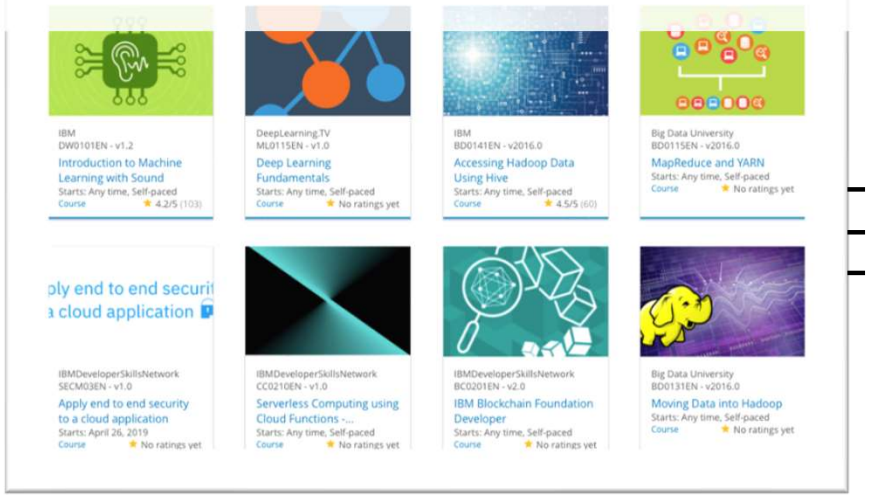
Build a Personalized Online Course Recommender System with Machine Learning

