Course4U
Build a Personalized Online Course Recommender System with Machine Learning

Presented by: Wahyu Ardhitama

Last Updated: May 31st. 2024

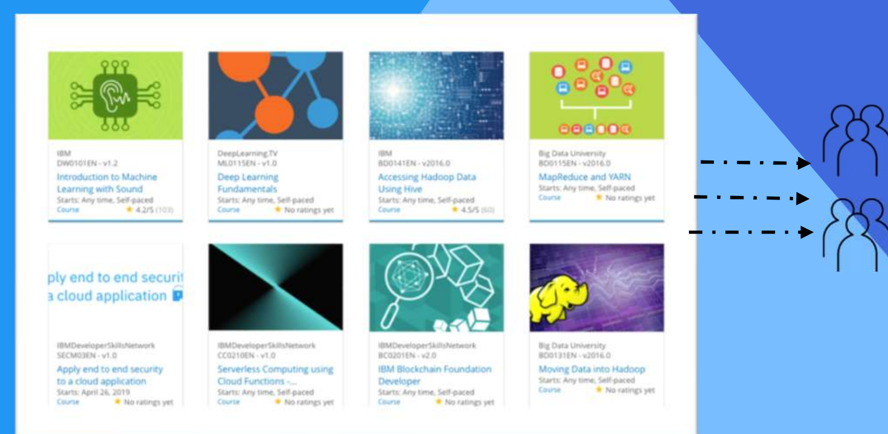


Table of Contents

Executive summary

➢ We use PACE framework

Data Analysis PACE Steps:

1. Plan/Prepare - import the relevant libraries and data

   Align project with business needs, requirements and contraints. Select an approriate machine learning model based on the problem and business context. KDD: Selection, Data Wrangling (Pre-processing and Transformation).

2. Analyze - Explarotary Data Analysis (EDA)

   Understanding data for accurate predictions, focus on the response variable (what the model predicts) and leverage exploratory data analysis to uncover patterns and address irregularities. KDD: Data Mining.

3. Construct - model

   Construct and evaluate model. KDD: Evaluation.

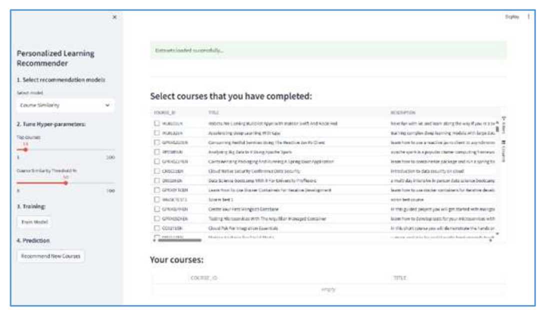
4. Execute - share

   Interpret model and share the story. KDD: Communicate to stakeholders.

CO-V-FAST Principles: **C**lear/**C**lean/**C**ommunication/**C**ollaboration/**C**orrection, **O**bjective, **V**aluable, **F**ocus, **A**gile, **S**cientific and **T**ime-bound/**T**rustworthiness

➢ The recommender system apps

1. Recommender system with supervised and unsupervised learning



Project Management Flowchart

Project Initiation (PI)

1. OKR and SMART
2. Stakeholder Analysis
3. Project Charter

1. VOC
2. CTQ – CTP – CTR - COPQ
3. SIPOC

Project Planning (PP)

1. WBS and RACI
2. Gantt Chart
3. Risk Management
4. Project Budget
5. Communication Plan – Team Profile

1. KPI Trees
2. Lean Measures
3. Operational Definitions

Project Execution (PE)

1. Customer Survey
2. ROAM Analysis
3. Status Report
4. Hours Burndown

1. Data Collection and Sample
2. MSA
3. Process Capability

Agile Scrum Project

1. Sprint Backlog
2. Sprint Retrospective

1. The Process Door
2. EDA
3. Experimentation Platform

Lean Six Sigma Project

1. Yellow Belt
2. Green Belt
3. Black Belt
4. Master Black Belt

1. SPC

---

1. Team Meeting Agenda
2. Email management (Escalation, Invitation, Release Plan, etc.)

Project Initiation : Business case and Objective

Project Management Flowchart

Project Initiation (PI)

1. OKR and SMART
2. Stakeholder Analysis
3. Project Charter

1. VOC
2. CTQ – CTP – CTR - COPQ
3. SIPOC

Project Planning (PP)

1. WBS and RACI
2. Gantt Chart
3. Risk Management
4. Project Budget
5. Communication Plan – Team Profile

1. KPI Trees
2. Lean Measures
3. Operational Definitions

Project Execution (PE)

1. Customer Survey
2. ROAM Analysis
3. Status Report
4. Hours Burndown

1. Data Collection and Sample
2. MSA
3. Process Capability

Agile Scrum Project

1. Sprint Backlog
2. Sprint Retrospective

1. The Process Door
2. EDA
3. Experimentation Platform

Lean Six Sigma Project

1. Yellow Belt
2. Green Belt
3. Black Belt
4. Master Black Belt

1. SPC

---

1. Team Meeting Agenda
2. Email management (Escalation, Invitation, Release Plan, etc.)

Plan

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%)

Voice of Customer (VoC)

- What **features** are important to our customers (B2C, B2B)?
- How do our customers **prioritize** their courses selection?
- Reasons behind the courses selection?
- Reasons behind the payment selection?
- What courses are considered good quality and bad quality?

VoC :

We do not have a formal budget for customer research, we utilize the customers feedbacks and the vast amounts of market research available in the internet.

Action Plan:

Meet the customers, generate insights from feedbacks (e.g. course reviews, course ratings) and online questionnaire.

Critical to Quality (CTQ)

Plan

Voice of Customer

Required for careers

The right courses

More Options

Provide Personalized and Relevant Course Recommendations

Ease to navigate

Required for careers

Updated

Courses can be accessed with no delay

Courses can be accessed anywhere and anytime

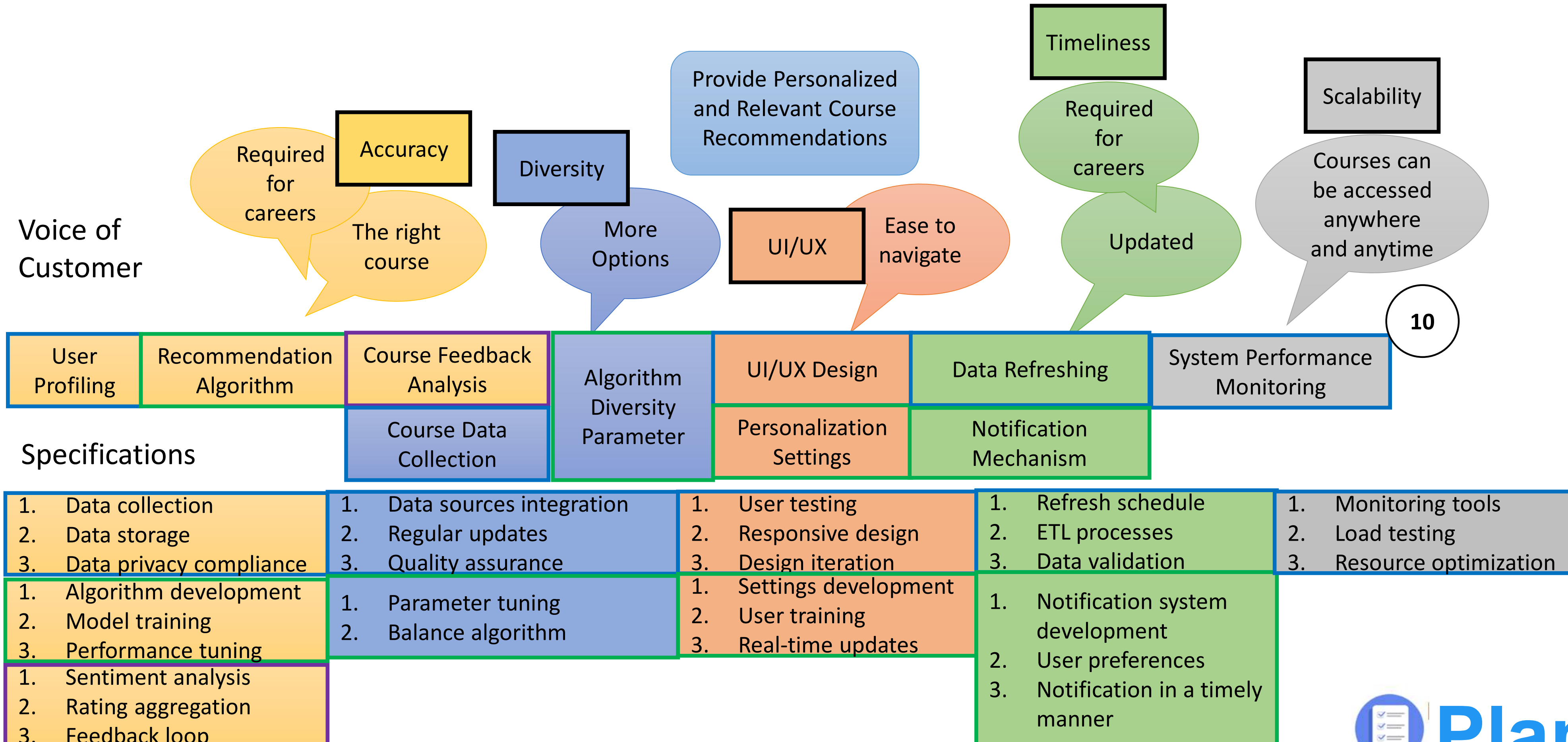
Relevance | Engagement | Success Rate | Difficulty Levels | Interface Usability | Feedback Mechanism | Notification System | Performance | Expandable Data Handling

Specifications

Course Variety | Specialized Courses | Personalization | Real-time Updates | Adaptability | Resource Efficiency

# Recommend courses that are precise or in accordance to careers advancement	# Recommend other courses that are similar	UI/UX: a clean and intuitive layout – easy to navigate	If a new courses is added it appears	Increase # users, maintain speed and accuracy
Time completion; completion rate	# Specialized courses	UI/UX: set preferences(careers, interests, etc.)	Adapt to the changes of the users behaviours	Optimize use of computational resources (e.g. memory, CPU)
Completion rate	# Courses for beginners, intermediate and advance	UI/UX: rate courses, provide comments, interact with trainers, discussions and feedback – refine the recommender system	Prompt notification of new relevant and updates courses	Integrate with new data source and handle additional data type

Critical to Process (CTP) – in Production



Plan

Critical to Resources (CTR) – in Production

C4U is establishing the Data Analytics Dept (e.g. Analyst(Data, Business, Supply Chain), Data Engineer, Data Science, Machine Learning Engineer and AI Engineer)

Voice of Customer

Accuracy
Required for careers
The right course

Diversity
More Options

Provide Personalized and Relevant Course Recommendations

UI/UX
Ease to navigate

Timeliness
Required for careers
Updated

Scalability
Courses can be accessed anywhere and anytime

9

Data Science (DS) TBH
1. Data collection
2. Data analysis
3. Modelling user profiles

Account/Content Manager
1. Content integration
2. Quality assurance
3. Course catalog management

UI/UX
1. Interface design
2. User testing
3. Responsive design

Data Engineer TBH
1. ETL processes
2. Data pipeline maintenance
3. Data validation

DevOps Engineer
1. Monitoring tools
2. Performance optimization
3. Incident response

Machine Learning Engineer (ML) TBH
1. Algorithm development
2. Model training
3. Performance tuning

Machine Learning Engineer
1. Parameter design
2. Testing and validation
3. Continous improvement

Product Manager
1. Feature development
2. User education
3. Customization option

Software Engineer
1. Notification system development
2. User preferences integration
3. Timeliness Assurance

Data/Marketing Analyst
1. Feedback collection
2. Sentiment analysis
3. Insights generation

*TBH = To Be Hired

Plan

OKR – Strategic (QSCP)

Plan

O1 Q	**Make it easy to choose course in our C4U Web and Apps via course recommender system.**
KR1	Increase the enrolments from current users with more than 10 enrolments from 30% to 35%, from 25,000 users with fewer than 10 courses, 1,250 (5%) users have more than 10 courses.
KR2	Revenue increase by 10% compare last year, min $ 280,500/month (Q3). Active users from 2,040 to 2,250 users/month.
KR3	New users increase by 10%. By the end of this year we have 3,500 new users.
O2 S	**Provide a reliable and consistent course recommender system and its service.**
KR1	90% of recommender system meet security standards at monthly audits
KR2	95% of integration with the web/app backend to serve recommendations to users.
KR3	Wait times decrease by 10% within two months of launch

O3 C	**Promote course recommender system as to assist user in their study journey.**
KR1	Generate conversion from marketing ads, active users min. 2,250 user/month (average price $125)
KR2	% click-through rate from banner ads on social media (A/B Testing) increase 2 times compared control group.
KR3	6 press pieces published in relevant print and online publications
KR4	Employ collaborative filtering or content-based filtering algorithms. Batch processing to provide recommendations. Later, upgrade to real-time from stream processing to continuously update recommendations based on the latest user behaviour when the users reach 100K.
O4 P	**Actively and meaningfully engage the user to generate buy-in and project support.**
KR1	Establish a YouTube channel, Total 400 attendees of 3 live YouTube focused on courses awareness and transit talks introducing recommendation system.
KR2	75% of Users surveyed before launch
KR3	70% of top users participate in user outreach program (e.g. webinars, YouTube Live Sessions, etc.)

OKR - Prod

Accuracy | DS, ML and DA

O1 **Improve the relevance and engagement of course recommendations**

- KR1 Increase the **relevance score of recommendations to 90%** by the end of Q4.
- KR2 **Achieve a 15% increase in the average time users spend** on recommended courses by the end of Q3.
- KR3 Boost **the course completion rate** for recommended courses to 75% by the end of Q3.

Diversity | CM and ML

O2 **Enhance the diversity of courses offered to users.**

- KR1 Increase **the Course Diversity Index to an average of 6 distinct topics** per user by Q3.
- KR2 Ensure that **at least 25% of recommendations include niche courses** by the end of Q2.
- KR3 Achieve **a balanced recommendation mix** with 30% beginner, 40% intermediate, and 30% advanced courses by Q4.

O3 **Optimize the user experience of the course recommendation interface.**

UI/UX | UI/UX and PM

- KR1 Increase **the User Satisfaction Score** to 4.7/5 by Q4.
- KR2 Ensure **95% of recommendations align** with user preferences by Q3.
- KR3 Utilize **80% of user feedback to make iterative improvements** to the system by the end of Q4.

Plan

Timeliness | DE and SE

O4 **Improve the timeliness and adaptability of the recommendation system.**

- KR1 **Reduce the Recommendation Update Time** to under 30 minutes by Q2.
- KR2 Achieve **a Notification Response Rate of 60%** within 24 hours by Q3.
- KR3 **Ensure 95% of user behaviour changes** are reflected in updated recommendations within 12 hours by Q4.

O5 **Ensure the recommender system scales efficiently with growing demand.**

- KR1 Maintain 99.9% **System Uptime** as user base doubles by the end of Q3.
- KR2 **Keep Latency** under 1.5 seconds per query even as the course library expands by 50% by Q4.
- KR3 Successfully **support up to 40,000 users** with no performance degradation by the end of Q4.

Scalability | DOE

SIPOC Analysis

Plan

S Supplier	I Input	P Process	O Output	C Customer
1. C4U 2. Course providers or instructor(s) 3. Technology providers	1. Course (materials. subject and level) 2. Coursework / test 3. Personal information	Registration 1. Course offered/ searched 2. Subscribed 3. Paid 4. Learning 5. Assessed 6. Reviewed	1. Assignment 2. Certification 3. Invoice 4. Email report 5. Course ratings	1. User 2. Course providers or instructor(s)

Project Team

Plan

Project Team

Project Sponsor: Director of Customer Data

Project Lead: Head of Data Science and Machine Learning

Project Team: Director of Procurement, API Strategist, Data Warehousing Specialist, Data Governance Manager, Data Analyst, Director of IT, Project Manager, Product Manager, Marketing Promotion Manager, Financial Analyst, HR Recruitment and Training

Additional Stakeholders: Account Manager, Sales and Marketing Director, Investors

Data Science

1. Data collection
2. Data analysis
3. Modelling user profiles

Machine Learning Engineer

1. Algorithm development
2. Model training
3. Performance tuning

Data/Marketing Analyst

1. Feedback collection
2. Sentiment analysis
3. Insights generation

Data Engineer

1. ETL processes
2. Data pipeline maintenance
3. Data validation

Software Engineer

1. Notification system development
2. User preferences integration
3. Timeliness Assurance

Account/Content Manager

1. Content integration
2. Quality assurance
3. Course catalog management

Machine Learning Engineer

1. Parameter design
2. Testing and validation
3. Continous improvement

UI/UX

1. Interface design
2. User testing
3. Responsive design

Product Manager

1. Feature development
2. User education
3. Customization option

DevOps Engineer

1. Monitoring tools
2. Performance optimization
3. Incident response

Understanding stakeholders (stakeholder analysis)

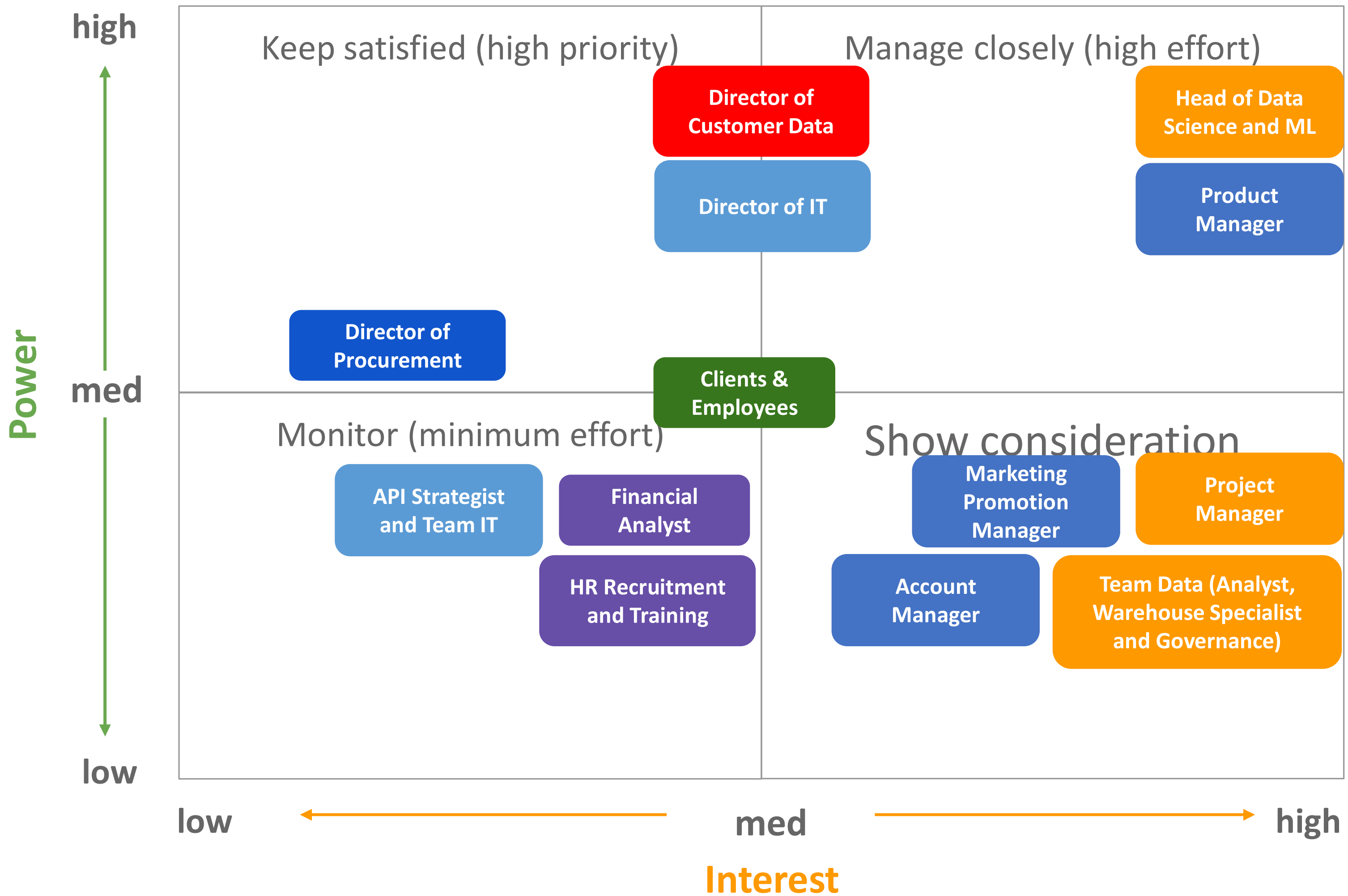
Stakeholder	Role (Related to project)	Involvement	Impact	Power or Influence (H/M/L)	Interest (H/M/L)	Engagement
Director of Customer Data	Project sponsor	Makes high-level decisions; serves as team resource	Wants the project to succeed. No resistance.	H	M	Communicate regularly, but not daily. Ask questions and give updates.
Head of Data Science and Machine Learning	Project leader	Knowledge of data science and machine learning	Invested in the project as a team leader. Possible resistance if data science and machine learning role is affected.	H	H	Communicate daily as project leader
Existing Clients and Employees	Office C4U customer	Can give feedback on the customer experience	Some highly interested; others less so. Resistance only if Plant Pals affects main product line.	M	M	Communicate as needed to inform and get feedback.
Director of Procurement	Project team member	Procurement support	Invested in the project as a team member. Little impact at present. Project could affect their investment budget allocation.	M	L	Communicate daily as project team member.
Director of IT	Project team member	IT support	Invested in the project as a team member. Little impact at present. Project could affect their IT infrastructure	H	M	Communicate daily as project team member.

