Course4U
Build a Personalized Online Course Recommender System with Machine Learning

Presented by: Wahyu Ardhitama

Last Updated: December 1st, 2023

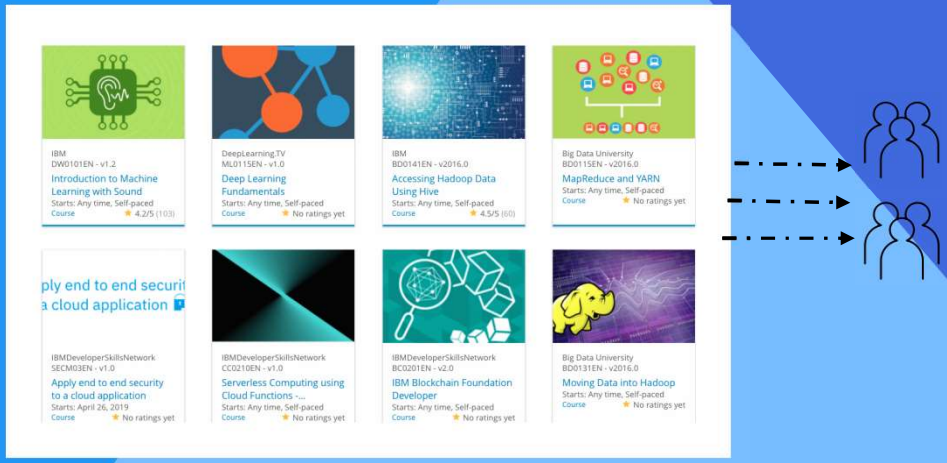


Table of Contents

What are we talking about?

Did you complete the Introduction slide? (4 pts)

Business Case

Course4U growing , having reached ~**34,000 users** and over **233,000 enrollments** in **a year**.

Opportunity/Problem Statement:

- 25,000 users (**70%**) who have enrolled in **fewer than 10 courses**.
- Among them, **8,000** users have enrolled in only a single course**.**
- Only **less than 45%** of the total courses have been chosen by users.
- Encourage existing users to enroll in more than 10 courses.
- Acquiring new users.

Goals

Maximize user engagement, increase revenue streams, and solidify Course4U's position in the online education market.
257,500 enrollments next year.

Campaign Objective:

Conversion/Enrollments

KPI:

Number of enrollments

(Tracked via online conversions and mobile - SDK)

Primary metric:

- Increase course **enrollments** by 10% by identifying and offering more engaging and relevant courses to **learners** . (courses enrolled in the list from 45% to > 50%)

Analytics Objective

Explore and compare **various machine learning models** and **find one** with **the best performance** to improve learners' learning experience

C4U Recommender Systems :

- Quickly find new interested courses
- Better paving learning paths
- More learners interacting with more courses

Hypothesis :

Recommender system delivers **more incremental value** of **enrollments** relative to **the current systems.**

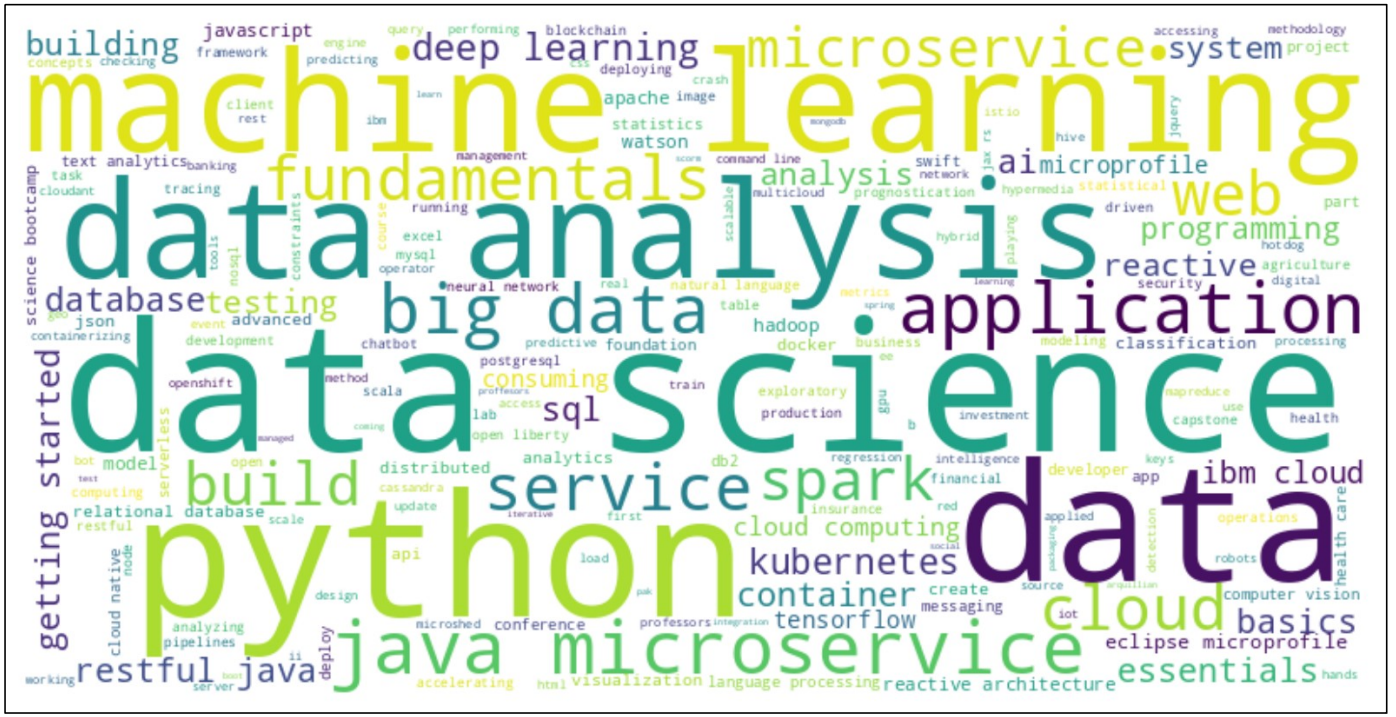
Models and Findings

Exploratory Data Analysis

Did you complete the 4 Exploratory Data Analysis slides? (8 pts)

Keywords

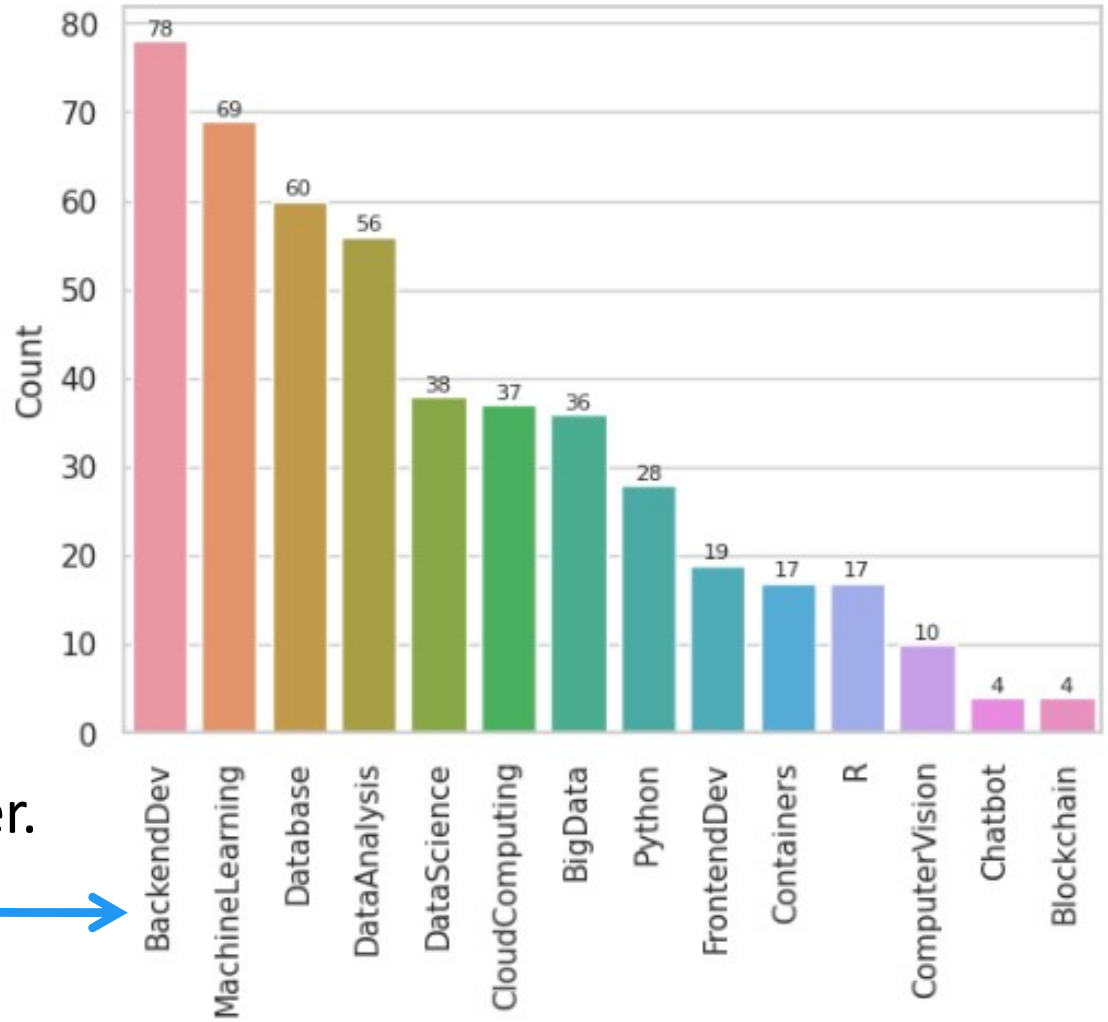
In general, the courses are focused on **demanding IT skills**



Course Genres

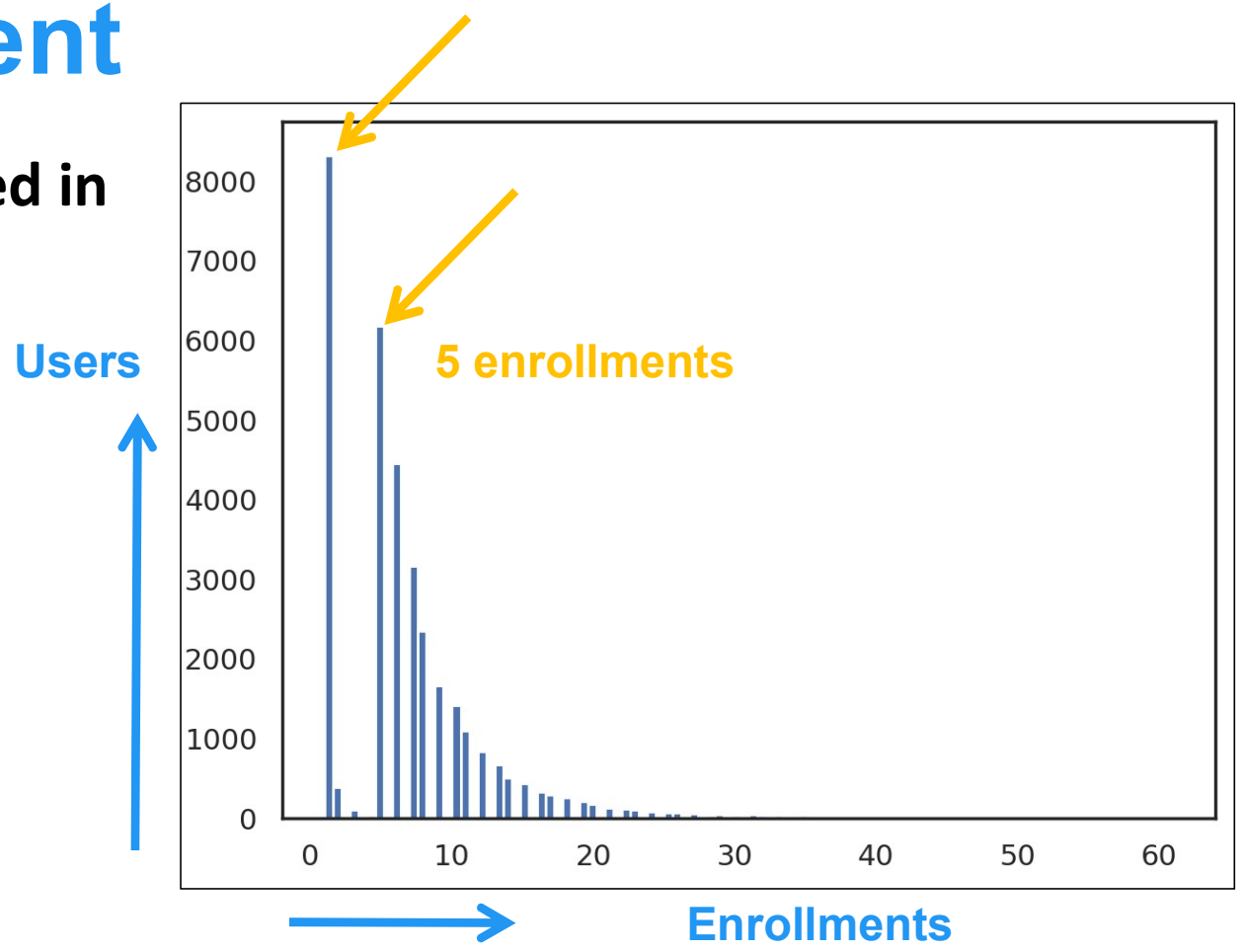
307 Total of courses offered

- Mostly related to **backend development**, **machine learning**, **database** and so on.
- Courses related to chatbot and blockchain are comparatively fewer.



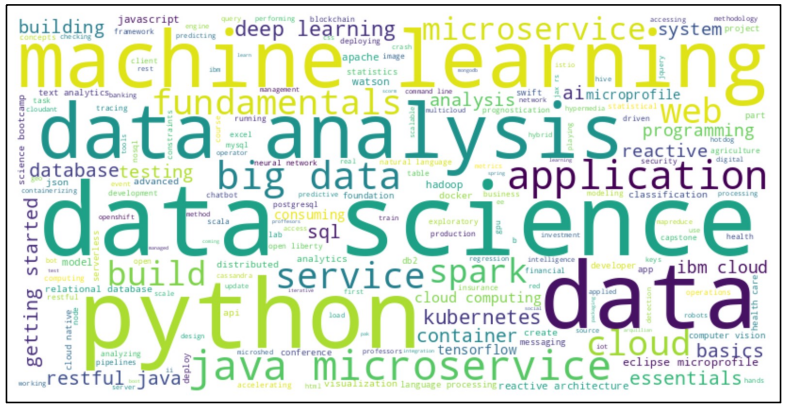
Course Enrollment

- **Over 8,000 users** have **enrolled in only one course.**
- The enrollment distribution is continuously declining, with **fewer users** as the number of **enrollments increases**.



Top 20 Courses

- **Just over 60%** of enrollment
- Related to data science, python, machine learning and so on as depicted on keywords
- Only **6.5 %** of total courses offered



	TITLE	Enrolls
0	python for data science	14936.0
1	introduction to data science	14477.0
2	big data 101	13291.0
3	hadoop 101	10599.0
4	data analysis with python	8303.0
5	data science methodology	7719.0
6	machine learning with python	7644.0
7	spark fundamentals i	7551.0
8	data science hands on with open source tools	7199.0
9	blockchain essentials	6719.0
10	data visualization with python	6709.0
11	deep learning 101	6323.0
12	build your own chatbot	5512.0
13	r for data science	5237.0
14	statistics 101	5015.0
15	introduction to cloud	4983.0
16	docker essentials a developer introduction	4480.0
17	sql and relational databases 101	3697.0
18	mapreduce and yarn	3670.0
19	data privacy fundamentals	3624.0

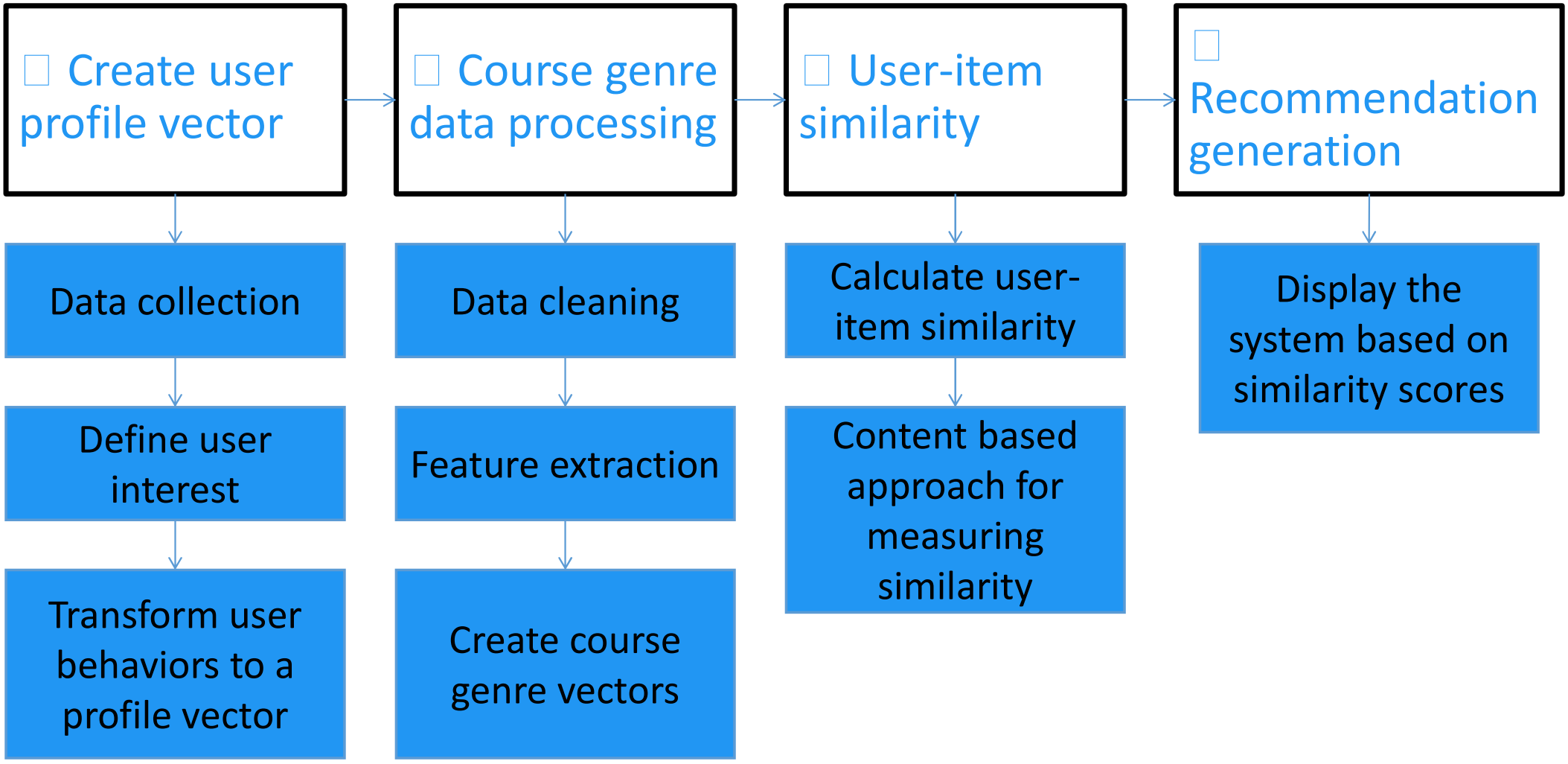
Content-Based Recommender System using Unsupervised Learning

Models and Findings

Content-Based Recommender System using User Profile and Course Genres

Did you complete the slides related to content-based recommender system using user profile and course genres? (6 pts)

Flowchart


Create user profile vector
Course genre data processing
User-item similarity
Recommendation generation
Data collection
Define user interest
Transform user behaviors to a profile vector
Data cleaning
Feature extraction
Create course genre vectors
Calculate user-item similarity
Content based approach for measuring similarity
Display the system based on similarity scores

Recommendation Generation

User and Genre

User	Database	...	Machine Learning	Blockchain
2	52.00		33.0	6.0
...				

