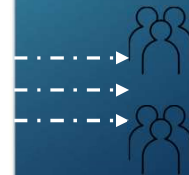
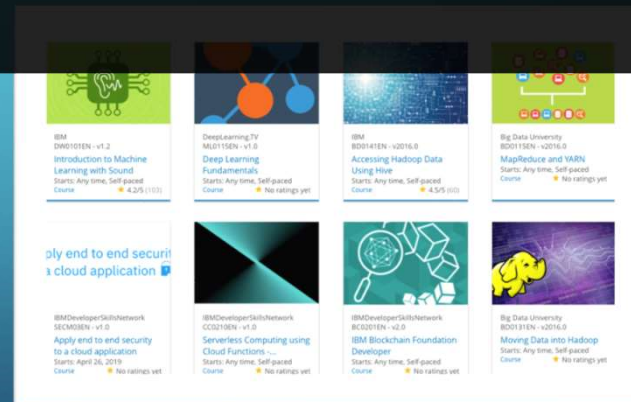


Building a Personalized Online Course Recommender System with Machine Learning

Nyasha Eysenck Gandah

06/07/2024





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

Project Background and Context:

- AI Training Room machine learning team started a recommender system project this year.
- Goal: Enhance learning experience and potentially increase revenue.
- Focus: Proof of Concept (PoC) phase, exploring and comparing machine learning models.

Problem Statement and Hypotheses:

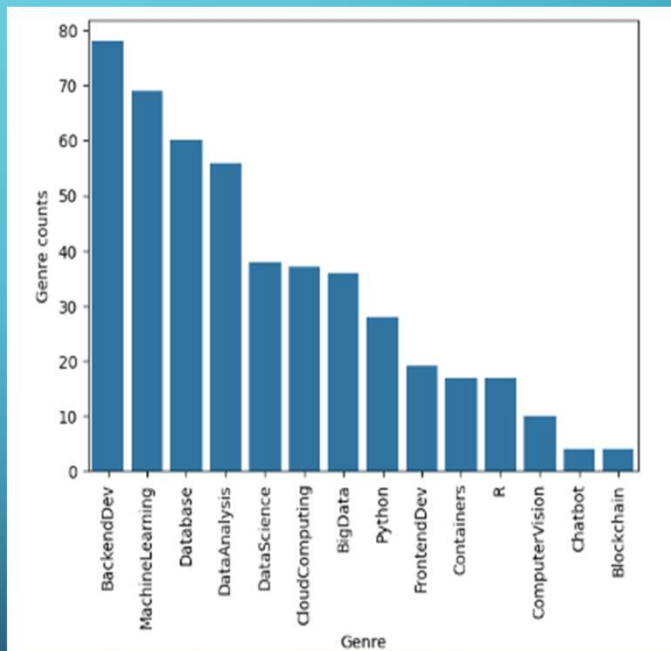
- Build the best recommender system possible during the Proof of Concept (PoC) phase by testing and comparing different machine learning models offline.
- A successful recommender system, chosen through thorough offline testing, will enhance how users find courses and navigate their learning journey, possibly leading to higher satisfaction and revenue

EXPLORATORY DATA ANALYSIS



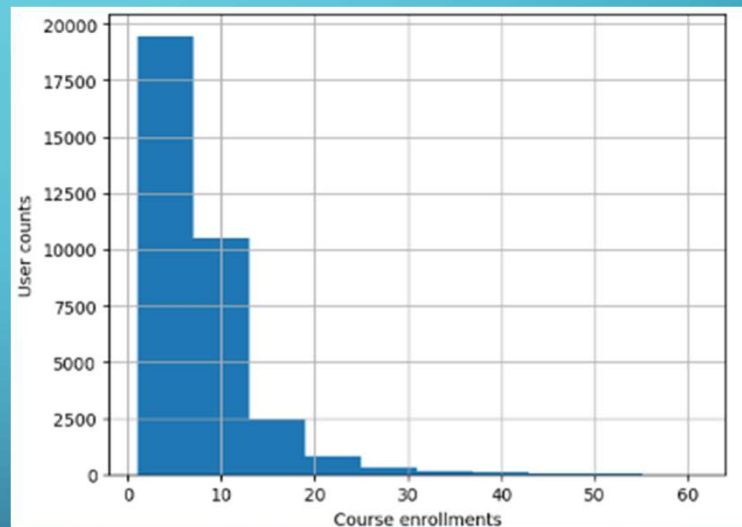
COURSE COUNTS PER GENRE

- ❑ The bar chart illustrates the distribution of courses by genre at AI Training Room.
- ❑ Backend Development emerged as the most prevalent genre, followed by Machine Learning, Database, and Data Analysis.
- ❑ This insight informed us of the popularity of Backend Development suggesting strong industry demand for those skills.