Understanding stakeholders (stakeholder analysis)

Stakeholder	Role (Related to project)	Involvement	Impact	Power or Influence (H/M/L)	Interest (H/M/L)	Engagement
Product Manager	Project team member	Knowledge of website design and plants; strong relationships with OG employees	Invested in the project as a team member. Possible resistance if Product Manager role is affected.	H	H	Communicate daily as project team member.
Team Data	Project team member	Knowledge of Big Data; strong relationships with OG employees	Invested in the project as a team member. Possible resistance if Team Data role is affected.	H	H	Communicate daily as project team member.
Marketing Promotion Manager and Account Manager	Project team member	Knowledge of Big Data; Establish marketing plan and campaign for project	Invested in the project as a team member. Possible resistance if Marketing role is affected.	H	H	Communicate daily as project team member.
API Strategis and Team IT (UI/UX, DevOps and Sofware Engineer)	Project team member	API support	Invested in the project as a team member. Little impact at present. Project could affect the API management.	L	L	Not directly involved. Keep updated on progress and performance.
Financial Analyst	Project team member	Financial support	Invested in the project as a team member. Little impact at present.	L	L	Not directly involved. Keep updated on progress and performance.
HR Recruitment and Training	Project team member	Recruitment and Training support	Invested in the project as a team member. Little impact at present.	L	L	Not directly involved, but should be updated before launch

Prioritizing stakeholders (power grid)

Plan



Project Planning: Gantt Chart and Budget

Plan

Project Management Flowchart

Project Initiation (PI)	Project Planning (PP)	Project Execution (PE)	Agile Scrum Project	Lean Six Sigma Project
1. OKR and SMART 2. Stakeholder Analysis 3. Project Charter	1. WBS and RACI 2. Gantt Chart 3. Risk Management 4. Project Budget 5. Communication Plan – Team Profile	1. Customer Survey 2. ROAM Analysis 3. Status Report 4. Hours Burndown	1. Sprint Backlog 2. Sprint Retrospective	1. Yellow Belt 2. Green Belt 3. Black Belt 4. Master Black Belt
1. VOC 2. CTQ – CTP – CTR - COPQ 3. SIPOC	1. KPI Trees 2. Lean Measures 3. Operational Definitions	1. Data Collection and Sample 2. MSA 3. Process Capability	1. The Process Door 2. EDA 3. Experimentation Platform	1. SPC

1. Team Meeting Agenda
2. Email management (Escalation, Invitation, Release Plan, etc.)

Analysis and Experimentation Team

Mission:

- Build a platform that is easy to integrate
- Foster a culture towards more data-driven decisions
- Accelerate innovation through trustworthy analysis and experimentation
- Empower the HiPPO (Highest Paid Person's Opinion) with valuable data

Team :

- Developers: Build the experimentation platform and the analysis tools
- Data Scientists – ML – Project (Program) Managers
- Admin

Analytics Objective

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

C4U Recommender Systems :

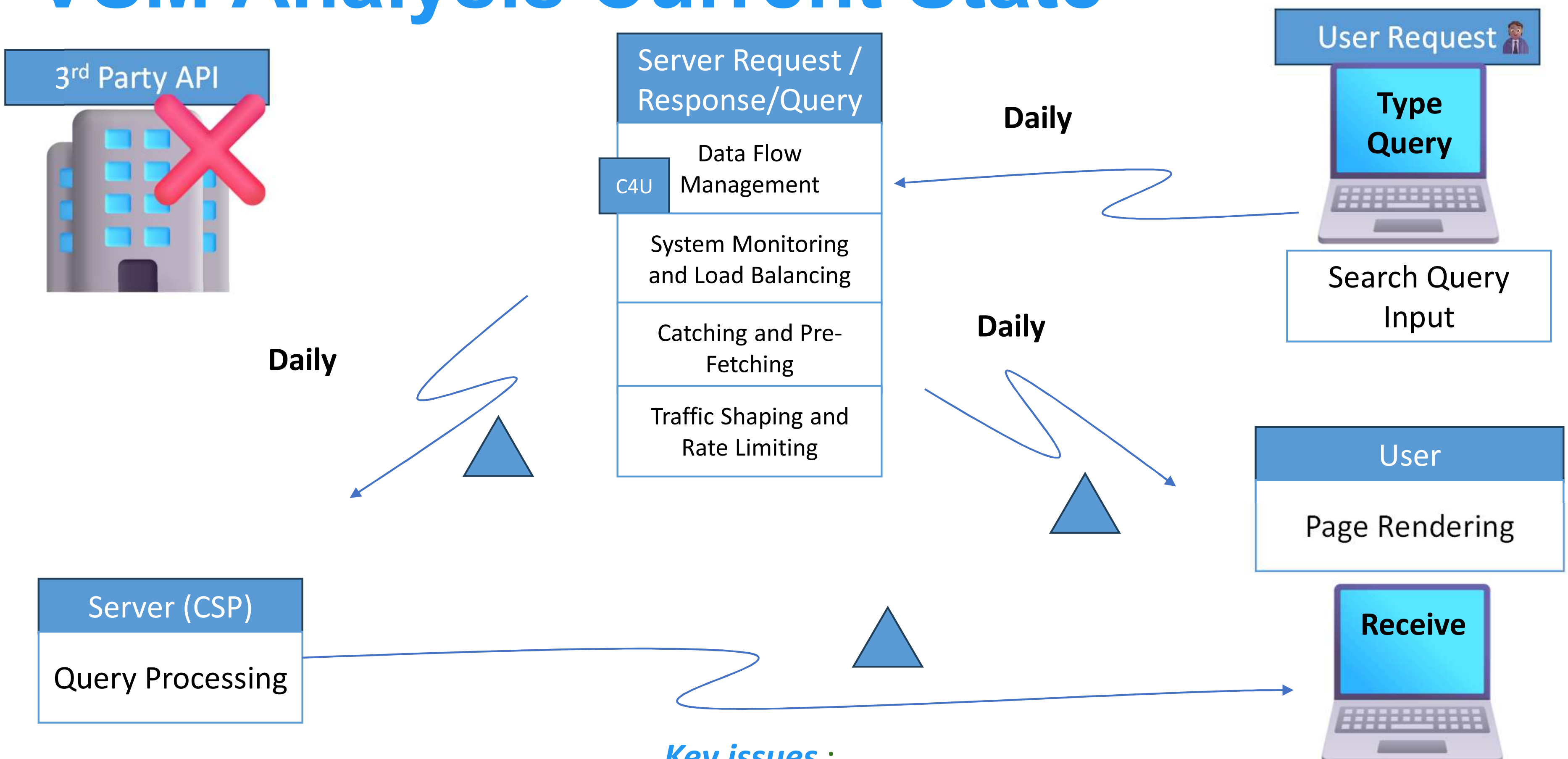
- Quickly find new interested courses
- Better paving learning paths
- More learners interacting with more courses

Hypothesis :

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

Plan

VSM Analysis Current State



Key issues :

Inefficiency, Lack of personalization and Bad user experience

The 7 Wastes Plus

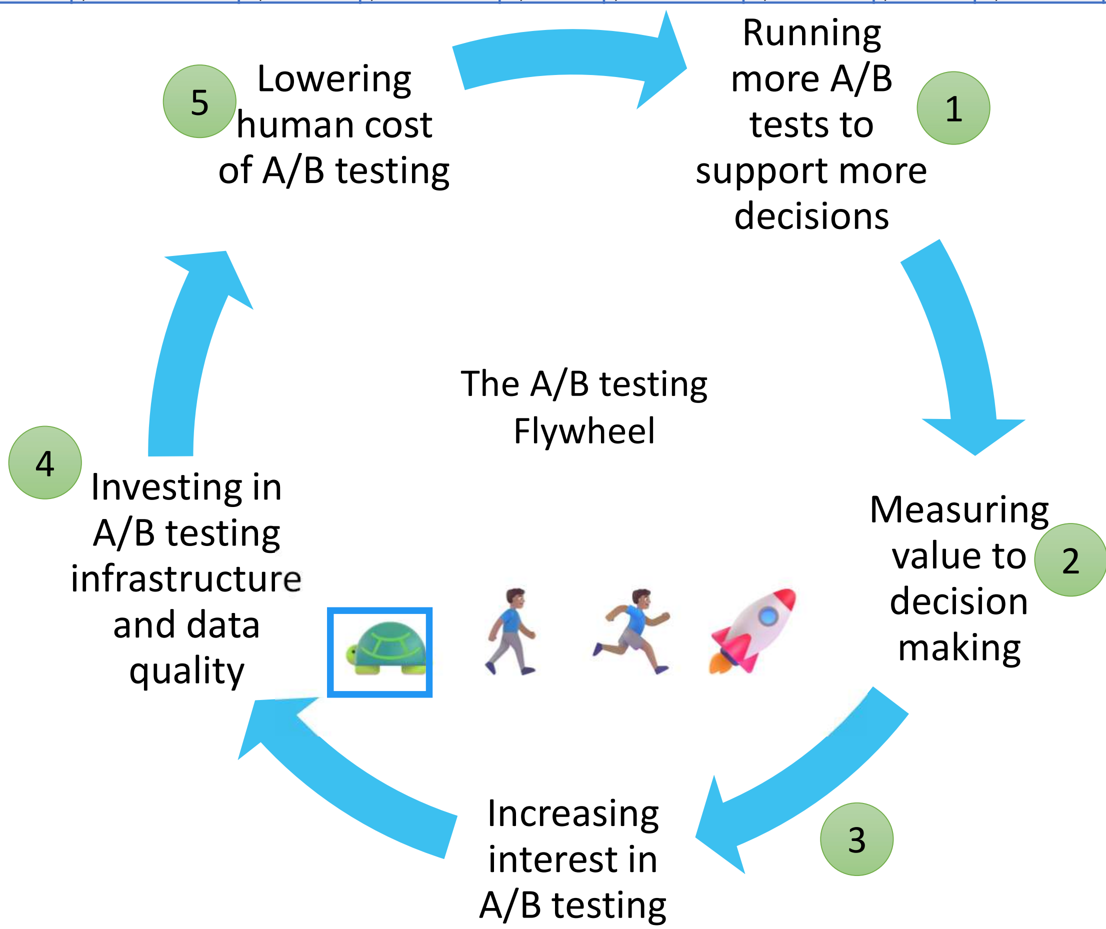
Plan

WIT MODE U		
7 Waste + In Lean (Muda)	**Findings**	**Mitigate Strategy**
Waiting	Delays such as slow load times for course recommendations or search results frustrate users and hinder engagement	Optimizing system performance and using caching techniques
Inventory	Irrelevant or unused courses sitting in the system without aligning with	Regular audits and clean-ups of course offerings
Transport	Inefficient data transfers between systems slow down processes, espe	Streamline data integration and ensure that communication between services is efficient through API optimization and faster data pipelines.
Motion	Excessive steps or clicks required by users to find the right course due to poor UI design	Improve UI/UX design to minimize user effort
Overproduction	Generate more content or features than users actually need	Regular user feedback surveys and data analysis
Defects	Errors in course recommendations or system bugs	Continuous testing and monitoring of the recommender system
Extra Processing	Overly complex algorithms or redundant steps that don't add value to the recommendations.	Simplify algorithms and focus on using only those processes that enhance recommendation relevance and user experience
Unused Talent	Underutilization of the team's expertise	Actively involve all team members in decision-making and problem-solving

Experimentations Life Cycle

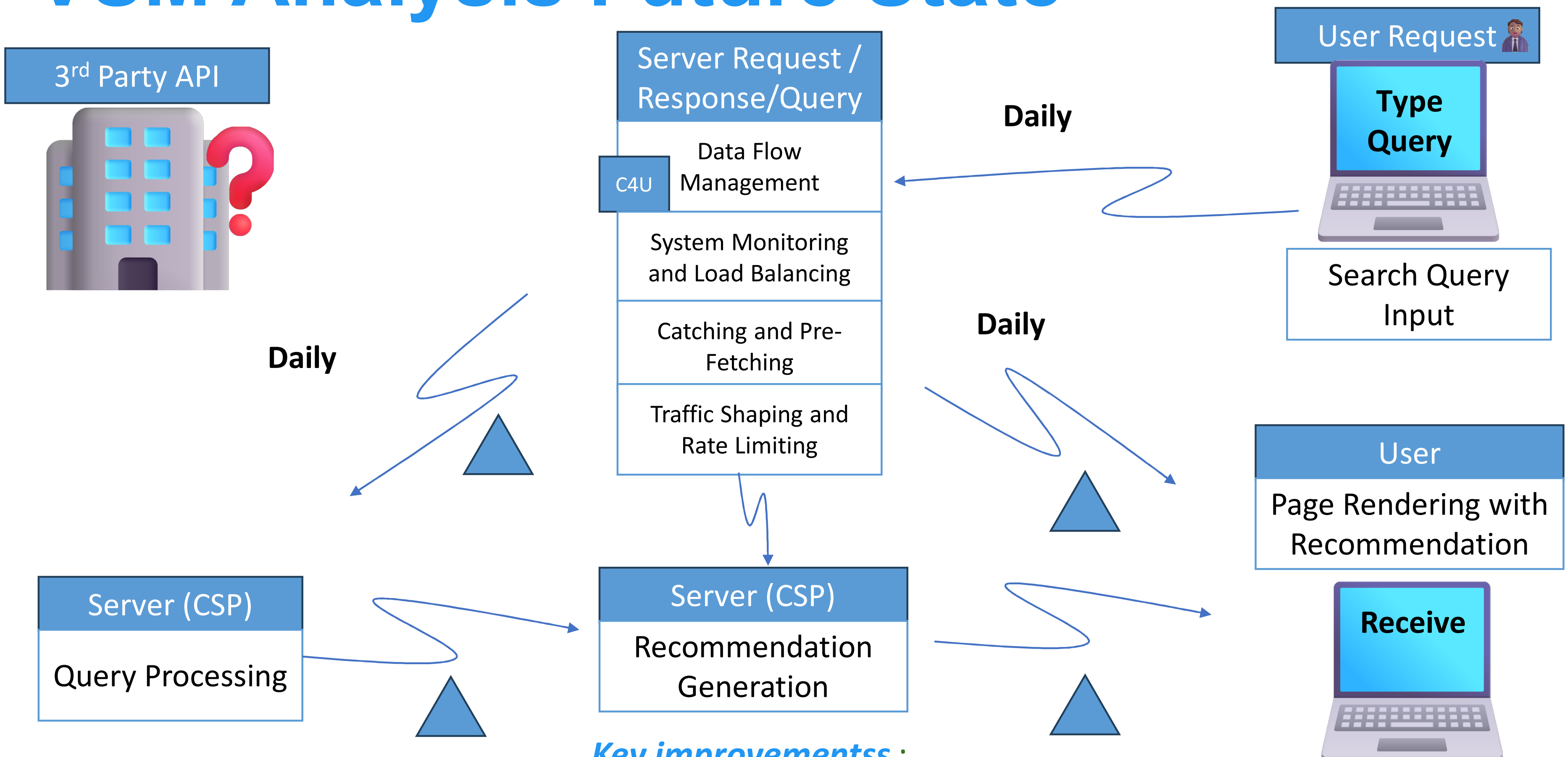
Plan

Monthly Revenue Forecast									
Revenue Y1	Increase Revenue Y1	Revenue Y2	Increase Revenue Y2	Revenue Y3	Increase Revenue Y3	Revenue Y4	Increase Revenue Y4	Revenue Y5	Increase Revenue Y5
$ 280,500.00	$ 25,500.00	$ 308,550.00	$ 28,050.00	$ 339,405.00	$ 30,855.00	$ 373,345.50	$ 33,940.50	$ 410,680.05	$ 37,334.55
User Y1	Increase User Y1	User Y2	Increase User Y2	User Y3	Increase User Y3	User Y4	Increase User Y4	User Y5	Increase User Y5
37,400	3,400	41,140	3,740	45,254	4,114	49,779	4,525	54,757	4,978



Key takeaways : C4U is a small company that has recently started generating monthly revenue in the six-figure range, exceeding a quarter of a million USD. The organization is still in the early stages, with a limited budget and resources..

