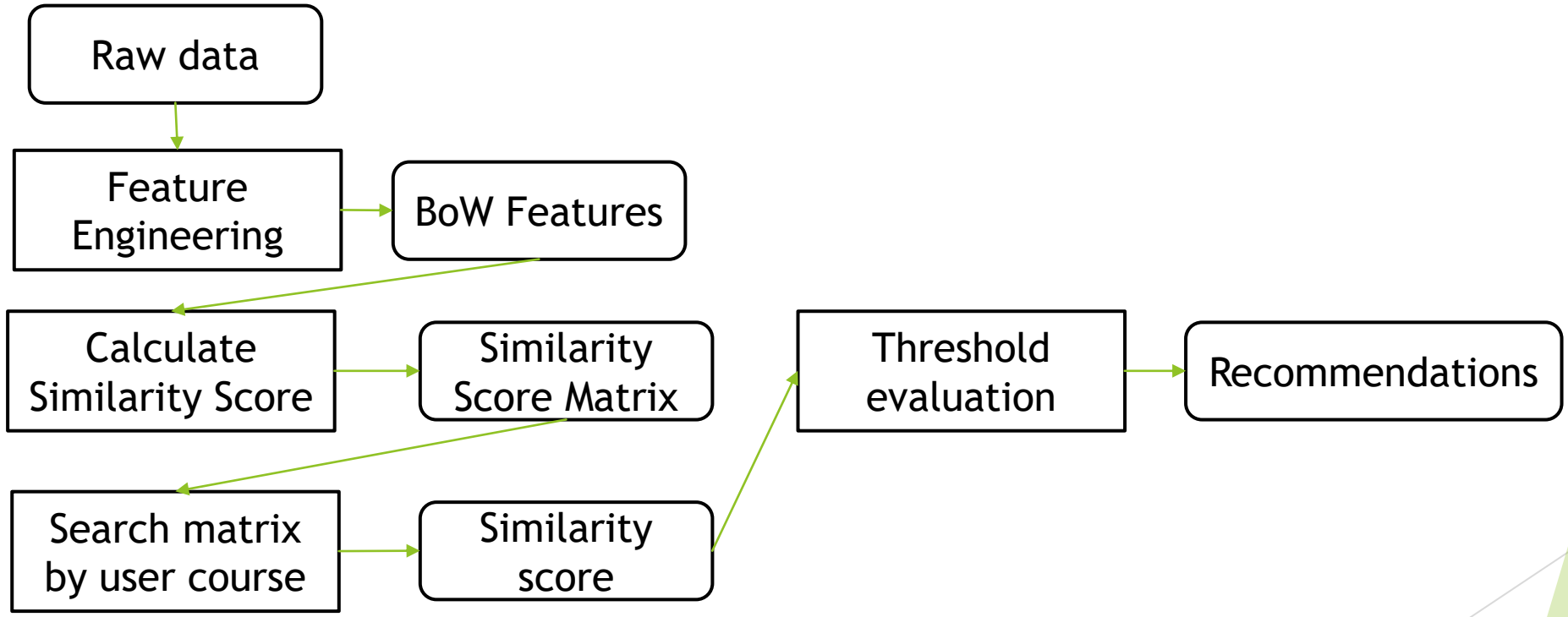


Top 10 courses

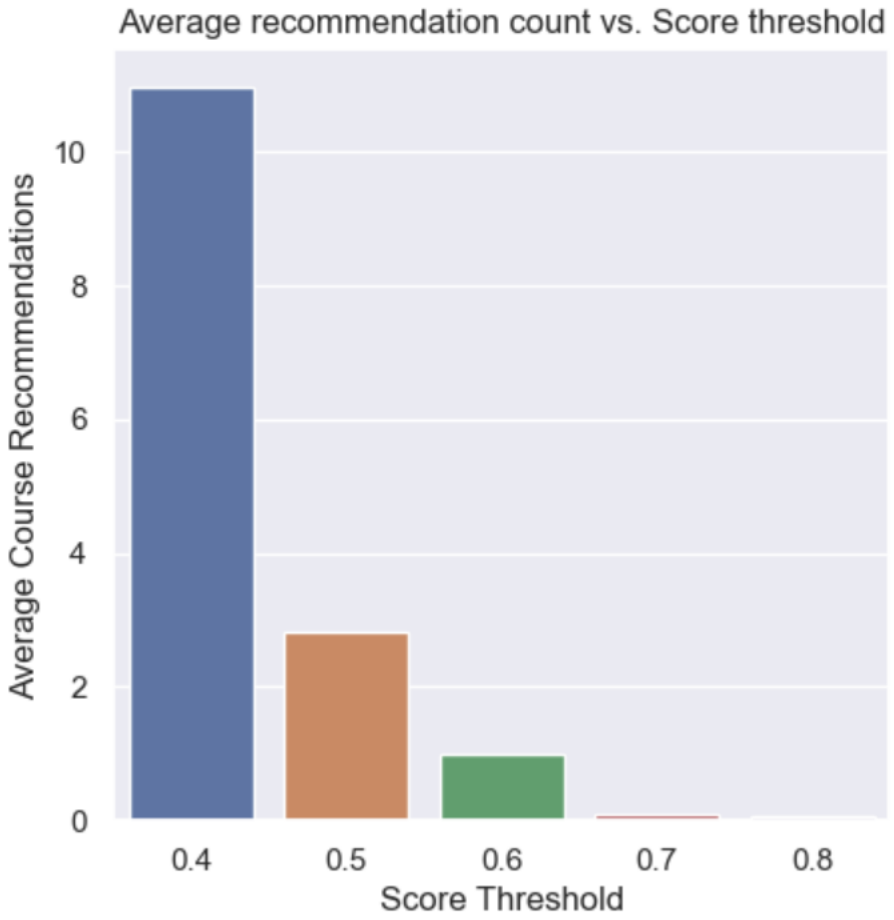
threshold: 10

COURSE_ID	count
TA0106EN	608
GPXX0IBEN	548
excourse22	547
excourse21	547
ML0122EN	544
GPXX0TY1EN	533
excourse04	533