COURSE ENROLLMENT DISTRIBUTION

- ❑ The histogram illustrates the distribution of course enrollments among users.
- ❑ Most users enroll in only a few courses, with nearly 19,000 users enrolling in just 1 course.
- ❑ User count declines as course enrollments increase, with very few users enrolling in more than 10 courses.
- ❑ Long tail effect: Almost no users have more than 30 course enrollments.



20 MOST POPULAR COURSES

	TITLE	Ratings
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

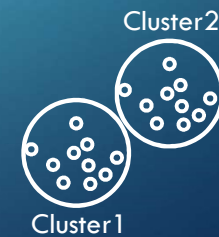
- ❑ The top 20 most popular courses by amount of user ratings.
- ❑ Highlights high-demand areas: Python, data science, big data, and cloud.
- ❑ The top course enrollments cover about 63.3% of all course enrollments

WORD CLOUD OF COURSE TITLES



- ❑ The word cloud highlights key topics in course titles for the course recommender system.
- ❑ Major topics include "python," "data science," "machine learning," "data analysis," and "big data."
- ❑ Other significant areas are "microservice," "application," "web," "cloud," and "java."
- ❑ This indicates high user interest and a strong focus on programming, data-related fields, and modern software development technologies.

CONTENT-BASED RECOMMENDER SYSTEM USING UNSUPERVISED LEARNING



FLOWCHART OF CONTENT-BASED RECOMMENDER SYSTEM USING USER PROFILE AND COURSE GENRES



FLOWCHART OF CONTENT-BASED RECOMMENDER SYSTEM USING USER PROFILE AND COURSE GENRES

1. **Load Data:** Loading the necessary data, including course similarity matrix, course content, and Bag of Words (BoW) features.
2. **Preprocess Data:** Preparing the data for similarity calculation, including creating mappings between course IDs and indices.
3. **Calculate Similarities:** Computing the similarity between courses based on their content and features using dot product.
4. **Filter Courses:** Applying a similarity threshold to filter out courses that are not sufficiently similar to the user's enrolled courses.
5. **Generate Recommendations:** Generating a list of recommended courses based on the filtered results.

EVALUATION RESULTS OF USER PROFILE-BASED RECOMMENDER SYSTEM

HYPERPARAMETERS:

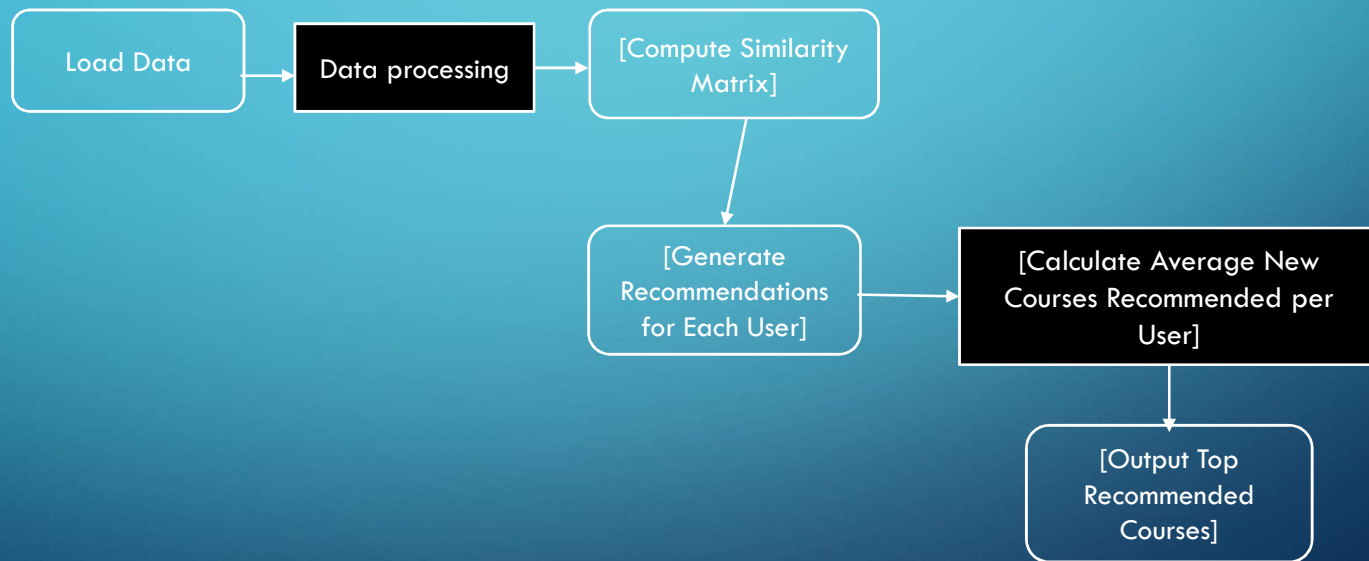
- ❑ SCORE_THRESHOLD: 10
- ❑ SIMILARITY METRIC: DOT PRODUCT

- ✓ Number of Users : 16554
- ✓ Total New Courses : 479131
- ✓ Average New Courses per User:29

TOP 10 RECOMMENDED COURSES

TITLE	RECOMMENDATIONS
foundations for big data analysis with sql	9138
analyzing big data with sql	9138
getting started with the data apache spark ma...	8954
analyzing big data in r using apache spark	8769
park overview for scala analytics	7970
cloud computing applications part 2 big data...	7853
applied machine learning in python	7671
introduction to data science in python	7671
accelerating deep learning with gpu	7633
spark fundamentals ii	7203

FLOWCHART OF CONTENT-BASED RECOMMENDER SYSTEM USING COURSE SIMILARITY



FLOWCHART OF CONTENT-BASED RECOMMENDER SYSTEM USING COURSE SIMILARITY

1. **Load Course Data:** Import the dataset containing information about courses.
2. **Preprocess Course Bag of Words (BoW) Features:** Prepare the Bag of Words features for each course.
3. **Computing Similarity Matrix:** Calculate the similarity matrix between courses based on their features.
4. **Generate Recommendations for Each User:** Utilize the similarity matrix to recommend courses for each user based on their enrolled courses.
5. **Calculate Average New Courses Recommended per User:** Determine the average number of new courses recommended to each user.
6. **Identify Top Recommended Courses:** Find the most frequently recommended courses across all users.
7. **Output Top Recommended Courses:** Display the top recommended courses to users

EVALUATION RESULTS OF COURSE SIMILARITY BASED RECOMMENDER SYSTEM

HYPERPARAMETERS:

❑ SIMILARITY_THRESHOLD: 0.65

❑ SIMILARITY_METRIC: COSINE

✓ Number of Users : 8595

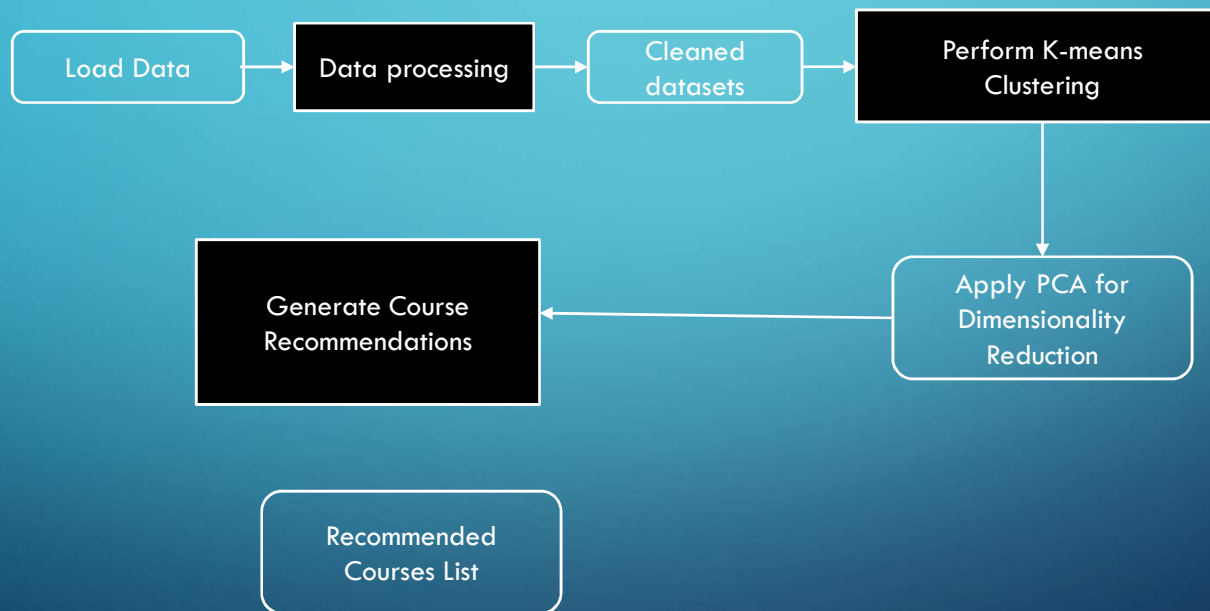
✓ Total New Courses : 14195

✓ Average New Courses per User:1.65

TOP 10 RECOMMENDED COURSES

TITLE	RECOMMENDATIONS
Build Your Own Chatbots	1429
Data Science with Open Data	1326
Deep Learning with TensorFlow	979
Deep Learning with TensorFlow	899
Deep Learning with TensorFlow	873
IBM Cloud Essentials v3	517
Accelerating Deep Learning with GPUs	433
Text Analysis	402
Introduction to Big Data Accelerating Deep Learning with GPU	358
Accelerating deep learning with gpu	252

FLOWCHART OF CLUSTERING-BASED RECOMMENDER SYSTEM



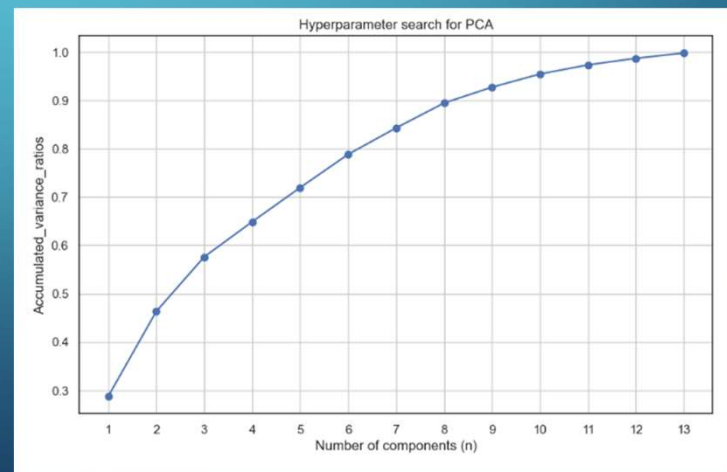
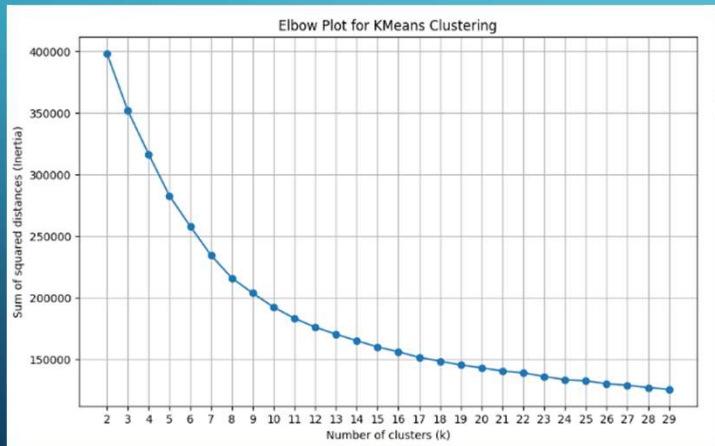
FLOWCHART OF CLUSTERING-BASED RECOMMENDER SYSTEM