VSM Analysis Future State

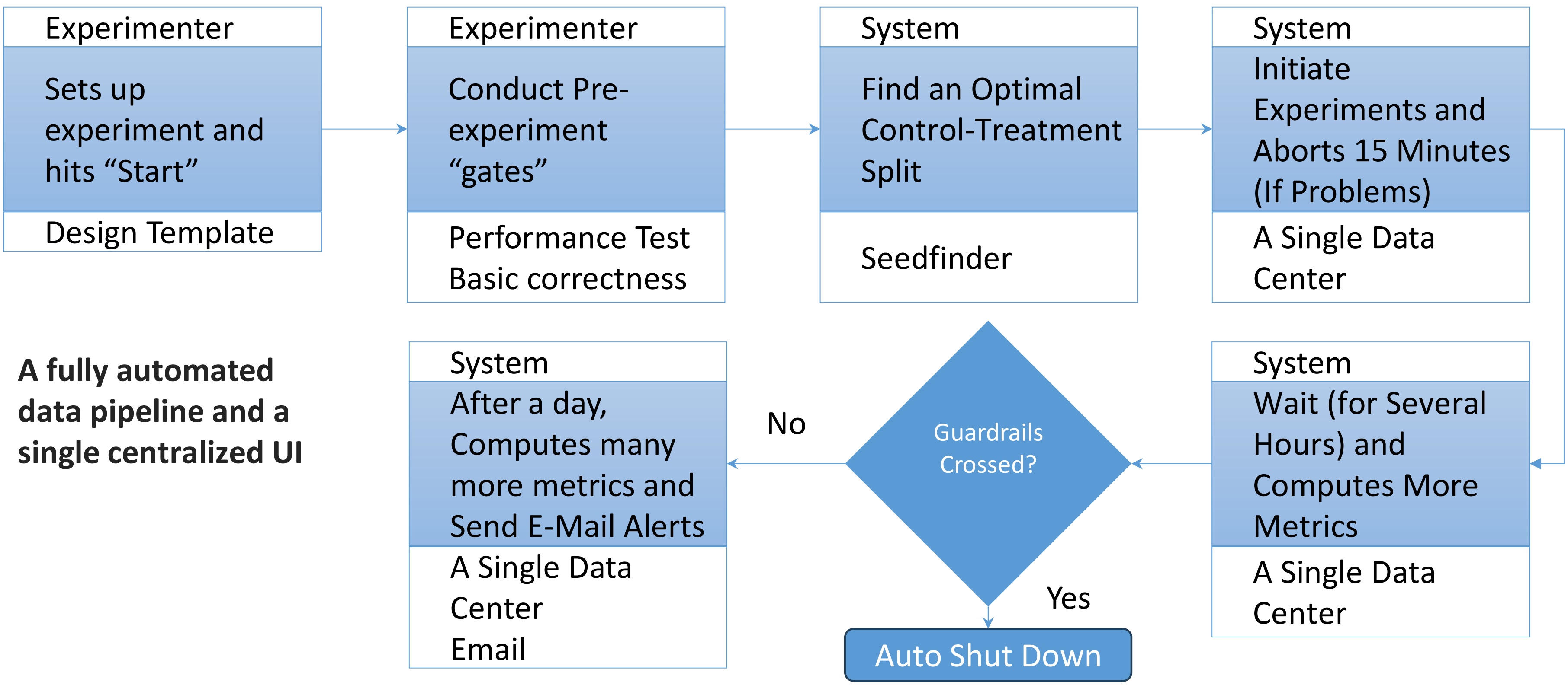


Key improvementss :
Efficiency, Personalization and Enhanced user experience

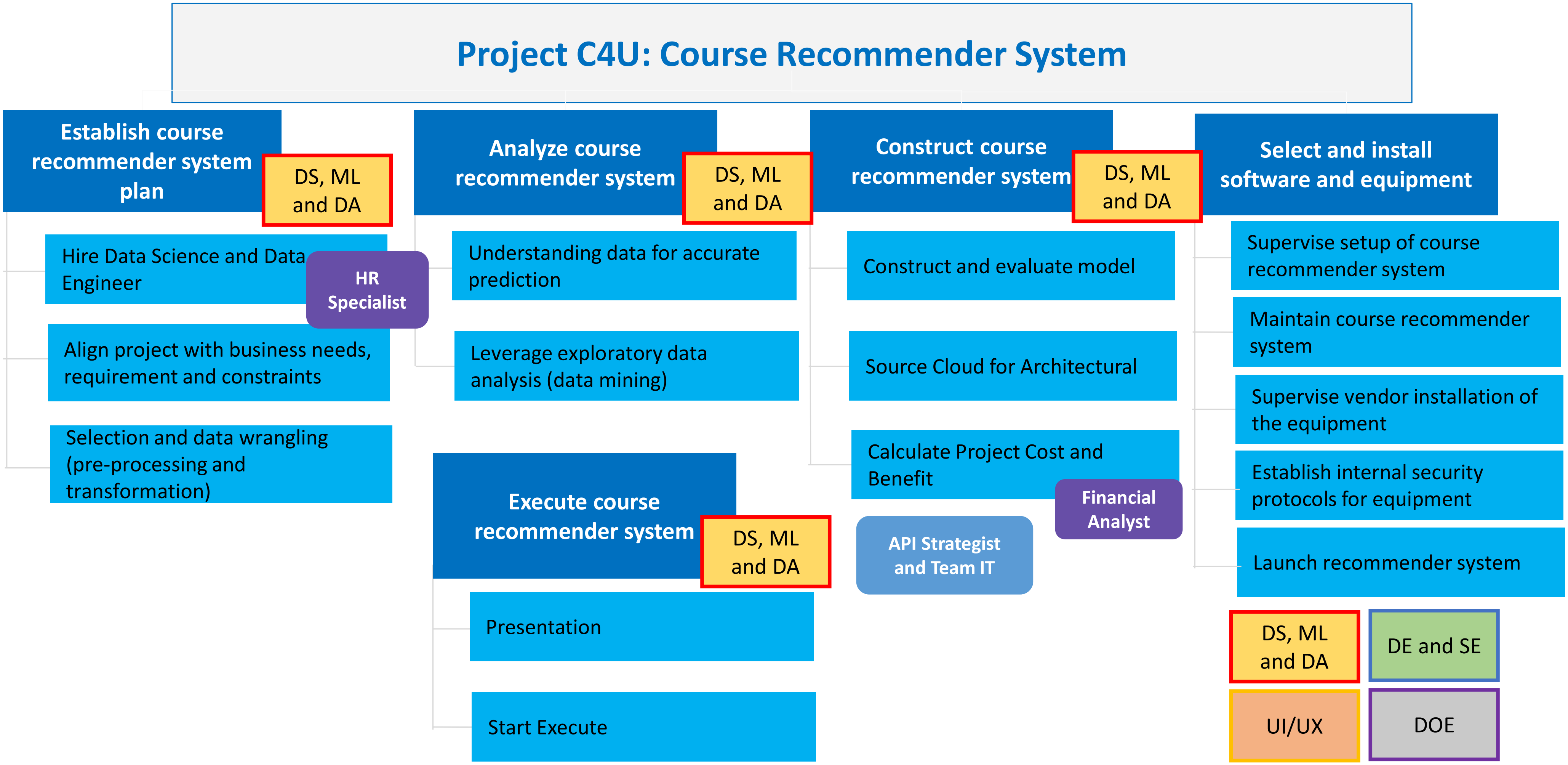
The Experimentation Platform (Future State)

Plan

- Experimentation Platform provides full experiment-lifecycle management



A fully automated data pipeline and a single centralized UI


Project C4U: Course Recommender System
Establish course recommender system plan
DS, ML and DA
Hire Data Science and Data Engineer
HR Specialist
Align project with business needs, requirement and constraints
Selection and data wrangling (pre-processing and transformation)
Analyze course recommender system
DS, ML and DA
Understanding data for accurate prediction
Leverage exploratory data analysis (data mining)
Execute course recommender system
DS, ML and DA
Presentation
Start Execute
Construct course recommender system
DS, ML and DA
Construct and evaluate model
Source Cloud for Architectural
Calculate Project Cost and Benefit
Financial Analyst
API Strategist and Team IT
Select and install software and equipment
Supervise setup of course recommender system
Maintain course recommender system
Supervise vendor installation of the equipment
Establish internal security protocols for equipment
Launch recommender system
DS, ML and DA
DE and SE
UI/UX
DOE

Project C4U: Course Recommender System

Promote course recommender system as to assist user in their study journey.

- Banner Ads on social media
- 3 press pieces published in relevant print and online publications

Marketing Promotion Manager

Actively and meaningfully engage the user to generate buy-in and project support.

- 3 live YouTube focused on transit talks
- 75% of Users surveyed before launch
- 70% of top users participate in user outreach program

UI/UX and PM

Establish an experimentation platform - software

- Implement A/B Testing
- Apply Multiarmed Bandit and Bayesian Optimization

DS, ML and DA

Marketing and TV Ads Campaign

- Marketing Ads
- Television Ads

Marketing Promotion Manager

WBS and RACI

Plan

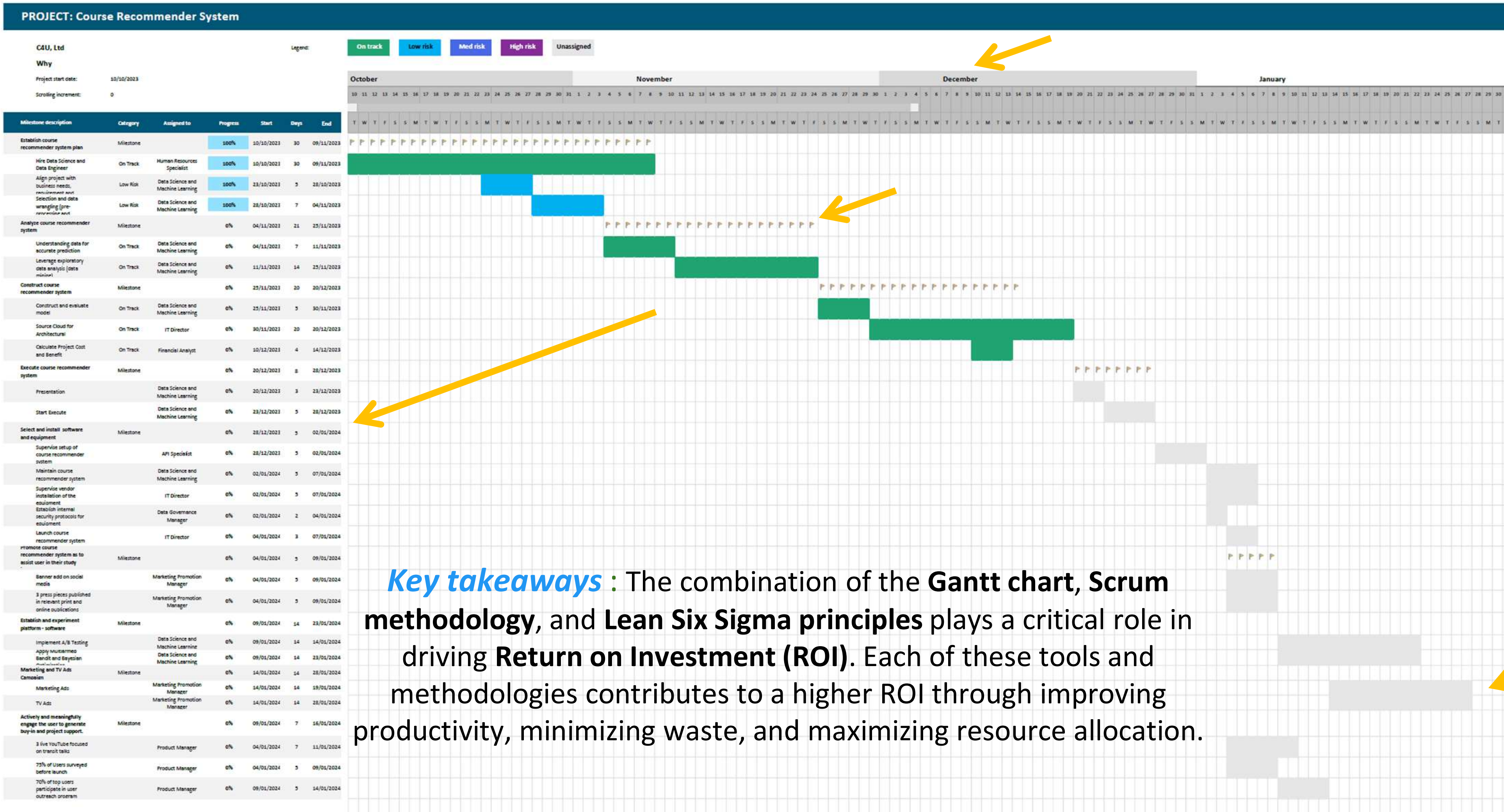
Course Rec Sys WBS Spreadsheet

R - Responsible	Completes the deliverable or task.
A - Accountable	Makes final decisions and signs off on task completion. Only 1 per task.
C - Consulted	An advisor, stakeholder, or subject matter expert who offers guidance before an action is taken.
I - Informed	Kept up to date on decisions made.

Milestone	ID	Task	Owner	Duration (days)	Notes	Director of Customer Data	Head of Data Science and ML	Existing Clients and Employees	Director of Procurement	IT Director	Product Manager	Team Data	MPM and AM	Team IT	Financial Analyst	HR
Establish course recommender system plan	1.1	Hire Data Science and Data Engineer	Human Resources Specialist	30	The Human Resource Specialist oversees hiring	I	A			I	I	C	I	I		R
	1.2	Align project with business needs, requirement and constraints	Data Science and Machine Learning	5	The Data Science and Data Engineer collect data	A	R		I	I	I	I	C	I	I	I
	1.3	Selection and data wrangling (pre-processing and transformation)	Data Science and Machine Learning	7	The Data Science and Machine Learning pre-process the data. Clean/clear, objective and valuable data acquired.	C	A		I	I	I	R	I	I	I	I
Analyze course recommender system	2.1	Understanding data for accurate prediction	Data Science and Machine Learning	7	The Data Science and Machine Learning focus and elaborate data statistically to understand more.	A	R		I	I	C	I	I	I	I	I
	2.2	Leverage exploratory data analysis (data mining)	Data Science and Machine Learning	14	The Data Science and Machine Learning focus and elaborate data statistically to uncover pattern and address ireggularities (data mining)	A	R		I	I	C	I	I	I	I	I
Construct course recommender system	3.1	Construct and evaluate model	Data Science and Machine Learning	5	The Data Science and Machine Learning choose the best model according to the metrics.	A	R		I	I	C	I	I	I	I	I
	3.2	Source Cloud for Architectural	IT Director	20	The IT Director sources materials	I	C		I	A	C	I	I	R	I	I
	3.3	Calculate Project Cost and Benefit	Financial Analyst	4	The Financial Analyst tracks costs and revenue	C	A		I	I	C	I	I	I	R	I
Execute course recommender system	4.1	Presentation	Data Science and Machine Learning	3	Share and present project result to the project team	A	R		I	I	C	I	I	I	I	I
	4.2	Start Execute	Data Science and Machine Learning	5	Initiate execution course recommender system after approval	A	R		I	I	C	I	I	I	I	I
Select and install software and equipment	5.1	Supervise setup of course recommender system	API Specialist	5	The Data Science and Machine Leraning setup fullstack with course recommender system and oversees software installation	I	C		I	A	C	I	I	R	I	I
	5.2	Maintain course recommender system	Data Science and Machine Learning	5	The Data Science and Machine Learning maintain course recommender system	A	R		I	I	C	I	I	I	I	I
	5.3	Supervise vendor installation of the equipment	IT Director	5	The IT Director ensures equipment is functional with the help of the Data Governance Manager/Quality Assurance Tester	I	C		I	A	C	I	I	R	I	I
	5.4	Establish internal security protocols for equipment	Data Governance Manager	2	The Data Governance Manager/Quality Assurance Tester ensures product quality and determines security protocols and best practices	I	C		I	A	C	R	I	I	I	I
	5.5	Launch course recommender system	IT Director	3	The IT Director launch the course recommender system											
Promote course recommender system as to assist user in their study journey.	6.1	Banner add on social media	Marketing Promotion Manager	5	The Marketing Promotion Manager develop click through banner add	A	C		I	I	C	I	R	I	I	I
	6.2	3 press pieces published in relevant print and online publications	Marketing Promotion Manager	5	The Marketing Promotion Manager publish relevan print and online publication	A	C		I	I	C	I	R	I	I	I
Establish and experiment platform - software	7.1	Implement A/B Testing	Data Science and Machine Learning	14	Test the Recommender System	I	C		I	I	R	I	A	I	I	C
	7.2	Apply Multiarmed Bandit and Bayesian Optimization	Data Science and Machine Learning	14	Test the Recommender System	I	C		I	I	R	I	A	I	I	C
Marketing and TV Ads Campaign	8.1	Marketing Ads	Marketing Promotion Manager	14	The Marketing Promotion Manager develop Product Service Announcement (PSA)	A	C		I	I	C	I	R	I	I	I
	8.2	TV Ads	Marketing Promotion Manager	14	The Marketing Promotion Manager develop TV Ads Campaign	A	C		I	I	C	I	R	I	I	I
Actively and meaningfully engage the user to generate buy-in and project support.	9.1	3 live YouTube focused on transit talks	Product Manager	7	The Product Manager develops the training sessions with the help of the Training Manager	I	C		I	I	R	I	A	I	I	C
	9.2	75% of Users surveyed before launch	Product Manager	5	The Training Manager runs the training program on the established protocols	I	C		I	I	R	I	A	I	I	C
	9.3	70% of top users participate in user outreach program	Product Manager	5	The Training Manager refines training processes	I	C		I	I	R	I	A	I	I	C

Key takeaways : The **RACI Matrix** is a responsibility assignment chart that outlines key activities and designates who is Responsible (R), Accountable (A), Consulted (C), and Informed (I).

Project Gantt Chart



Milestone description	Category	Assigned to	Progress	Start	Days	End
Establish course recommender system plan	Milestone		100%	10/10/2023	30	09/11/2023
Hire Data Science and Data Engineer	On Track	Human Resources Specialist	100%	10/10/2023	30	09/11/2023
Align project with business needs, requirement and	Low Risk	Data Science and Machine Learning	100%	23/10/2023	5	28/10/2023
Selection and data wrangling (pre-processing and	Low Risk	Data Science and Machine Learning	100%	28/10/2023	7	04/11/2023
Analyze course recommender system	Milestone		0%	04/11/2023	21	25/11/2023
Understanding data for accurate prediction	On Track	Data Science and Machine Learning	0%	04/11/2023	7	11/11/2023
Leverage exploratory data analysis (data mining)	On Track	Data Science and Machine Learning	0%	11/11/2023	14	25/11/2023
Construct course recommender system	Milestone		0%	25/11/2023	20	20/12/2023
Construct and evaluate model	On Track	Data Science and Machine Learning	0%	25/11/2023	5	30/11/2023
Source Cloud for Architectural	On Track	IT Director	0%	30/11/2023	20	20/12/2023
Calculate Project Cost and Benefit	On Track	Financial Analyst	0%	10/12/2023	4	14/12/2023
Execute course recommender system	Milestone		0%	20/12/2023	8	28/12/2023
Presentation		Data Science and Machine Learning	0%	20/12/2023	3	23/12/2023
Start Execute		Data Science and Machine Learning	0%	23/12/2023	5	28/12/2023
Select and install software and equipment	Milestone		0%	28/12/2023	5	02/01/2024
Supervise setup of course recommender system		API Specialist	0%	28/12/2023	5	02/01/2024
Maintain course recommender system		Data Science and Machine Learning	0%	02/01/2024	5	07/01/2024
Supervise vendor installation of the equipment		IT Director	0%	02/01/2024	5	07/01/2024
Establish internal security protocols for equipment		Data Governance Manager	0%	02/01/2024	2	04/01/2024
Launch course recommender system		IT Director	0%	04/01/2024	3	07/01/2024
Promote course recommender system as to assist user in their study	Milestone		0%	04/01/2024	5	09/01/2024
Banner add on social media		Marketing Promotion Manager	0%	04/01/2024	5	09/01/2024
3 press pieces published in relevant print and online publications		Marketing Promotion Manager	0%	04/01/2024	5	09/01/2024
Establish and experiment platform - software	Milestone		0%	09/01/2024	14	23/01/2024
Implement A/B Testing		Data Science and Machine Learning	0%	09/01/2024	14	14/01/2024
Apply Multiarmed Bandit and Bayesian		Data Science and Machine Learning	0%	09/01/2024	14	23/01/2024
Marketing and TV Ads Campaign	Milestone		0%	14/01/2024	14	28/01/2024
Marketing Ads		Marketing Promotion Manager	0%	14/01/2024	14	19/01/2024
TV Ads		Marketing Promotion Manager	0%	14/01/2024	14	28/01/2024
Actively and meaningfully engage the user to generate buy-in and project support.	Milestone		0%	09/01/2024	7	16/01/2024
3 live YouTube focused on transit talks		Product Manager	0%	04/01/2024	7	11/01/2024
75% of Users surveyed before launch		Product Manager	0%	04/01/2024	5	09/01/2024
70% of top users participate in user outreach program		Product Manager	0%	09/01/2024	5	14/01/2024

Key takeaways : The combination of the **Gantt chart**, **Scrum methodology**, and **Lean Six Sigma principles** plays a critical role in driving **Return on Investment (ROI)**. Each of these tools and methodologies contributes to a higher ROI through improving productivity, minimizing waste, and maximizing resource allocation.