excourse06	533
excourse31	524
excourse73	516

Flowchart of content-based recommender system using course similarity



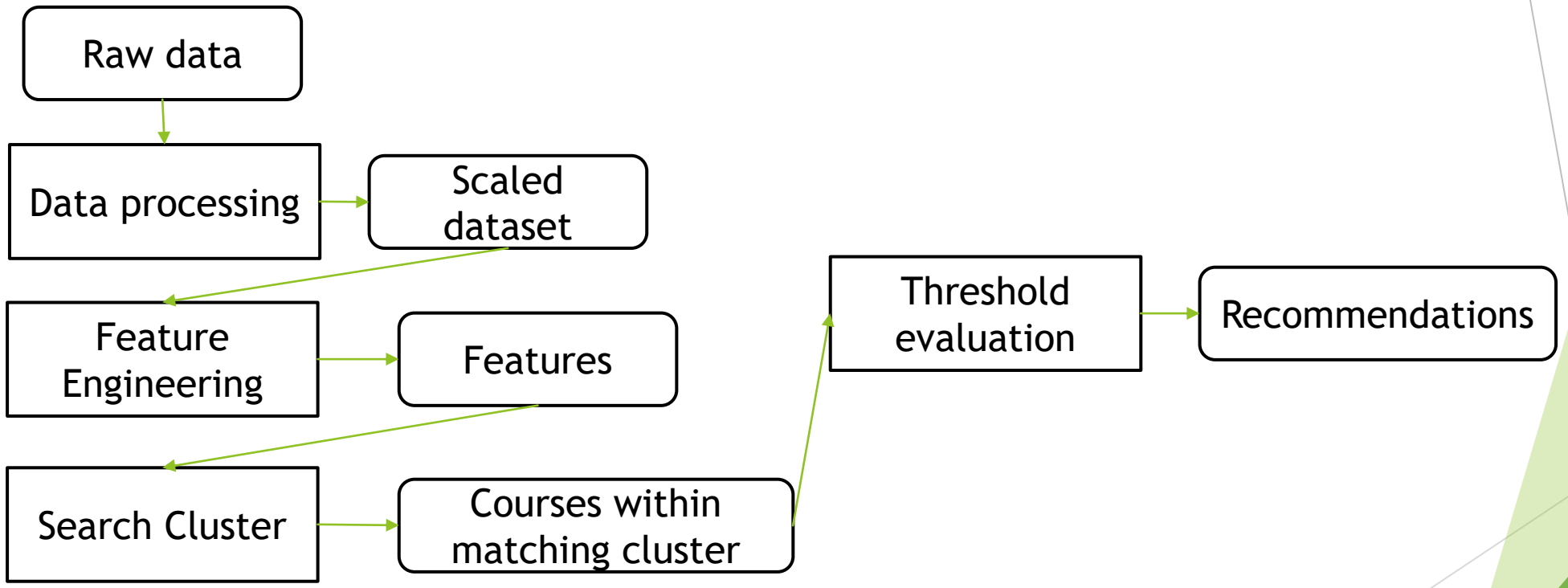
Evaluation results of course similarity based recommender system