User rating (1000 Users)

User	Item	Rating
2	ML0201EN	3
...		

Course genre data processing

course_matrix

Course and Genre

Course_ID	Title	Database	...	Machine Learning	Blockchain
ML0201EN	robots are ...	0		1.0	
...					

Create user profile vector

test_user_vector

dot product

User-item similarity

Recommendation generation

	USER	COURSE_ID	SCORE
14636	733707	RP0105EN	99.0
14655	733707	SC0103EN	90.0
14796	733707	excourse73	90.0
11608	674939	RP0105EN	90.0
11		SC0103EN	90.0
11		excourse72	90.0
14795	733707	excourse72	90.0
11701	674939	excourse73	90.0
14649	733707	TMP0105EN	90.0
36771	1312255	TMP0105EN	87.0

90

Evaluation Results

score_threshold = 10

score_threshold = 10

	USER	COURSE_ID	SCORE
14636	733707	RP0105EN	99.0
14655	733707	SC0103EN	90.0
14796	733707	excourse73	90.0
11608	674939	RP0105EN	90.0
11618	674939	SC0103EN	90.0
...	...	...	...
3053	435051	excourse42	10.0
2680	418401	BD0131EN	10.0
3051	435051	excourse10	10.0
3050	435051	excourse05	10.0
2690	418401	GPXX0M6UEN	10.0

53411 rows × 3 columns

53411 recommendations for 864 users of 1,000 users(86.4%)

Top 10 recomended all users courses

	USER	COURSE_ID	SCORE	TITLE
0	733707	RP0105EN	99.0	analyzing big data in r using apache spark
1	733707	SC0103EN	90.0	spark overview for scala analytics
2	733707	excourse73	90.0	analyzing big data with sql
3	674939	RP0105EN	90.0	analyzing big data in r using apache spark
4	674939	SC0103EN	90.0	spark overview for scala analytics
5	674939	excourse72	90.0	foundations for big data analysis with sql
6	733707	excourse72	90.0	foundations for big data analysis with sql
7	674939	excourse73	90.0	analyzing big data with sql
8	733707	TMP0105EN	90.0	getting started with the data apache spark ma...
9	1312255	TMP0105EN	87.0	getting started with the data apache spark ma...

→ Scores >= 99

→ Scores >= 87

3 users 5 courses
733707, 674939,1312255

- Big data (data analysis) sql
- Foundation
- Apache spark

Recommendation based on users and courses

User profile 1078030

	user	item	rating	COURSE_ID	TITLE
0	1078030	DA0101EN	3.0	DA0101EN	data analysis with python
1	1078030	ST0101EN	3.0	ST0101EN	statistics 101
2	1078030	ML0122ENv1	3.0	ML0122ENv1	accelerating deep learning with gpu
3	1078030	ML0120ENv2	3.0	ML0120ENv2	deep learning with tensorflow
4	1078030	DV0101EN	3.0	DV0101EN	data visualization with python
5	1078030	ML0115EN	3.0	ML0115EN	deep learning 101
6	1078030	ML0101ENv3	3.0	ML0101ENv3	machine learning with python
7	1078030	PY0101EN	3.0	PY0101EN	python for data science

Participate in 8 courses

- **Data analysis**
- **Deep learning**
- **Python**

Evaluation Results

score_threshold = 10

Top 10 recomended score's for 1078030

	COURSE_ID	SCORE	TITLE
0	ML0122EN	30.0	accelerating deep learning with gpu
1	excourse21	30.0	applied machine learning in python
2	excourse22	30.0	introduction to data science in python
3	ML0101EN	30.0	machine learning with python
4	GPXX0IBEN	27.0	data science in insurance basic statistical a...
5	excourse49	24.0	applied machine learning in python
6	GPXX0D14EN	24.0	build a personal movie recommender with django
7	GPXX0YMEEN	24.0	launch an ai hotdog detector as a serverless p...
8	excourse54	21.0	exploratory data analysis for machine learning
9	excourse20	21.0	python and statistics for financial analysis

→ Score 30

→ Score 21

85 recommendations
lowest score 12

Participate in 8 courses

- Data analysis
- Deep learning
- Python

10 recommended courses

- Machine learning
- Deep learning
- Python
- Data analysis
- Data science

Recommendation based on users and courses

User profile 733707

	user	item	rating	COURSE_ID	TITLE
0	733707	RP0103	3.0	RP0103	using r with databases
1	733707	BD0212EN	3.0	BD0212EN	spark fundamentals ii
2	733707	BD0211EN	3.0	BD0211EN	spark fundamentals i
3	733707	ST0101EN	3.0	ST0101EN	statistics 101
4	733707	BD0115EN	3.0	BD0115EN	mapreduce and yarn
18	733707	DV0101EN	3.0	DV0101EN	data visualization with python
19	733707	SC0105EN	3.0	SC0105EN	data science with scala
20	733707	BD0145EN	3.0	BD0145EN	sql access for hadoop
21	733707	DB0151EN	3.0	DB0151EN	nosql and dbaas 101
22	733707	BD0131EN	3.0	BD0131EN	moving data into hadoop

Participate in 23 courses

- **Spark**
- **Sql**
- **Python**

Evaluation Results

score_threshold = 10

Top 10 recomended courses for 733707

	COURSE_ID	SCORE	TITLE
0	RP0105EN	99.0	analyzing big data in r using apache spark
1	SC0103EN	90.0	spark overview for scala analytics
2	TMP0105EN	90.0	getting started with the data apache spark ma...
3	excourse73	90.0	analyzing big data with sql
4	excourse72	90.0	foundations for big data analysis with sql
5	excourse31	69.0	cloud computing applications part 2 big data...
6	excourse71	69.0	big data essentials hdfs mapreduce and spark...
7	excourse05	69.0	\r\ndistributed computing with spark sql
8	excourse03	69.0	nosql systems
9	BD0143EN	69.0	using hbase for real time access to your big data

→ Scores 99

→ Scores 69

172 recommendations
lowest score 12

Participate in 23 courses

- Spark
- Sql
- Python

10 recommended courses

- Big data (data analysis) sql
- Apache spark
- Foundation

Recommendation based on users and courses

User profile 674939

	user	item	rating	COURSE_ID	TITLE
0	674939	BD0111EN	3.0	BD0111EN	hadoop 101
1	674939	BD0211EN	3.0	BD0211EN	spark fundamentals i
2	674939	BD0101EN	3.0	BD0101EN	big data 101
3	674939	BD0135EN	3.0	BD0135EN	developing distributed applications using zook...
4	674939	BD0143EN	3.0	BD0143EN	using hbase for real time access to your big data
5	674939	BD0121EN	3.0	BD0121EN	apache pig 101
6	674939	BD0131EN	3.0	BD0131EN	moving data into hadoop
7	674939	BD0221EN	3.0	BD0221EN	spark mllib
8	674939	BD0115EN	3.0	BD0115EN	mapreduce and yarn
9	674939	BD0212EN	3.0	BD0212EN	spark fundamentals ii
10	674939	TMP0105EN	3.0	TMP0105EN	getting started with the data apache spark ma...
11	674939	BD0133EN	3.0	BD0133EN	controlling hadoop jobs using oozie
12	674939	BD0223EN	3.0	BD0223EN	exploring spark s graphx
13	674939	BD0141EN	3.0	BD0141EN	accessing hadoop data using hive
14	674939	BD0145EN	3.0	BD0145EN	sql access for hadoop

Participate in 15 courses

- Spark
- Hadoop
- Big data

Evaluation Results

score_threshold = 10

Top 10 recomended courses for 674939

	USER	COURSE_ID	SCORE	TITLE
3	674939	RP0105EN	90.0	analyzing big data in r using apache spark
4	674939	SC0103EN	90.0	spark overview for scala analytics
5	674939	excourse72	90.0	foundations for big data analysis with sql
7	674939	excourse73	90.0	analyzing big data with sql
30	674939	excourse10	78.0	database architecture scale and nosql with e...
34	674939	excourse05	78.0	\r\ndistributed computing with spark sql
36	674939	excourse03	78.0	nosql systems
39	674939	DB0151EN	78.0	nosql and dbaas 101
41	674939	GPXX0M6UEN	78.0	using the cql shell to execute keyspace operat...
42	674939	GPXX097UEN	78.0	performing table and crud operations with cass...

→ Scores 90

→ Scores 78

Participate in 15 courses

- Spark
- Hadoop
- Big data

10 recommended courses

- Big data (data analysis) sql
- Nosql
- Spark

101 recommendations
lowest score 12

Rating 2 Users

score_threshold = 10

User 2057052 courses enrollment

	user	item	rating	COURSE_ID	TITLE
0	2057052	DS0132EN	2.0	DS0132EN	data ai jumpstart your journey
1	2057052	DS0101EN	3.0	DS0101EN	introduction to data science
2	2057052	ML0101ENv3	3.0	ML0101ENv3	machine learning with python
3	2057052	PY0101EN	3.0	PY0101EN	python for data science
4	2057052	DB0101EN	3.0	DB0101EN	sql and relational databases 101

Participate in 5 courses

Recomended courses for 2057052 & 1871627

USER	COURSE_ID	SCORE	TITLE

No recommendations;
their highest scores are 2

User 1871627 courses enrollment

	user	item	rating	COURSE_ID	TITLE
0	1871627	CC0103EN	3.0	CC0103EN	ibm cloud essentials v3
1	1871627	ML0101ENv3	3.0	ML0101ENv3	machine learning with python
2	1871627	ML0103EN	3.0	ML0103EN	digital analytics regression
3	1871627	ST0101EN	3.0	ST0101EN	statistics 101
4	1871627	PY0101EN	3.0	PY0101EN	python for data science
5	1871627	DV0151EN	3.0	DV0151EN	data visualization with r
6	1871627	DS0101EN	3.0	DS0101EN	introduction to data science
7	1871627	DS0103EN	3.0	DS0103EN	data science methodology
8	1871627	CC0101EN	3.0	CC0101EN	introduction to cloud
9	1871627	ML0115EN	3.0	ML0115EN	deep learning 101
10	1871627	DB0101EN	3.0	DB0101EN	sql and relational databases 101
11	1871627	OS0101EN	3.0	OS0101EN	introduction to open source
12	1871627	CB0103EN	3.0	CB0103EN	build your own chatbot
13	1871627	DS0132EN	2.0	DS0132EN	data ai jumpstart your journey

Participate in 14 courses

Summary

The idea

- **Interested in certain topics**
- **a course has a similar topics**

Past interactions

- User-item similarity (a dot product) capture it

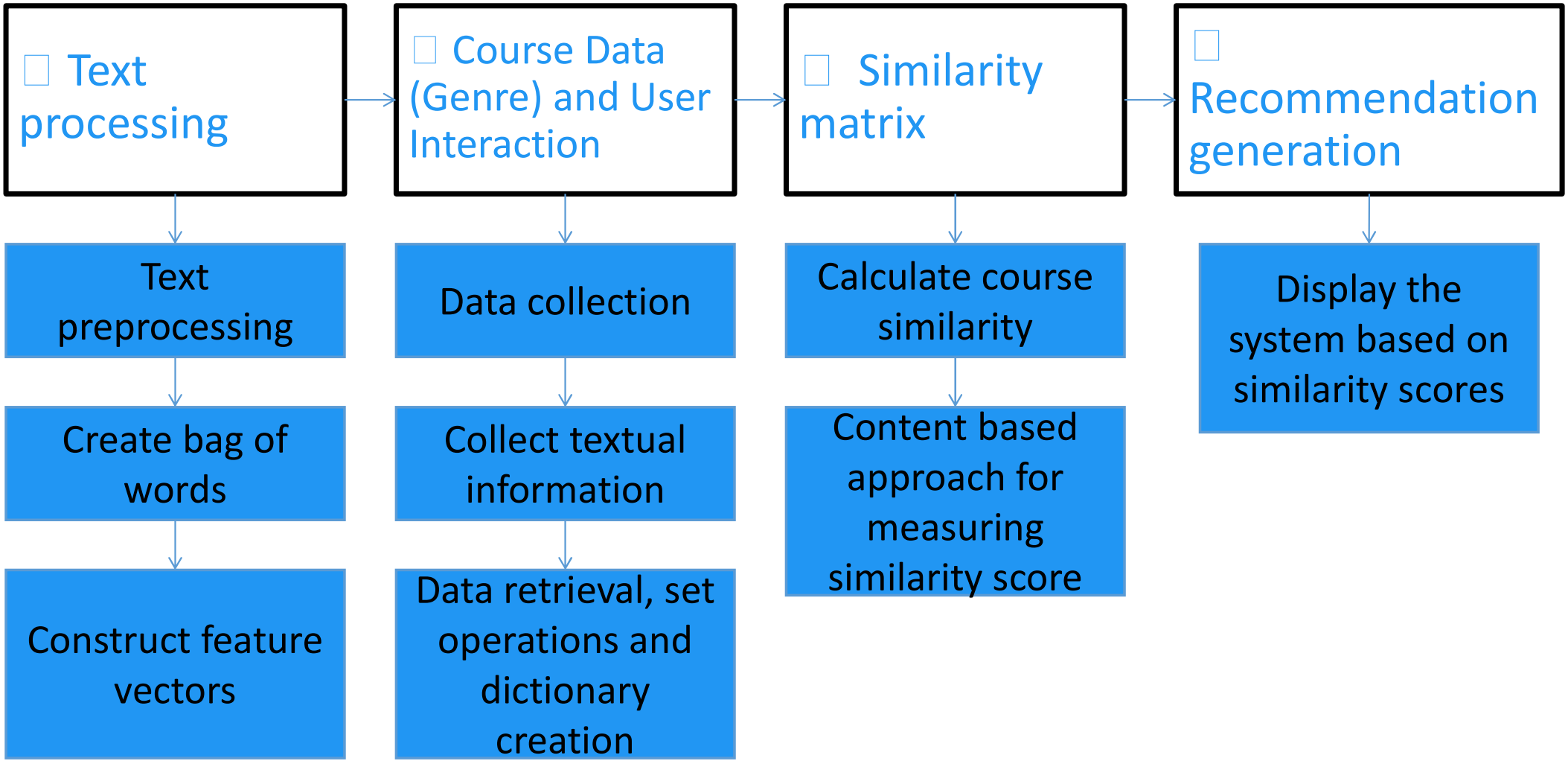
1. **Personalized suggestions based on a user's preferences and past interactions.**
2. **Has insights into a user's preferences and can recommend courses that align with their interests.**
3. **Can make recommendations even for new users with limited interaction history.**
4. **Based on explicit features (e.g., genres) that users can understand.**
5. **Adjust the threshold for users who have courses with a rating of 2.**

Models and Findings

Content-Based Recommender System using Course Similarity

Did you complete the slides related to content-based recommender system using course similarity? (6 pts)

Flowchart


Text processing
Course Data (Genre) and User Interaction
Similarity matrix
Recommendation generation
Text preprocessing
Create bag of words
Construct feature vectors
Data collection
Collect textual information
Data retrieval, set operations and dictionary creation
Calculate course similarity
Content based approach for measuring similarity score
Display the system based on similarity scores

Files

sim_df.head()

Course Similarity

0	1	2	...	305	306
1.000000	0.088889	0.088475		0.039276	0.121113
...					

bow_df.head()

Bag of Words

doc_index	doc_id	token	bow
0	ML0201EN	ai	2
...			

course_df.head()

Course 1

Course_ID	Title	Description
ML0151EN	machine learning ...	this machine learning...

Recommendation Generation

User and Course Data

Text processing

Recommendation generation

Course 1(index=200)

Course_ID	Title	Description
ML0151EN	machine learning ...	this machine learning...

Course 2(index=158)

Course_ID	Title	Description
ML0101ENv3	machine learning with ..	machine learning can be..