Project Budget vs Cashflow

Monthly Revenue Forecast									
Revenue Y1	Increase Revenue Y1	Revenue Y2	Increase Revenue Y2	Revenue Y3	Increase Revenue Y3	Revenue Y4	Increase Revenue Y4	Revenue Y5	Increase Revenue Y5
$ 280,500.00	$ 25,500.00	$ 308,550.00	$ 28,050.00	$ 339,405.00	$ 30,855.00	$ 373,345.50	$ 33,940.50	$ 410,680.05	$ 37,334.55
User Y1	Increase User Y1	User Y2	Increase User Y2	User Y3	Increase User Y3	User Y4	Increase User Y4	User Y5	Increase User Y5
37,400	3,400	41,140	3,740	45,254	4,114	49,779	4,525	54,757	4,978

Increase Revenue | 10% | $280,500 per month

Budget: Project Course Recommender System **Operations & Training**

	TARGET BUDGET	ACTUAL/FINAL SPEND	UNDER/ OVER
	$ 154.936,00	$ -	$ 154.936,00

MILESTONES & TASKS	EMPLOYEE (LABOR)	HOURS (LABOR)	RATE (LABOR)	UNITS (MATERIALS)	$/UNIT(S) (MATERIALS)	TRAVEL (OTHER)	EQUIPMENT / SPACE (OTHER)	FIXED (OTHER)	MISC. (OTHER)	TOTAL OTHER (OTHER)	BUDGET	ACTUAL	UNDER/OVER
Milestone 1: Establish course recommender system plan													
Task 1: Hire Data Science and Data Engineer	Human Resources Specialist	80	$ 40,00			$ -	$ -	$ -	$ -	$ -	$ 3.200,00		$ (3.200,00)
Task 2: Align project with business needs, requirement and constraints	Data Science and Machine Learning	20	$ 65,00			$ -	$ -	$ -	$ -	$ -	$ 1.300,00		$ (1.300,00)
Task 3: Selection and data wrangling (pre-processing and transformation)	Data Science and Machine Learning	28	$ 65,00			$ -	$ -	$ -	$ -	$ -	$ 1.820,00		$ (1.820,00)
										Total	$ 6.320,00	$ -	
Milestone 2: Analyze course recommender system													
Task 1: Understanding data for accurate prediction	Data Science and Machine Learning	28	$ 65,00			$ -	$ -	$ -	$ -	$ -	$ 1.820,00		$ (1.820,00)
Task 2: Leverage exploratory data analysis (data mining)	Data Science and Machine Learning	56	$ 65,00			$ -	$ -	$ -	$ -	$ -	$ 3.640,00		$ (3.640,00)
										Total	$ 5.460,00	$ -	
Milestone 3: Construct course recommender system													
Task 1: Construct and evaluate model	Data Science and Machine Learning	20	$ 50,00			$ -	$ -	$ -	$ -	$ -	$ 1.000,00		$ (1.000,00)
Task 2: Source Cloud for Architectural	IT Director	8	$ 130,00			$ -	$ -	$ -	$ -	$ -	$ 1.040,00		$ (1.040,00)
Task 3: Calculate Project Cost and Benefit	Financial Analyst	16	$ 30,00			$ -	$ -	$ -	$ -	$ -	$ 480,00		$ (480,00)
										Total	$ 2.040,00	$ -	
Milestone 4: Execute course recommender system													
Task 1: Presentation										-	$ 780,00		$ (780,00)
Task 2: Start Execute										-	$ 1.300,00		$ (1.300,00)
										Total	$ 2.080,00	$ -	
Milestone 5: Select and install software and equipment													
Task 1: Supervise setup of course recommender system										-	$ 800,00		$ (800,00)
Task 2: Maintain course recommender system										-	$ 1.300,00		$ (1.300,00)
Task 3: Supervise vendor installation of the equipment										-	$ 2.600,00		$ (2.600,00)
Task 4: Establish internal security protocols for equipment	Data Governance Manager	8	$ 45,00			$ -	$ -	$ -	$ -	$ -	$ 360,00		$ (360,00)
Task 5: Launch course recommender system										$ 8.000,00	$ 9.560,00		$ (9.560,00)
										Total	$ 14.620,00	$ -	
Milestone 6: Promote course recommender system as to assist													
Task 1: Banner add on social media										8.000,00	$ 8.900,00		$ (8.900,00)
Task 2: 3 press pieces published in relevant print and online publicati										5.500,00	$ 6.400,00		$ (6.400,00)
										Total	$ 15.300,00	$ -	
Milestone 7: Establish and experiment platform - software													
Task 1: Implement A/B Testing										$ 13.000,00	$ 16.640,00		$ (16.640,00)
Task 2: Apply Multiarmed Bandit and Bayesian Optimization	Data Science and Machine Learning	56	$ 65,00			$ -	$ -	$ -	$ 2.000,00	$ 2.000,00	$ 5.640,00		$ (5.640,00)
										Total	$ 22.280,00	$ -	
Milestone 8: Marketing and TV Ads Campaign													
Task 1: Marketing Ads										12.500,00	$ 15.020,00		$ (15.020,00)
Task 2: TV Ads										25.000,00	$ 27.520,00		$ (27.520,00)
										Total	$ 42.540,00	$ -	
Milestone 9: Actively and meaningfully engage the user to genera project support.													
Task 1: 3 live YouTube focused on transit talks										$ 9.000,00	$ 10.316,00		$ (10.316,00)
Task 2: 75% of Users surveyed before launch	Product Manager	20	$ 47,00			$				$ 10.000,00	$ 10.940,00		$ (10.940,00)
Task 3: 70% of top users participate in user outreach program	Product Manager	20	$ 47,00			$				$ 18.500,00	$ 19.440,00		$ (19.440,00)
										Total	$ 40.696,00	$ -	
Reserve buffer											$ 3.600,00		
TOTAL											$ 154.936,00	$ -	

Launch recommender system | $8,000

Promote course recommender system | $13,500

Establish experiment platform | $15,000

Marketing and TV ads campaign | $27,500

Actively engage with users | $37,500

$155,000

Monthly Expenses	Expenses	Percent per Revenue
Developer Salaries and Other Overheads	$ 37.950.00	14.88%
Development and maintenance costs	$ 10.000.00	3.92%
Frontend Cloud Cost	$ 87.30	0.03%
Backend Framework Cost	$ 59.95	0.02%
Backend Cloud Cost	$ 173.95	0.07%
Backend Caching Cost	$ 12.96	0.01%
Backend Message Queues Cost	$ 50.00	0.02%
		2%
		4%
Total Monthly Expenses	**$ 49.272.74**	**19.32%**
Cost per user	$ 1.45	

Baseline C4U Apps (Monthly) | $50,000

Additional Monthly Expenses		Percent per Revenue
Data Analytics Salaries and Other		
Total Additional Monthly Expenses from Recommender System	**$ 24,341.70**	**9.55%**

Additional monthly expenses from Recommender Systems | $25,000

Increase Expense All Year	10%
Discount Rate	10%
Initial Investment	$ 87.833,40
Year 1 Cash Flow	[illegible]
Year 4 Cash Flow	[illegible]
IRR	**14,36%**

NPV Expected | $12,000

IRR Expected | 14%

Root Causes and Risk Mitigation

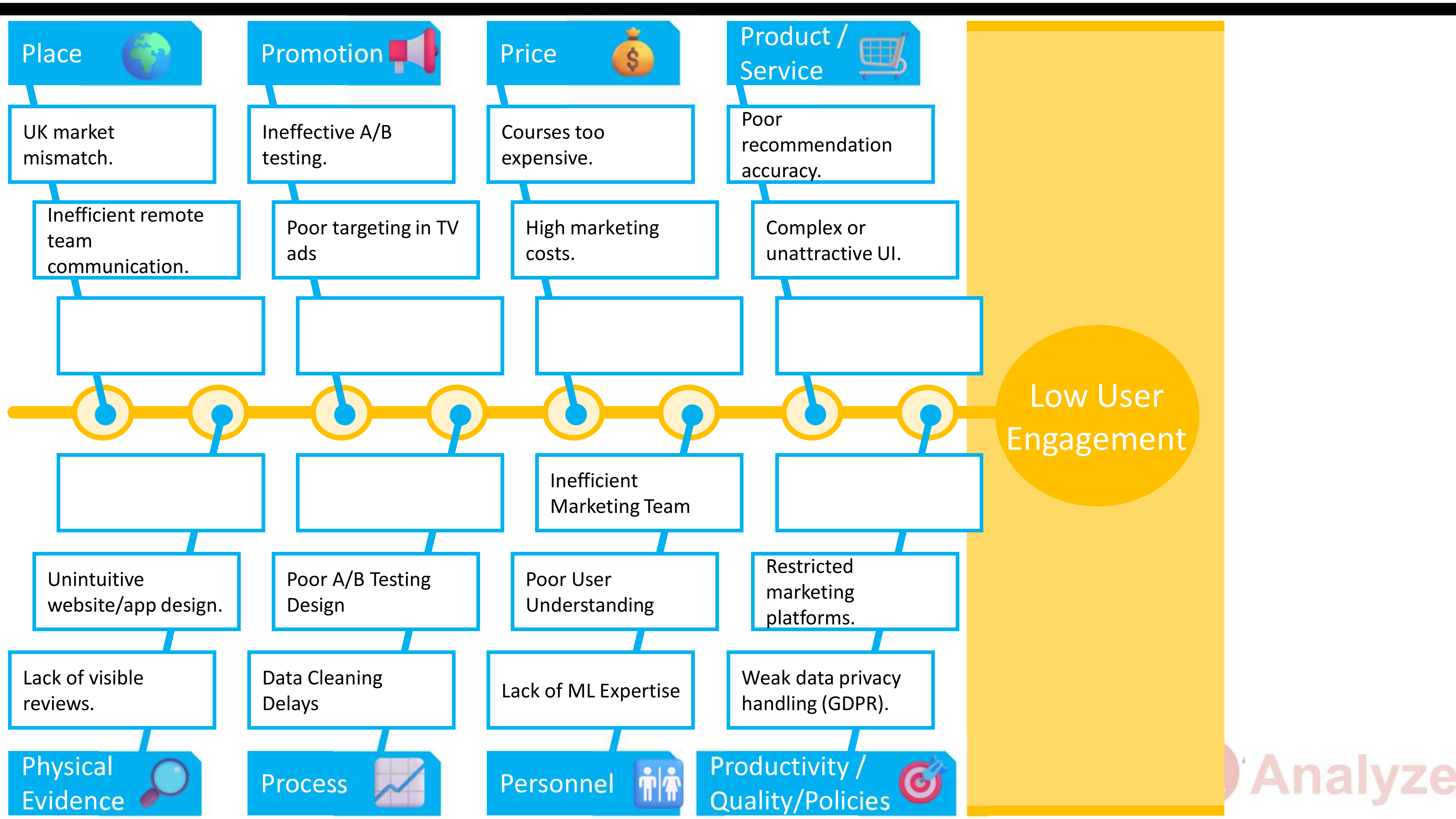
Project Management Flowchart

Project Initiation (PI)	Project Planning (PP)	Project Execution (PE)	Agile Scrum Project	Lean Six Sigma Project
1. OKR and SMART 2. Stakeholder Analysis 3. Project Charter	1. WBS and RACI 2. Gantt Chart 3. Risk Management 4. Project Budget 5. Communication Plan – Team Profile	1. Customer Survey 2. ROAM Analysis 3. Status Report 4. Hours Burndown	1. Sprint Backlog 2. Sprint Retrospective	1. Yellow Belt 2. Green Belt 3. Black Belt 4. Master Black Belt
1. VOC 2. CTQ – CTP – CTR - COPQ 3. SIPOC	1. KPI Trees 2. Lean Measures 3. Operational Definitions	1. Data Collection and Sample 2. MSA 3. Process Capability	1. The Process Door 2. EDA 3. Experimentation Platform	1. SPC

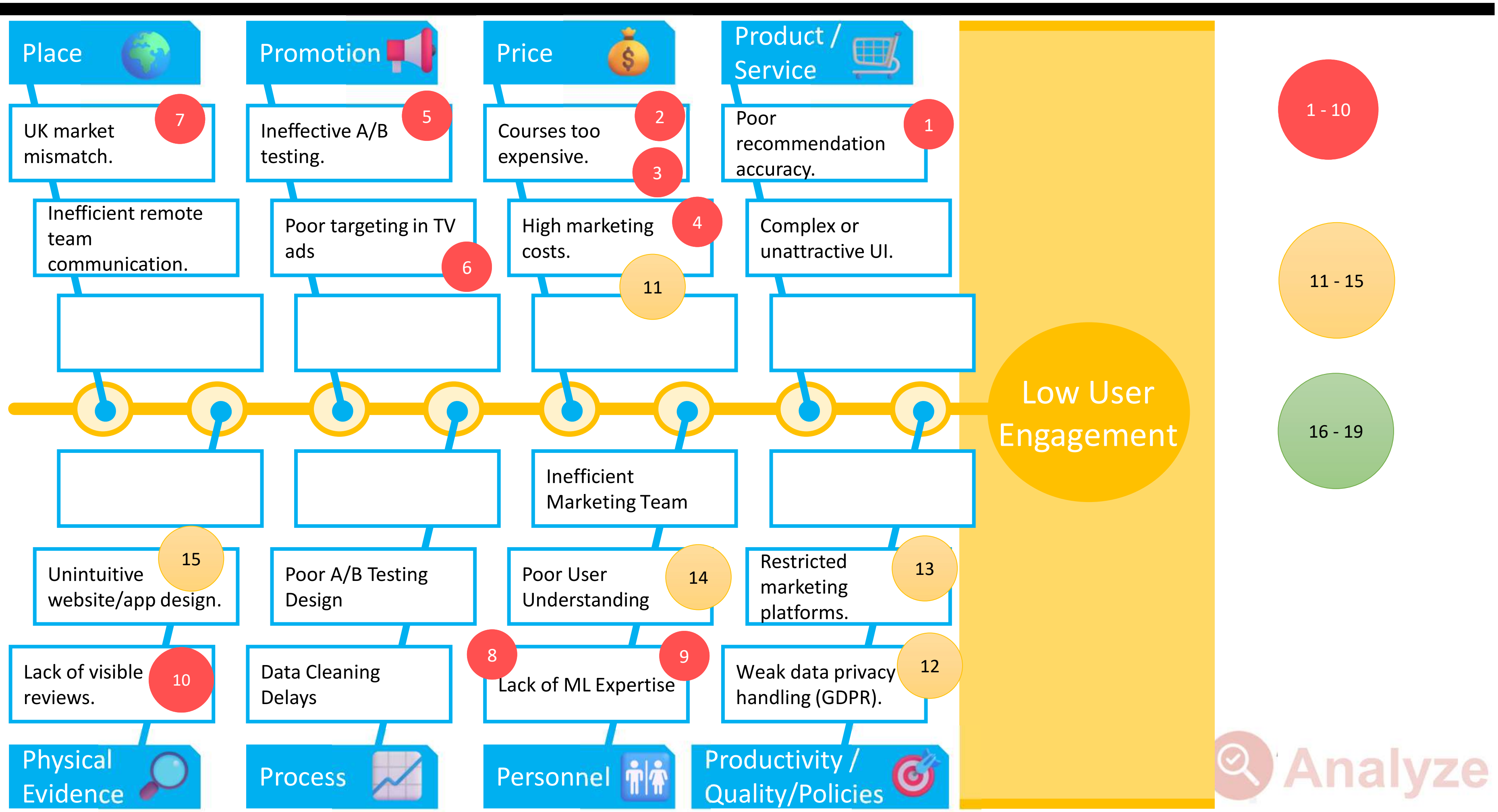
1. Team Meeting Agenda
2. Email management (Escalation, Invitation, Release Plan, etc.)

Plan

Potential Causes

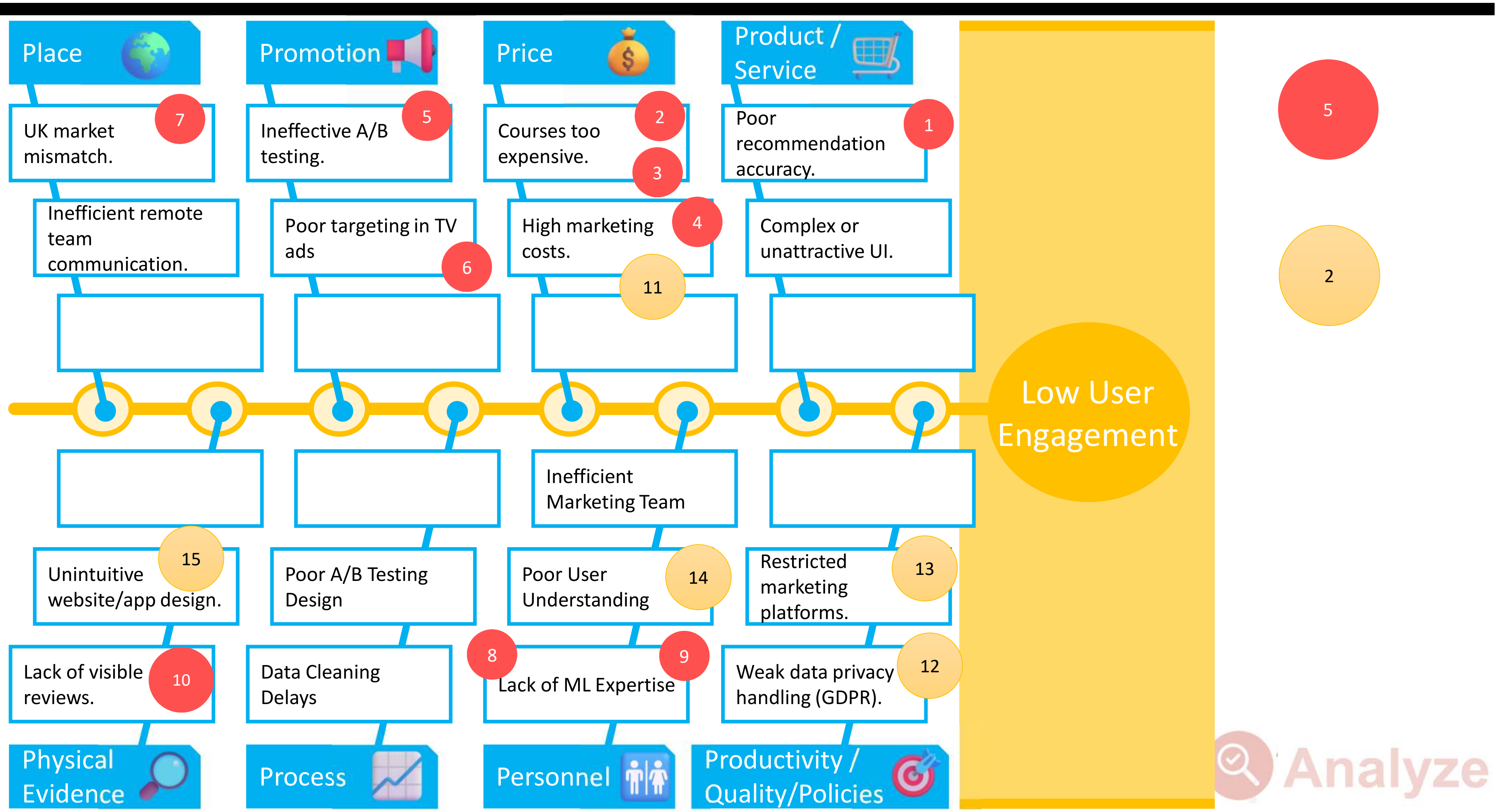


FMEA



Place
Promotion
Price
Product / Service
UK market mismatch.
7
Ineffective A/B testing.
5
Courses too expensive.
2
3
Poor recommendation accuracy.
1
Inefficient remote team communication.
Poor targeting in TV ads
6
High marketing costs.
4
11
Complex or unattractive UI.
Low User Engagement
Inefficient Marketing Team
Unintuitive website/app design.
15
Poor A/B Testing Design
Poor User Understanding
14
Restricted marketing platforms.
13
Lack of visible reviews.
10
Data Cleaning Delays
8
Lack of ML Expertise
9
Weak data privacy handling (GDPR).
12
Physical Evidence
Process
Personnel
Productivity / Quality/Policies
1 - 10
11 - 15
16 - 19
Analyze

Data Science and Machine Learning



Data Science and Machine Learning FMEA

FMEA

Process/Product Name: Course Recommendation System
Responsible: Data Science and Machine Learning
Prepared By: Wahyu
FMEA Date (Orig.):