Tao Shan

2022-08-10



Outline

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

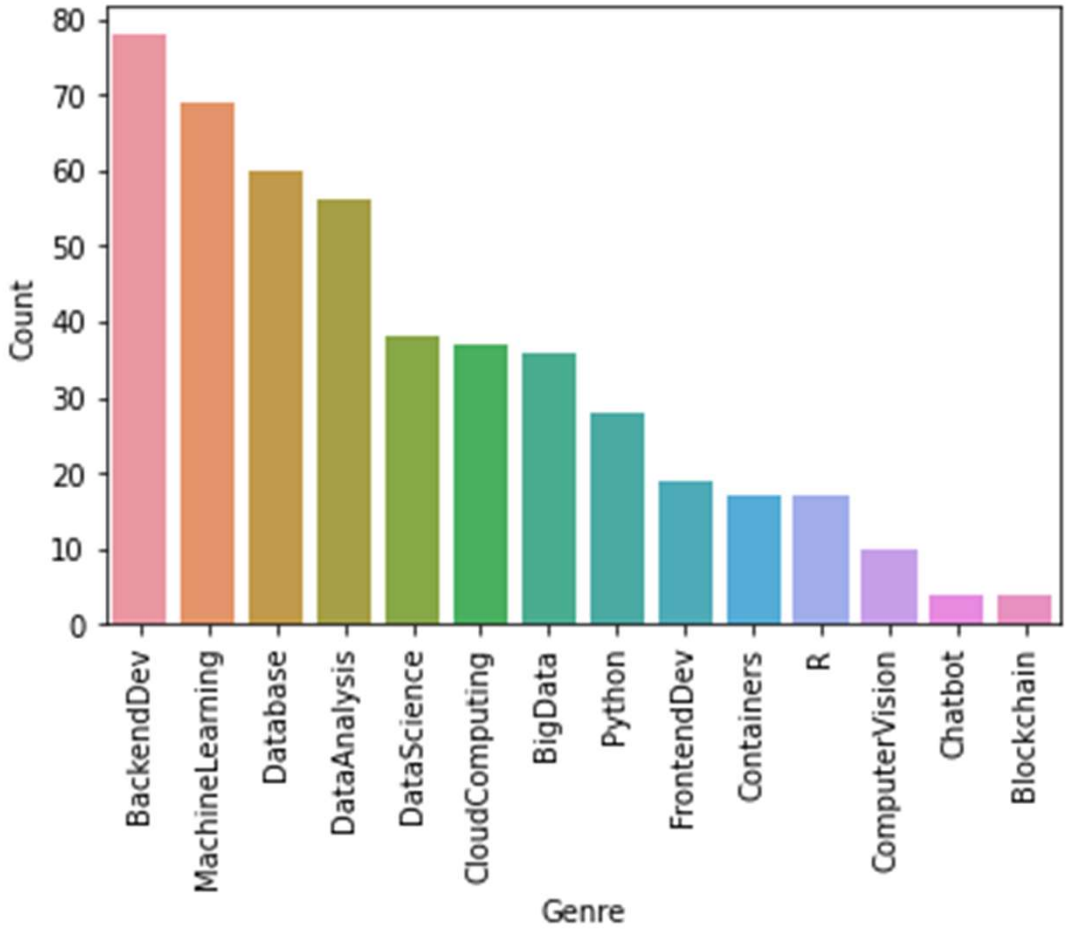
Introduction

- Massive Open Online Courses (MOOCs) startup - AI Training Room
- Recommendation system for people select courses
- Assume we know people's interest such as python, Cloud Computing

Exploratory Data Analysis

ts0

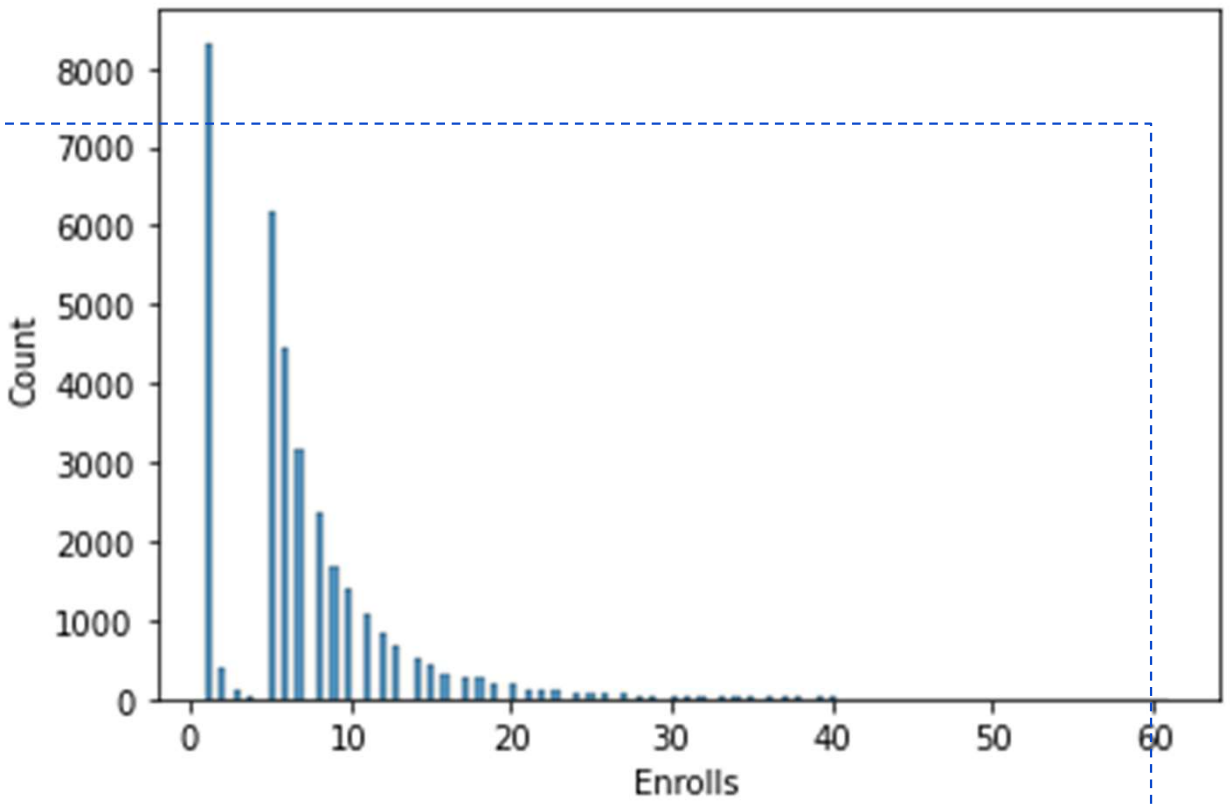
Course counts per genre



Slide 5

ts0 Barchart: We found backend develop, machine learning, database, data analysis has a high value for course count per genre. This means many course
relate to these topics.
t s, 2022-08-08T21:07:36.842