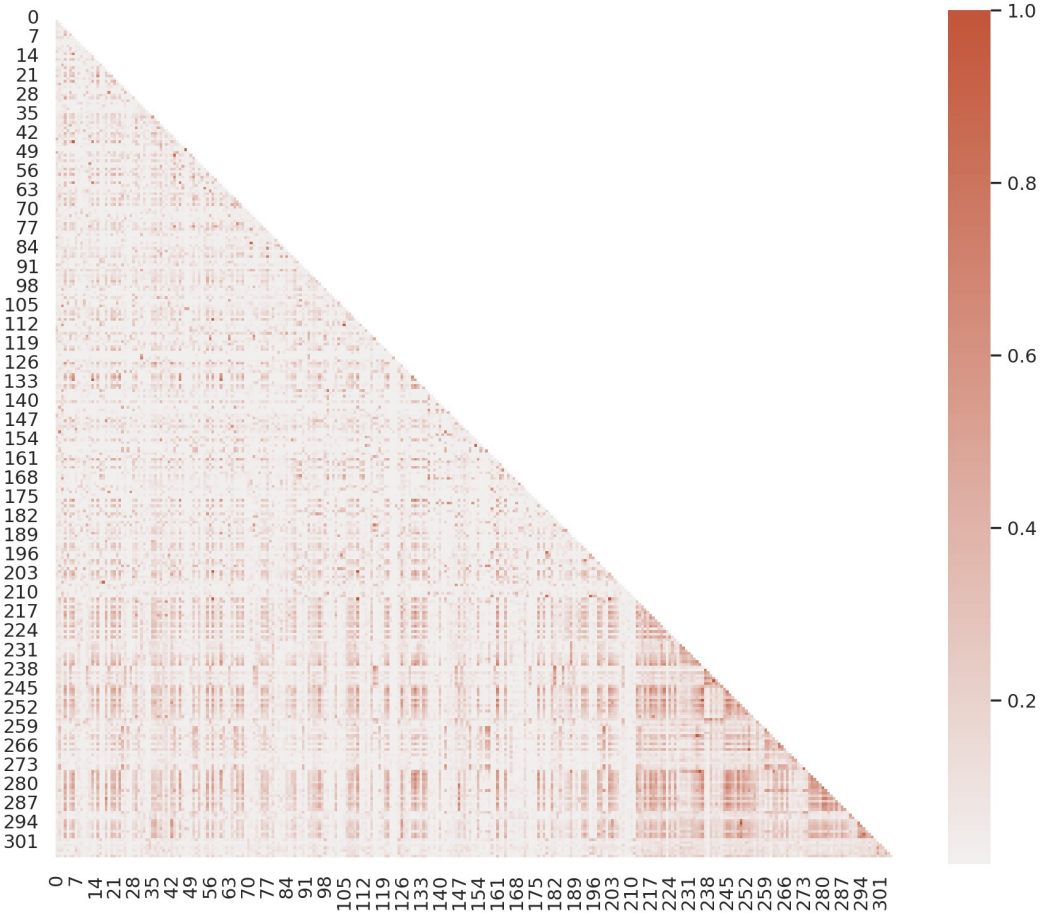
Similarity matrix

Similiraty calculation:
Cosine, Euclidean, Jaccard index,..

	USER	COURSE_ID	SCORE
0	37465	ML0120EN	1.000000
1	37465	ML0120ENv3	1.000000
2	37465	excourse36	0.739704
3	37465	excourse23	0.739704
4	37465	DV0151EN	0.723536
...	...	...	...
15995	2087663	excourse62	0.647502
		excourse47	0.634755
		excourse60	0.615568
15998	2087663	excourse46	0.612054
15999	2087663	excourse09	0.608330

75%

Similarity matrix



Hot spots shown. Possible to build a recommender system based on course similarities.

Evaluation Results

Machine learning courses
ML0151EN & ML0101ENv3

score_threshold = 0.6

	USER	COURSE_ID	SCORE
0	37465	ML0120EN	1.000000
1	37465	ML0120ENv3	1.000000
2	37465	excourse36	0.739704
3	37465	excourse23	0.739704
4	37465	DV0151EN	0.723536
...	...	...	...
15995	2087663	excourse62	0.647502
15996	2087663	excourse47	0.634755
15997	2087663	excourse60	0.615568
15998	2087663	excourse46	0.612054
15999	2087663	excourse09	0.608330

16000 rows × 3 columns

16000 recommendations for 1000 users of 1,000 users(100%)

Top 10 recomended all users courses

	USER	COURSE_ID	SCORE	TITLE
0	37465	ML0120EN	1.000000	deep learning with tensorflow
1	37465	ML0120ENv3	1.000000	deep learning with tensorflow
2	37465	excourse36	0.739704	data analysis using python
3	37465	excourse23	0.739704	data analysis using python
4	37465	DV0151EN	0.723536	data visualization with r
5	37465	excourse32	0.722018	introduction to data analytics
6	37465	ML0122ENv3	0.707107	accelerating deep learning with gpus
7	37465	excourse38	0.681638	data analysis with python
8	37465	excourse33	0.664509	excel basics for data analysis
9	37465	ML0151EN	0.662622	machine learning with r

→ **Scores >= 100%**

→ **Scores >= 66%**

1 users **10 courses**
User 37465

- **Data analysis**
- **Deep learning**
- **Python**

Recommendation based course similarity

User profile 1078030

	user	item	rating	COURSE_ID	TITLE
0	1078030	DA0101EN	3.0	DA0101EN	data analysis with python
1	1078030	ST0101EN	3.0	ST0101EN	statistics 101
2	1078030	ML0122ENv1	3.0	ML0122ENv1	accelerating deep learning with gpu
3	1078030	ML0120ENv2	3.0	ML0120ENv2	deep learning with tensorflow
4	1078030	DV0101EN	3.0	DV0101EN	data visualization with python
5	1078030	ML0115EN	3.0	ML0115EN	deep learning 101
6	1078030	ML0101ENv3	3.0	ML0101ENv3	machine learning with python
7	1078030	PY0101EN	3.0	PY0101EN	python for data science

Participate in 8 courses

- **Data analysis**
- **Deep learning**
- **Python**

Recommendation based on course similarity

User profile 733707

	user	item	rating	COURSE_ID	TITLE
0	733707	RP0103	3.0	RP0103	using r with databases
1	733707	BD0212EN	3.0	BD0212EN	spark fundamentals ii
2	733707	BD0211EN	3.0	BD0211EN	spark fundamentals i
3	733707	ST0101EN	3.0	ST0101EN	statistics 101
4	733707	BD0115EN	3.0	BD0115EN	mapreduce and yarn
18	733707	DV0101EN	3.0	DV0101EN	data visualization with python
19	733707	SC0105EN	3.0	SC0105EN	data science with scala
20	733707	BD0145EN	3.0	BD0145EN	sql access for hadoop
21	733707	DB0151EN	3.0	DB0151EN	nosql and dbaas 101
22	733707	BD0131EN	3.0	BD0131EN	moving data into hadoop

Participate in 23 courses

- **Spark**
- **Sql**
- **Python**

Recommendation based on users and courses

User profile 674939

	user	item	rating	COURSE_ID	TITLE
0	674939	BD0111EN	3.0	BD0111EN	hadoop 101
1	674939	BD0211EN	3.0	BD0211EN	spark fundamentals i
2	674939	BD0101EN	3.0	BD0101EN	big data 101
3	674939	BD0135EN	3.0	BD0135EN	developing distributed applications using zook...
4	674939	BD0143EN	3.0	BD0143EN	using hbase for real time access to your big data
5	674939	BD0121EN	3.0	BD0121EN	apache pig 101
6	674939	BD0131EN	3.0	BD0131EN	moving data into hadoop
7	674939	BD0221EN	3.0	BD0221EN	spark mllib
8	674939	BD0115EN	3.0	BD0115EN	mapreduce and yarn
9	674939	BD0212EN	3.0	BD0212EN	spark fundamentals ii
10	674939	TMP0105EN	3.0	TMP0105EN	getting started with the data apache spark ma...
11	674939	BD0133EN	3.0	BD0133EN	controlling hadoop jobs using oozie
12	674939	BD0223EN	3.0	BD0223EN	exploring spark s graphx
13	674939	BD0141EN	3.0	BD0141EN	accessing hadoop data using hive
14	674939	BD0145EN	3.0	BD0145EN	sql access for hadoop

Participate in 15 courses

- Spark
- Hadoop
- Big data

Recommendation based on users and courses

User profile 1078030

Participate in 8 courses

- Data analysis
- Deep learning
- Python

User profile 733707

Participate in 23 courses

- Spark
- Sql
- Python

User profile 674939

Participate in 15 courses

- Spark
- Hadoop
- Big data

Evaluation Results

Machine learning courses
ML0151EN & ML0101ENv3

- **Top 10 recomended score's for 1078030**

	USER	COURSE_ID	SCORE	TITLE
7984	1078030	ML0120EN	1.000000	deep learning with tensorflow
7985	1078030	ML0120ENv3	1.000000	deep learning with tensorflow
7986	1078030	excourse36	0.739704	data analysis using python
7987	1078030	excourse23	0.739704	data analysis using python
7988	1078030	DV0151EN	0.723536	data visualization with r
7989	1078030	excourse32	0.722018	introduction to data analytics
7990	1078030	ML0122ENv3	0.707107	accelerating deep learning with gpus
7991	1078030	excourse38	0.681638	data analysis with python
7992	1078030	excourse33	0.664509	excel basics for data analysis
7993	1078030	ML0151EN	0.662622	machine learning with r

→ Score 100%

→ Score 66%

16 recommendations
lowest score 60%

10 recommended courses

- **Data analysis**
- **Deep learning**
- **Python**

For user 733707 & 674939

Evaluation Results

Machine learning courses
ML0151EN & ML0101ENv3

Top 10 recomended score's for 733707

	USER	COURSE_ID	SCORE	TITLE
3870	733707	ML0120ENv3	1.000000	deep learning with tensorflow
3871	733707	ML0120ENv2	1.000000	deep learning with tensorflow
3872	733707	ML0122ENv1	0.982873	accelerating deep learning with gpu
3873	733707	DS0110EN	0.732941	data science with open data
3874	733707	excourse67	0.708214	introduction to big data
3875	733707	excourse72	0.703648	foundations for big data analysis with sql
3876	733707	excourse63	0.694563	a crash course in data science
3877	733707	excourse46	0.689253	machine learning
3878	733707	excourse47	0.680065	machine learning for all
3879	733707	ML0101ENv3	0.662622	machine learning with python

→ Score 100%

→ Score 66%

15 recommendations
lowest score 60%

10 recommended courses

- Deep learning
- Data science
- Big data
- Machine learning

Evaluation Results

Machine learning courses
ML0151EN & ML0101ENv3

Top 10 recomended score's for 674939

	USER	COURSE_ID	SCORE	TITLE
832	674939	excourse67	0.708214	introduction to big data
833	674939	excourse72	0.703648	foundations for big data analysis with sql
834	674939	excourse74	0.650071	fundamentals of big data
835	674939	excourse68	0.616759	big data modeling and management systems

Score 70%

Score 61%

4 recommendations
lowest score 61%

4 recommended courses

- **Big data**

Recommendation based on course similarity

User profile 1078030

Participate in 8 courses

- Data analysis
- Deep learning
- Python

10 recommended courses (16)

- Data analysis
- Deep learning
- Python

User profile 733707

Participate in 23 courses

- Spark
- Sql
- Python

10 recommended courses (15)

- Deep learning
- Data science
- Big data
- Machine learning

User profile 674939

Participate in 15 courses

- Spark
- Hadoop
- Big data

4 recommended courses

- Big data

Evaluation Results

Machine learning courses
ML0151EN & ML0101ENv3

Top 10 recomended score's for 2057052 (rating 2)

	USER	COURSE_ID	SCORE	TITLE
8946	2057052	DS0110EN	0.732941	data science with open data
8947	2057052	excourse63	0.694563	a crash course in data science
8948	2057052	DAI101EN	0.668994	data ai essentials
8949	2057052	ML0151EN	0.662622	machine learning with r
8950	2057052	excourse22	0.647502	introduction to data science in python
8951	2057052	excourse62	0.647502	introduction to data science in python
8952	2057052	excourse65	0.638641	data science fundamentals for data analysts
8953	2057052	excourse47	0.634755	machine learning for all
8954	2057052	excourse46	0.612054	machine learning

Score 73%

Score 61%

9 recommendations
lowest score 61%

9 recommended courses

- Data science
- Data ai
- Machine learning

Summary

The idea

- **Interested in certain topics** — Feature vectors — **a course has a similar topics**
  - Similarity matrix capture it

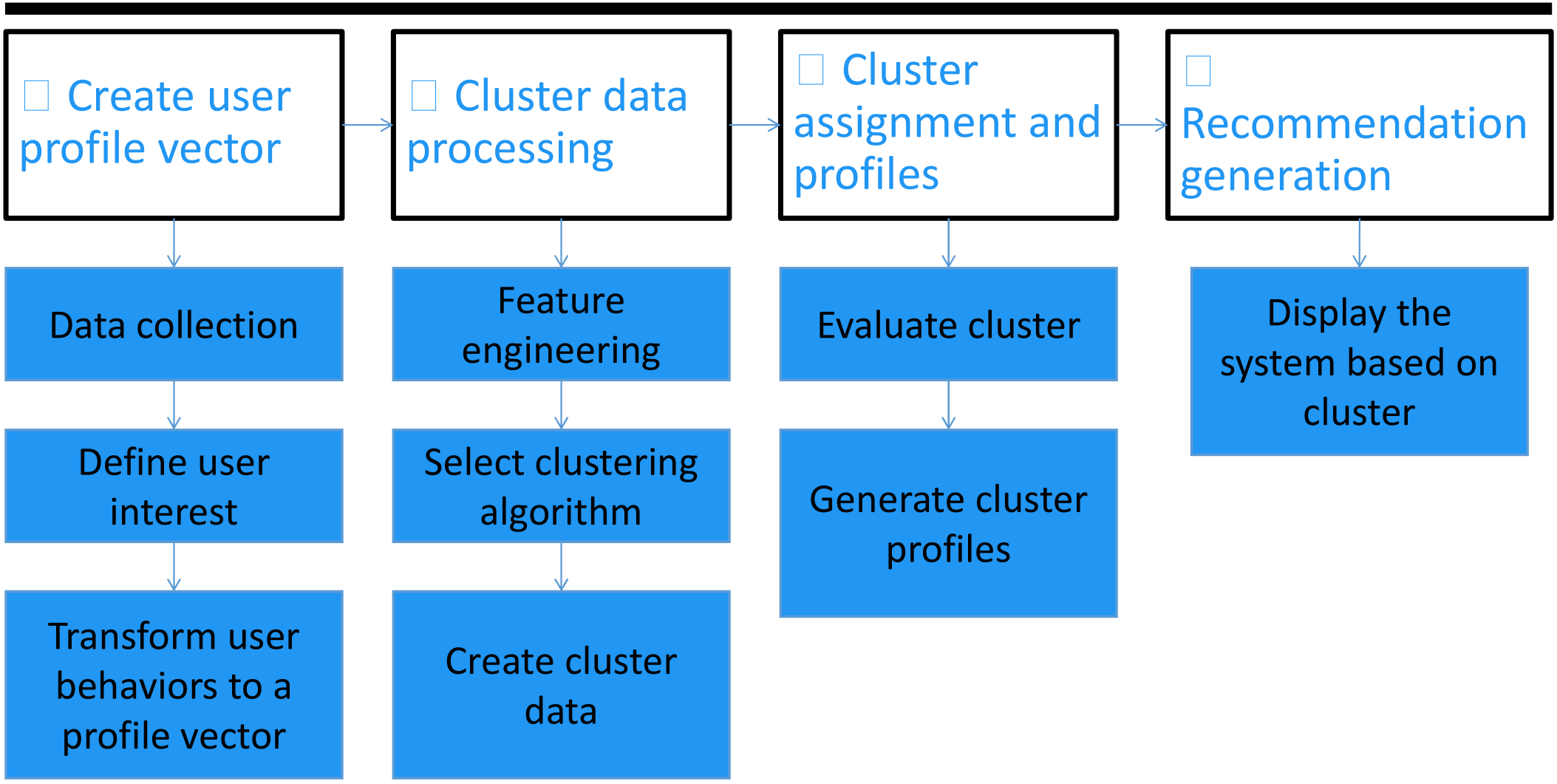
1. **Offer personalized suggestions based on the intrinsic characteristics of courses.**
2. **Based on explicit features (e.g., genres) that users can understand.**
3. **Can make recommendations even for new users with limited interaction history.**
4. **Based on specific features of courses, allowing users to interpret and understand the reasons behind each recommendation.**

Models and Findings

Clustering-Based Recommender System

Did you complete the slides related to content-based recommender system using user profile clustering? (6 pts)

Flowchart


Create user profile vector
Cluster data processing
Cluster assignment and profiles
Recommendation generation
Data collection
Define user interest
Transform user behaviors to a profile vector
Feature engineering
Select clustering algorithm
Create cluster data
Evaluate cluster
Generate cluster profiles
Display the system based on cluster

Recommendation Generation

 Create user profile vector

 Cluster data processing

 Cluster assignment and profiles

 Recommendation generation

 User and Genre

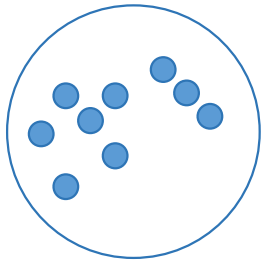
User	Python	Database	...	ML	Blockchain
1		52.00		33.0	6.0
...					

User profile standard scaler

User	Python	Database	...	ML	Blockchain
1	-0.3533	4.52998		2.3685	0.519419
...					