Process Step/Input	Potential Failure Mode	Potential Failure Effects	SEVERITY (1 - 10)	Potential Causes	OCCURRENCE (1 - 10)	Current Controls	DETECTION (1 - 10)	RPN	Action Recommended
What is the process step, change or feature under investigation?	In what ways could the step, change or feature go wrong?	What is the impact on the customer if this failure is not prevented or corrected?		What causes the step, change or feature to go wrong? (how could it occur?)		What controls exist that either prevent or detect the failure?			What are the recommended actions for reducing the occurrence of the cause or improving detection?
Data Collection & Processing	*Missing or inaccurate data collected from users*	*Poor quality recommendations, leading to low engagement*	*8*	*Incomplete user profiles, faulty data sources*	*6*	*Data validation scripts that check for missing or malformed values; daily data integrity checks*	*4*	*192*	*Implement real-time data validation processes that flag and fix anomalies before they propagate into the system*
Recommendation Algorithm	*Algorithm fails to provide relevant course recommendations*	*Users may not find the courses they want, leading to churn*	*9*	*Poor algorithm tuning, lack of diversity in recommendations*	*6*	*Periodic retraining of the model using new data; recommendation model is tested with sample users before production*	*5*	*270*	*Regularly evaluate the algorithm with A/B testing on live users to ensure relevance, and introduce diversity in recommendation logic (e.g., collaborative filtering + content-based)*
User Interface & Experience	*Poor or confusing UI/UX makes it difficult for users to find recommendations*	*Frustration, drop-off in user engagement, and poor user retention*	*7*	*Unintuitive design, poorly placed recommendation widgets*	*4*	*Basic UI/UX testing during development; user focus groups conducted to get feedback*	*6*	*168*	*Perform extensive A/B testing on various UI designs; refine UI layout based on heatmaps to ensure ease of navigation for recommendation widgets*
System Performance	*System lags or crashes during high-traffic periods*	*Users will abandon the system, leading to dissatisfaction*	*10*	*Recommendations don't improve over time, leading to static system*	*4*	*Load testing is done periodically to simulate traffic spikes; basic server monitoring (CPU, RAM, disk space) is in place*	*6*	*240*	*Scale infrastructure automatically based on user traffic; Implement load balancing across multiple servers to handle traffic surges*
Feedback Loop	*User feedback on recommendations is not incorporated*	*Late response to underperformance*	*6*	*Lack of a proper feedback mechanism, or ignoring feedback*	*6*	*User feedback is collected through star ratings, but this is not tied directly to algorithm improvement*	*4*	*144*	*Develop mechanisms for users to provide more granular feedback (e.g., thumbs up/down on specific courses); Use feedback to fine-tune the algorithm*
Data Security & Privacy	*User data is compromised or used inappropriately*	*Loss of user trust, legal issues, penalties*	*10*	*Inadequate data encryption, lack of compliance with regulations*	*3*	*Basic data encryption (SSL for transmission); periodic security audits*	*5*	*150*	*Ensure end-to-end encryption, perform more frequent security audits, and ensure compliance with GDPR, CCPA, and other data protection laws*

Key takeaways & action items: To achieve 10% increase in revenue we regularly evaluate algorithm with **A/B testing.**

Key takeaways & action items: We also conduct **A/B testing** on various UI designs

Marketing FMEA

FMEA

Process/Product Name: Marketing Campaign
Responsible: Marketing Promotion Manager

Prepared By: Wahyu
FMEA Date (Orig.): (Rev.):

Process Step/Input	Potential Failure Mode	Potential Failure Effects	SEVERITY (1 - 10)	Potential Causes	OCCURRENCE (1 - 10)	Current Controls	DETECTION (1 - 10)	RPN	Action Recommended	Resp.	Actions Taken	VERITY (1 - 10)	CURRENCE (1 - 10)	ETECTION (1 - 10)	RPN
What is the process step, change or feature under investigation?	In what ways could the step, change or feature go wrong?	What is the impact on the customer if this failure is not prevented or corrected?		What causes the step, change or feature to go wrong? (how could it occur?)		What controls exist that either prevent or detect the failure?			What are the recommended actions for reducing the occurrence of the cause or improving detection?	Who is responsible for making sure the actions are completed?	What actions were completed (and when) with respect to the RPN?				
Ad Design & Creation	Ad message is unclear or not compelling	Low engagement with the ads, low CTR (low customer interest, reduced clicks)	8	Poor copywriting, unclear message, Lack of understanding of user needs or inaccurate audience segmentation	5	Initial audience research done using basic analytics tools; Marketing personas created based on existing users' data	4	160	Conduct in-depth surveys and interviews to refine personas; Use A/B testing on multiple ad copies to see which messaging resonates best. Ensure clear, value-driven messaging that speaks to user pain points.	Marketing Promotion Manager	Conducted user surveys and interviews to understand needs. A/B tests on multiple ad copies led to optimized messaging that increased click-through rate (CTR) by 15% compared to the initial ads.	8	4	2	64
Audience Identification	Wrong target audience is selected	High bounce rate, low conversions	9	Inaccurate customer segmentation	6	Monthly review	5	270	Conduct thorough customer analysis; run small-scale A/B tests on different audience segments before full campaign.	Marketing Promotion Manager	Collaborate with data science and machine learning dept running AB Testing	8	4	3	96
Social Media Strategy	Social media posts are inconsistent or irrelevant	Reduced engagement, poor brand visibility	7	Poor content planning, mismatch with audience	4	Monthly review	5	140	Develop a social media content calendar aligned with target audience interests and campaign objectives.	Marketing Promotion Manager	Social media calender - target audiences interest - campaign objectives	7	4	3	84
Budget Allocation	Over-budget in one channel (e.g., TV)	Financial overspend, reduced ROI	8	Inaccurate budget forecasting or resource waste	5	Tracking by monthly financial report	3	120	Use a marketing budget planner; allocate resources based on channel performance metrics.	Marketing Promotion Manager	Project budget and project controller work with finance and accounting and monitor ROAS	8	4	3	96
Performance Monitoring	Campaign metrics are not tracked in real-time	Late response to underperformance	6	Inadequate use of tracking tools	6	Monthly review	4	144	Implement real-time campaign tracking tools; set up dashboards for key metrics (CTR, conversion rate).	Marketing Promotion Manager	Campaign tracking tools	6	4	3	72
Customer Engagement	Leads are not followed up promptly	Low conversion from leads to customers	9	Poor lead nurturing, slow response time	6	Manual follow up leads	6	324	Schedule lead nurturing emails immediately after capturing leads using autamatef.	Course Manager	Implement automated lead follow-up through CRM	9	4	3	108
Ad Placement Strategy	Inefficient ad placement, missing key platforms	Reduced reach and visibility, lower conversion rates	7	Failure to choose the right platforms where the target audience is active	6	Ads placed on general platforms (Google, Facebook); Basic analytics used to track impressions and clicks	5	210	Conduct audience behavior analysis to identify underutilized platforms (e.g., LinkedIn for professionals, YouTube for tutorials); Allocate ad spend based on platform performance	Ads Manager	Analyzed user behavior and found underutilized platforms like LinkedIn and YouTube. Reallocated budget towards those platforms, improving overall reach by 20% and boosting conversions from professionals in particular.	7	4	3	84
Press Release & Media	Delayed or poorly distributed press releases	Reduced media coverage and brand awareness	6	Inefficient coordination with media outlets, delays in approval	4	Press releases sent to a few key outlets manually	7	168	Build relationships with a larger network of journalists, influencers, and bloggers; Use automated tools to schedule timely releases	PR Manager	Automated press release distribution and built a media relationship pipeline. Press coverage increased by 25%, and timely publication ensured visibility aligned with campaign	8	4	3	96
Social Media Ads	Low engagement rates on social media campaigns	Lower-than-expected ROI, wasted budget	8	Unattractive visuals, weak call-to-action (CTA)	6	Basic social media engagement tracking (likes, shares, and comments)	6	288	Enhance visuals and CTA with A/B testing; Use interactive elements like polls and quizzes; Tailor ads for each platform (e.g., Instagram vs LinkedIn)	Ads Manager	Revamped visuals and CTAs with a focus on A/B testing. Engagement on social media improved by 18%, with polls and interactive quizzes increasing direct interaction and ad performance on platforms like Instagram and Facebook.	8	4	3	96
Budget Management	Budget overrun due to poor cost tracking	Depletion of marketing funds, leading to campaign cutback	9	Inefficient allocation, unexpected expenses	3	Weekly budget reviews by the finance team, using basic expense tracking software	6	162	Implement real-time budget tracking and automated alerts for over-budget items; Forecast using dynamic spend analysis based on real-time ad performance	Marketing Promotion Manager	[illegible] tracking software that sent real-time alerts when costs approached set limits. This led to a 10% reduction in unexpected expenses and better control of the [illegible]	9	3	3	81
User Conversion Rate	Low conversion rate from ad click to signing up for the course	Low ROI, fewer course sign-ups	8	Poor ad targeting, ineffective landing pages	7	Basic conversion tracking using Google Analytics; Weekly reviews of conversion data	4	224	Conduct A/B tests on landing pages to improve conversion rates; Optimize the user journey from ad click to course sign-up with personalized landing pages	Marketing Promotion Manager	A/B testing of landing pages resulted in a 12% improvement in conversion rates. Personalized landing pages for different user personas were created, aligning better with the audience's needs and journey.	8	4	3	96
Timing of Campaign	Campaign launched at a suboptimal time	Low user interest, low traffic	7	Lack of coordination with key seasonal trends or competing events	5	Ads are timed based on previous marketing schedules and intuition	6	210	Analyze user behavior and seasonality data to schedule campaigns during peak times (e.g., exam prep seasons for courses); Optimize ad frequency	Marketing Promotion Manager	Adjusted campaign schedules to [illegible] (e.g., before exams). Increased sign-up rates by 20% during high-demand periods, compared to campaigns launched at random times in the past.	7	4	3	84
Competitor Activity	Competitor runs simultaneous campaigns	Marketing noise and reduced effectiveness of the course campaign	6	Lack of awareness of competitors' marketing schedules	4	Basic competitor monitoring using general tools like Google Alerts; React to competitor campaigns after they've been noticed	6	144	Set up proactive monitoring for competitor activities and market trends; Plan campaigns around or in anticipation of competitor launches	Competitive Intelligence Analyst	Implemented competitor activity monitoring using specialized tools. Planned the launch of our ads one week before a major competitor's campaign, leading to 10% higher engagement during that competitive period.	6	4	3	72

Key takeaways & action items: We apply prediction model on marketing campaign and TV ads.

Key takeaways & action items: **A/B testing** was apply in Ads copies, social media Ads and landing pages.

Experimentations at C4U

Randomized Experiments
(With randomization over treatment and control group)

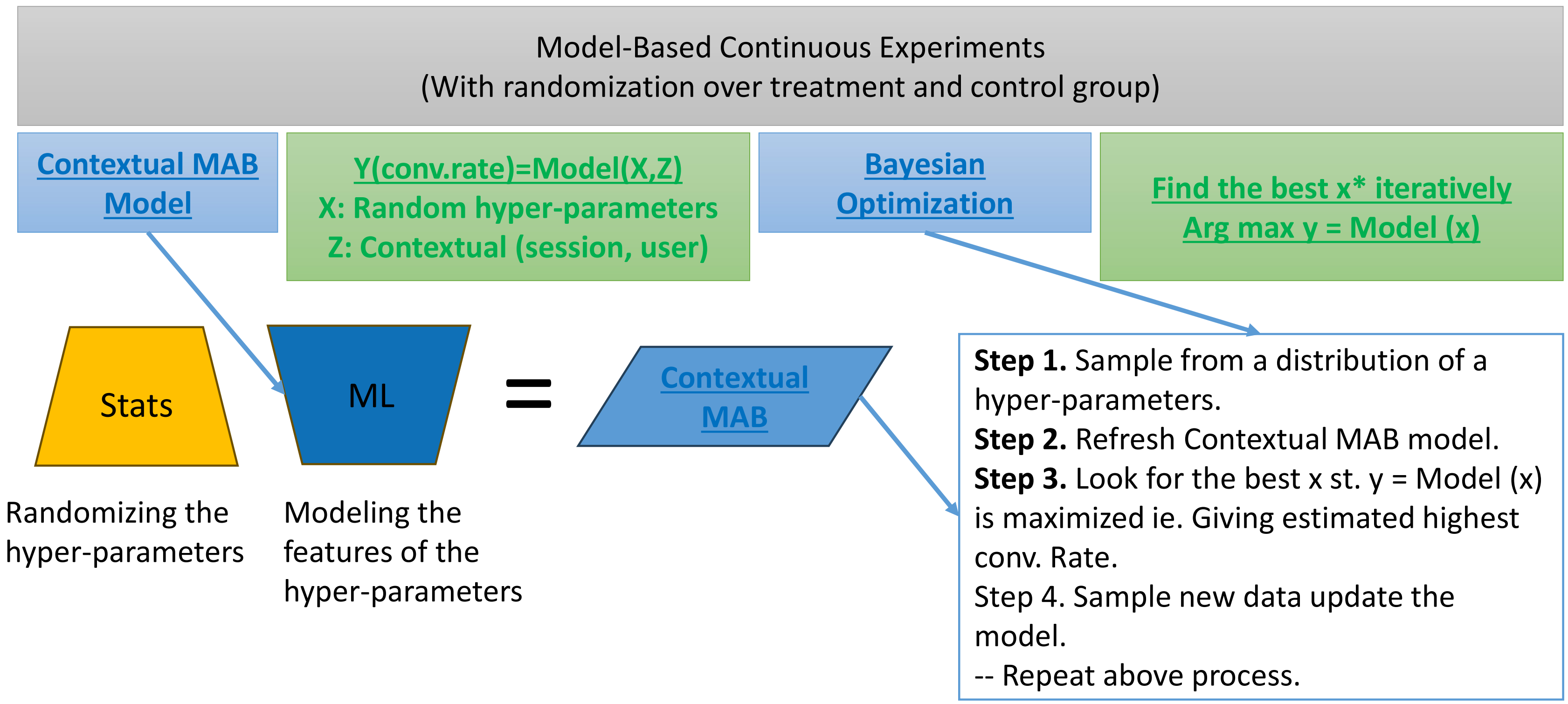
- Classic Experiments (Non-Recurring)
  - Univariate Tests
    - A/B Tests
- Continuous Experiments (Recurring)
  - Statistical Techniques
    - Allocation % — **Eg. Thompson Sampling**
    - Rollout % — **Eg. Power Based via Sequential Tests** — **Eg. Risk Based**
  - Model-Based Techniques
    - Contextual MAB
    - Bayesian Optimization

Observational Studies
(Pure Observation with no randomization)

- Different Methods to Estimate Associated Lifts
- Synthetic Control
- A/B-like: How to construct a weighted “control” group
- Regression

Plan

Experimentations at C4U



A/B Testing

- To get the probability that we need to apply Bayes Rule, which requires a prior $P(H_0)$
- FPR, or False Positive Risk/Rate, is a proxy: the probability that is H_0 is true (no real effect) when the test was statistically significant
- Assuming the experiment was properly run with 80% power, here is a usefull table relative to reported success rates by companies and p-value threshold (alpha) of 0.05 for statistical significance

Company/Source	Success Rate	FPR	Revenue/Benefit Estimate from A/B Testing
Microsoft	33%	5.9%	$517 million increase in revenue (from various A/B tests improving features like Bing search speed and Office product optimization)
Bing	15%	15%	
Avinash Kaushnik	20%	11.1%	- (Revenue estimates not directly available)
Booking.com, Google Ads, Netflix	10%	22%	Booking.com: Reportedly added $1 billion in annual revenue from continuous experimentation improvements Google generating over $224 billion in ad revenue in 2022 Netflix increased its subscriber base by 30 million in 2022. The new user model, designed based on A/B test results, reduced churn and enhanced customer satisfaction, significantly contributing to its revenue and market value growth
Airbnb Search	8%	26.4%	In 2018, Airbnb's A/B testing on search algorithms reportedly resulted in $80 million in additional revenue

Project Execution: Customer Survey, ROAM Analysis and Scrum

Project Management Flowchart

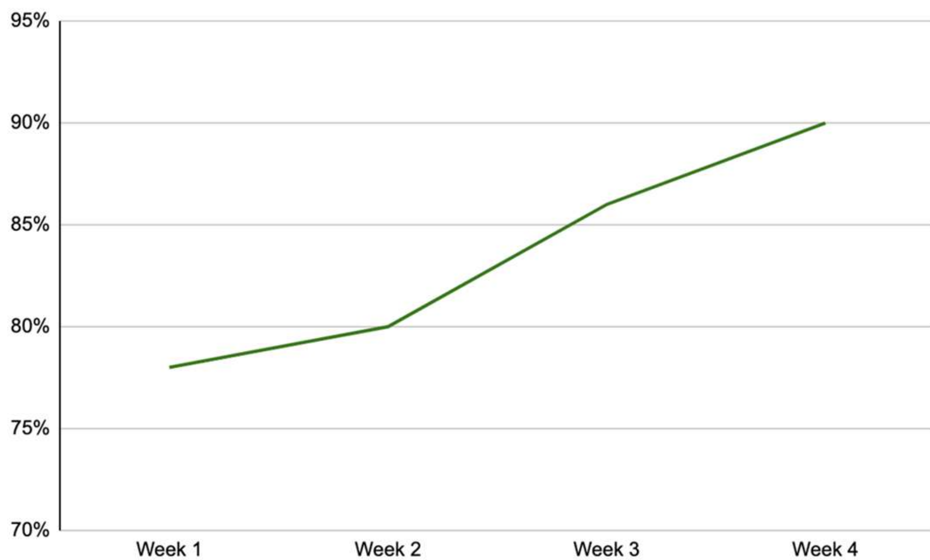
Project Initiation (PI)	Project Planning (PP)	Project Execution (PE)	Agile Scrum Project	Lean Six Sigma Project
1. OKR and SMART 2. Stakeholder Analysis 3. Project Charter	1. WBS and RACI 2. Gantt Chart 3. Risk Management 4. Project Budget 5. Communication Plan – Team Profile	1. Customer Survey 2. ROAM Analysis 3. Status Report 4. Hours Burndown	1. Sprint Backlog 2. Sprint Retrospective	1. Yellow Belt 2. Green Belt 3. Black Belt 4. Master Black Belt
1. VOC 2. CTQ – CTP – CTR - COPQ 3. SIPOC	1. KPI Trees 2. Lean Measures 3. Operational Definitions	1. Data Collection and Sample 2. MSA 3. Process Capability	1. The Process Door 2. EDA 3. Experimentation Platform	1. SPC

1. Team Meeting Agenda
2. Email management (Escalation, Invitation, Release Plan, etc.)

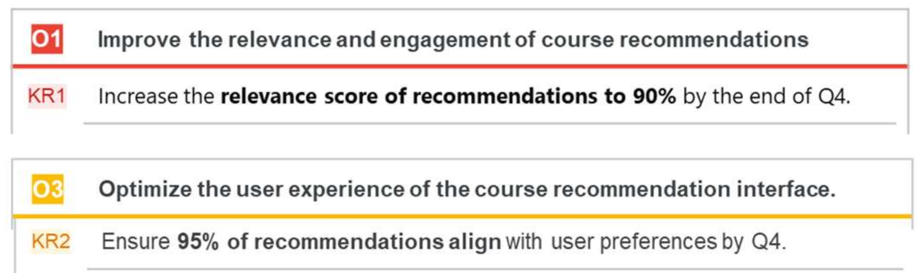
Plan

Customer Survey

We surveyed 50 C4U Pals test batch customers over a four-week period to learn about their satisfaction with the product, delivery process, and customer support.



Did you participate in the course we recommend for you?