Course enrollment distribution



Slide 6

ts0 Many courses has close to 0 value of enrollments. When the value of enroll becomes higher, the number of courses decreases.

t s, 2022-08-08T21:09:24.747

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

ts0

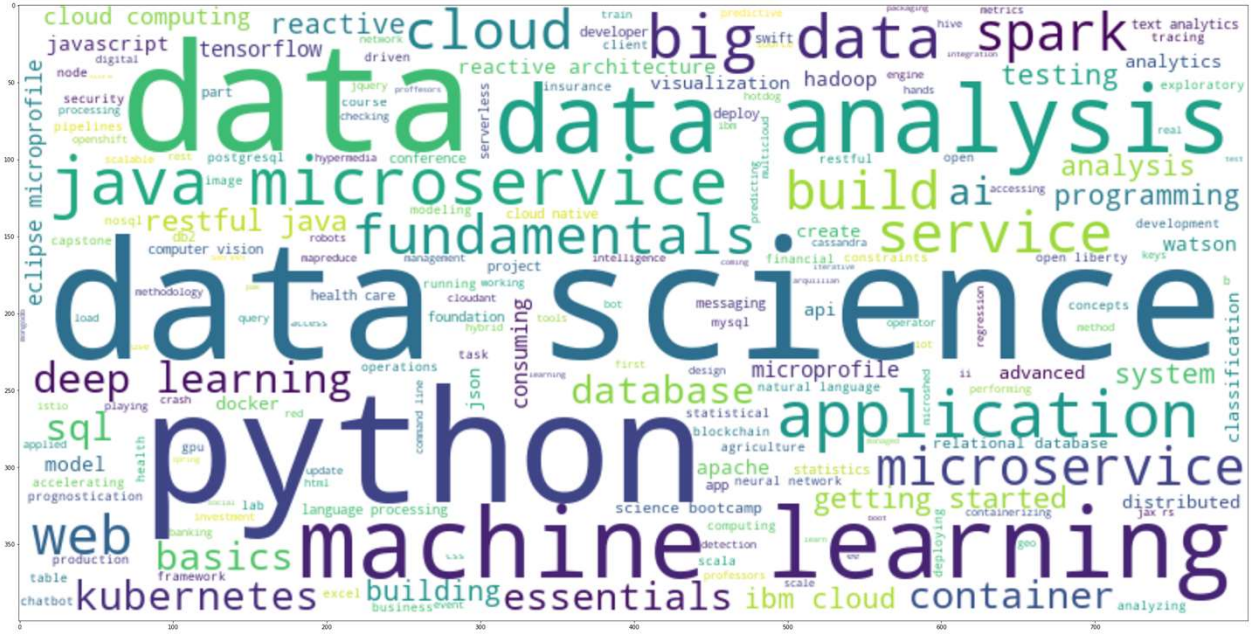
Slide 7

ts0 These twenty courses are the most popular courses that has the highest course enrollment. We found the topics are relate to data science, big data o
related topics (R, Python, Hadoop, Statistics)

t s, 2022-08-08T21:13:00.444

Word cloud of course titles

ts0

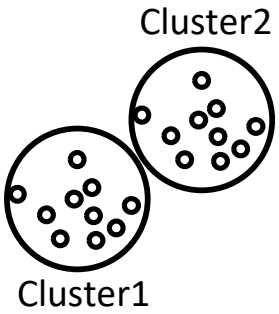


Slide 8

ts0 From the word cloud, the most biggest words are python, data science, data, data analysis and machine learning. We found these words are commor
in the course titles, it shows the trend that course relates to data science/ ML are very popular.