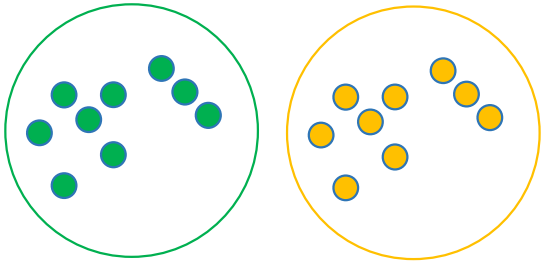
Machine learning (ML) learners:

- ML 101
- ML with python



Database learners:

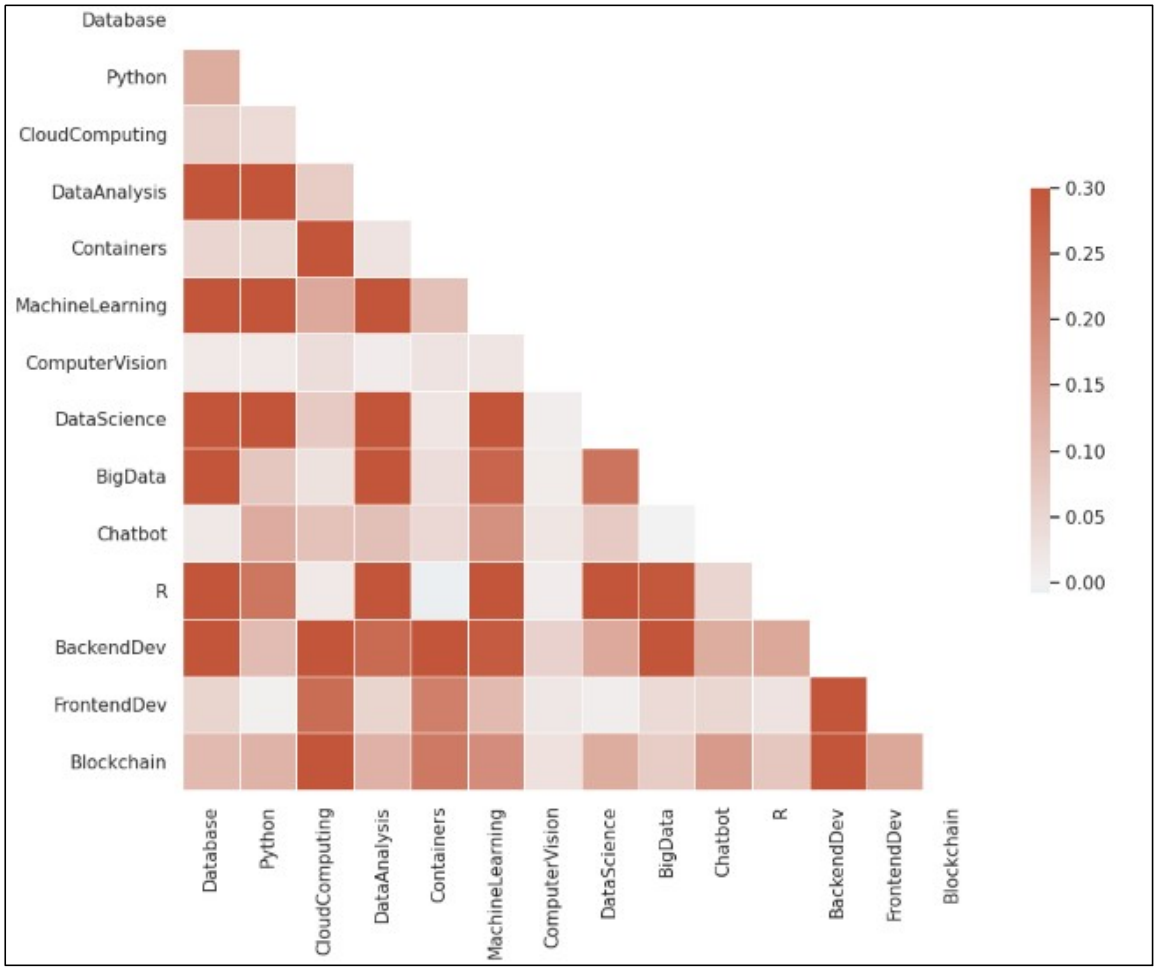
- SQL 101
- SQL with python



Python:

- Python 101
- Python for analysis

Covariance matrix of the user profile feature vectors with 14 features

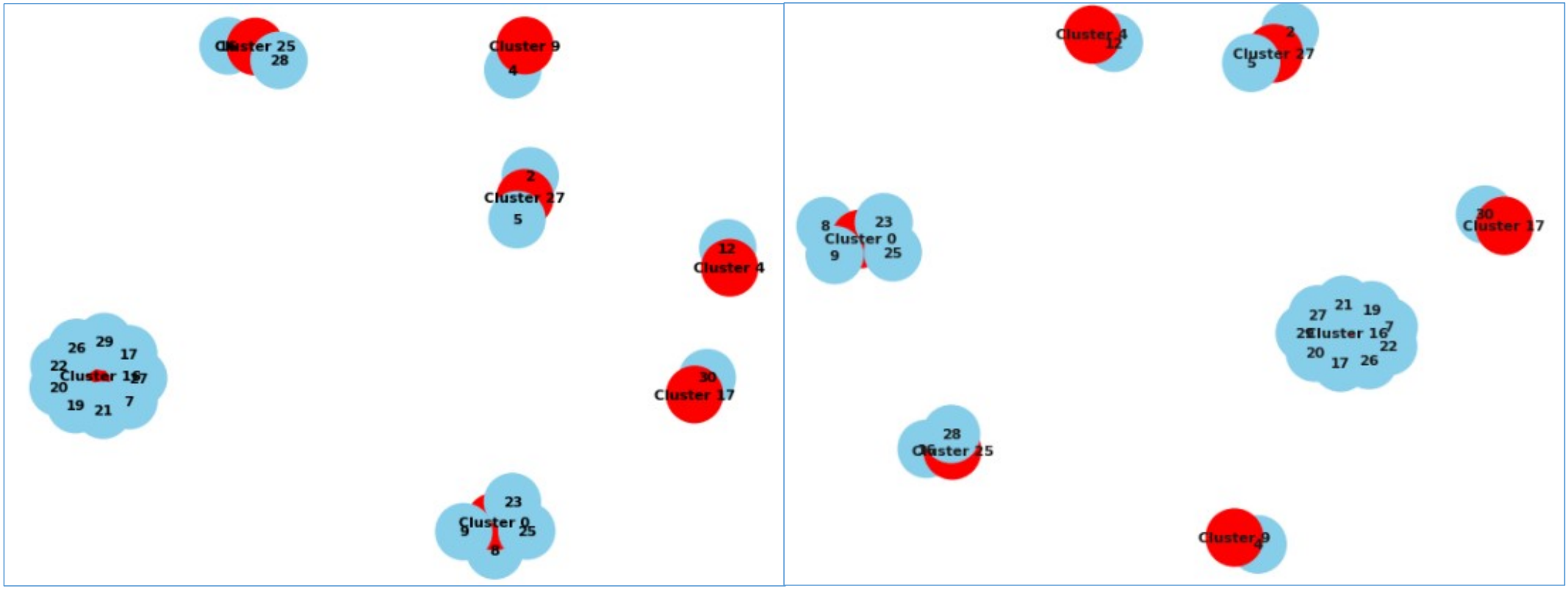


Hot spots shown. Possible to build a recommender system based on cluster.

Evaluation Results

- Top 20 recomended course based on cluster user profiler feature vectors

PCA



Recommendation based on users and courses

User profile 1078030

Participate in 8 courses

- Data analysis
- Deep learning
- Python

User profile 733707

Participate in 23 courses

- Spark
- Sql
- Python

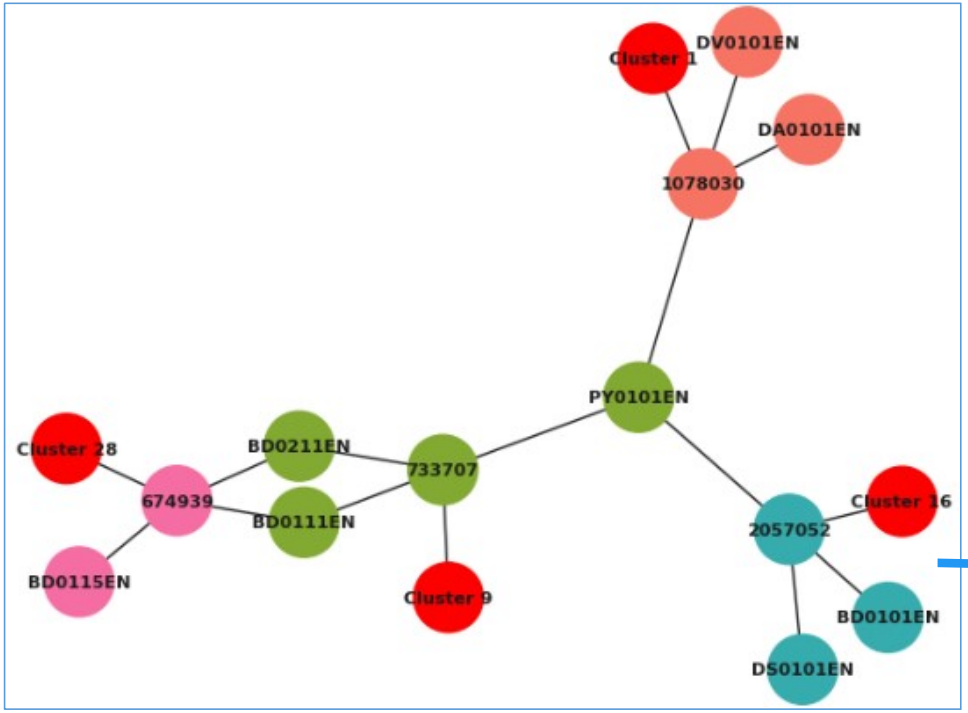
User profile 674939

Participate in 15 courses

- Spark
- Hadoop
- Big data

Course recommendations based on the popular courses in the same cluster

	user	cluster	rec_1	rec_2	rec_3	title_1	title_2	title_3
102	1078030	1	PY0101EN	DA0101EN	DV0101EN	python for data science	data analysis with python	data visualization with python
151	674939	28	BD0111EN	BD0115EN	BD0211EN	hadoop 101	mapreduce and yarn	spark fundamentals i
221	2057052	16	DS0101EN	BD0101EN	PY0101EN	introduction to data science	big data 101	python for data science
298	733707	9	BD0111EN	PY0101EN	BD0211EN	hadoop 101	python for data science	spark fundamentals i



→ Rating 2

Recommendation based on cluster

User profile 1078030

Participate in 8 courses

- Data analysis
- Deep learning
- Python

- Python
- Data analysis
- Data science
- Data visualization

User profile 733707

Participate in 23 courses

- Spark
- Sql
- Python

3 recommended courses

- Python
- Data science
- Hadoop
- Spark

User profile 674939

Participate in 15 courses

- Spark
- Hadoop
- Big data

- Hadoop
- Maproduce
- Spark

Summary

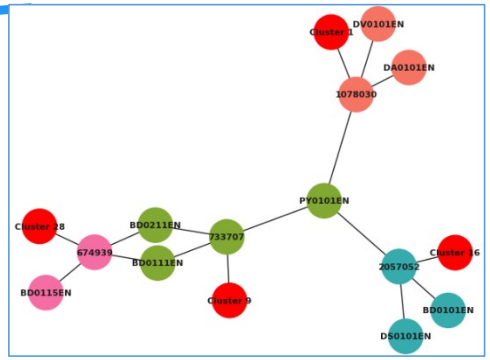
The idea

- **user profile vector**

Cluster data processing

- Cluster assignment and profile

- **cluster profile**



1. **Identify groups of users with similar preferences within the same cluster.**
2. **Users within the same cluster typically share common characteristics or preferences.**
3. **The system focuses on clusters, reducing the complexity from considering every user.**
4. **Recommendations remain relevant from changes.**

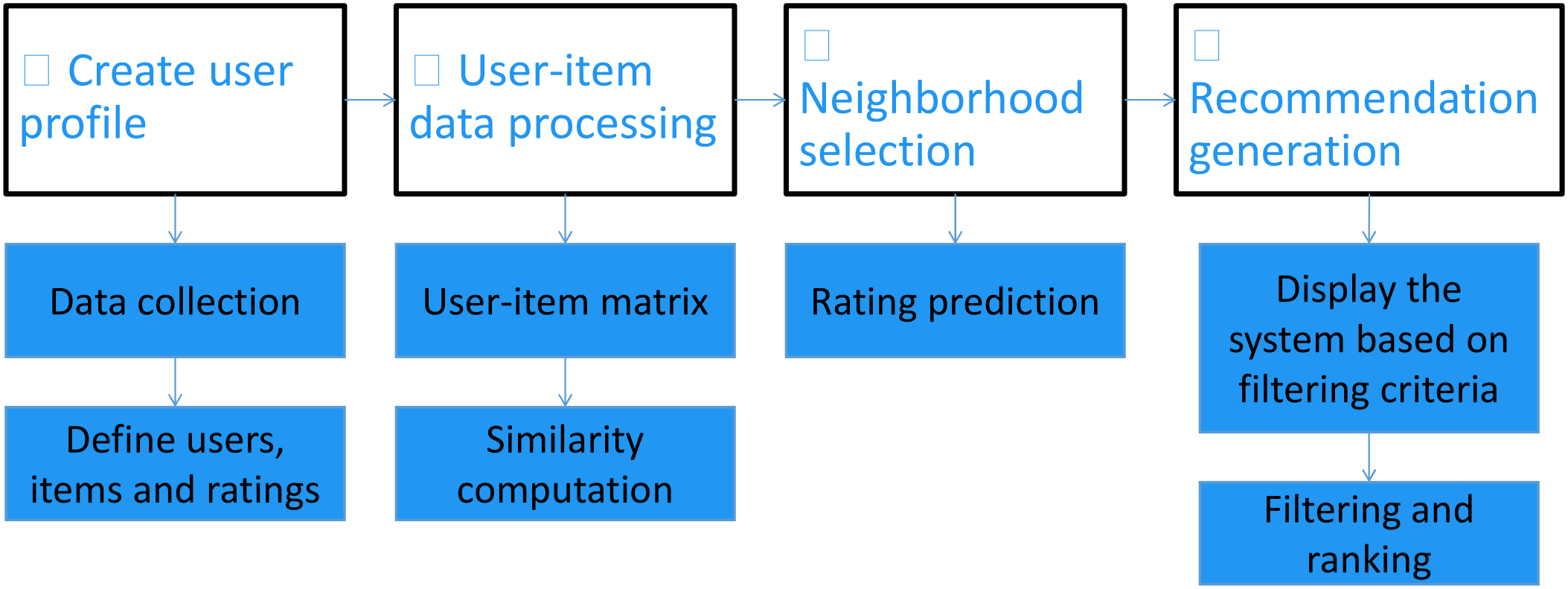
Collaborative-Filtering Recommender System using Supervised Learning

Models and Findings

KKN-Based Collaborative Filtering

Did you complete the slide related to KNN-based collaborative filtering slide? (6 pts)

Flowchart


Create user profile
User-item data processing
Neighborhood selection
Recommendation generation
Data collection
User-item matrix
Rating prediction
Display the system based on filtering criteria
Define users, items and ratings
Similarity computation
Filtering and ranking

Matrix

Collaborative filtering is probably the most commonly used recommendation algorithm, there are two main types of methods:

User-based collaborative filtering is based on the **user similarity or neighborhood**
Item-based collaborative filtering is based on **similarity among items**

User-item matrix

	Machine Learning with Python	Machine Learning 101	Machine Learning Capstone	SQL with Python	Python 101
…	…	…	…	…	…
user2	3.0	3.0	3.0	3.0	3.0
user3	2.0	3.0	3.0	2.0	
user4	3.0	3.0	2.0	2.0	3.0
user5	2.0	3.0	3.0		
user6	3.0	3.0	?		3.0
…	…	…	…	…	…

Predict the rating of the user user6 to item Machine Learning Capstone

Evaluation Results (display 15)

	User	Item	Predicted Rating	TITLE
0	1078030	ML0122ENv1	2.900	accelerating deep learning with gpu
1	1078030	DV0101EN	3.000	data visualization with python
2	733707	DS0101EN	3.000	introduction to data science
3	733707	ML0120EN	3.000	deep learning with tensorflow
4	733707	BD0101EN	3.000	big data 101
5	733707	BD0115EN	3.000	mapreduce and yarn
6	733707	ST0101EN	2.975	statistics 101
7	733707	DB0151EN	2.975	nosql and dbaas 101
8	733707	BD0212EN	3.000	spark fundamentals ii
9	733707	DV0151EN	3.000	data visualization with r
10	733707	ML0101EN	3.000	machine learning with python
11	733707	BD0135EN	3.000	developing distributed applications using zook...
12	674939	BD0141EN	3.000	accessing hadoop data using hive
13	674939	TMP0105EN	2.800	getting started with the data apache spark ma...
14	674939	BD0223EN	3.000	exploring spark s graphx
15	674939	BD0133EN	3.000	controlling hadoop jobs using oozie
16	674939	BD0115EN	3.000	mapreduce and yarn
17	674939	BD0145EN	3.000	sql access for hadoop

User profile 1078030

Participate in 8 courses

- Data analysis
- Deep learning
- Python

User profile 733707

Participate in 23 courses

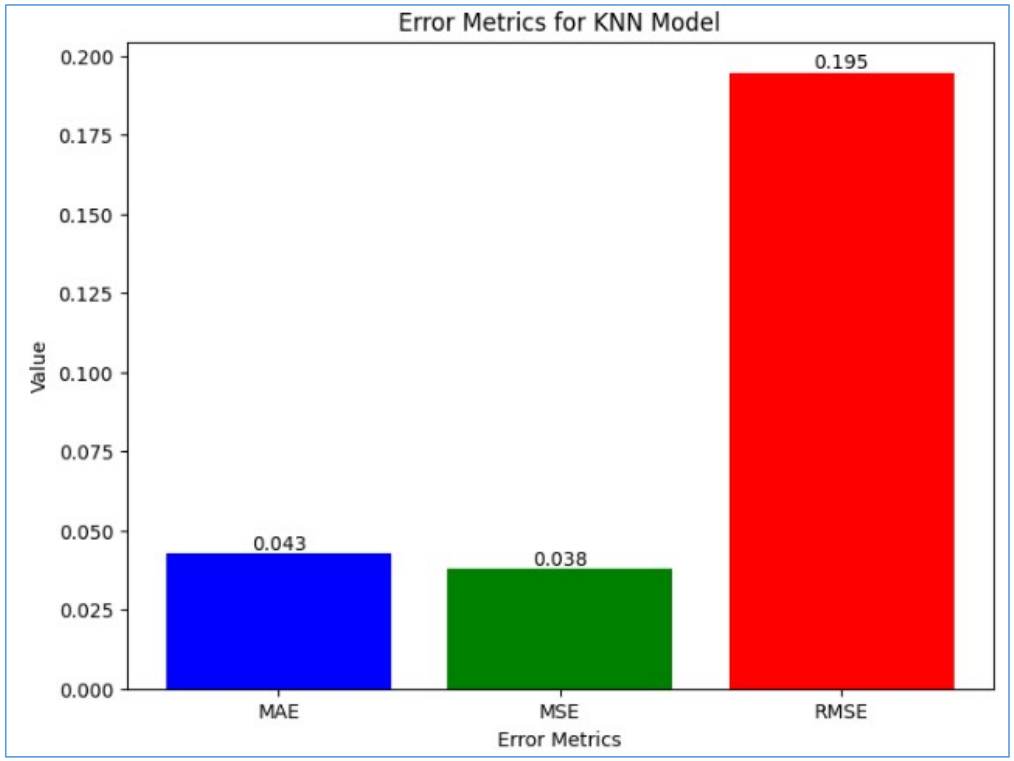
- Spark
- Sql
- Python

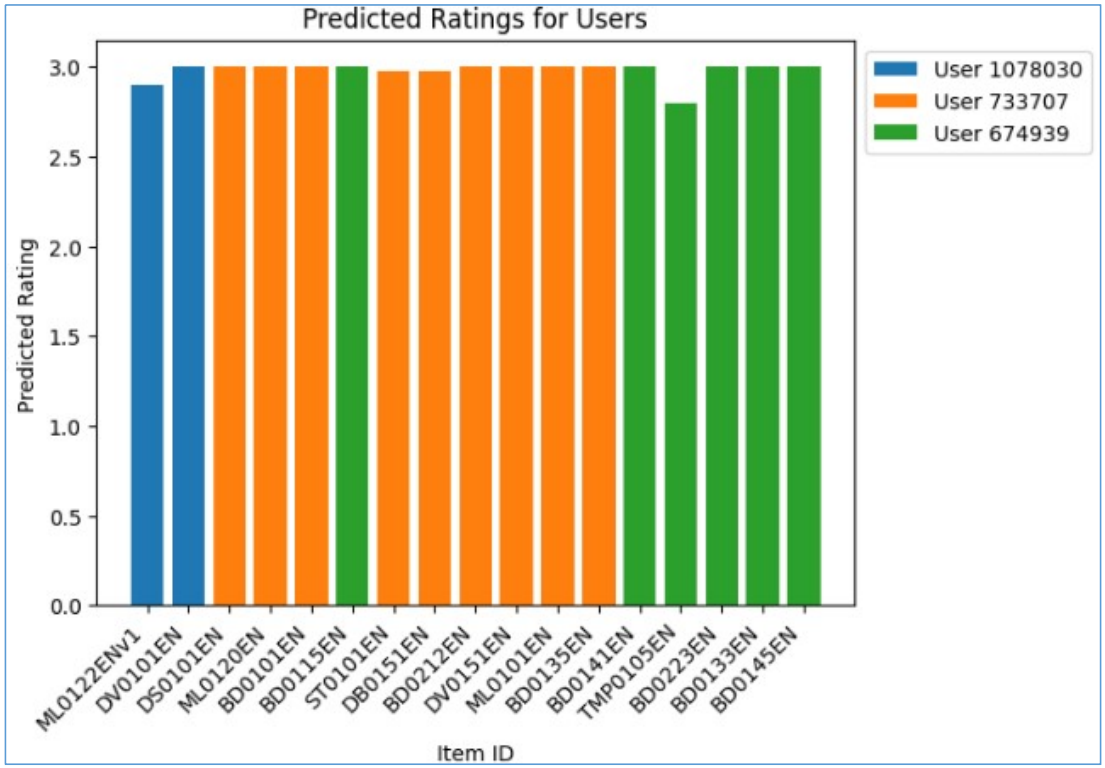
User profile 674939

Participate in 15 courses

- Spark
- Hadoop
- Big data

Evaluation Results





Recommendation based on neighborhood

User profile 1078030

Participate in 8 courses

- Data analysis
- Deep learning
- Python

- Python
- Data visualization

User profile 733707

Participate in 23 courses

- Spark
- Sql
- Python

Recommended courses

- Data science
- Deep learning
- Big data
- Spark
- Data visualization
- Machine learning
- Statistics
- Nosql