1. **Load Data:** Load the user profile dataset containing user interests and course ratings.
2. **Data Preprocessing:** Standardize the user profile features., Prepare the data for clustering analysis.
3. **Perform K-means Clustering:** Group users based on their shared interests, Determine the optimal number of clusters, Assign cluster labels to each user.
4. **Apply PCA for Dimensionality Reduction:** Reduce the dimensions of the user profile feature vectors., Preserve the essential information while reducing complexity.
5. **Perform K-means Clustering on PCA Components:**Cluster users based on the reduced feature set., Determine the optimal number of clusters., Assign cluster labels to each user.
6. **Generate Course Recommendations:** Identify popular courses within each cluster., Recommend courses to users based on their cluster's preferences.

EVALUATION RESULTS OF CLUSTERING-BASED RECOMMENDER SYSTEM

Hyperparameters:

- ❑ Kmeans n_clusters: 14 (Chosen via elbow method)
- ❑ PCA n_components: 9 (Chosen via hyperparameter search the smallest n_components value for which accumulated_variance_ratios ≥ 0.9)



EVALUATION RESULTS OF CLUSTERING-BASED RECOMMENDER SYSTEM

Hyperparameters:

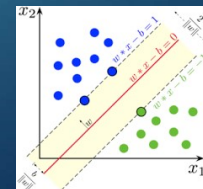
- ❑ Kmeans n_clusters: 14
- ❑ Similarity metric: dot product

- ✓ Number of Users : 33901
- ✓ Total New Courses : 3305152
- ✓ Average New Courses per User:97

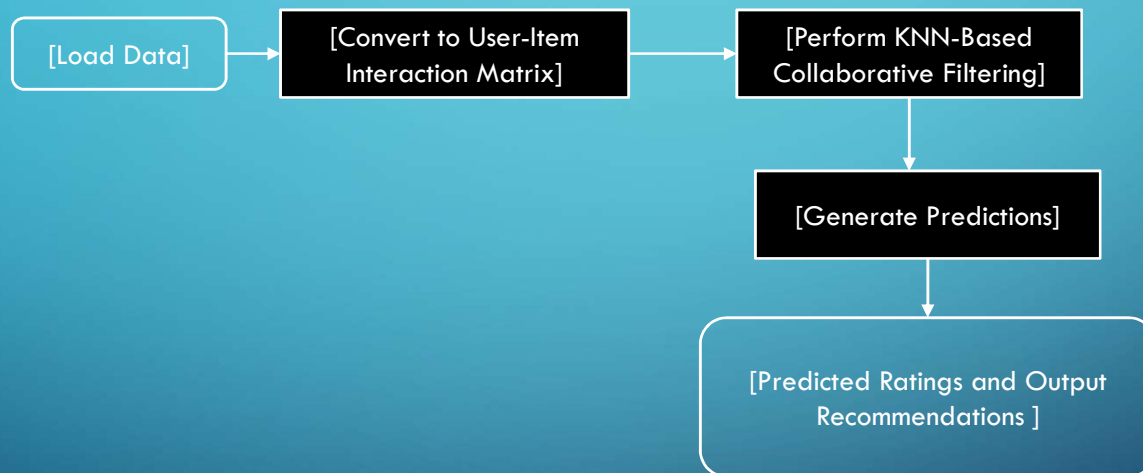
TOP 10 RECOMMENDED COURSES

TITLE	RECOMMENDATIONS
accelerating deep learning with gpus	33728
apply end to end security to a cloud application	33659
deep learning with tensorflow	33596
uild swift mobile apps with watson ai services	33579
how to build watson ai and swift apis and make...	33569
serverless computing using cloud functions d...	33550
getting started with the data apache spark ma..	33534
ata journalism first steps skills and tools	33518
Deep Lgame playing ai with swift for tensorflow s4tf earning with GPU	33498
Aend to end data science on cloudpak for data	33495