t s, 2022-08-12T14:40:01.832

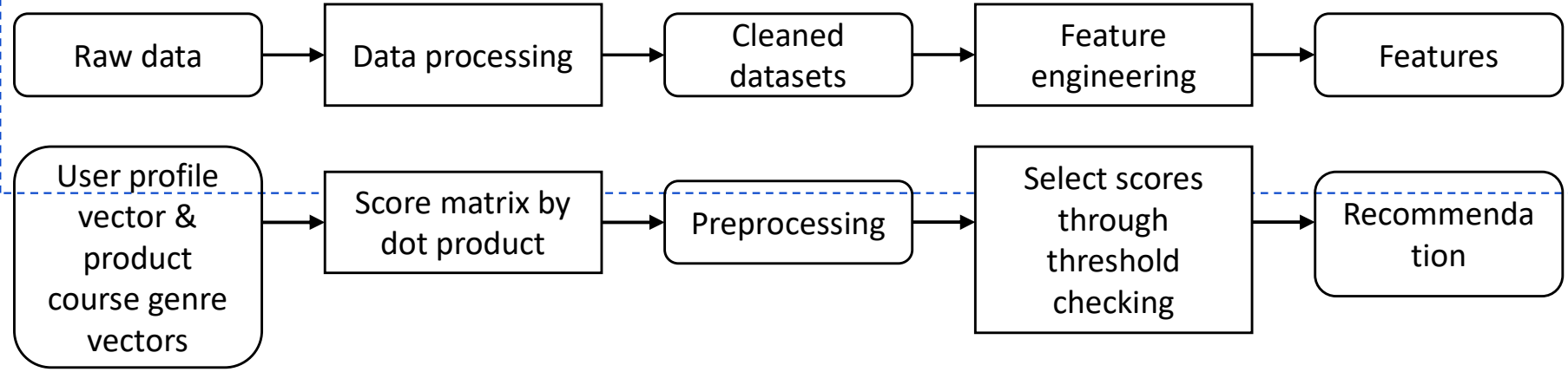
Content-based Recommender System using Unsupervised Learning



ts0

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

- Raw data: Course textual data and user profile data
- Data preprocessing: word tokenizer
- Cleaned datasets: stop word removal, upper case, stemming, lemmatization
- Feature engineering: Bow of words (doc2bow)
- Features: user profile dot product with course genre vector



Slide 10

ts0

For content based recommender system, we focused on the original value of description on user profile and course genre vector. The original raw data was course textual data and user profile data. Then for preprocessing steps, I was using word tokenizer, stop word removal, upper case, stemming and lemmatization, and using bow of words to generate more meaningful features. At last, in content-based method I use dot product between user profile dot product with course genre vector.

t s, 2022-08-12T14:40:14.604

ts0

Evaluation results of user profile-based recommender system

Dot product value threshold for recommendation: I tried threshold value = 10,15,20,40 found 20 should be great.

Average number of recommendation

Threshold value = 10: 61.8 recommendations

15: 28.3 recommendations

20: 19.1 recommendations

40: 2.0 recommendations

What are the most frequently recommended courses?

For Threshold value = 20:

	TITLE	Count
0	analyzing big data with sql	322
1	foundations for big data analysis with sql	322
2	getting started with the data apache spark ma...	318
3	analyzing big data in r using apache spark	309
4	spark overview for scala analytics	292
5	cloud computing applications part 2 big data...	283
6	introduction to data science in python	270
7	applied machine learning in python	270
8	spark fundamentals ii	268
9	accelerating deep learning with gpu	267

ts0 If the score of any course is above the threshold, we may recommend that course to the user. A lower score threshold yields more recommended courses but with smaller confidence so that some test users may receive very long course recommendation lists and feel overwhelmed.

t s, 2022-08-12T14:40:23.301

ts0

Flowchart of content-based recommender system using course similarity

- Raw data: user rated history, and which course user enrolled or not
- We need to find the cosine similarity (or using other distance matrics) between chosen course and unchosen course, to check if there is any unknown course can recommend to the person.
- Preprocessing method similar as before