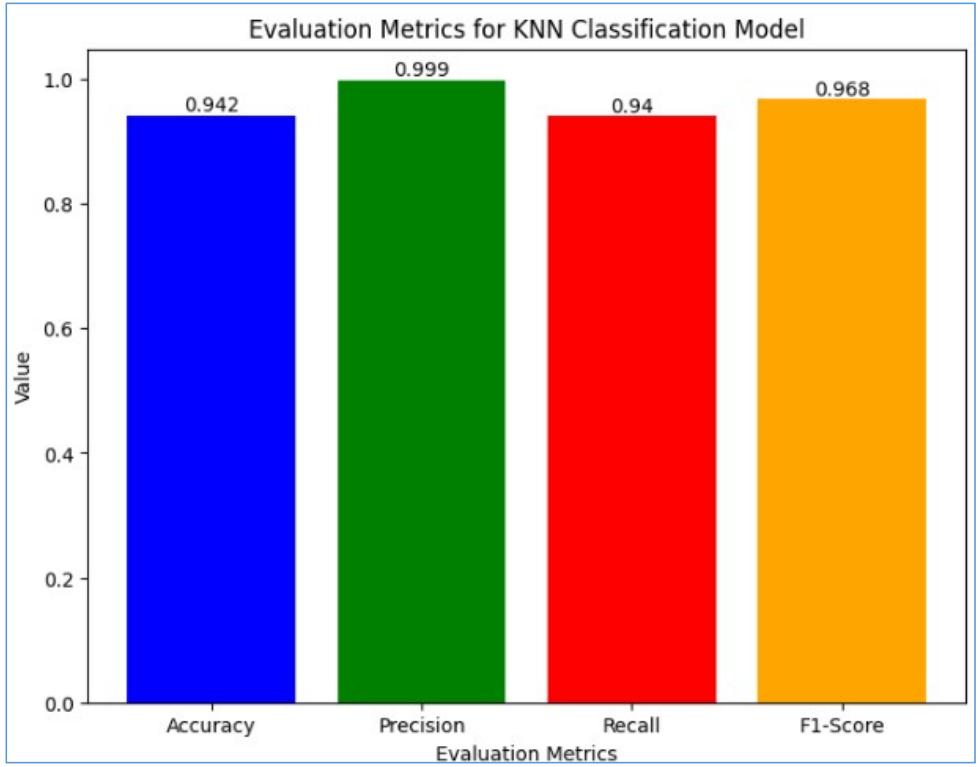
User profile 674939

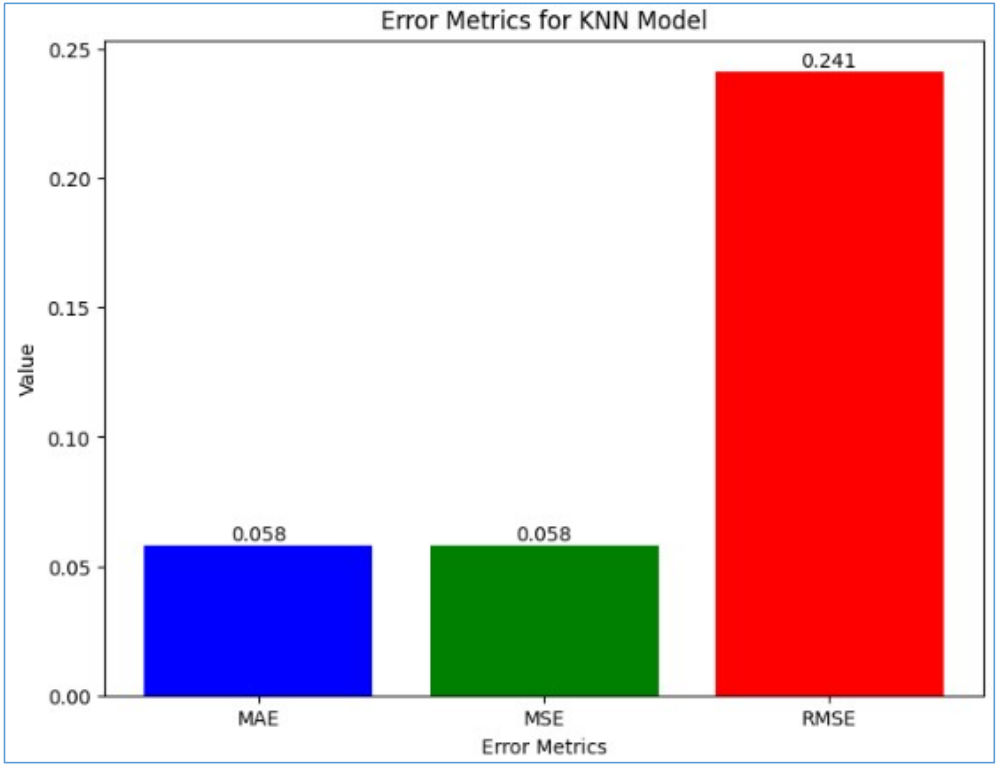
Participate in 15 courses

- Spark
- Hadoop
- Big data

- Spark
- Hadoop

Evaluation Results (binary labels)





Summary

The idea

- Create user profile
- User-item matrix
- Neighborhood selection

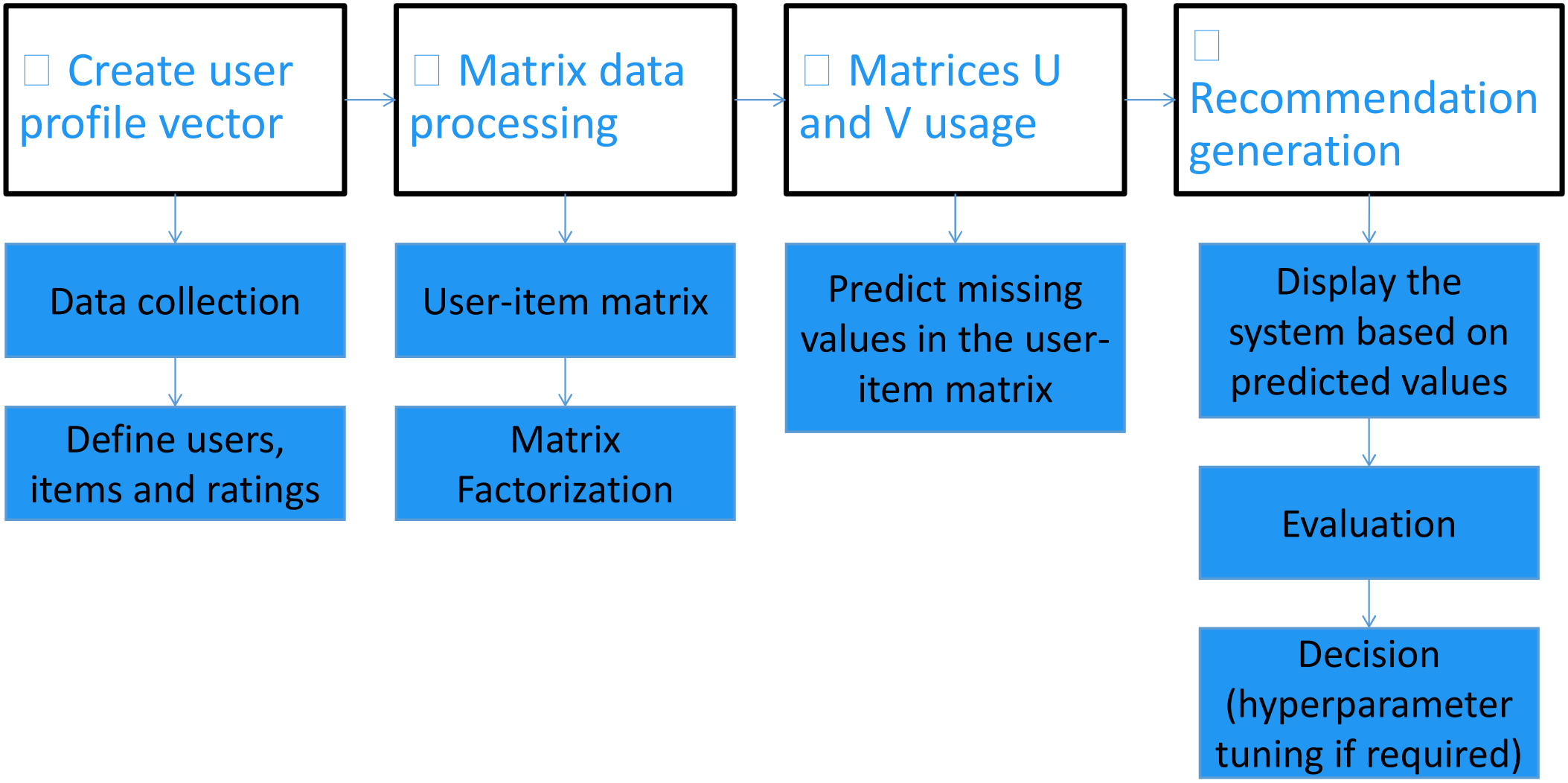
1. **Personalized recommendations by considering the preferences of similar users or items.**
2. **The similarity between users or items, using a straightforward nearest-neighbor approach.**
3. **New users or items based on the preferences of similar entities. Similar users or items are used to infer preferences for new entities.**
4. **Relies on the local neighborhood of users or items, and it can find meaningful connections.**

Models and Findings

NMF-Based Collaborative Filtering

Did you complete the slide related to NMF-based collaborative filtering slide? (6 pts)

Flowchart


Create user profile vector
Matrix data processing
Matrices U and V usage
Recommendation generation
Data collection
Define users, items and ratings
User-item matrix
Matrix Factorization
Predict missing values in the user-item matrix
Display the system based on predicted values
Evaluation
Decision (hyperparameter tuning if required)

Matrix

- **Non-negative matrix factorization (NMF**), decomposes a **big sparse matrix** into two **smaller** and **dense** matrices.
- **User features** and another represents the transformed **item features**.
- Non-negative matrix factorization can be one **solution to big matrix issues**.

Non-negative Matrix Factorization

User-item matrix: A 1000 x 100

	Item1	ML101	...	Item100
...	...	...	...	...
user2	3.0	3.0	3.0	3.0
user3	2.0	3.0	2.0	
user4	3.0	3.0	2.0	3.0
user5	2.0	3.0		
user6	3.0	3.0		3.0
...	...	...	...	...

≈

User matrix: U 1000 x 16

	Feature1	Feature2	...	Feature16
...	...	...	...	...
user2	....	....	....	....
user3	....	....	....	....
user4	....	....	....	....
user5	....	....	....	....
user6	....	....	....	....
...	...	...	...	...

X

Item matrix: I 16 x 100

	Item1	ML101	...	Item100
...	...	...	...	...
Feature1	....	....	....	....
Feature2	....	....	....	....
Feature3	....	....	....	....
....	....	....	....	....
Feature16	....	....	....	....
...	...	...	...	...

Evaluation Results

	User	Item	Predicted_Rating	TITLE
0	1078030	ML0122ENv1	2.798383	accelerating deep learning with gpu
1	1078030	DV0101EN	2.982449	data visualization with python
2	733707	DS0101EN	2.998862	introduction to data science
3	733707	ML0120EN	3.000000	deep learning with tensorflow
4	733707	BD0101EN	2.975825	big data 101
5	733707	BD0115EN	2.975243	mapreduce and yarn
6	733707	ST0101EN	2.963512	statistics 101
7	733707	DB0151EN	3.000000	nosql and dbaas 101
8	733707	BD0212EN	2.982911	spark fundamentals ii
9	733707	DV0151EN	2.896570	data visualization with r
10	733707	ML0101EN	2.898463	machine learning with python
11	733707	BD0135EN	2.979971	developing distributed applications using zook...
12	674939	BD0141EN	3.000000	accessing hadoop data using hive
13	674939	TMP0105EN	2.956095	getting started with the data apache spark ma...
14	674939	BD0223EN	2.965433	exploring spark s graphx
15	674939	BD0133EN	2.851486	controlling hadoop jobs using oozie
16	674939	BD0115EN	3.000000	mapreduce and yarn
17	674939	BD0145EN	3.000000	sql access for hadoop

User profile 1078030

Participate in 8 courses

- **Data analysis**
- **Deep learning**
- **Python**

User profile 733707

Participate in 23 courses

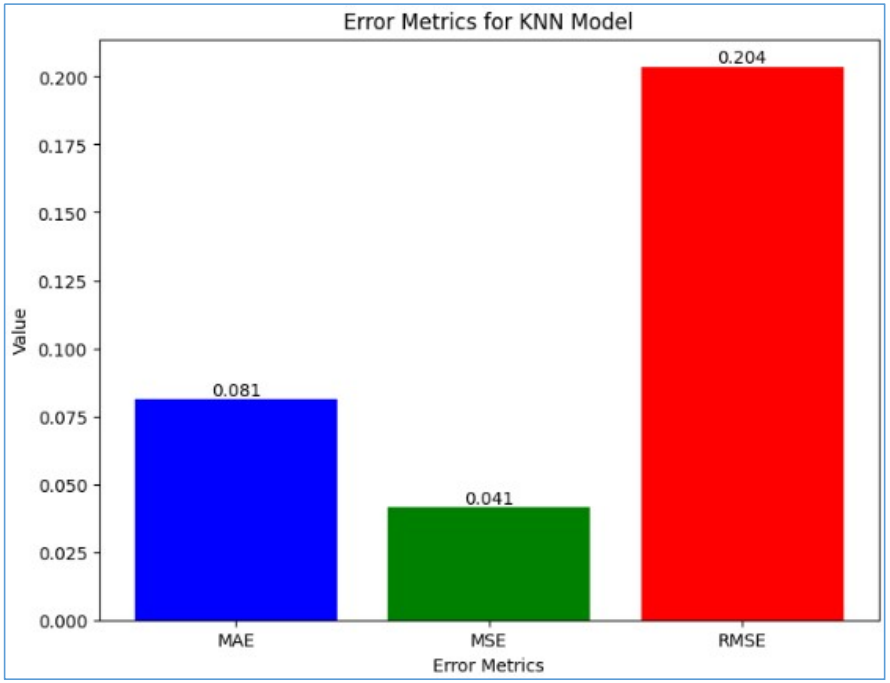
- **Spark**
- **Sql**
- **Python**

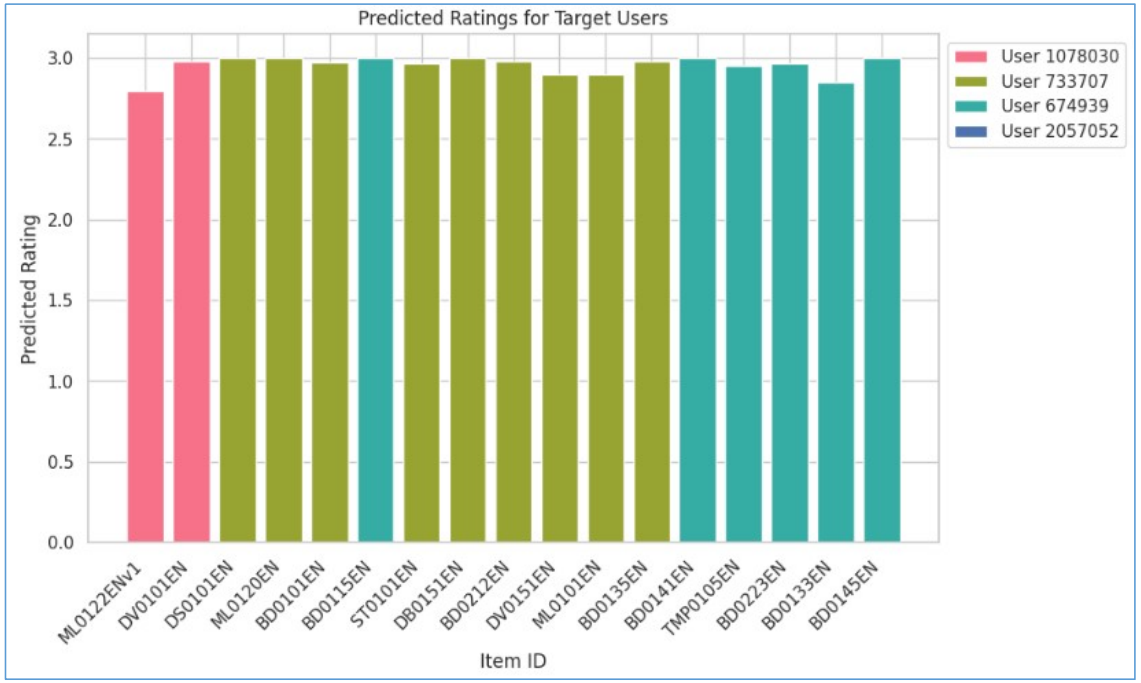
User profile 674939

Participate in 15 courses

- **Spark**
- **Hadoop**
- **Big data**

Evaluation Results





Recommendation based on NMF

User profile 1078030

Participate in 8 courses

- Data analysis
- Deep learning
- Python

Recommended courses

- Data visualization
- Deep learning

User profile 733707

Participate in 23 courses

- Spark
- Sql
- Python

Recommended courses

- Data science
- Deep learning
- Big data
- Spark
- Data visualization
- Machine learning
- Statistics
- Nosql