O1 **Improve the relevance and engagement of course recommendations**

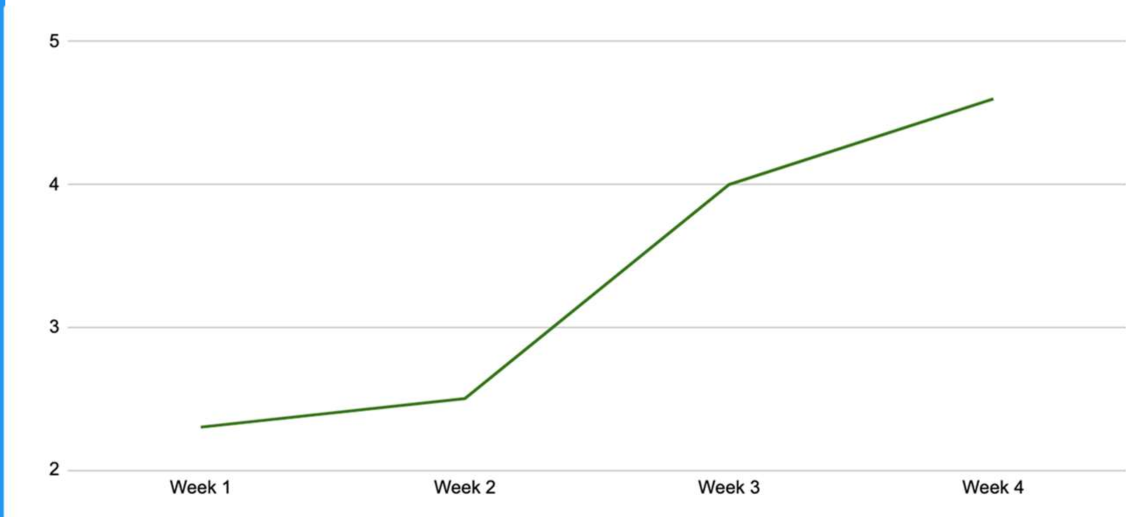
KR1 Increase the **relevance score of recommendations to 90%** by the end of Q4.

O3 **Optimize the user experience of the course recommendation interface.**

KR2 Ensure **95% of recommendations align** with user preferences by Q4.

Key takeaways & action items: Users participate to 90% by the end of the survey—a solid improvement, but still short of our 95% target. Investigate additional reasons for courses participation.

Customer Survey

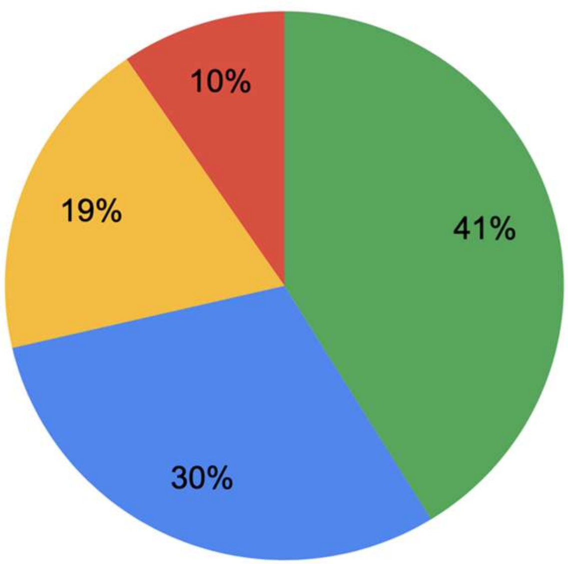


In general, how do you suggest we improve our customer support?

O3	**Optimize the user experience of the course recommendation interface.**
KR1	Increase **the User Satisfaction Score** to 4.7/5 by Q4.

Key takeaways & action items: Satisfaction with support increased once we fixed the customer service software problem. There is still room for improvement, so continue to monitor responses and solutions to support tickets.

Customer Survey



In general, how do you suggest we improve our customer support?

O3	**Optimize the user experience of the course recommendation interface.**
KR3	Utilize **80% of user feedback to make iterative improvements** to the system by the end of Q4.

Key takeaways & action items: A number of customers volunteered that a live chat option would improve customer support. Also, many respondents found the guides and tutorials helpful. Research expanding these offerings for specific courses.

Data Science and Machine Learning FMEA

FMEA

Process/Product Name: Course Recommendation System

Responsible: Data Science and Machine Learning

Prepared By: Wahyu

FMEA Date (Orig.):

(Rev.):

Process Step/Input	Potential Failure Mode	Potential Failure Effects	SEVERITY (1 - 10)	Potential Causes	OCCURRENCE (1 - 10)	Current Controls	DETECTION (1 - 10)	RPN	Action Recommended
What is the process step, change or feature under investigation?	In what ways could the step, change or feature go wrong?	What is the impact on the customer if this failure is not prevented or corrected?		What causes the step, change or feature to go wrong? (how could it occur?)		What controls exist that either prevent or detect the failure?			What are the recommended actions for reducing the occurrence of the cause or improving detection?
Data Collection & Processing	*Missing or inaccurate data collected from users*	*Poor quality recommendations, leading to low engagement*	*8*	*Incomplete user profiles, faulty data sources*	*6*	*Data validation scripts that check for missing or malformed values; daily data integrity checks*	*4*	*192*	*Implement real-time data validation processes that flag and fix anomalies before they propagate into the system*
Recommendation Algorithm	*Algorithm fails to provide relevant course recommendations*	*Users may not find the courses they want, leading to churn*	*9*	*Poor algorithm tuning, lack of diversity in recommendations*	*6*	*Periodic retraining of the model using new data; recommendation model is tested with sample users before production*	*5*	*270*	*Regularly evaluate the algorithm with A/B testing on live users to ensure relevance, and introduce diversity in recommendation logic (e.g., collaborative filtering + content-based)*
User Interface & Experience	*Poor or confusing UI/UX makes it difficult for users to find recommendations*	*Frustration, drop-off in user engagement, and poor user retention*	*7*	*Unintuitive design, poorly placed recommendation widgets*	*4*	*Basic UI/UX testing during development; user focus groups conducted to get feedback*	*6*	*168*	*Perform extensive A/B testing on various UI designs; refine UI layout based on heatmaps to ensure ease of navigation for recommendation widgets*
System Performance	*System lags or crashes during high-traffic periods*	*Users will abandon the system, leading to dissatisfaction*	*10*	*Recommendations don't improve over time, leading to static system*	*4*	*Load testing is done periodically to simulate traffic spikes; basic server monitoring (CPU, RAM, disk space) is in place*	*6*	*240*	*Scale infrastructure automatically based on user traffic; Implement load balancing across multiple servers to handle traffic surges*
Feedback Loop	*User feedback on recommendations is not incorporated*	*Late response to underperformance*	*6*	*Lack of a proper feedback mechanism, or ignoring feedback*	*6*	*User feedback is collected through star ratings, but this is not tied directly to algorithm improvement*	*4*	*144*	*Develop mechanisms for users to provide more granular feedback (e.g., thumbs up/down on specific courses); Use feedback to fine-tune the algorithm*
Data Security & Privacy	*User data is compromised or used inappropriately*	*Loss of user trust, legal issues, penalties*	*10*	*Inadequate data encryption, lack of compliance with regulations*	*3*	*Basic data encryption (SSL for transmission); periodic security audits*	*5*	*150*	*Ensure end-to-end encryption, perform more frequent security audits, and ensure compliance with GDPR, CCPA, and other data protection laws*

Resp.	Actions Taken	SEVERITY (1 - 10)	OCCURRENCE (1 - 10)	DETECTION (1 - 10)	RPN
Who is responsible for making sure the actions are completed?	What actions were completed (and when) with respect to the RPN?				
Data Science and Machine Learning	*Implemented a real-time data validation pipeline, where missing or incorrect data is flagged and rectified within minutes, improving data accuracy and preventing downstream issues in recommendations.*	*8*	*4*	*2*	*64*
Data Science and Machine Learning	*A/B testing was performed with 10% of users. Diversified the recommendations based on content preferences and collaborative filtering. Algorithm adjustments reduced churn by 15% for new users.*	*9*	*4*	*3*	*108*
UI/UX	*Implemented A/B tests for multiple UI designs and improved navigation based on click-tracking heatmaps, reducing bounce rates by 12% and increasing engagement with recommendations by 35%.*	*7*	*4*	*3*	*84*
DevOps	*Implemented auto-scaling on the cloud server infrastructure, which dynamically adjusts resources based on real-time demand. Load balancing reduced crash incidents by 90% during high-traffic periods.*	*10*	*4*	*3*	*120*
Software Engineers	*Integrated a thumbs-up/down feature for individual courses, tied to immediate algorithm adjustments. Feedback utilization increased relevance scores, and user satisfaction improved by 18%.*	*6*	*4*	*3*	*72*
Data Govern	*Enhanced security by implementing AES-256 encryption for data storage, and scheduled quarterly security audits. Compliance with GDPR and CCPA was improved, reducing potential legal risks by 85%.*	*10*	*3*	*4*	*120*

Marketing FMEA

FMEA

Process/Product Name: Marketing Campaign
Responsible: Marketing Promotion Manager

Prepared By: Wahyu
FMEA Date (Orig.):
(Rev.):

Process Step/Input	Potential Failure Mode	Potential Failure Effects	SEVERITY (1 - 10)	Potential Causes	OCCURRENCE (1 - 10)	Current Controls	DETECTION (1 - 10)	RPN	Action Recommended	Resp.	Actions Taken	SEVERITY (1 - 10)	OCCURRENCE (1 - 10)	DETECTION (1 - 10)	RPN
What is the process step, change or feature under investigation?	In what ways could the step, change or feature go wrong?	What is the impact on the customer if this failure is not prevented or corrected?		What causes the step, change or feature to go wrong? (how could it occur?)		What controls exist that either prevent or detect the failure?			What are the recommended actions for reducing the occurrence of the cause or improving detection?	Who is responsible for making sure the actions are completed?	What actions were completed (and when) with respect to the RPN?				
Ad Design & Creation	*Ad message is unclear or not compelling*	*Low engagement with the ads, low CTR (low customer interest, reduced clicks)*	8	*Poor copywriting, unclear message, Lack of understanding of user needs or inaccurate audience segmentation*	5	*Initial audience research done using basic analytics tools; Marketing personas created based on existing users' data*	4	160	*Conduct in-depth surveys and interviews to refine personas; Use A/B testing on multiple ad copies to see which messaging resonates best. Ensure clear, value-driven messaging that speaks to user pain points.*	*Marketing Promotion Manager*	*Conducted user surveys and interviews to understand needs. A/B tests on multiple ad copies led to optimized messaging that increased click-through rate (CTR) by 15% compared to the initial ads.*	8	4	2	64
Audience Identification	*Wrong target audience is selected*	*High bounce rate, low conversions*	9	*Inaccurate customer segmentation*	6	*Monthly review*	5	270	*Conduct thorough customer analysis; run small-scale A/B tests on different audience segments before full campaign.*	*Marketing Promotion Manager*	*Collaborate with data science and machine learning dept running AB Testing*	8	4	3	96
Social Media Strategy	*Social media posts are inconsistent or irrelevant*	*Reduced engagement, poor brand visibility*	7	*Poor content planning, mismatch with audience*	4	*Monthly review*	5	140	*Develop a social media content calendar aligned with target audience interests and campaign objectives.*	*Marketing Promotion Manager*	*Social media calender - target audiences interest - campaign objectives*	7	4	3	84
Budget Allocation	*Over-budget in one channel (e.g., TV)*	*Financial overspend, reduced ROI*	8	*Inaccurate budget forecasting or resource waste*	5	*Tracking by monthly financial report*	3	120	*Use a marketing budget planner; allocate resources based on channel performance metrics.*	*Marketing Promotion Manager*	*Project budget and project controller work with finance and accounting and monitor ROAS*	8	4	3	96
Performance Monitoring	*Campaign metrics are not tracked in real-time*	*Late response to underperformance*	6	*Inadequate use of tracking tools*	6	*Monthly review*	4	144	*Implement real-time campaign tracking tools; set up dashboards for key metrics (CTR, conversion rate).*	*Marketing Promotion Manager*	*Campaign tracking tools*	6	4	3	72
Customer Engagement	*Leads are not followed up promptly*	*Low conversion from leads to customers*	9	*Poor lead nurturing, slow response time*	6	*Manual follow up leads*	6	324	*Schedule lead nurturing emails immediately after capturing leads using autamatef.*	*Course Manager*	*Implement automated lead follow-up through CRM*	9	4	3	108
Ad Placement Strategy	*Inefficient ad placement, missing key platforms*	*Reduced reach and visibility, lower conversion rates*	7	*Failure to choose the right platforms where the target audience is active*	6	*Ads placed on general platforms (Google, Facebook); Basic analytics used to track impressions and clicks*	5	210	*Conduct audience behavior analysis to identify underutilized platforms (e.g., LinkedIn for professionals, YouTube for tutorials); Allocate ad spend based on platform performance*	*Ads Manager*	*Analyzed user behavior and found underutilized platforms like LinkedIn and YouTube. Reallocated budget towards those platforms, improving overall reach by 20% and boosting conversions from professionals in particular.*	7	4	3	84
Press Release & Media	*Delayed or poorly distributed press releases*	*Reduced media coverage and brand awareness*	6	*Inefficient coordination with media outlets, delays in approval*	4	*Press releases sent to a few key outlets manually*	7	168	*Build relationships with a larger network of journalists, influencers, and bloggers; Use automated tools to schedule timely releases*	*PR Manager*	*Automated press release distribution and built a media relationship pipeline. Press coverage increased by 25%, and timely publication ensured visibility aligned with campaign peaks.*	8	4	3	96
Social Media Ads	*Low engagement rates on social media campaigns*	*Lower-than-expected ROI, wasted budget*	8	*Unattractive visuals, weak call-to-action (CTA)*	6	*Basic social media engagement tracking (likes, shares, and comments)*	6	288	*Enhance visuals and CTA with A/B testing; Use interactive elements like polls and quizzes; Tailor ads for each platform (e.g., Instagram vs LinkedIn)*	*Ads Manager*	*Revamped visuals and CTAs with a focus on A/B testing. Engagement on social media improved by 18%, with polls and interactive quizzes increasing direct interaction and ad performance on platforms like Instagram and Facebook.*	8	4	3	96
Budget Management	*Budget overrun due to poor cost tracking*	*Depletion of marketing funds, leading to campaign cutback*	9	*Inefficient allocation, unexpected expenses*	3	*Weekly budget reviews by the finance team, using basic expense tracking software*	6	162	*Implement real-time budget tracking and automated alerts for over-budget items; Forecast using dynamic spend analysis based on real-time ad performance*	*Marketing Promotion Manager*	*Introduced automated budget tracking software that sent real-time alerts when costs approached set limits. This led to a 10% reduction in unexpected expenses and better control of the overall marketing spend.*	9	3	3	81
User Conversion Rate	*Low conversion rate from ad click to signing up for the course*	*Low ROI, fewer course sign-ups*	8	*Poor ad targeting, ineffective landing pages*	7	*Basic conversion tracking using Google Analytics; Weekly reviews of conversion data*	4	224	*Conduct A/B tests on landing pages to improve conversion rates; Optimize the user journey from ad click to course sign-up with personalized landing pages*	*Marketing Promotion Manager*	*A/B testing of landing pages resulted in a 12% improvement in conversion rates. Personalized landing pages for different user personas were created, aligning better with the audience's needs and journey.*	8	4	3	96
Timing of Campaign	*Campaign launched at a suboptimal time*	*Low user interest, low traffic*	7	*Lack of coordination with key seasonal trends or competing events*	5	*Ads are timed based on previous marketing schedules and intuition*	6	210	*Analyze user behavior and seasonality data to schedule campaigns during peak times (e.g., exam prep seasons for courses); Optimize ad frequency*	*Marketing Promotion Manager*	*Adjusted campaign schedules to target key user interest periods (e.g., before exams). Increased sign-up rates by 20% during high-demand periods, compared to campaigns launched at random times in the past.*	7	4	3	84
Competitor Activity	*Competitor runs simultaneous campaigns*	*Marketing noise and reduced effectiveness of the course campaign*	6	*Lack of awareness of competitors' marketing schedules*	4	*Basic competitor monitoring using general tools like Google Alerts; React to competitor campaigns after they've been noticed*	6	144	*Set up proactive monitoring for competitor activities and market trends; Plan campaigns around or in anticipation of competitor launches*	*Competitive Intelligence Analyst*	*Implemented competitor activity monitoring using specialized tools. Planned the launch of our ads one week before a major competitor's campaign, leading to 10% higher engagement during that competitive period.*	6	4	3	72

Key takeaways : Data science and machine learning will assist on providing the analysis and reporting results from the **experimentation pla**tform and the prediction model for the Marketing and Product Manager.

ROAM Analysis

ROAM Analysis

	Issues	Actions	ROAM designation
Issue 1	10% of courses are not being updated	Contacted the courses manager and have had no issues for two weeks.	Resolved
Issue 2	Customers are complaining.	Reached out to the customers, listened to their issues, and offered to send them a free coupon. Most customers were satisfied with this solution.	Mitigated
Issue 3	There is a software issue preventing your customer relations team receiving all requests and complaints.	Assign your IT Specialist to fix the problem as soon as possible.	Owned
Issue 4	Some customers are canceling their subscriptions.	Checks in with the customers and, despite offering them a subscription promotion, they still want to cancel. There is nothing you can do, but the impact is minimal.	Accepted
Issue 5	The cost HR per revenue ratio to recruit data engineer is above target (over budget)	Assign your HR Specialist to postpone hiring data engineer	Owned
Issue 6	Courses are being increasing.	Reassessed the model making it more relatable and importan for users.	Mitigated
Issue 7	Your budget is tightening.	Financial Analyst reassessed project spending and was able to increase the budget. The additional money was enough to offset recent losses.	Resolved

Status Report

Project Name: C4U Course Recommender System

Today's date: July 5

Summary	Overall Status (RAG)
We have installed new course recommender system software to recommend a more personal courses and begun sending out the first test batch of c4U to customers. However, we have run into issues with product quality, customer communication, and the delivery process. Our next milestones include sending the test batch customers newsletters on plant upkeep and sending out the second batch of plants. This report also includes top risks and issues that have arisen and how we intend to take action.	Amber

Completed Milestones and Tasks

Description	Date	Status	Owner	Comments
Purchased and installed new software to keep track of incoming orders	June 15	Completed	IT Specialist	The installation took three days longer than expected.
Began sending test batches of C4U course recommender system	June 21	Completed	Head of Data Science and Machine Learning	The number of orders exceeded targets by 15%.