Slide 12

ts0 By the data with each person rated the course, and how they chosen the course, We need to find the cosine similarity between chosen course and unchosen course, to check if there is any unknown course can recommend to the person.

t s, 2022-08-12T14:40:32.824

Evaluation results of course similarity based recommender system

similarity threshold: 0.6

On average, 11.4 courses have been recommended per user (in the test user dataset)

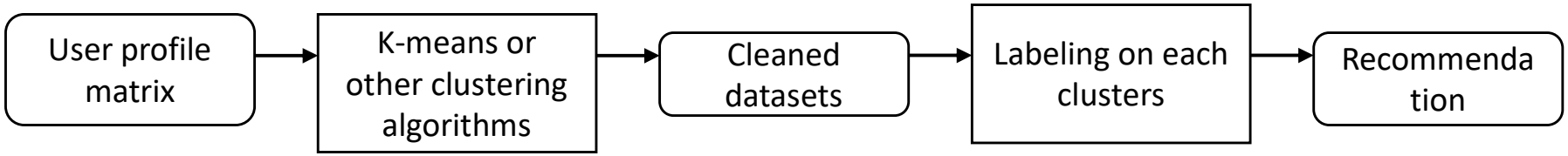
What are the most frequently recommended courses?

	TITLE	Count
0	introduction to data science in python	579
1	introduction to data science in python	579
2	data science with open data	562
3	a crash course in data science	555
4	data science fundamentals for data analysts	555
5	foundations for big data analysis with sql	551
6	big data modeling and management systems	550
7	fundamentals of big data	539
8	introduction to big data	539
9	sql access for hadoop	506

Flowchart of clustering-based recommender system

ts0

- Raw data: user rated history, and which course user enrolled or not
- We need to find the clusters among all the users, to find similar users to recommend new courses.
- Preprocessing method similar as before



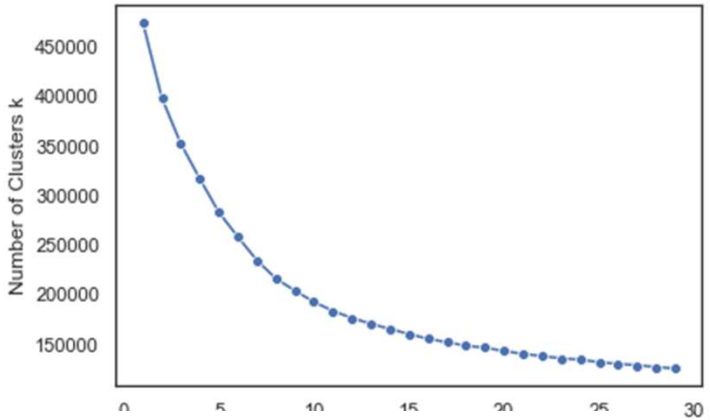
Slide 14

ts0 By the data with each person rated the course, and how they chosen the course, We need to find the cosine similarity between chosen course and unchosen course, to check if there is any unknown course can recommend to the person.

t s, 2022-08-12T14:40:45.158

Evaluation results of clustering-based recommender system

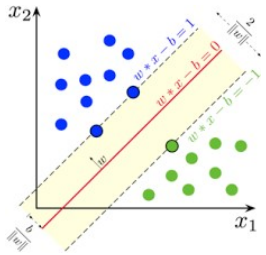
K means hyper parameter: choose k = 30, since sum of square distance is the smallest.



On average, 4 courses have been recommended per user (in the test user dataset)

	TITLE	Count
0	openrefine 101	8536
1	watson analytics for social media	7512
2	introduction to open source	7512
3	text analytics 101	7512
4	scalable web applications on kubernetes	7512
5	building robots with tjbot	5944
6	dataops methodology	5253
7	game playing ai with swift for tensorflow s4tf	5227
8	deep learning with tensorflow	5183
9	accelerating deep learning with gpu	5078