COLLABORATIVE-FILTERING RECOMMENDER SYSTEM USING SUPERVISED LEARNING



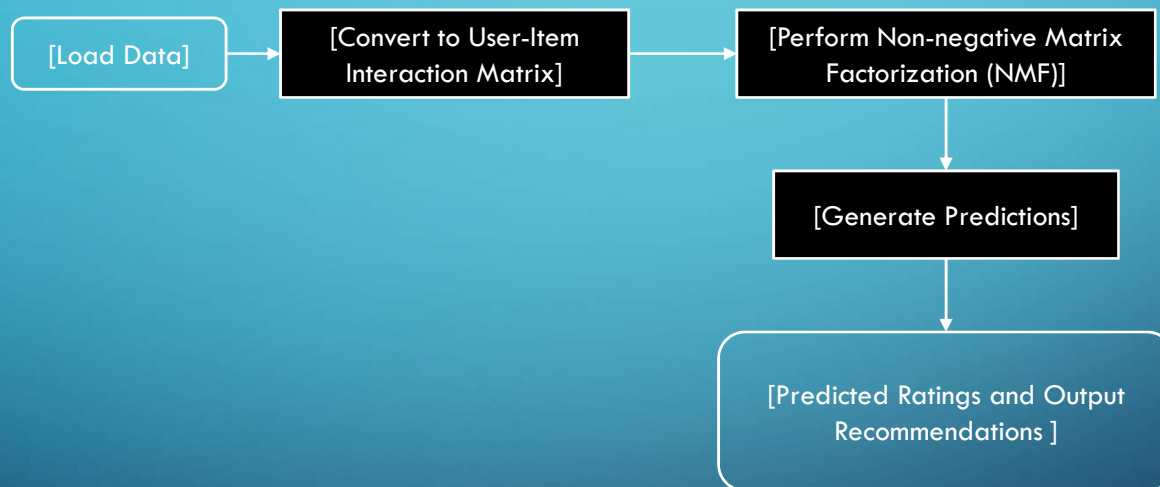
FLOWCHART OF KNN BASED RECOMMENDER SYSTEM



FLOWCHART OF KNN BASED RECOMMENDER SYSTEM

- 1. Load Data:** Load the user-item interaction data containing user IDs, item IDs, and ratings.**Convert to User-Item Interaction Matrix:** Transform the raw data into a user-item interaction matrix, where rows represent users, columns represent items, and each cell contains the rating given by a user to an item.
- 2. Perform KNN-Based Collaborative Filtering:** Utilize K Nearest Neighbor (KNN) algorithm to find similar users or items based on their rating patterns.
- 3. Generate Predictions:** Predict ratings for items that users have not interacted.
- 4. Output Recommendations and Predicted Ratings:** Output is the predicted ratings for items they might like.