Top 10 courses threshold: 0.6

COURSE_ID	count
excourse62	257
excourse22	257
WA0103EN	101
TA0105	41
DS0110EN	38
excourse47	24
excourse46	24
excourse65	23
excourse63	23
TMP0101EN	17

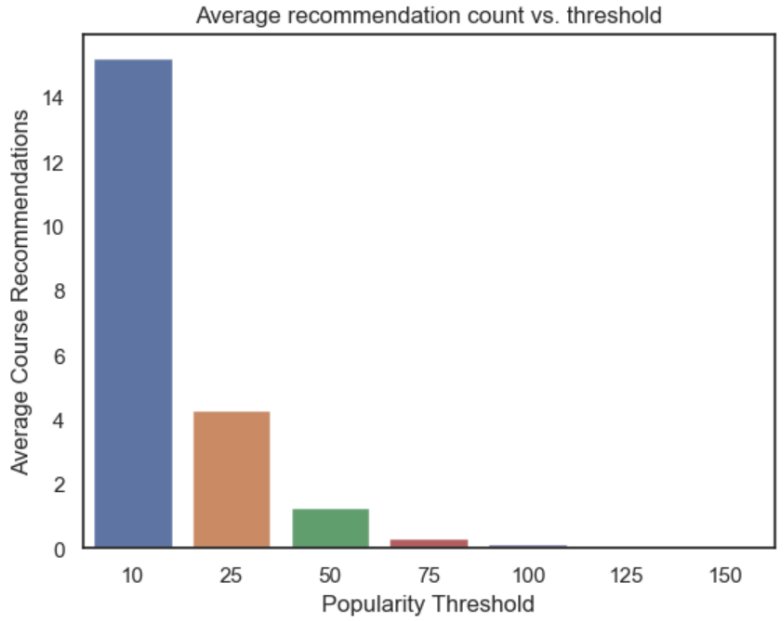
Flowchart of clustering-based recommender system