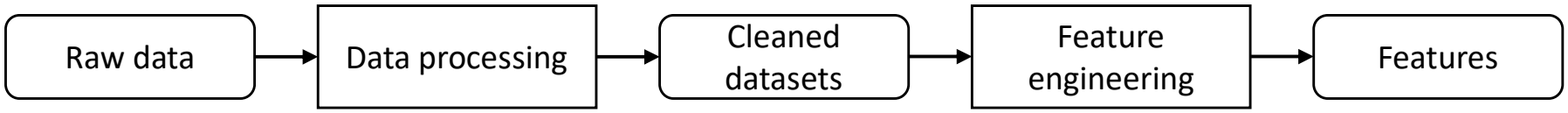
Collaborative-filtering Recommender System using Supervised Learning



Flowchart of KNN based recommender system

ts0

- Raw data: user preference profile
- We are using KNN to find the nearest neighbors for users. For these nearest neighbors for users, We can find the similar user's choice for courses, then recommend to our users
- Preprocessing method similar as before



Slide 17

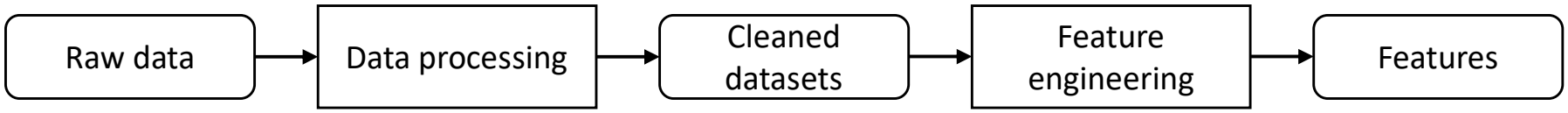
ts0 The raw data is user preference profile, which kind of course user prefers. We are using KNN to find the nearest neighbors for users. For these neares neighbors for users, We can find the similar user's choice for courses, then recommend to our users. The preprocessing steps are similar as before mentioned.

t s, 2022-08-12T14:40:55.581

Flowchart of NMF based recommender system

ts0

- Raw data: user preference profile
- We are using NMF to find the user and item matrixes for users. The advantages is to reduce the high dimension for previous user matrix in KNN.
- Preprocessing method similar as before



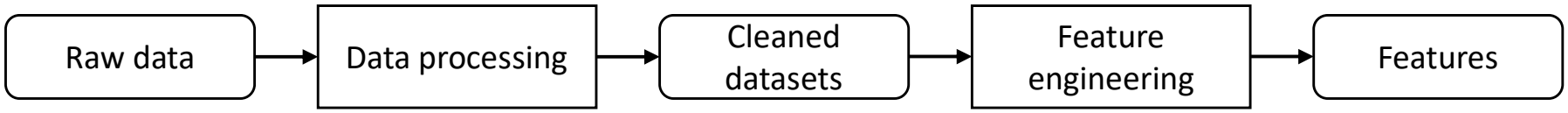
ts0 The raw data is user preference profile, which kind of course user prefers. We are using NMF to find the user and item matrixes for users. The advantages is to reduce the high dimension for previous user matrix in KNN.

t s, 2022-08-12T14:41:04.901

Flowchart of Neural Network Embedding based recommender system

ts0

- Raw data: user preference profile
- We are using neural networks to learning patterns from data and extract latent features.
- Preprocessing method similar as before