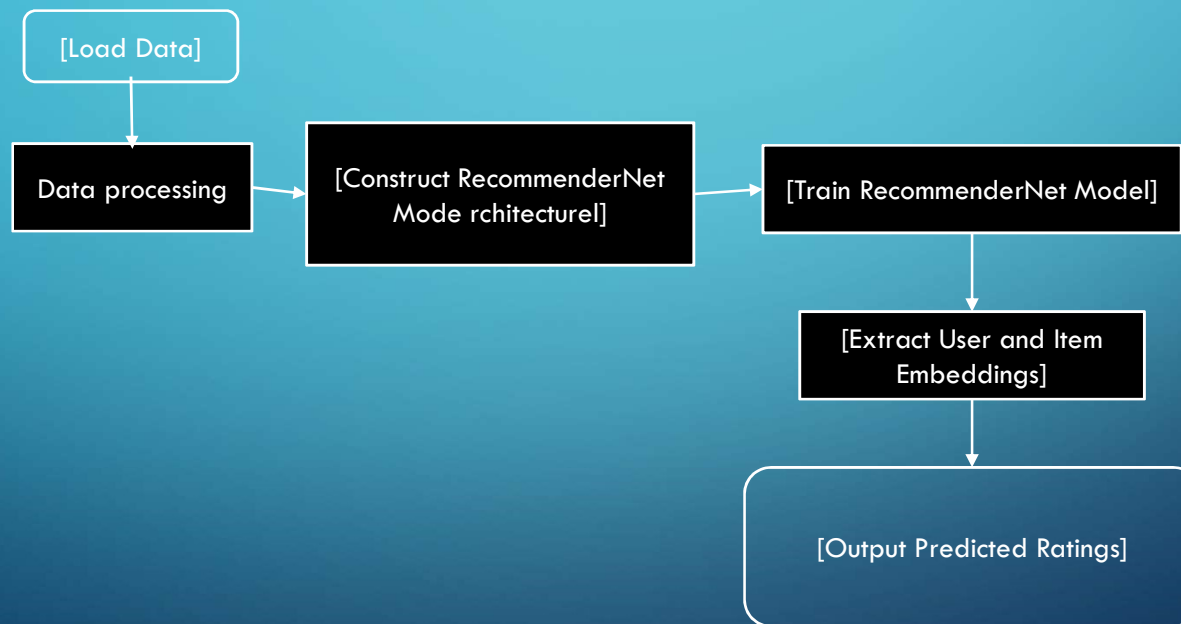
FLOWCHART OF NMF BASED RECOMMENDER SYSTEM



FLOWCHART OF NMF BASED RECOMMENDER SYSTEM

- 1. Load Data:** Load the user-item interaction data containing user IDs, item IDs, and ratings.**Convert to User-Item Interaction Matrix:** Transform the raw data into a user-item interaction matrix, where rows represent users, columns represent items, and each cell contains the rating given by a user to an item.
- 2. Perform Non-negative Matrix Factorization (NMF):** Utilize Non-negative Matrix Factorization (NMF) algorithm to decompose the user-item interaction matrix into two smaller matrices representing user and item features. This step aims to reduce dimensionality and capture latent features.
- 3. Generate Predictions:** Predict ratings for items that users have not interacted with based on the decomposed matrices using dot product operation.
- 4. Output Recommendations and Predicted Ratings:** Present the generated recommendations to users along with the predicted ratings for items they might like.

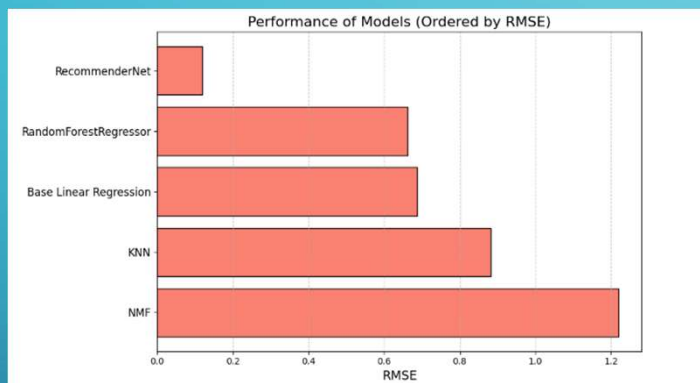
FLOWCHART OF NEURAL NETWORK EMBEDDING BASED RECOMMENDER SYSTEM



FLOWCHART OF NEURAL NETWORK EMBEDDING BASED RECOMMENDER SYSTEM

1. **Load and Process Data:** Load the rating dataset and preprocess it, converting user and item IDs into integer indices.
2. **Define RecommenderNet Model:** Implement the neural network model (RecommenderNet) using TensorFlow, with embedding layers for users and items.
3. **Train RecommenderNet Model:** Train the RecommenderNet model using the processed dataset, optimizing its parameters to minimize the loss function.
4. **Evaluate Model Performance:** Assess the performance of the trained model on a validation set to ensure its effectiveness in rating prediction.
5. **Extract User and Item Embeddings:** Extract the latent features (embeddings) learned by the model for both users and items.
6. **Output Ratings or Predictions:** Generate the final output ratings or predictions using the trained model and the extracted embeddings. These ratings represent the estimated user-item interactions.

COMPARE THE PERFORMANCE OF COLLABORATIVE-FILTERING MODELS



RMSE Bar Chart:

- "RecommenderNet" has the lowest RMSE, indicating superior predictive accuracy.
- "RandomForestRegressor" follows with moderate accuracy.
- "Base Linear Regression" and "KNN" have relatively higher RMSE values.