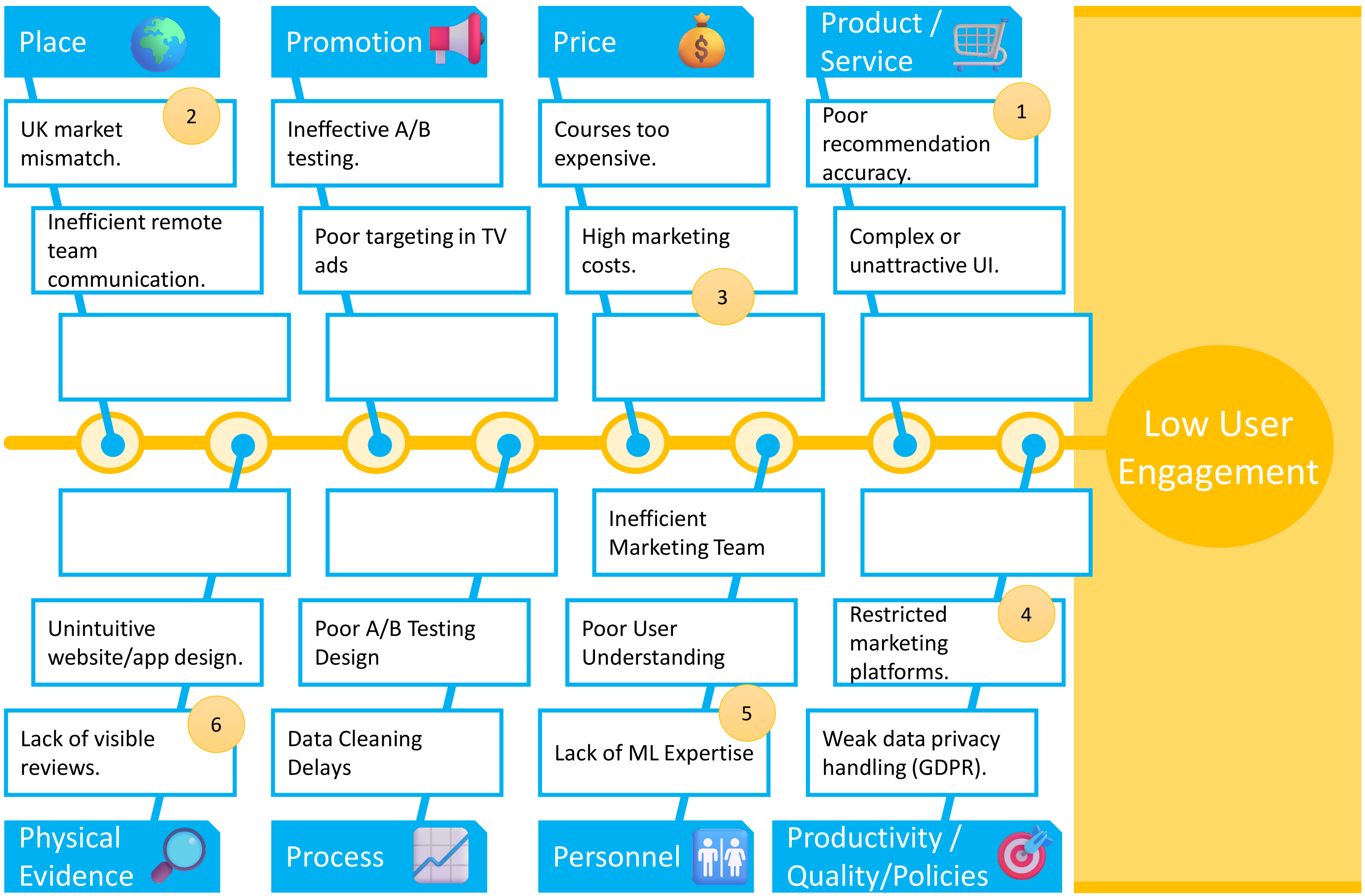
Upcoming Milestones and Tasks

Description	Date	Status	Owner	Comments
Send the first batch customers e-newsletters with a product announcement on our improved search engine	July 7	Upcoming	Marketing Promotion Manager	The newsletter must C4U brand design guidelines.
Hit at least 95% of recommendation align with user preference	July 19	Upcoming	Head of Data Science and Machine Learning	The error rate should be under 5%

Top Risks and Issues

Issue	Impact	Action	Owner
The data team reports that 10% of the courses were not available	Profit loss, complaints, and budget issues	Evaluate and adjusting the courses inventory and obselete	Courses Manager
The customer relations team is receiving only 30% of requests and complaints	Customer dissatisfaction	Fix problems with new customer service software	IT Specialist
The course completion rate is only 60%	Cancelled subscriptions	Evaluate, selecting and adjusting the courses to tailor the demand of the customer preference	Courses Manager

FMEA Updated



6

Key takeaways :

Data science and machine learning has been applied not only to Recommendation System but also in our sales pipeline, TV ads and marketing campaign.

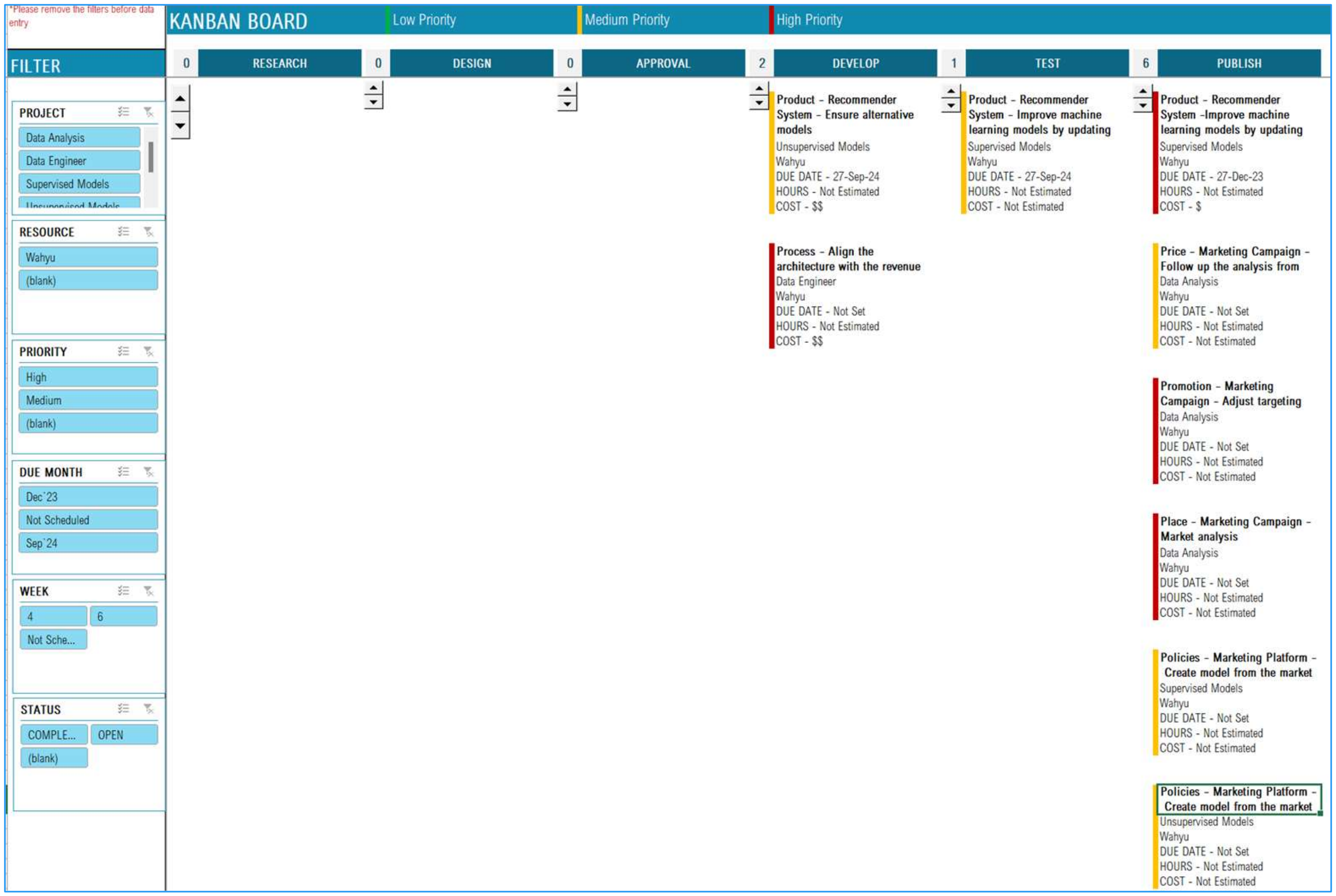
Sprint Backlog

Epic	User Story Title	Story	Acceptance Criteria	Value	Estimate (Story Points)	Sprint
Course Recommendations	Personalize Course Recommendations	As a user, I want to receive personalized course recommendations based on my previous activity and interests so that I can easily find relevant courses.	- Recommendations are shown based on user preferences (past purchases, ratings, and browsing history). - Recommendations are updated dynamically as user activity changes.	$$$	8	Current Sprint
	Course Recommendations Algorithm	As a data scientist, I want to implement a recommendation algorithm that considers user demographics, past activity, and course ratings.	- Algorithm uses collaborative filtering and content-based filtering. - The recommendations are shown in less than 3 seconds.	$$$	13	Current Sprint
	Integrate Feedback Mechanism	As a user, I want to provide feedback on recommended courses so the system can improve future recommendations.	- Users can rate recommended courses (1-5 stars). - The system adjusts future recommendations based on user feedback.	$$	5	Next Sprint
Course Search Functionality	Filter Courses by Category	As a user, I want to filter courses by category so I can easily find courses in specific areas of interest.	- Filters are available for categories (e.g., Technology, Business, Art). - Users can apply multiple filters. - Search results update in real-time based on the selected filters.	$$	8	Current Sprint
	Search Courses by Keywords	As a user, I want to search courses by keywords to quickly find relevant content.	- Users can search by keywords. - Results are relevant to the entered keywords. - Search results are returned within 2 seconds.	$$$	5	Current Sprint
User Profiles	User Data Collection	As a product manager, I want to collect user profile data (age, preferences) so that the system can deliver more personalized experiences.	- User profile fields include age, education, and interests. - The data is stored securely and can be updated at any time by the user.	$$$	5	Current Sprint
	Update User Profile	As a user, I want to update my profile information to reflect changes in my preferences or education.	- Users can update their personal information. - Changes are reflected in the system immediately and influence future recommendations.	$$	3	Next Sprint
Social Media Ads	Create Engaging Ad Copy	As a marketer, I want to create engaging ad copy to increase click-through rates (CTR) and conversions.	- Ad copy is tested and optimized using A/B testing. - The ad has a strong CTA. - CTR increases by at least 10% after optimization.	$$$	5	Current Sprint
	Design Visuals for Social Media	As a social media manager, I want to design eye-catching visuals for the campaign to attract more user attention.	- Visuals are platform-specific (Instagram, Facebook, LinkedIn). - Visuals lead to a 15% improvement in engagement (likes, comments, shares).	$$	8	Next Sprint
Email Marketing	Create Personalized Email Campaigns	As a user, I want to receive personalized emails about courses based on my previous interests so that I can easily find relevant content.	- Emails are personalized based on user profiles and past activity. - Email open rates increase by 12%. - Click-through rate increases by at least 10%.	$$$	8	Current Sprint
	A/B Testing for Email Subject Lines	As a marketer, I want to run A/B tests on email subject lines to see which ones generate the highest open rates.	- Two subject lines are tested in each campaign. - The winning subject line increases open rates by 15% compared to the control.	$$	5	Current Sprint
Influencer Partnerships	Partner with Industry Influencers	As a campaign manager, I want to collaborate with influencers in the education space to promote the new course recommender system.	- At least 3 influencers post about the campaign. - Influencer posts generate a 20% increase in user sign-ups. - Positive engagement on influencer posts is high (e.g., comments, likes).	$$$	13	Next Sprint
	Track Influencer Campaign Success	As a campaign manager, I want to track the success of influencer marketing to evaluate the ROI.	- Influencer posts are tracked using unique referral links. - ROI is calculated based on the conversion rate from the posts. - Campaign report is generated post-campaign.	$$	5	Next Sprint

Sprints		
Name	Current Sprint	Next Sprint
Start date	June 1	June 22
End date	June 19	July 9
Point Capacity	57	34
Points Assigned	60	60
Value Attributed	17	9

Key takeaways : A project manager will help to ensure the ROI is healthy. 2-weeks to a month of projects are listed which related to C4U’s course recommender system

Data Science and Machine Learning Scrum



Key takeaways : Kanban booard is use to visualize the flow of tasks within the sprint.

Models and Findings

Exploratory Data Analysis

Project Management Flowchart

Project Initiation (PI)	Project Planning (PP)	Project Execution (PE)	Agile Scrum Project	Lean Six Sigma Project
1. OKR and SMART 2. Stakeholder Analysis 3. Project Charter	1. WBS and RACI 2. Gantt Chart 3. Risk Management 4. Project Budget 5. Communication Plan – Team Profile	1. Customer Survey 2. ROAM Analysis 3. Status Report 4. Hours Burndown	1. Sprint Backlog 2. Sprint Retrospective	1. Yellow Belt 2. Green Belt 3. Black Belt 4. Master Black Belt
1. VOC 2. CTQ – CTP – CTR - COPQ 3. SIPOC	1. KPI Trees 2. Lean Measures 3. Operational Definitions	1. Data Collection and Sample 2. MSA 3. Process Capability	1. The Process Door 2. EDA 3. Experimentation Platform	1. SPC

1. Team Meeting Agenda
2. Email management (Escalation, Invitation, Release Plan, etc.)

Plan

Keywords

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

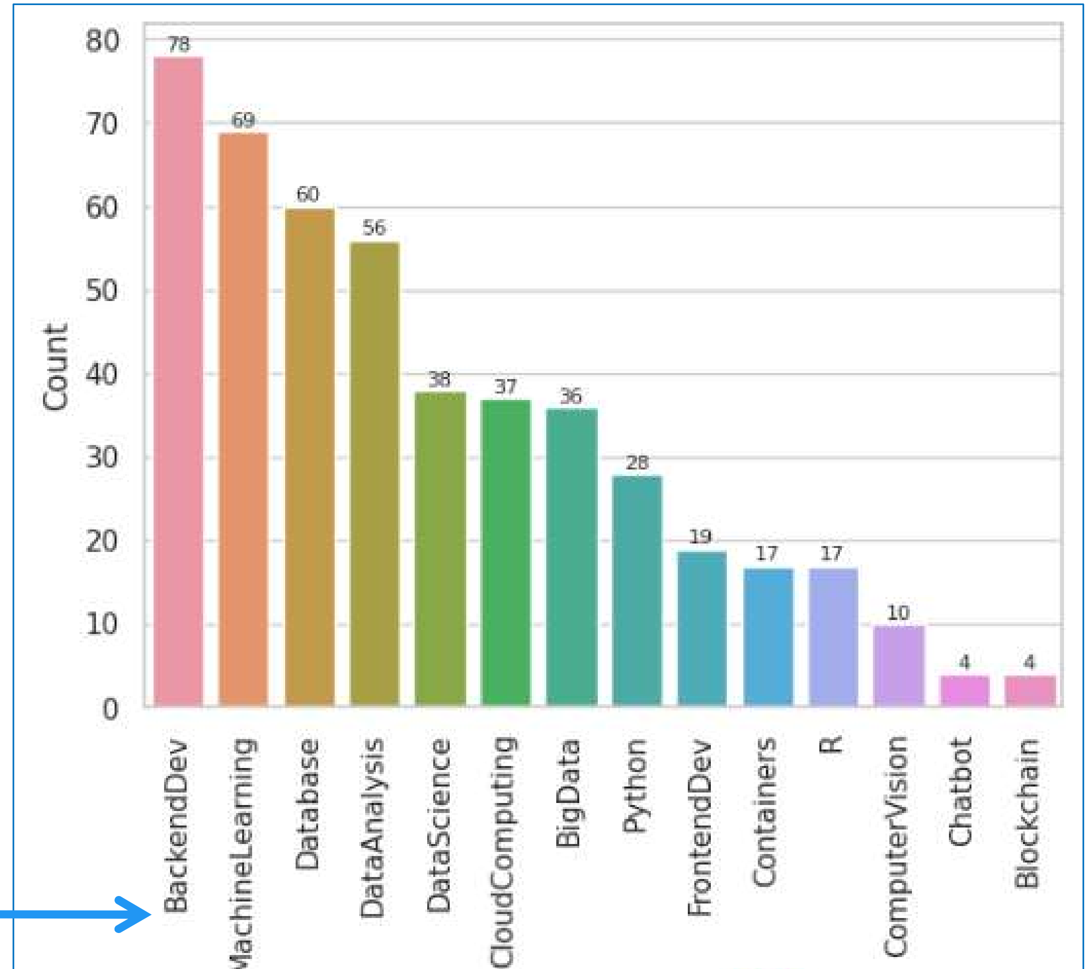
Plan

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.

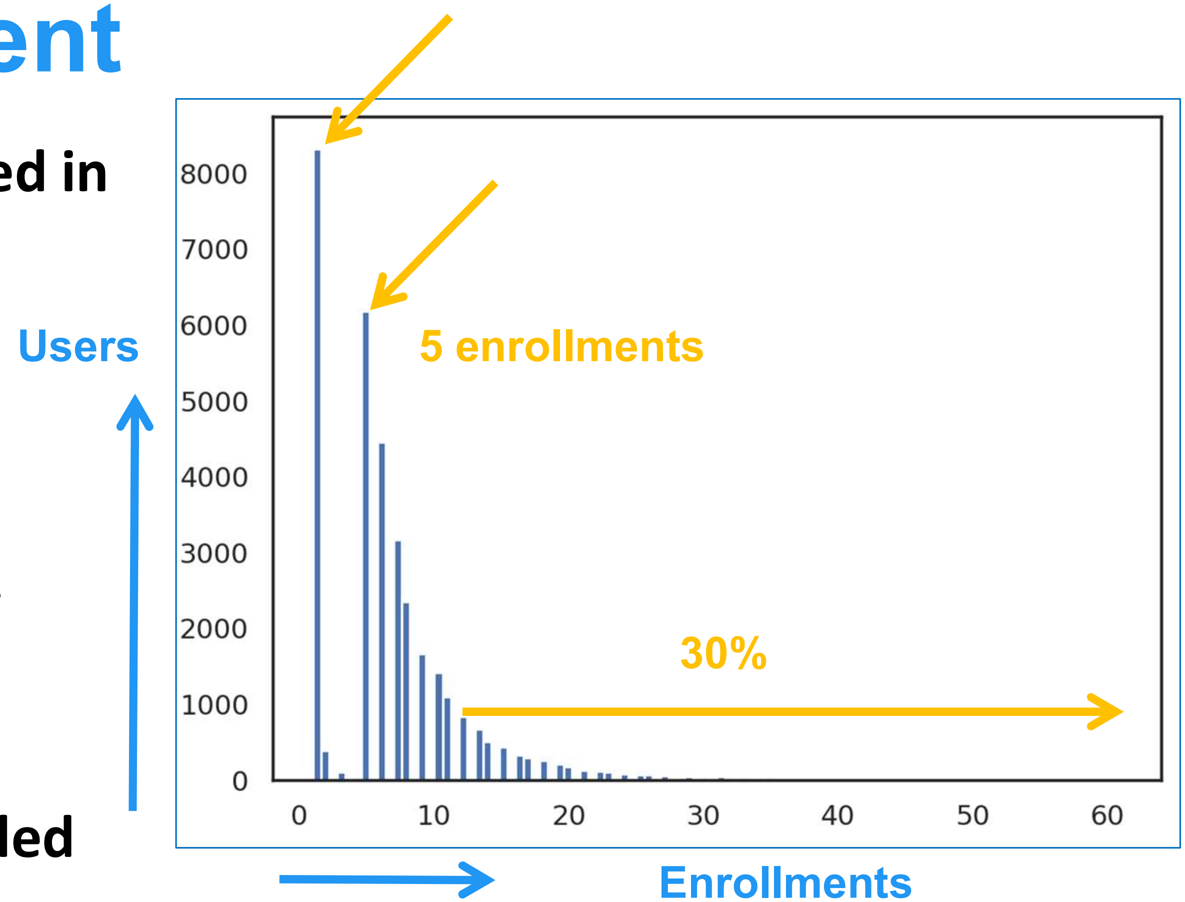
Genres



Analyze

Course Enrollment

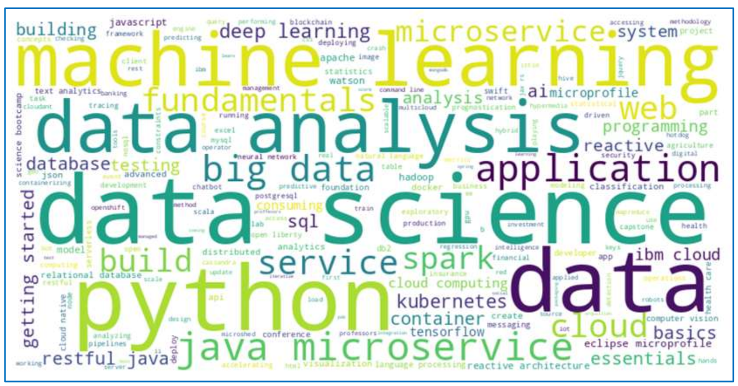
- **Over 8.000 users** have **enrolled in only one course.**
- The enrollment distribution is continuously declining. with **fewer users** as the number of **enrollments increases**.
- Only **30% of users have enrolled more than 10 courses.**



Analyze

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

Analyze

Content-Based Recommender System using Unsupervised Learning

Models and Findings

Content-Based Recommender System using User Profile and Course Genres

User Profile Flowchart

Create user profile vector
- Data collection
- Define user interest
- Transform user behaviors to a profile vector

Course genre data processing
- Data cleaning
- Feature extraction
- Create course genre vectors

User-item similarity
- Calculate user-item similarity
- Content based approach for measuring similarity

Recommendation generation
- Display the system based on similarity scores

Construct

Data

```
test_users_df.head()
```

	user	item	rating
0	1502801	RP0105EN	3.0
1	1609720	CNSC02EN	2.0
2	1347188	CO0301EN	3.0
3	755067	ML0103EN	3.0
4	538595	BD0115EN	3.0

test_users.df

	user	Database	Python	CloudComputing	DataAnalysis	Containers	MachineLearning	ComputerVision	DataScience	BigData	Chatbot	R	BackendDev	FrontendDev	Blockchain
0	2	52.0	14.0	6.0	43.0	3.0	33.0	0.0	29.0	41.0	2.0	18.0	34.0	9.0	6.0
1	4	40.0	2.0	4.0	28.0	0.0	14.0	0.0	20.0	24.0	0.0	6.0	6.0	0.0	2.0
2	5	24.0	8.0	18.0	24.0	0.0	30.0	0.0	22.0	14.0	2.0	14.0	26.0	4.0	6.0
3	7	2.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0
4	8	6.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	6.0	0.0	2.0	0.0	0.0	0.0

profile.df

	COURSE_ID	TITLE	Database	Python	CloudComputing	DataAnalysis	Containers	MachineLearning	ComputerVision	DataScience	BigData	Chatbot	R	BackendDev	FrontendDev	Blockchain
0	ML0201EN	robots are coming build iot apps with watson ...	0	0	0	0	0	0	0	0	0	0	0	1	1	0
1	ML0122EN	accelerating deep learning with gpu	0	1	0	0	0	1	0	1	0	0	0	0	0	0
2	GPXX0ZG0EN	consuming restful services using the reactive ...	0	0	0	0	0	0	0	0	0	0	0	1	1	0
3	RP0105EN	analyzing big data in r using apache spark	1	0	0	1	0	0	0	0	1	0	1	0	0	0
4	GPXX0Z2PEN	containerizing packaging and running a sprin...	0	0	0	0	1	0	0	0	0	0	0	1	0	0

course_genres.df

course_df

(course_genre.csv)

307 courses

14 Genres

Construct

Recommendation Generation

User and Genre profile.df

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

Create user profile vector

test_user_vector

User rating (1000 Users) test_users.df

User	Item	Rating
2	ML0201EN	3
...		