User profile 674939

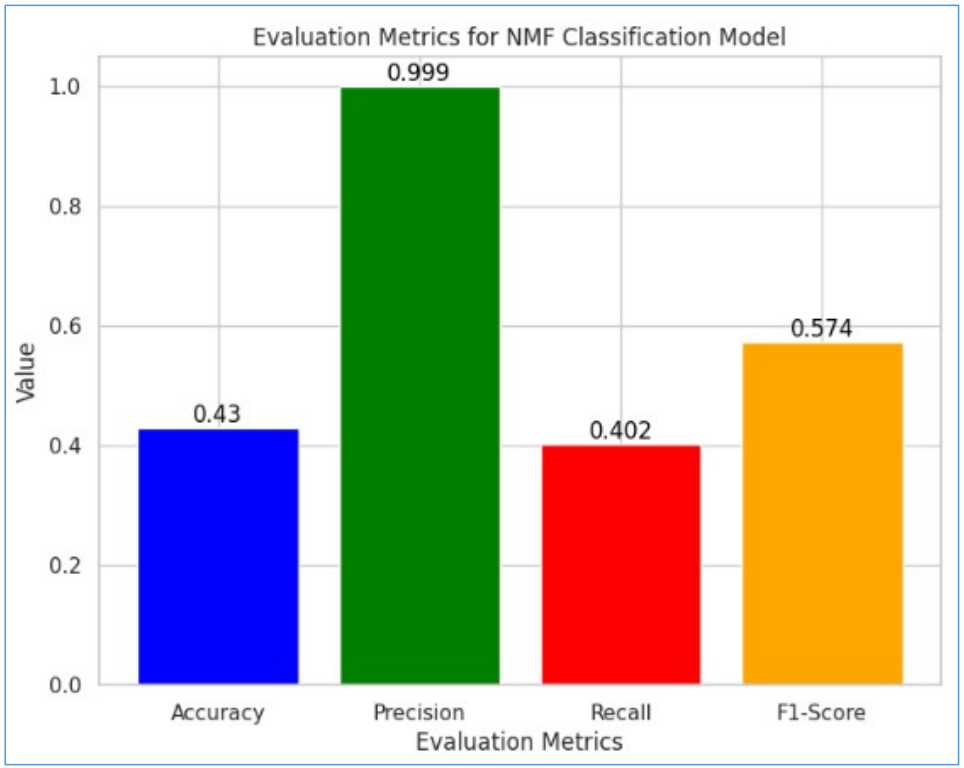
Participate in 15 courses

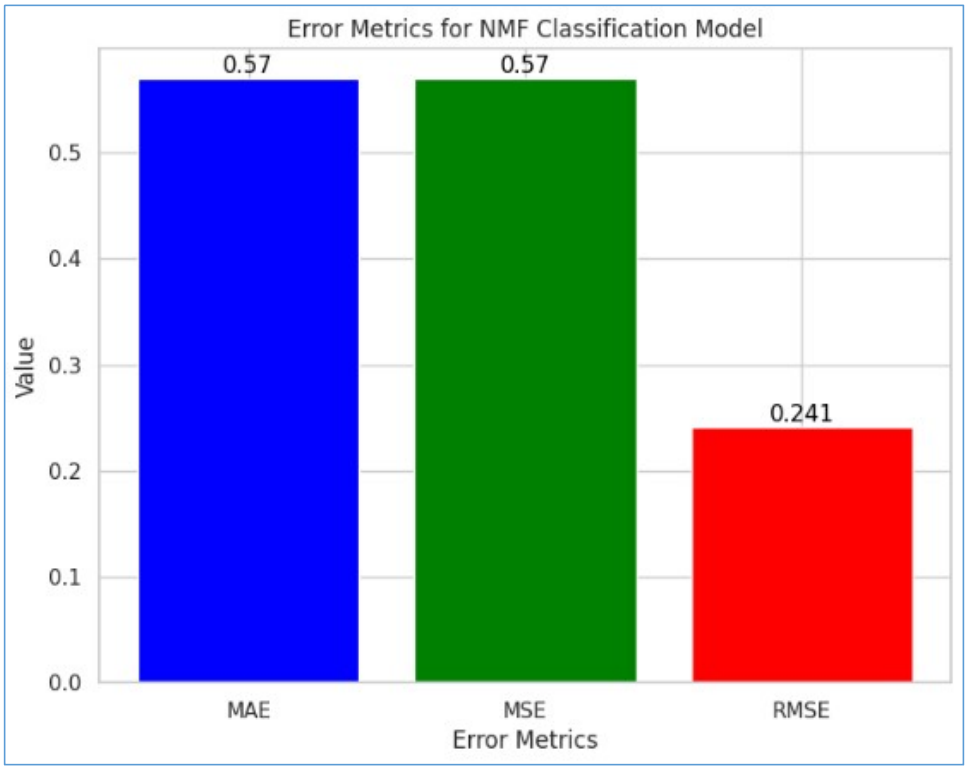
- Spark
- Hadoop
- Big data

Recommended courses

- Spark
- Hadoop
- Apache
- Sql

Evaluation Results (binary labels)





Summary

The idea

- Create user profile vector
- Matrix Factorization
- Matrix data processing
- Matrices U and V usage (Predict missing values)

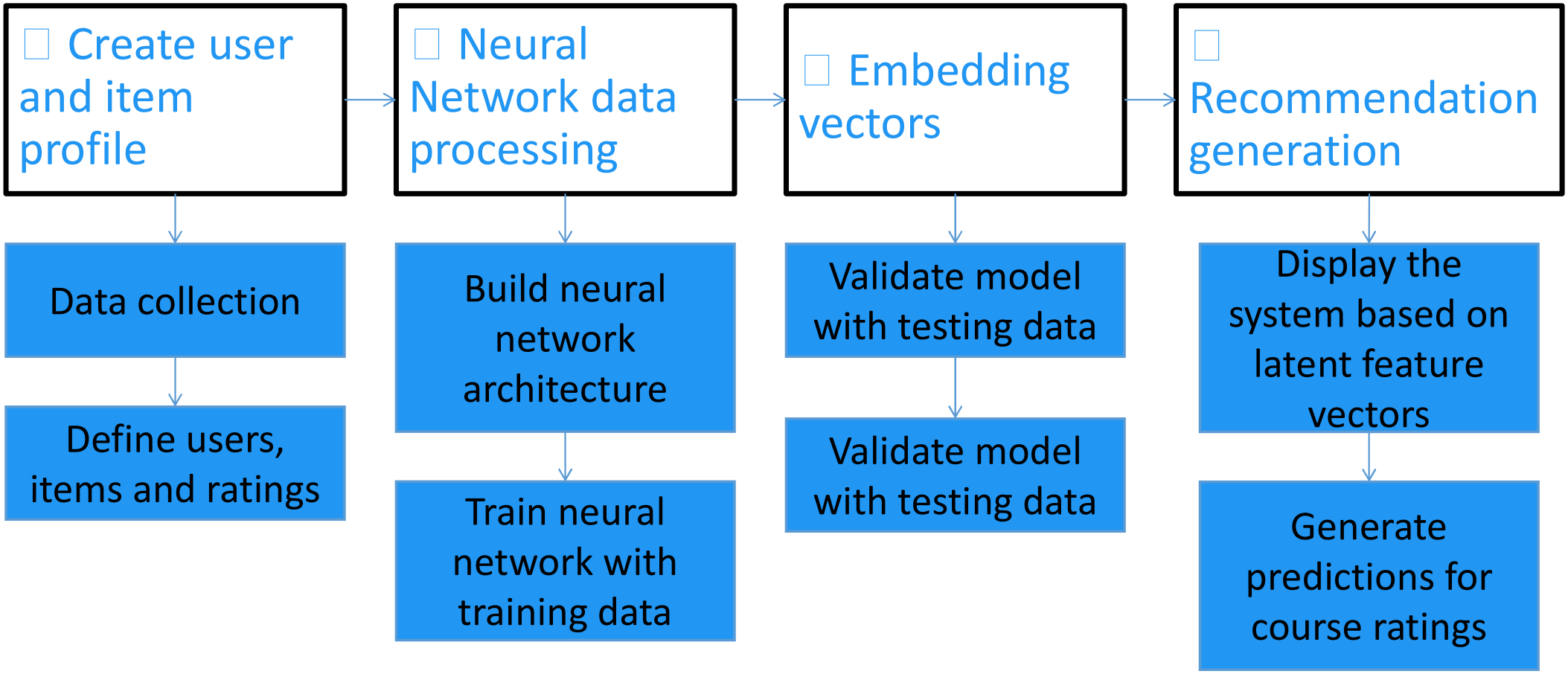
1. **Factorized matrices with non-negative values, providing an interpretable representation of users and items.**
2. **Relies on the underlying patterns and features present in the user-item interaction matrix.**
3. **New users or items based on the preferences of similar entities.Similar users or items are used to infer preferences for new entities.**
4. **Relies on the local neighborhood of users or items, and it can find meaningful connections.**

Models and Findings

Neural Network Embedding-Based Collaborative Filtering

Did you complete the slide related to Neural Network Embedding based collaborative filtering slide? (6 pts)

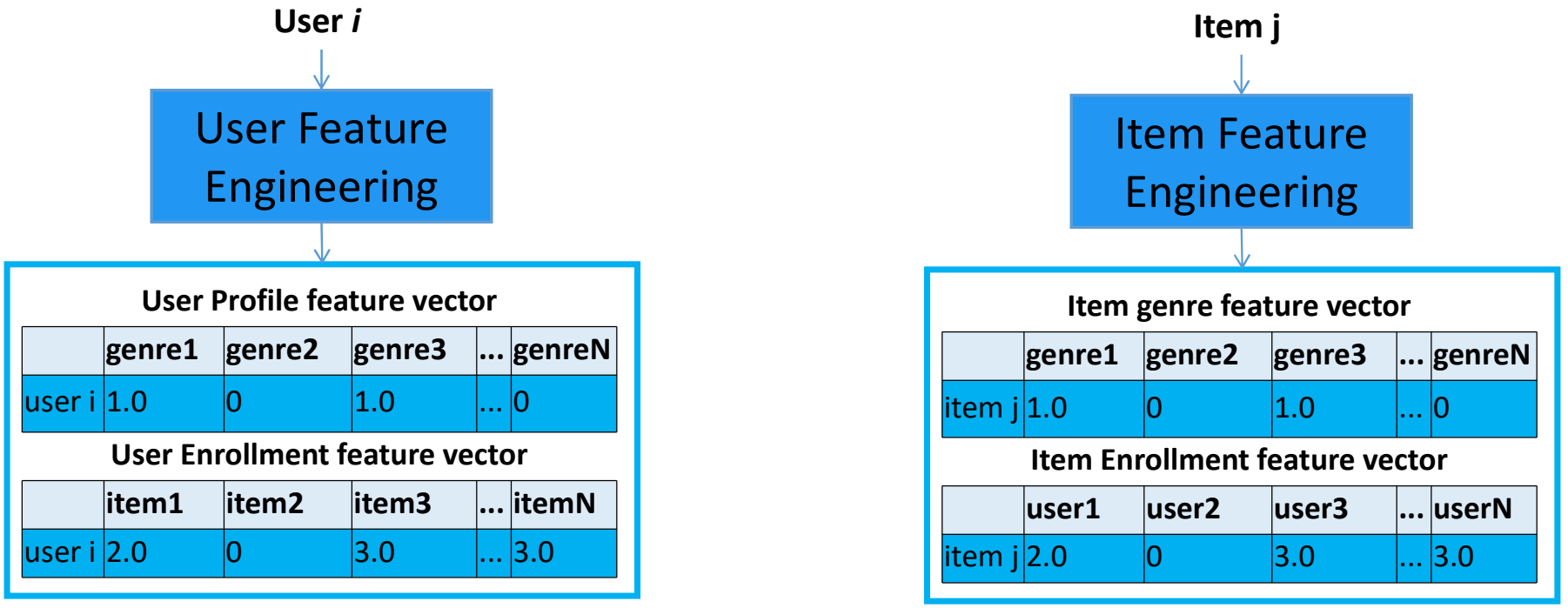
Flowchart


Create user and item profile
Neural Network data processing
Embedding vectors
Recommendation generation
Data collection
Define users, items and ratings
Build neural network architecture
Train neural network with training data
Validate model with testing data
Validate model with testing data
Display the system based on latent feature vectors
Generate predictions for course ratings

Matrix

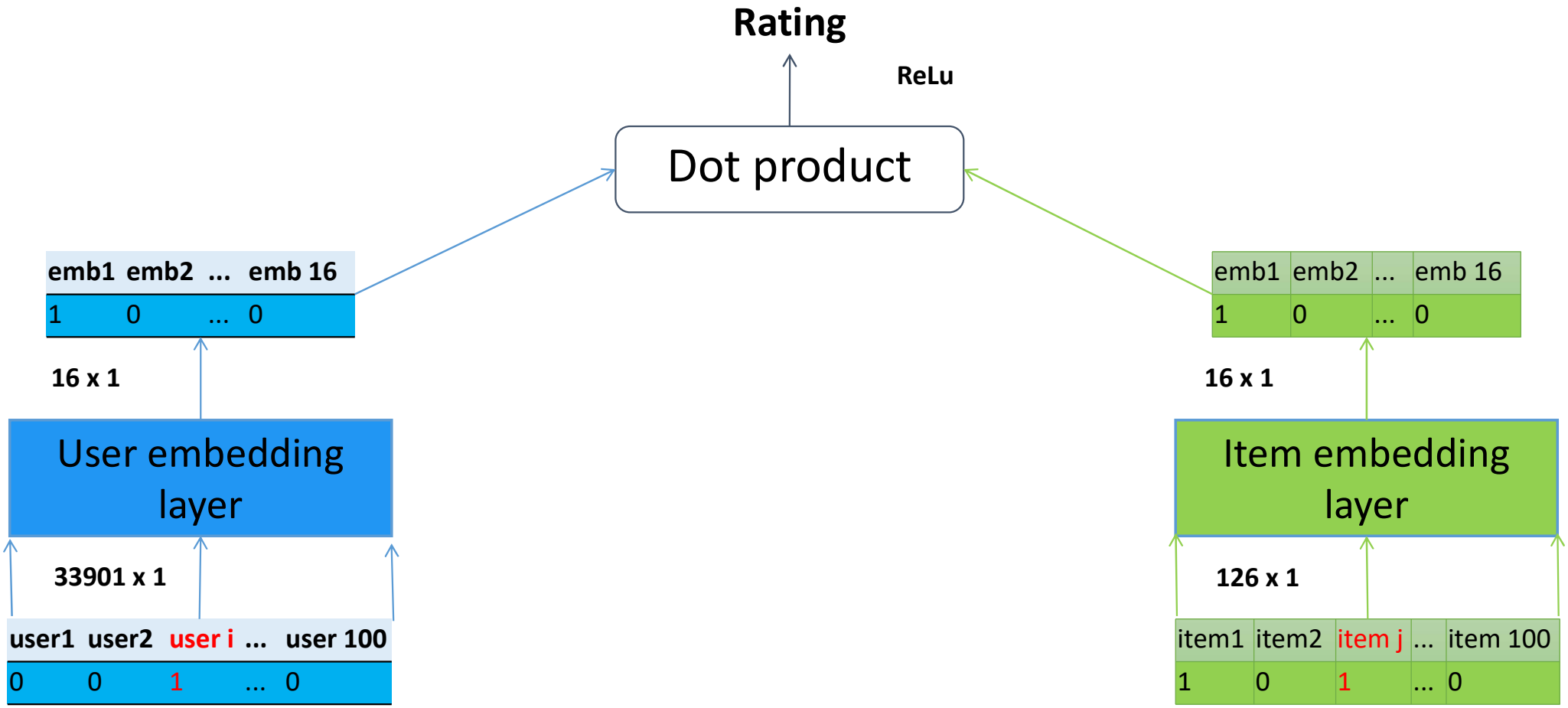
- Neural networks are very good at learning patterns from data and are widely used to extract latent features.
- Gradually captures and stores the features within its hidden layers as weight matrices and can be extracted to represent the original data.

Explicit User and Item Feature Engineering

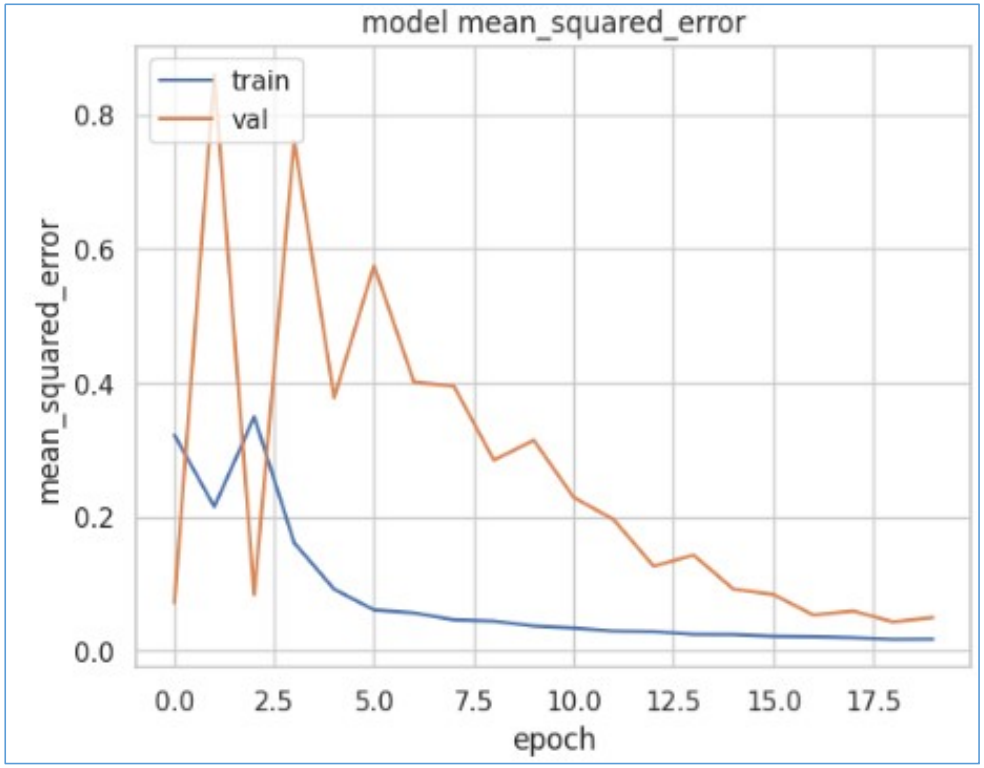


Latent feature vectors

Predict the **user-item interactions** while **simultaneously** extracting the **user** and **item embedding features**.