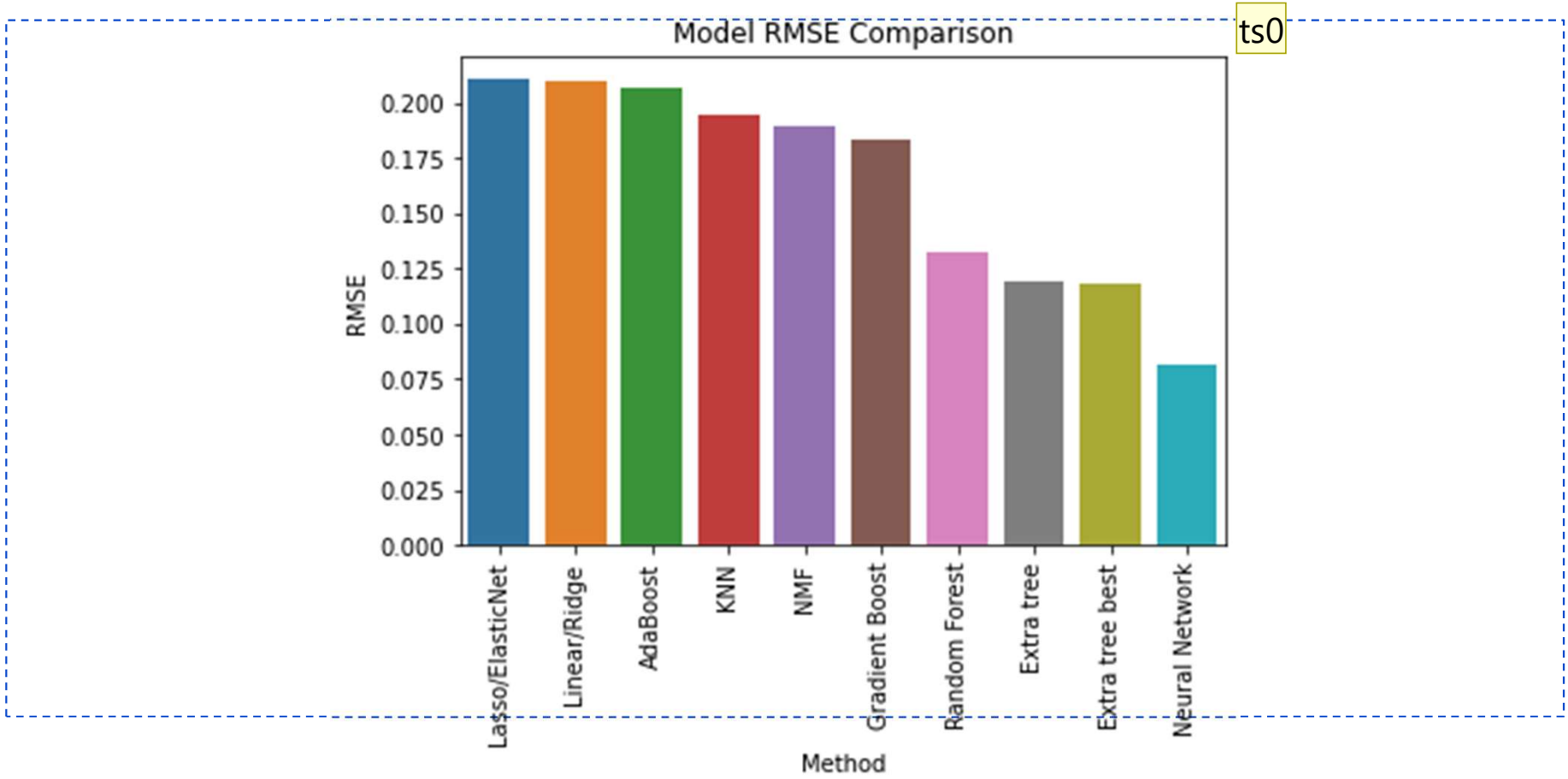


Slide 19

ts0 In addition to NMF, neural networks can also be used to extract the latent user and item features? In fact, neural networks are very good at learning patterns from data and are widely used to extract latent features. When training neural networks, it gradually captures and stores the features within its hidden layers as weight matrices and can be extracted to represent the original data.

t s, 2022-08-12T14:41:13.361

Compare the performance of collaborative-filtering models



ts0 Compare with these models, we found neural network has a low RMSE value, which means the recommendation is the most accurate. Also Extra tree and Random forest has great performance.

t s, 2022-08-12T14:41:22.488

Conclusions

- Popular Courses are all relates to data analytics
- content-based recommender system, user profile-based recommender system, course similarity based recommender system are all great recommendation systems. A essential number of courses are recommended to users
- Among KNN, NMF, Neural network, and other modeling strategies, Neural networks has the best performance, Extra Tree Regressor and Random forest also has great performance

Appendix

- Project Codes: https://github.com/davidshan0814/Recommendation-System