Evaluation results of clustering-based recommender system

K-Means: n_clusters = 15

PCA: n_components: 9



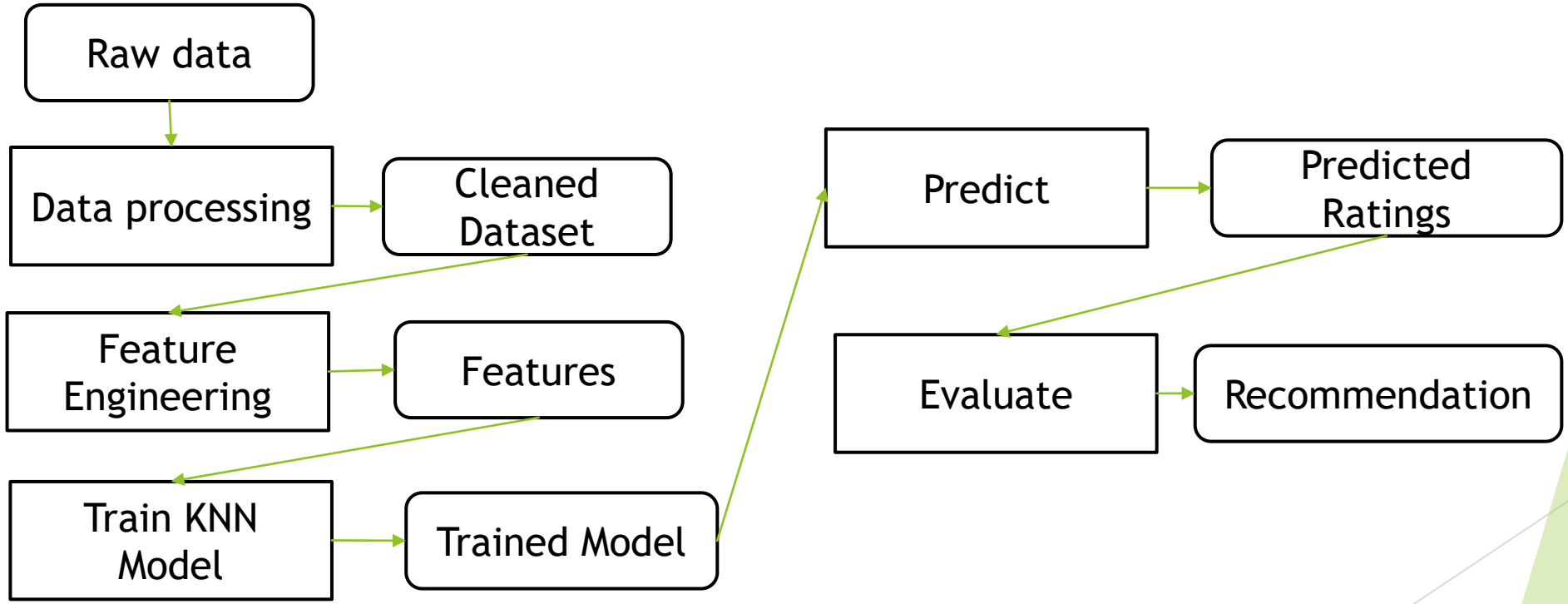
Top 10 courses

Threshold: 50

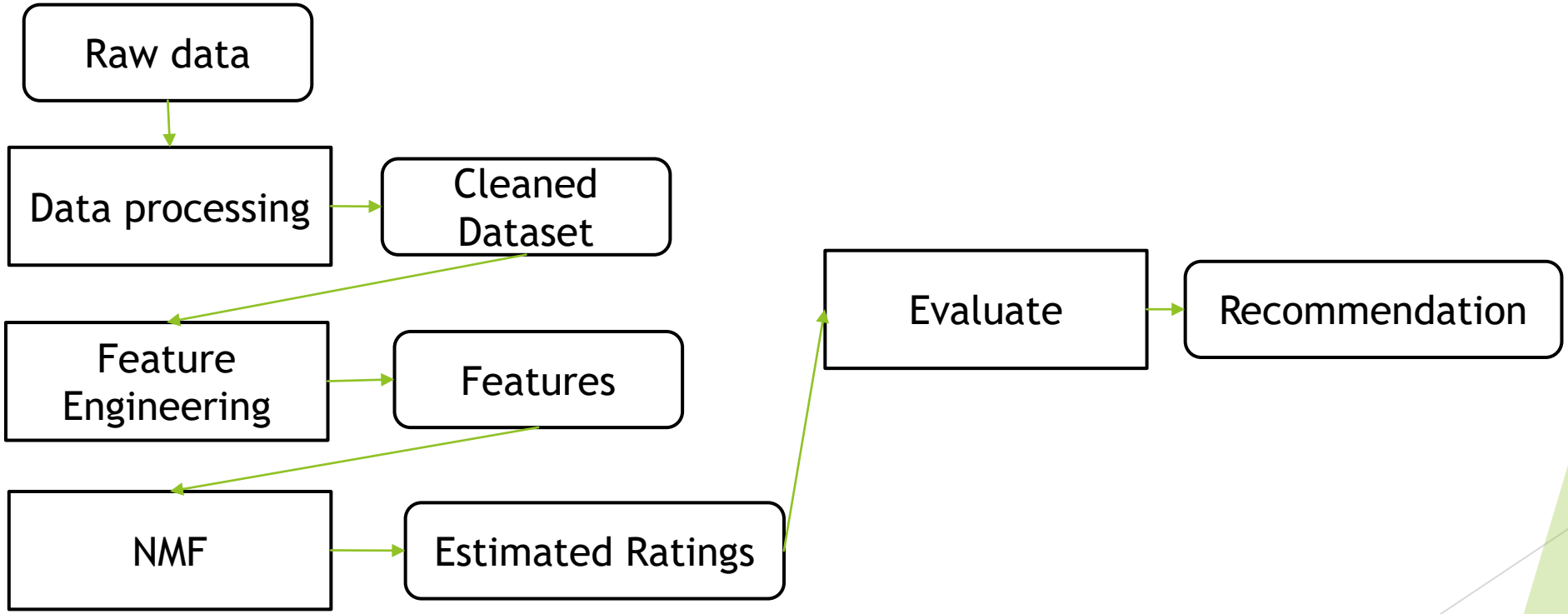
rec	count
DS0101EN	2326
BD0101EN	1958
PY0101EN	1123
DS0105EN	789
ML0115EN	762
DS0103EN	760
BD0115EN	594
DV0101EN	503
ML0101ENv3	464
DA0101EN	331