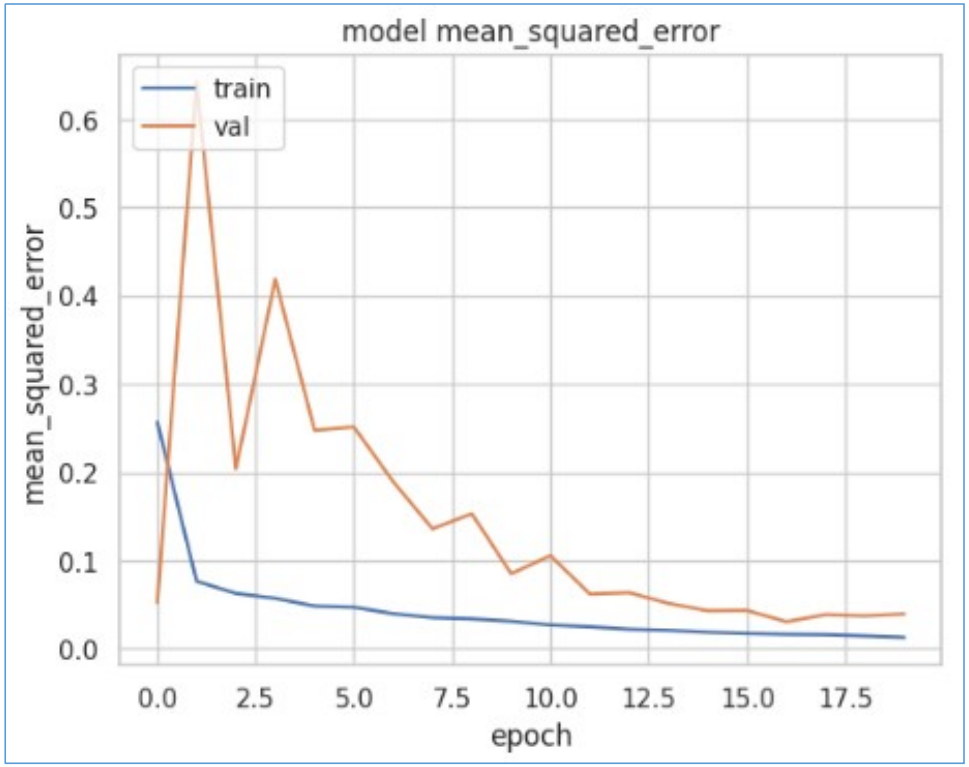


Evaluation Results



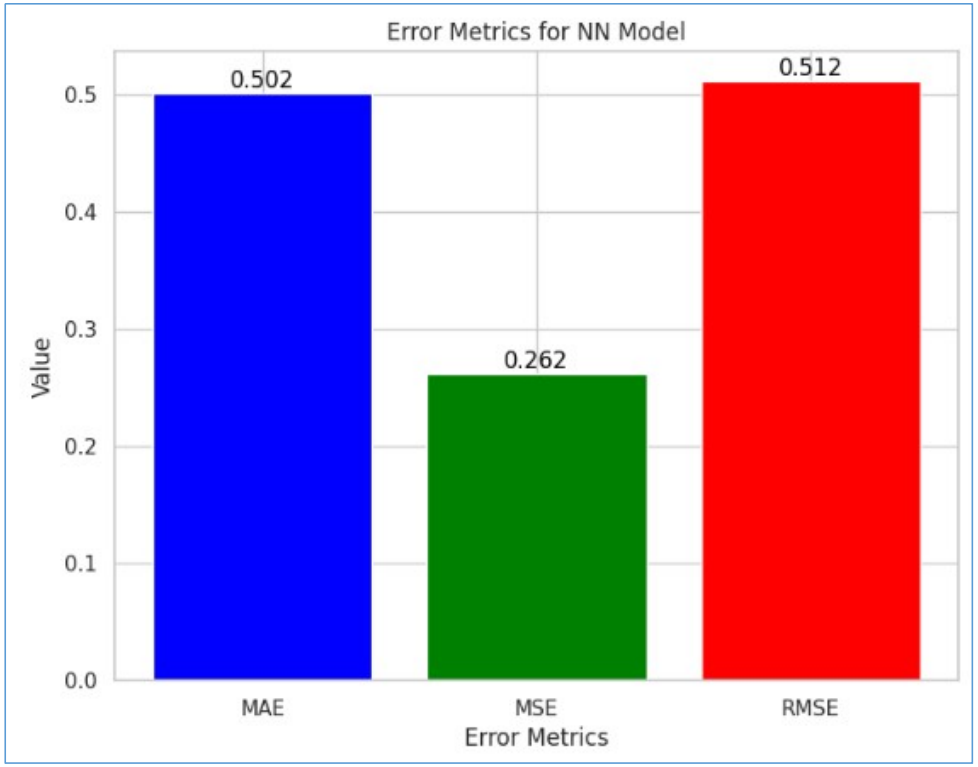

```
Model: "recommender_net"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 user_embedding_layer (Embed  multiple                 542416
 ding)

 user_bias (Embedding)       multiple                  33901

 item_embedding_layer (Embed  multiple                 2016
 ding)

 item_bias (Embedding)       multiple                  126

=================================================================
Total params: 578,459
Trainable params: 578,459
Non-trainable params: 0
_________________________________________________________________
```

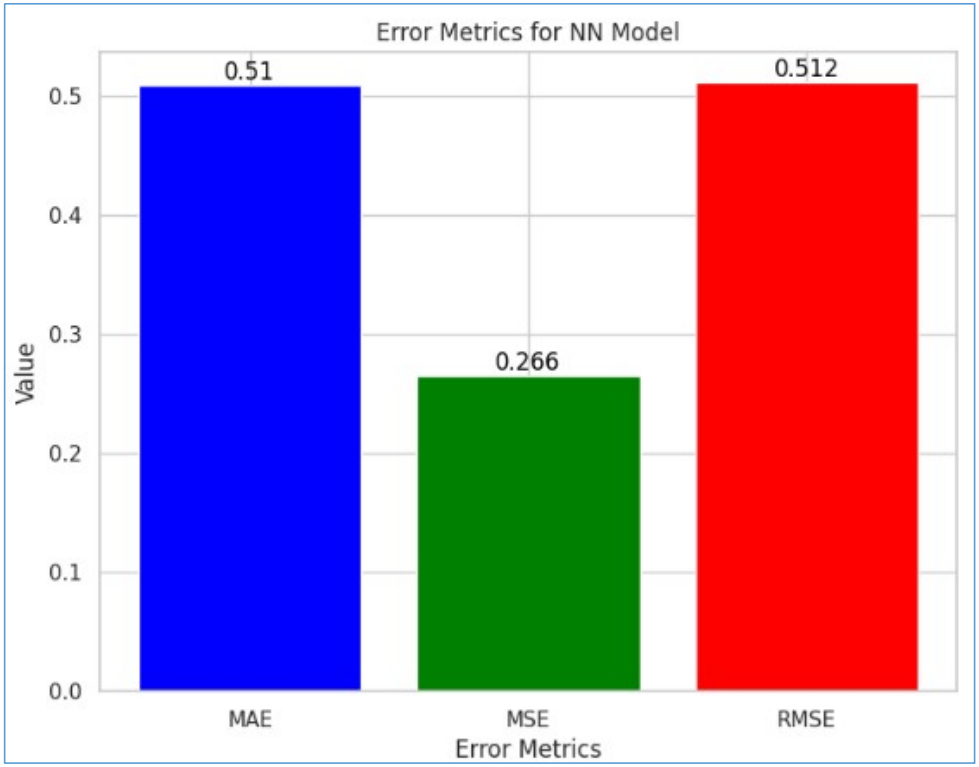
Evaluation Results (Improve performance)




```
Model: "recommender_net_plus"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 user_embedding_layer (Embed  multiple                 339010
 ding)

 user_bias (Embedding)       multiple                  33901

 item_embedding_layer (Embed  multiple                 1260
 ding)

 item_bias (Embedding)       multiple                  126

=================================================================
Total params: 374,297
Trainable params: 374,297
Non-trainable params: 0
_________________________________________________________________
```

Evaluation Results (3rd)





(Improve performance)

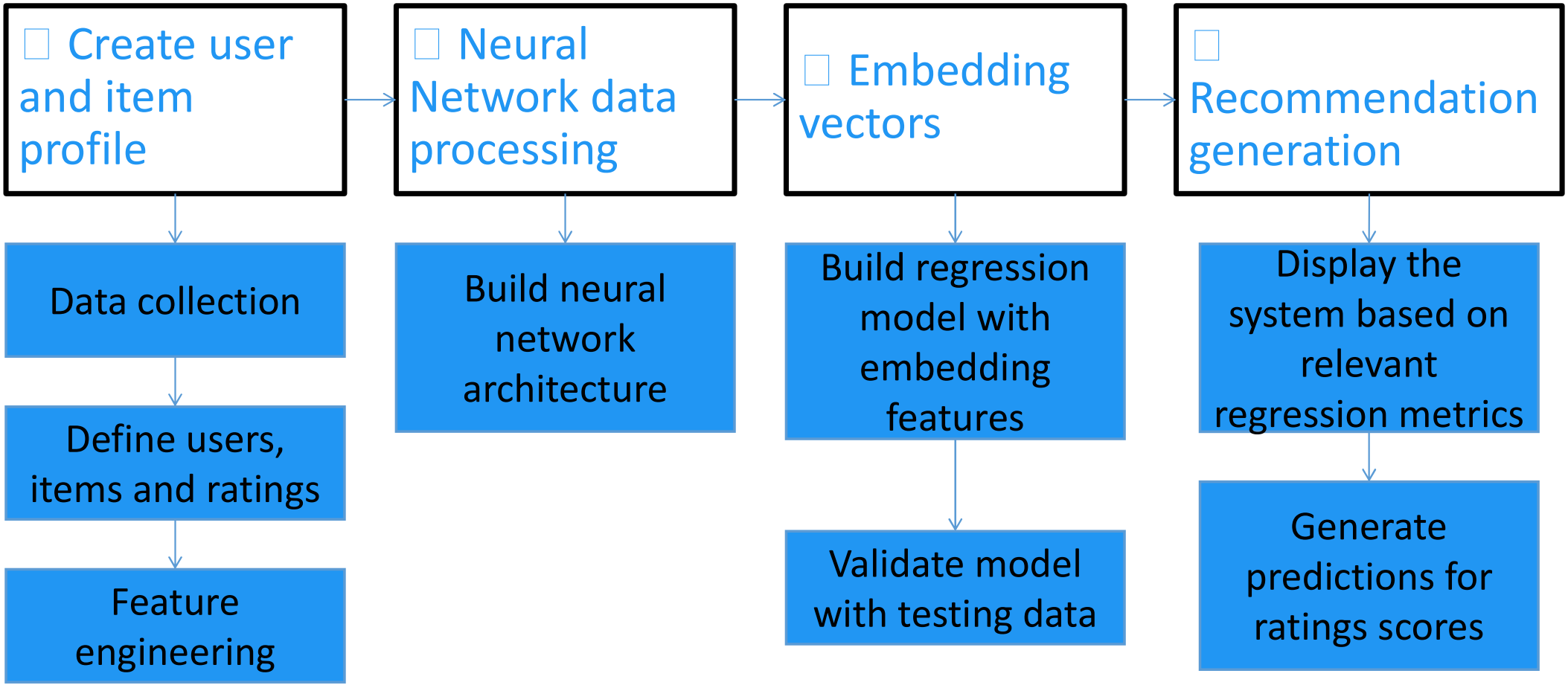
Summary

The idea

- Create user and item profile
- Neural network processing
- Embedding vectors

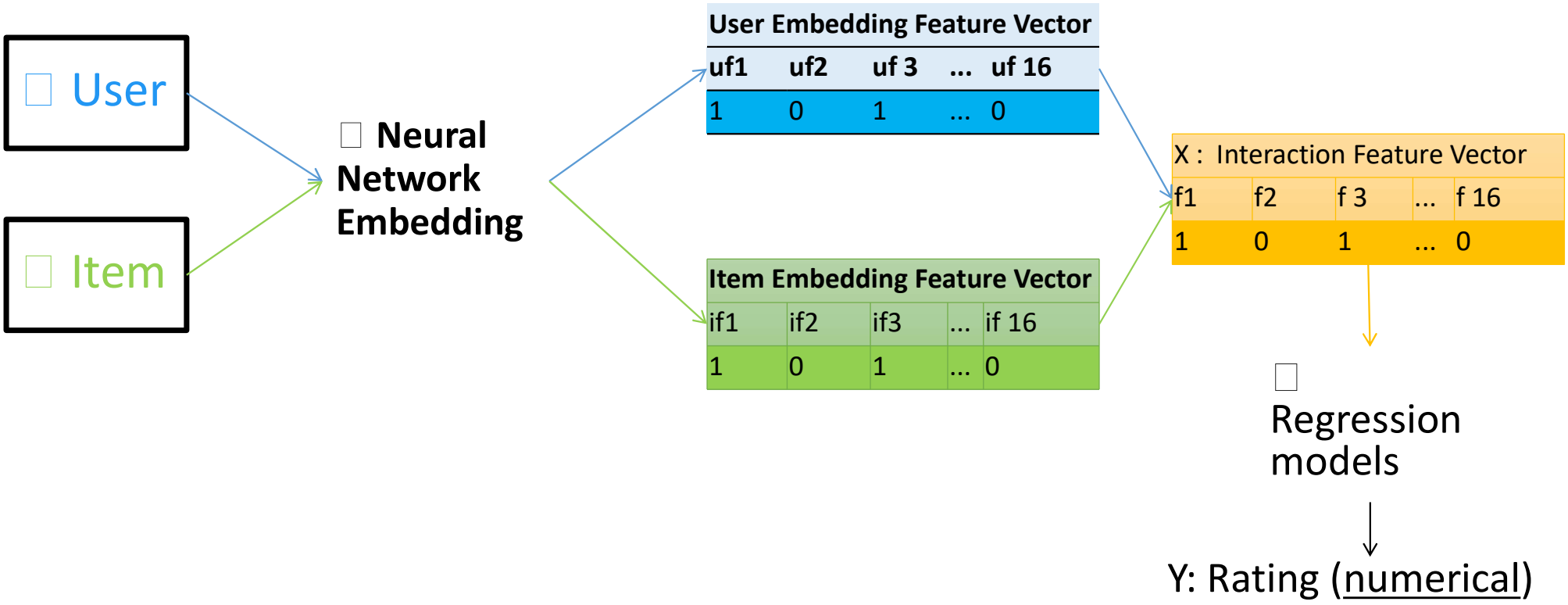
1. **Capture complex, non-linear relationships in course-rating data, allowing for more accurate modeling of user preferences.**
2. **The learned embeddings can capture latent factors and provide insights into the underlying features influencing course ratings.**
3. **Handling implicit feedback, such as user interactions and engagement, which may not be explicitly rated.**

Flowchart Regression Based



Neural Networks using Embedding Features

In the neural network, **extends** this by using **two embedding vectors as an input** into a **Neural Network** to **predict the rating**.

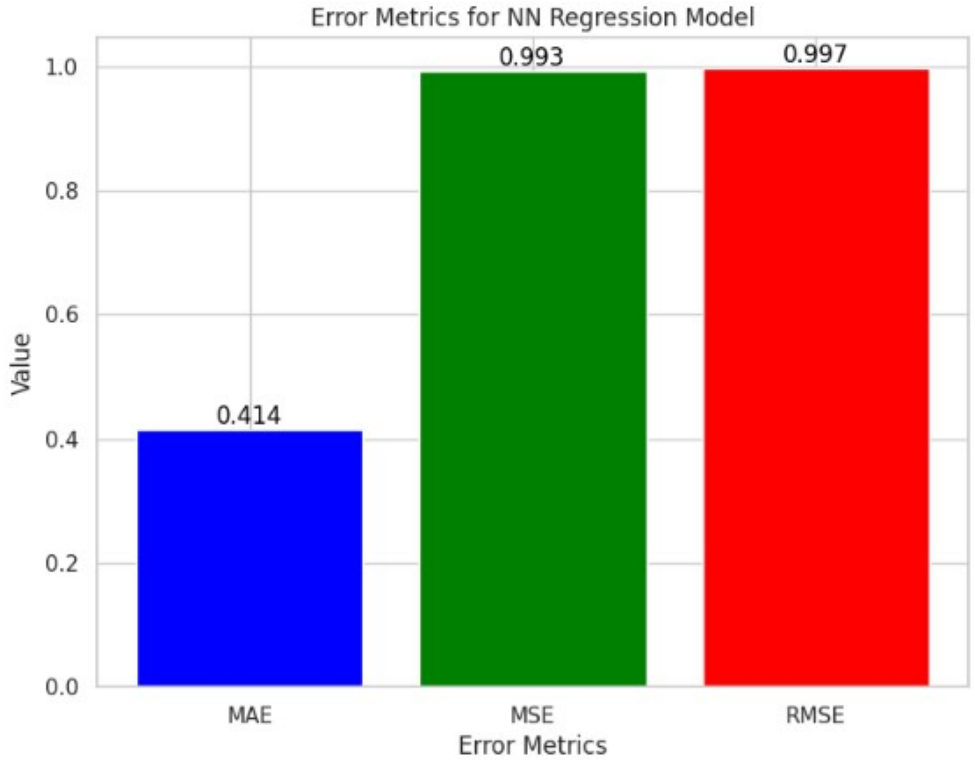


Evaluation Results (3rd)