Accuracy Bar Chart:

- "Bagging" achieves the highest accuracy.
- "Decision Tree" closely follows.
- "Logistic Regression" and "Boosting" exhibit slightly lower accuracy but remain competitive

COURSE RECOMMENDER SYSTEM APP WITH STREAMLIT

SELECTING TAKEN COURSES

Personalized Learning Recommender

1. Select recommendation models

Select model:

Course Similarity

2. Tune Hyper-parameters:

Top courses

Course Similarity Threshold %

3. Training:

Train Model

4. Prediction

Recommend New Courses

Select courses that you have audited or completed:

COURSE_ID	TITLE
<input type="checkbox"/> excourse47	Machine Learning For All
<input type="checkbox"/> excourse48	Introduction To Machine Learning Language Process
<input type="checkbox"/> excourse49	Applied Machine Learning In Python
<input type="checkbox"/> excourse50	Build Train And Deploy ML Pipelines Using Bert
<input type="checkbox"/> excourse51	Introduction To Machine Learning In Production
<input type="checkbox"/> excourse52	Machine Learning Data Lifecycle In Production
<input type="checkbox"/> excourse53	Deploying Machine Learning Models In Production
<input type="checkbox"/> excourse54	Exploratory Data Analysis For Machine Learning
<input type="checkbox"/> excourse55	Advanced Computer Vision With Tensorflow
<input checked="" type="checkbox"/> excourse56	Deep Learning Applications For Computer Vision
<input type="checkbox"/> excourse57	Deep Learning In Computer Vision
<input type="checkbox"/> excourse58	Computer Vision Basics
<input type="checkbox"/> excourse59	Fundamentals Of Digital Image And Video Processin

Your courses:

COURSE_ID	TITLE
202	ML0120ENV2 Deep Learning With Tensorflow
234	excourse21 Applied Machine Learning In Python
259	excourse46 Machine Learning
269	excourse56 Deep Learning Applications For Computer Vision

GETTING RECOMMENDATIONS

Personalized Learning Recommender

1. Select recommendation models

Select model:

Course Similarity

2. Tune Hyper-parameters:

Top courses

Course Similarity Threshold %

3. Training:

Train Model

4. Prediction

Recommend New Courses

Your courses:



COURSE_ID	TITLE
202	ML0120ENV2 Deep Learning With Tensorflow
234	excourse21 Applied Machine Learning In Python
259	excourse46 Machine Learning
269	excourse56 Deep Learning Applications For Computer Vision

Recommendations generated!

SCORE	TITLE	DESCRIPTION
0	1.0000 Deep Learning With Tensorflow	majority of data in the world are unlabeled and unstructured data for instance images sound and text data shallow neural networks cannot easily capture relevant structure in these kind of data but deep networks are capable of discovering hidden structures within--these data in this course you will use tensorflow library to apply deep learning on different data types to solve real world problems
		this course will introduce the learner to applied machine learning focusing more on the techniques and methods than on the statistics behind these methods the course will start with a discussion of how machine learning is different than descriptive statistics and introduce the scikit learn toolkit through a tutorial the



Other potential approaches

- **EDA with NLP:** Utilize Natural Language Processing (NLP) techniques to analyze course descriptions, providing deeper insights into user preferences.
 - **Transfer Learning for Training:** Explore the use of transfer learning techniques to pre-train the recommender system on a large dataset from a related domain before fine-tuning it on the specific course dataset.
 - **Ensemble Model Integration:** Combine multiple recommendation models using ensemble techniques like blending or stacking to leverage the strengths of individual models for improved accuracy.
- 
- 

Conclusions

- AI Training Room initiated a recommender system project to enhance learning experiences and potentially boost revenue.
- The Proof of Concept (PoC) phase focused on offline exploration and comparison of machine learning models to develop the best recommender system.
- Course genre distribution highlighted strong industry demand for Backend Development courses.
- Analysis revealed most users enroll in a few courses, with Python, data science, big data, and cloud courses being popular.
- Different recommendation techniques were used to generate personalized course recommendations, with RecommenderNet emerging as the best-performing model.
- The project achieved its goals, evidenced by superior predictive accuracy of RecommenderNet and improved user experience through a Streamlit app.