Collaborative-filtering Recommender System using Supervised Learning

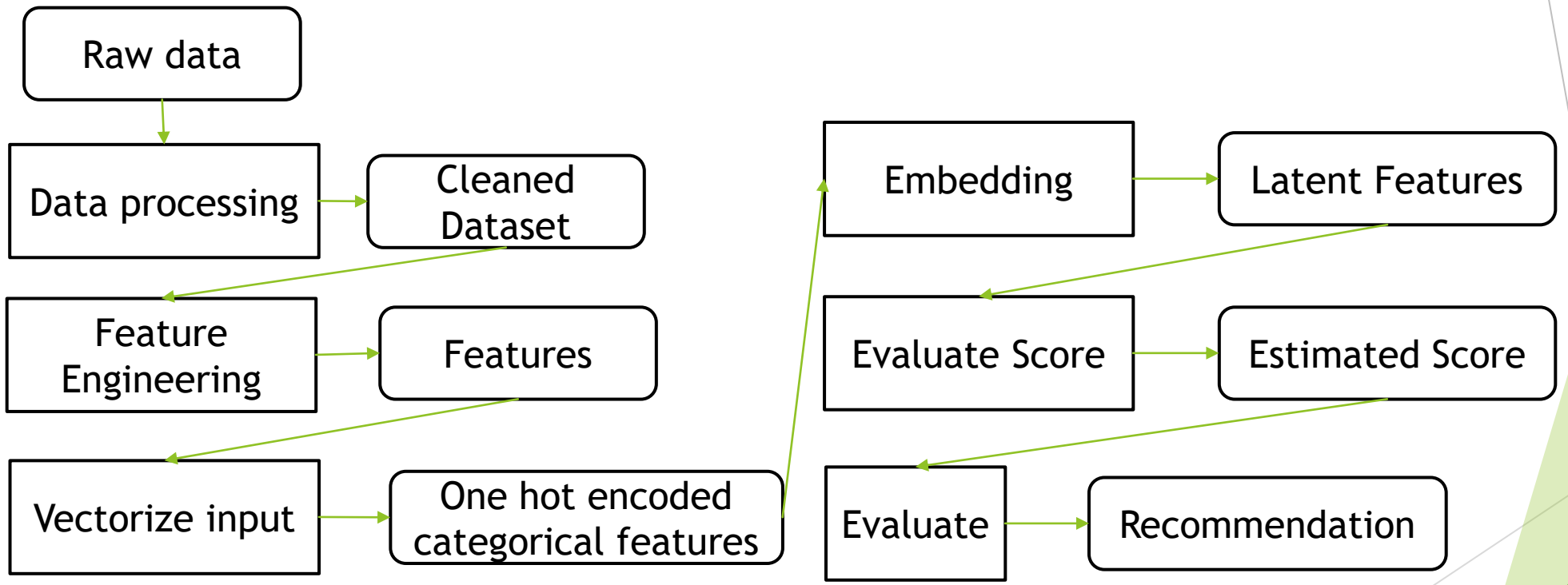
Flowchart of KNN based recommender system