MAE: 0.41428838083033687

MSE: 0.9932500760760065

RMSE: 0.9966193235513781



Summary

The idea

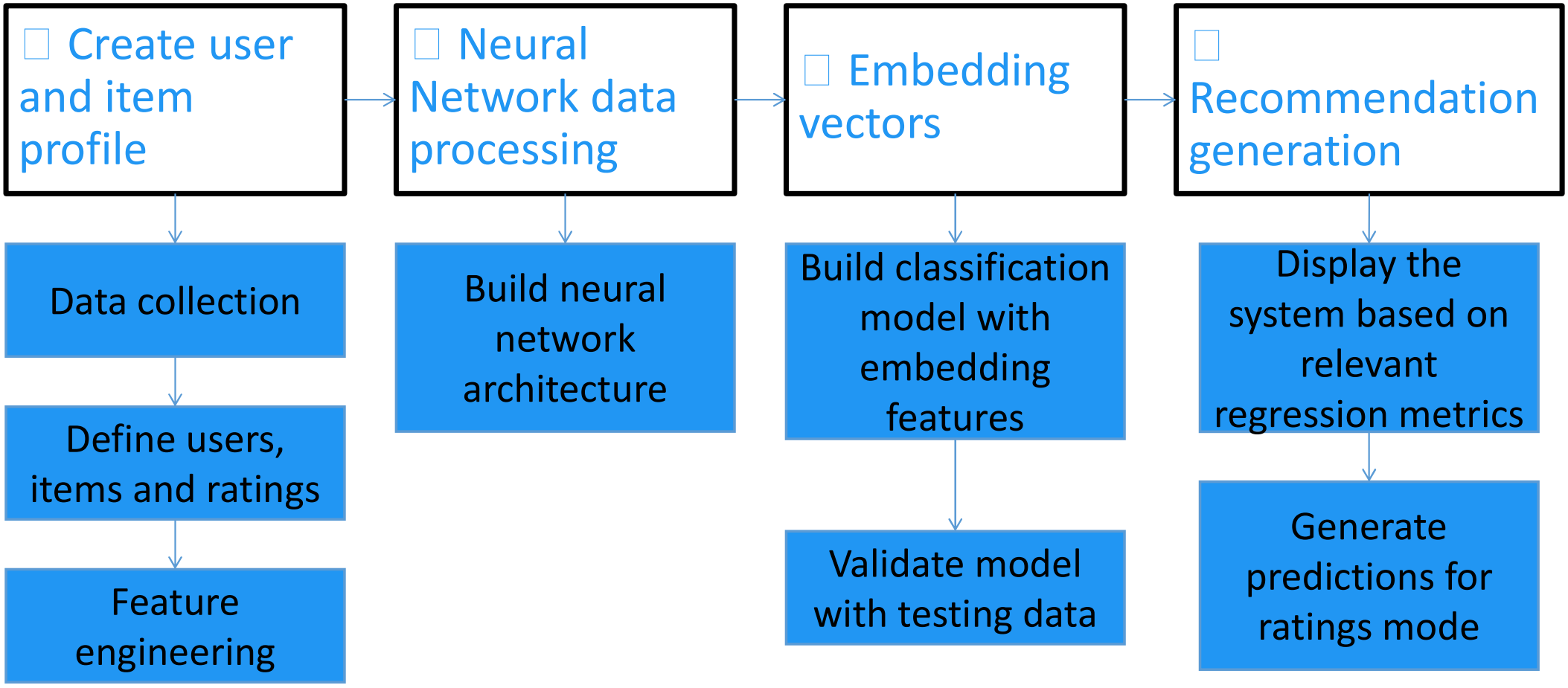
1. **Capture latent factors and relationships that contribute to the prediction of rating scores.**
2. **The embedding features implicitly learn latent factors without the need for explicit feature engineering. A more accurate representations of complex relationships in the data.**
3. **The learned embeddings enable the model to understand underlying patterns that contribute to rating scores.**
4. **Provide a dense representation that captures similarities between courses and users.**
5. **Users and courses with similar embeddings share common features, aiding interpretability.**

Models and Findings

Collaborative Filtering Algorithms Evaluation

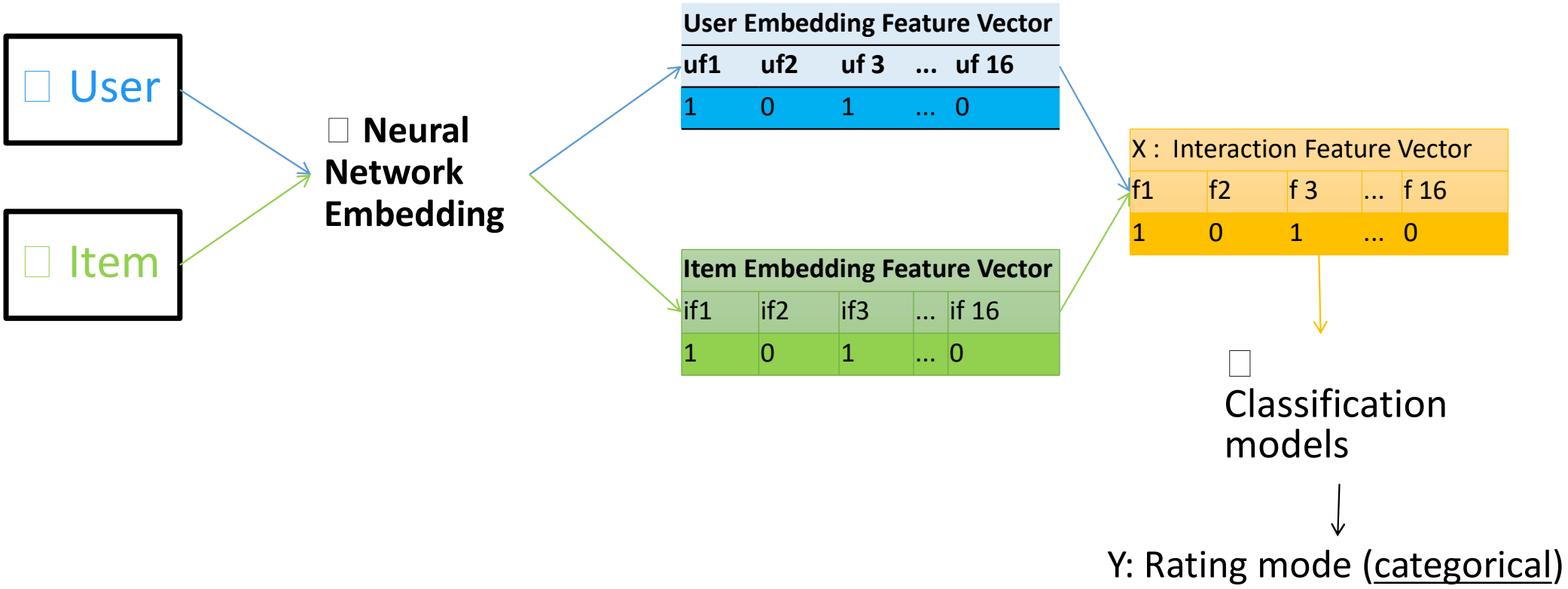
Did you complete the slide comparing the performance of the collaborative filtering models? (6 points)?

Flowchart Classification Based

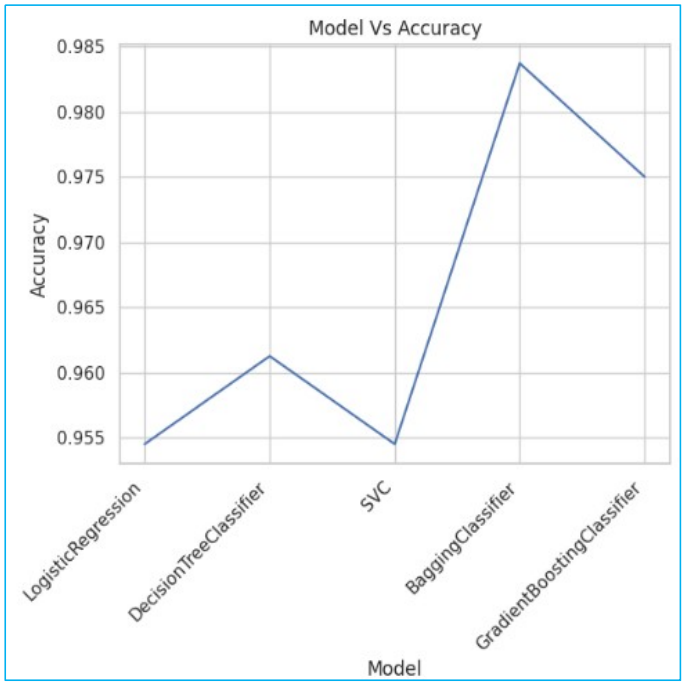


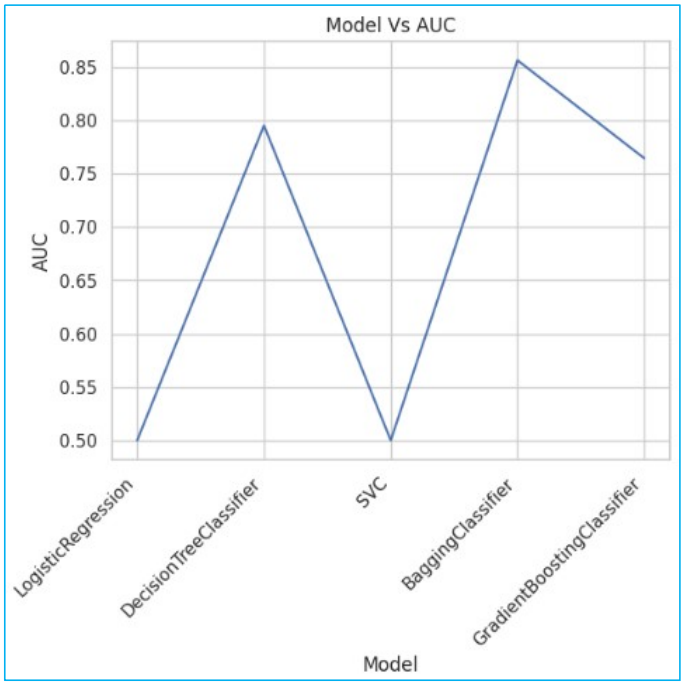
Neural Networks using Embedding Features

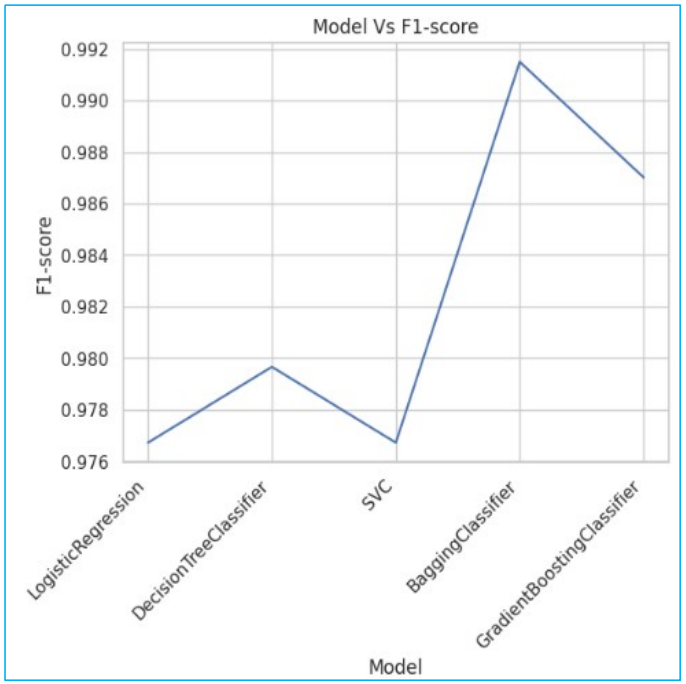
The prediction problem as **a classification problem** as rating only has **two categorical values (Aduit vs. Completion)**



Evaluation Results







	Model	Accuracy	Precision	Recall	F1-Score	AUC
0	LogisticRegression	0.954503	0.954503	1.000000	0.976722	0.500000
1	DecisionTreeClassifier	0.961253	0.981454	0.977885	0.979666	0.795113
2	SVC	0.954503	0.954503	1.000000	0.976722	0.500000
3	BaggingClassifier	0.983713	0.986596	0.996475	0.991511	0.856221
4	GradientBoostingClassifier	0.974990	0.978129	0.996071	0.987018	0.764404

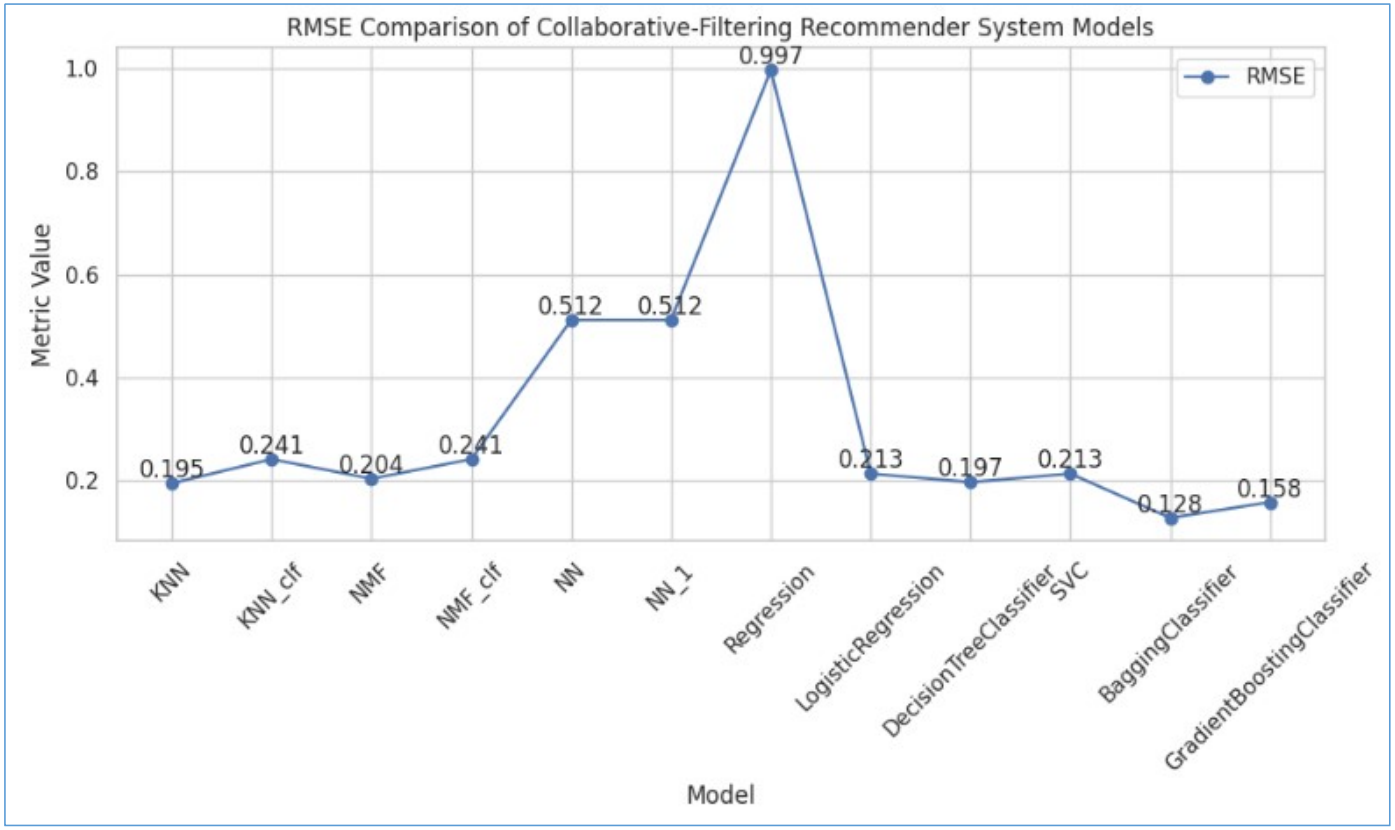
Summary

The idea

- Create user and item profile
- Build classification model
- Neural network processing
- Embedding vectors

1. **Provide insights into which features contribute to a specific rating class.**
2. **Easier to understand the distinctions between various user preferences.**
3. **Classification is well-suited for scenarios where the ratings are discrete and categorical, such as a system where users provide ratings on a scale (e.g., 1 to 5 stars).**

Summary



	Model	RMSE
0	KNN	0.194618
1	KNN_clf	0.241409
2	NMF	0.203553
3	NMF_clf	0.241409
4	NN	0.511761
5	NN_1	0.511761
6	Regression	0.996619
7	LogisticRegression	0.213301
8	DecisionTreeClassifier	0.196842
9	SVC	0.213301
10	BaggingClassifier	0.127622
11	GradientBoostingClassifier	0.158144

Conclusion

Did you complete the conclusion slide? (6 pts)
Did you apply your creativity to improve the presentation beyond the template? (4 pts)
Did you provide any innovative insights beyond the required tasks (in addition to the requirements in the slide template)? (4 pts)

Conclusion

1. The BaggingClassifier has the lowest RMSE (0.127622), indicating better performance in predicting ratings or recommendations among the provided models.
2. From the provided list, models such as DecisionTreeClassifier (RMSE: 0.196842) and BaggingClassifier (RMSE: 0.127622) are typically less computationally expensive compared to neural network models like NN and NN_1 (RMSE: 0.534776).
3. If you're exploring the structure or patterns within the data without labeled examples, unsupervised learning is more appropriate. It can help in understanding the underlying structure of the data and finding hidden patterns.
4. Unsupervised learning algorithms like k-means clustering can be useful for segmenting data into distinct groups based on similarities.
5. Techniques like Principal Component Analysis (PCA) or t-distributed Stochastic Neighbor Embedding (t-SNE) are used for dimensionality reduction and visualization, which can be valuable for understanding high-dimensional data.
6. Unsupervised learning is often used for anomaly detection where the goal is to identify rare events or outliers in the data.
7. If you have a sufficient amount of labeled data, supervised learning models can be a good choice. For example, in classification or regression tasks where you have labeled examples of input-output pairs, supervised learning can be effective.
8. When the objective is well-defined and can be framed as predicting an outcome based on input features, supervised learning is suitable. For example, predicting customer churn, spam detection, sentiment analysis, etc.
9. Supervised learning models are evaluated based on metrics like accuracy, precision, recall, F1-score, etc., which make it easier to assess model performance.
10. Sometimes, a combination of supervised and unsupervised learning techniques is used, known as semi-supervised learning. This can be beneficial when labeled data is limited but unlabeled data is abundant.

Appendix

Appendix

Documents:

- https://www.kaggle.com/code/wahyuardhitama/task003-p001-ml-dl-rec-sys-course-20231025
- https://www.kaggle.com/code/wahyuardhitama/task003-p002-ml-dl-rec-sys-course-20231029
- https://www.kaggle.com/code/wahyuardhitama/task003-p003-ml-dl-rec-sys-course-20231101
- https://github.com/whyzie/Task003-ML-DL-Rec-Sys-Course-20231201