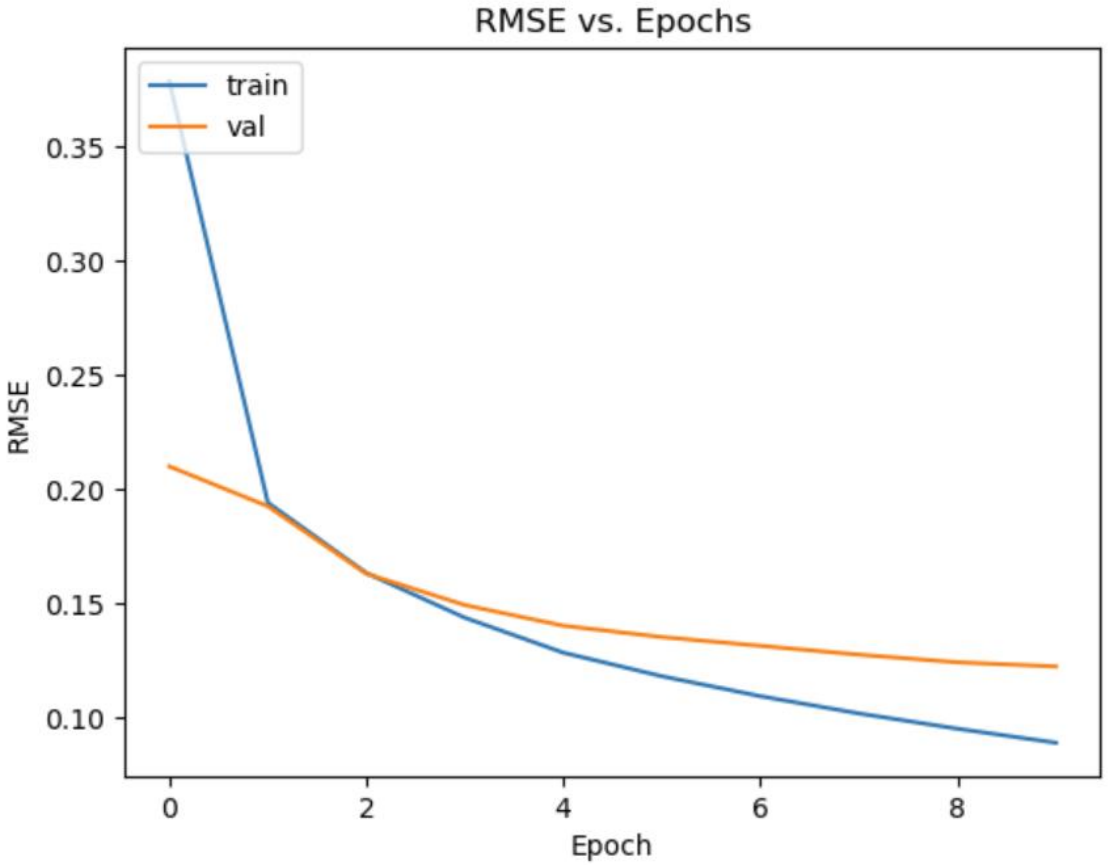
Flowchart of NMF based recommender system



Flowchart of Neural Network Embedding based recommender system



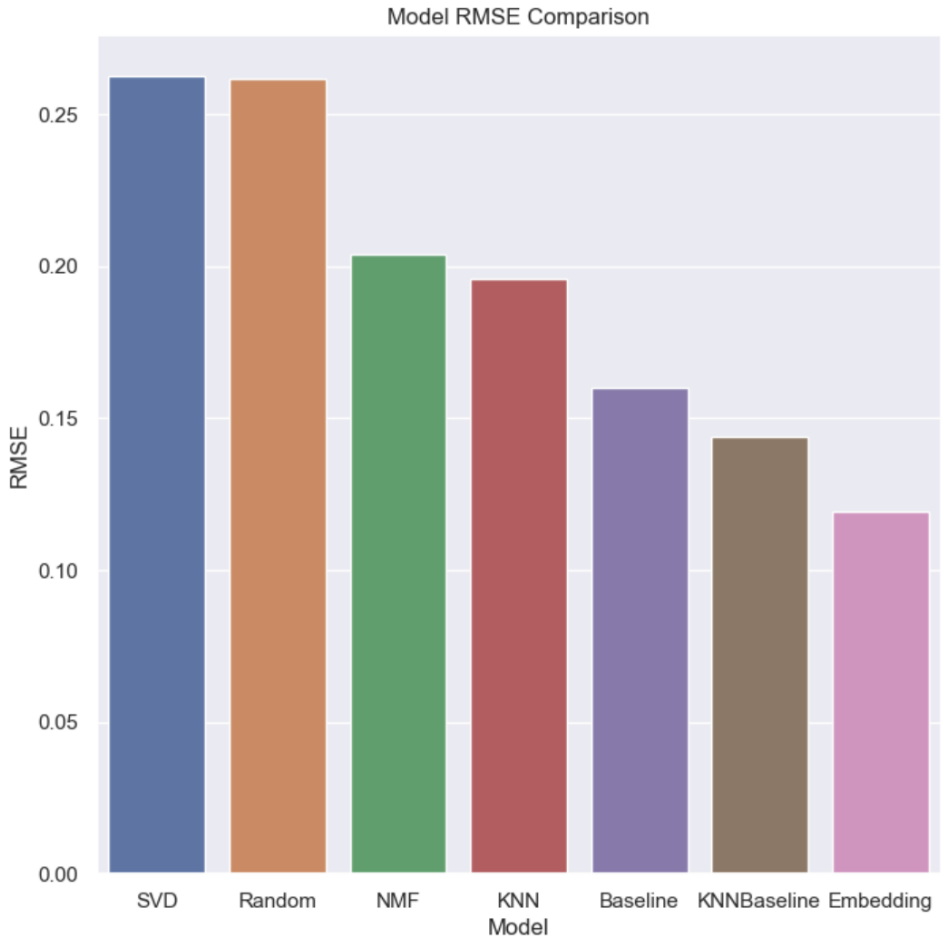
Validation result of Neural Network Embedding based recommender system



Hyperparameters

- Loss: MSE
- Optimizer: adam
- Metrics: RMSE
- Epochs: 10
- Batch size: 64
- Embedding layers: 16

Compare the performance of collaborative-filtering models



Conclusions

- Content-based recommendation methods
  - All 3 methods seem to return varying courses as their top recommendations. Suggesting that they capture different area of interest for the recommendation.
  - The results seem to heavily depend on the type of methods as well as the hyperparameters.
  - Choosing hyperparameters requires good insight into the dataset as well as the objective of the recommendation system.
  - Perhaps combining various content based recommendation methods may yield a better result as it may capture a larger scope of the user's interest. It may also be possible to capture the hit rate of each method and build a supervised learning model such as neural network to further improve the results.
- Collaborative filtering
  - Given the strong performance of the baseline method, the dataset doesn't seem to suffer from the overfit. It may be the reason why knn and nmf based methods performed worse than the baseline method as well. Perhaps the hyperparameters for knn and nmf require more tuning.
  - Embedding based method seems to have decreasing RMSE value at 10 epochs. The results may be further improved with fine tuning the hyperparameters.
  - While all methods seem to yield relatively low RMSE value, the embedding method seems to be the best method for the recommendation system.

Appendix

- Content-based clustering
  - Kmeans inertia vs. clusters
  - PCA acc. variance vs. n components