Course genre data processing

course_matrix

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
[illegible]	[illegible]	SC0103EN	90.0
[illegible]	[illegible]	excourse72	90.0
14795	733707	excourse72	90.0
11701	674939	excourse73	90.0
14649	733707	TMP0105EN	90.0
36771	1312255	TMP0105EN	87.0

90

Course and Genre course_genres.df

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

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

>5K recommendations for >850 users of 1.000 users(>85%)

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
733…, 674…,131…

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

Construct

Recommendation based on user profile

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

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

10 recommended courses

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

Construct

Recommendation based on user profile

User profile 1078030

Participate in 8 courses

- Data analysis
- Deep learning
- Python

10 recommendations
Highest score 30
Lowest score 21

10 recommended courses

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

User profile 733707

Participate in 23 courses

- Spark
- Sql
- Python

172 recommendations
Highest score 99
Lowest score 12

10 recommended courses

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

Lowest score 69

User profile 674939

Participate in 15 courses

- Spark
- Hadoop
- Big data

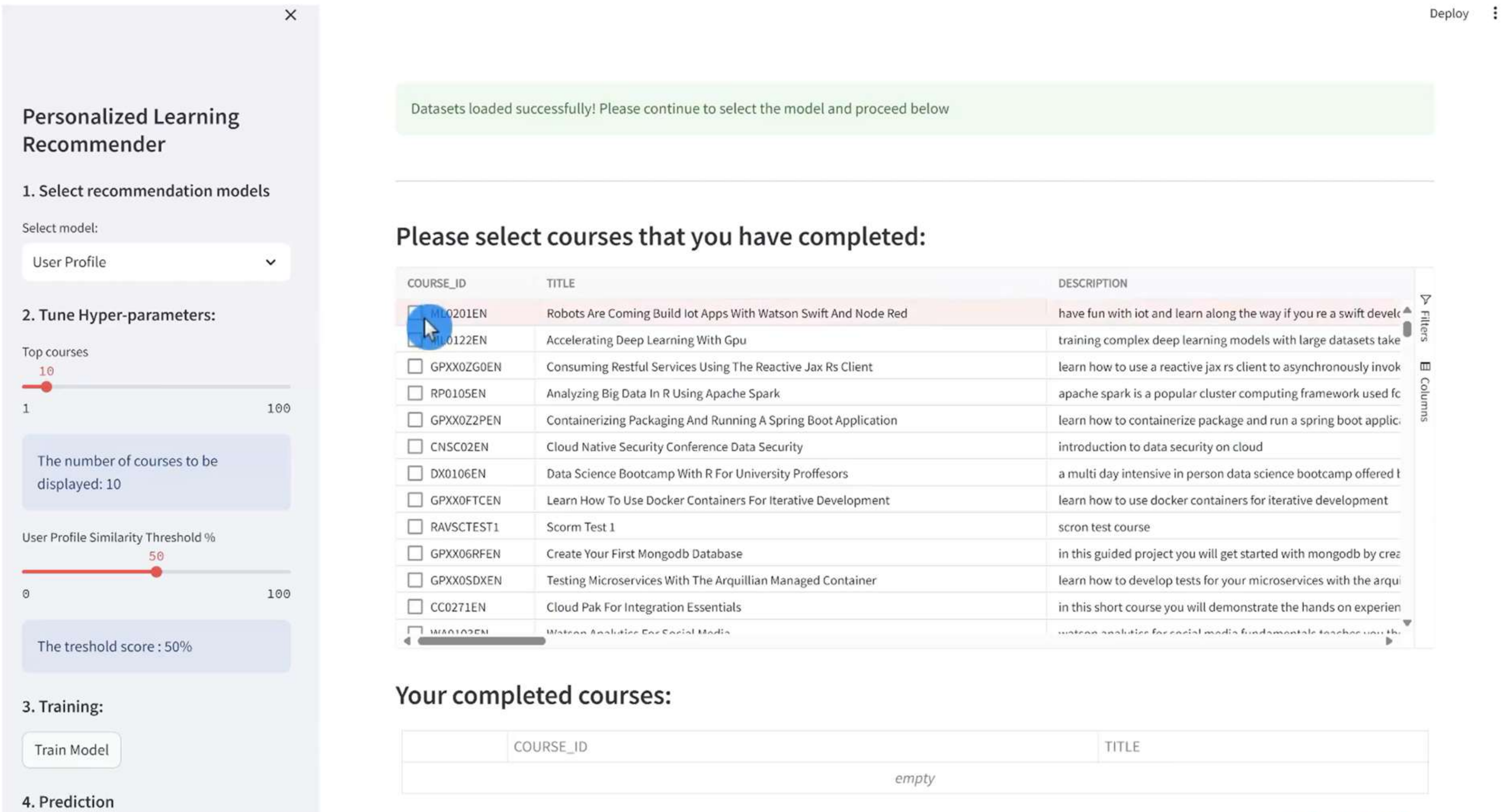
101 recommendations
Highest score 90
lowest score 12

10 recommended courses

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

Lowest score 78

Recommendation based on user profile



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; threshold 0; 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

Construct

Summary

The idea

🧑 **Interested in certain topics**

📖 **a course has a similar topics**

Past interactions

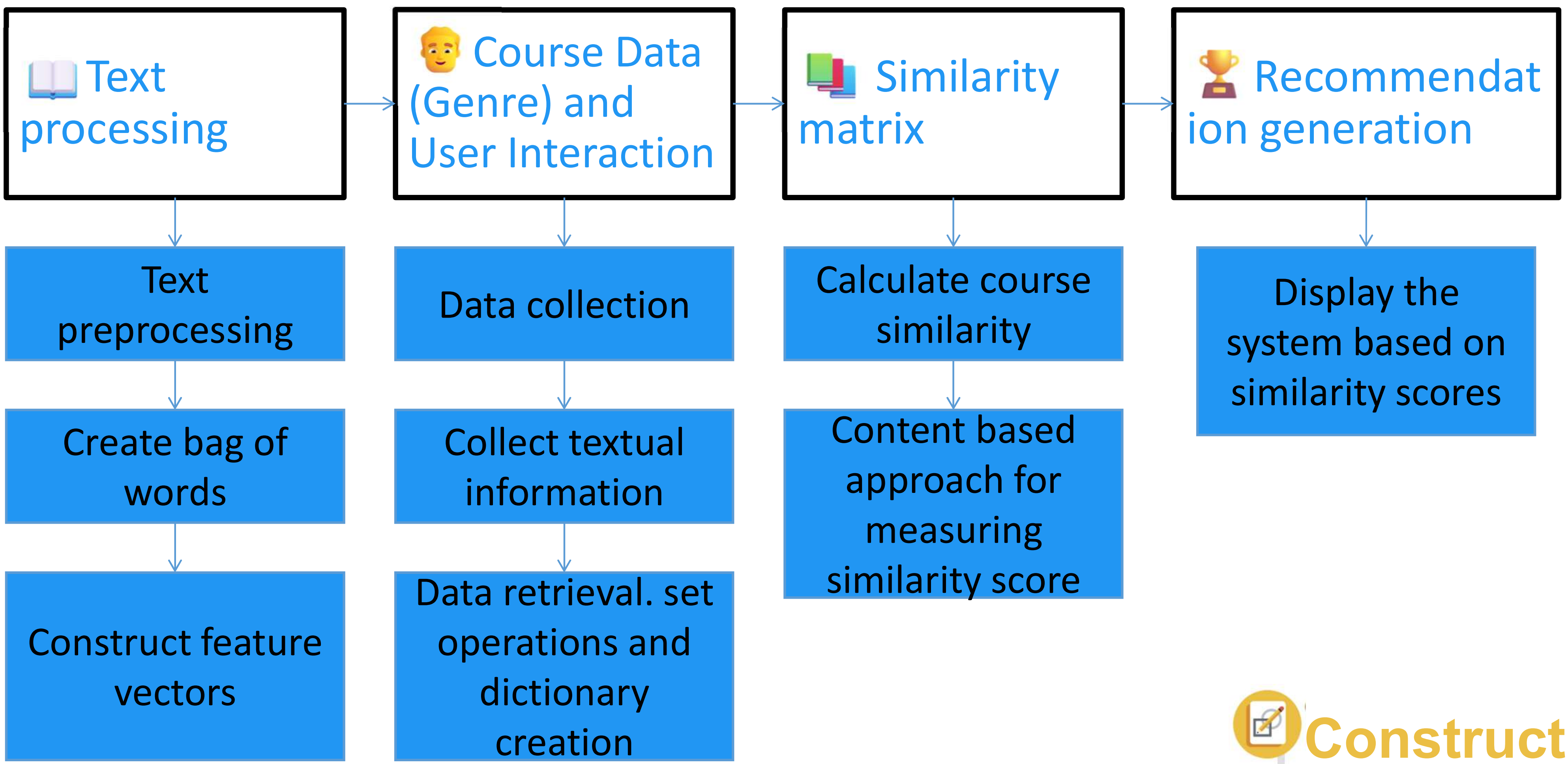
📚 User-item similarity (a dot product) capture it

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

Flowchart



Construct

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...

Construct

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

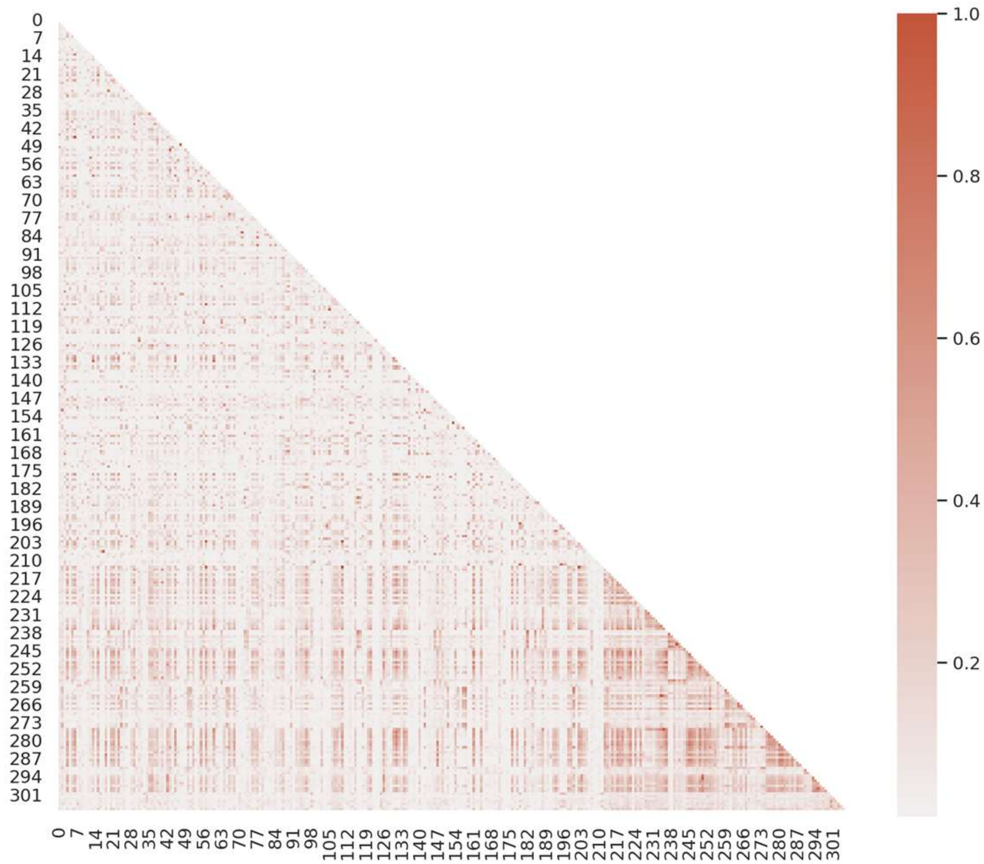
Similiraty calculation:
Cosine. Euclidean. Jaccard index...

	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
[illegible]	[illegible]	excourse47	0.634755
[illegible]	[illegible]	excourse60	0.615568
15998	2087663	excourse46	0.612054
15999	2087663	excourse09	0.608330

75%

Construct

Similarity matrix



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

Construct

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**

Construct

Recommendation based on course similarity

User profile 1078030

Participate in 8 courses

- Data analysis
- Deep learning
- Python

16 recommendations
Highest score 100%
Lowest score 60%

10 recommended courses

- Data analysis
- Deep learning
- Python

Lowest score 66%

User profile 733707

Participate in 23 courses

- Spark
- Sql
- Python

15 recommendations
Highest score 100%
Lowest score 60%

10 recommended courses

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

Lowest score 66%

User profile 674939

Participate in 15 courses

- Spark
- Hadoop
- Big data

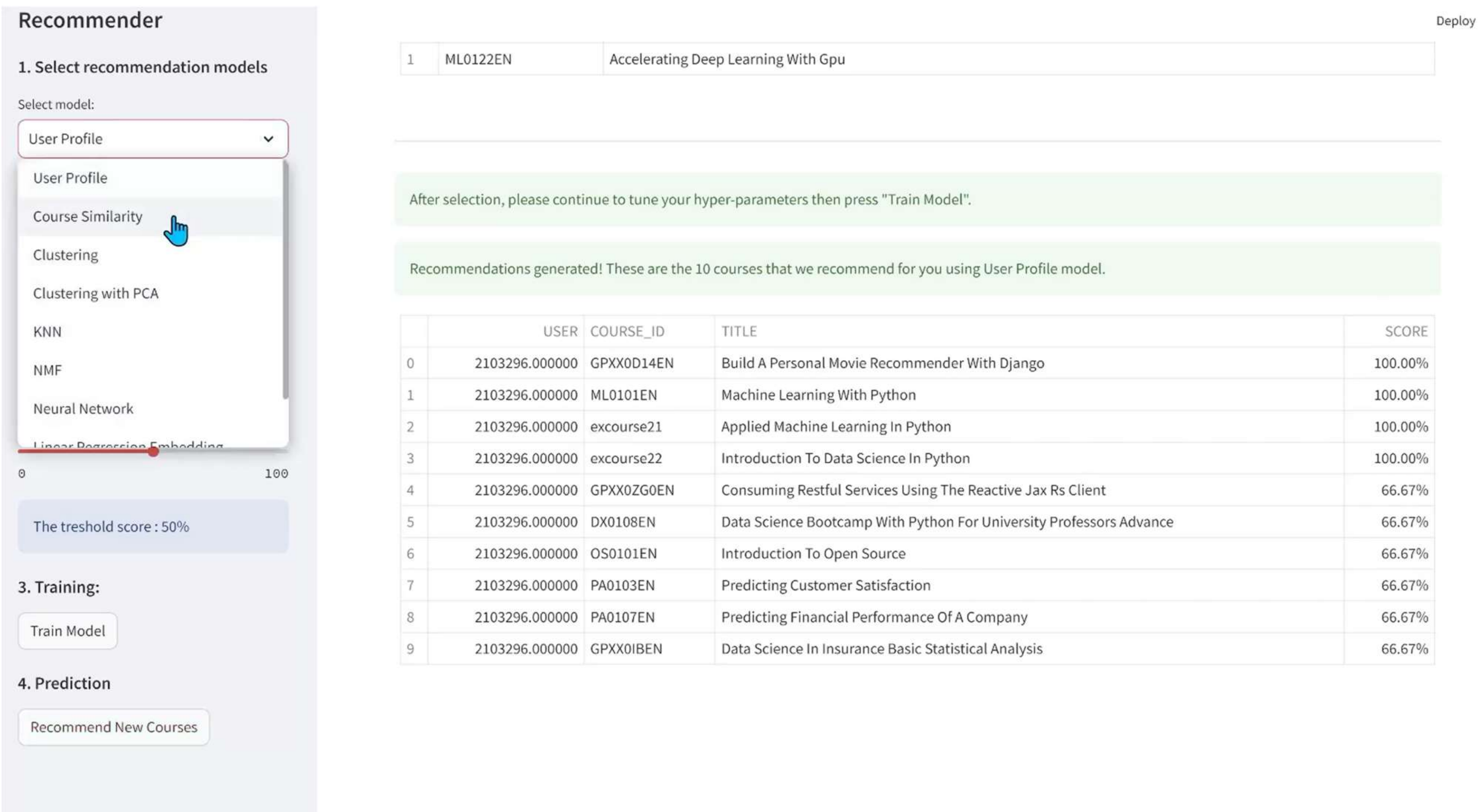
4 recommendations
Highest score 70%
Lowest score 61%

4 recommended courses

- Big data

Lowest score 61%

Recommendation based on course similarity



	USER	COURSE_ID	TITLE	SCORE
0	2103296.000000	GPXX0D14EN	Build A Personal Movie Recommender With Django	100.00%
1	2103296.000000	ML0101EN	Machine Learning With Python	100.00%
2	2103296.000000	excourse21	Applied Machine Learning In Python	100.00%
3	2103296.000000	excourse22	Introduction To Data Science In Python	100.00%
4	2103296.000000	GPXX0ZG0EN	Consuming Restful Services Using The Reactive Jax Rs Client	66.67%
5	2103296.000000	DX0108EN	Data Science Bootcamp With Python For University Professors Advance	66.67%
6	2103296.000000	OS0101EN	Introduction To Open Source	66.67%
7	2103296.000000	PA0103EN	Predicting Customer Satisfaction	66.67%
8	2103296.000000	PA0107EN	Predicting Financial Performance Of A Company	66.67%
9	2103296.000000	GPXX0IBEN	Data Science In Insurance Basic Statistical Analysis	66.67%

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**

Construct

Summary

The idea

🧑 **Interested in certain topics** Feature vectors 📖 **a course has a similar topics**

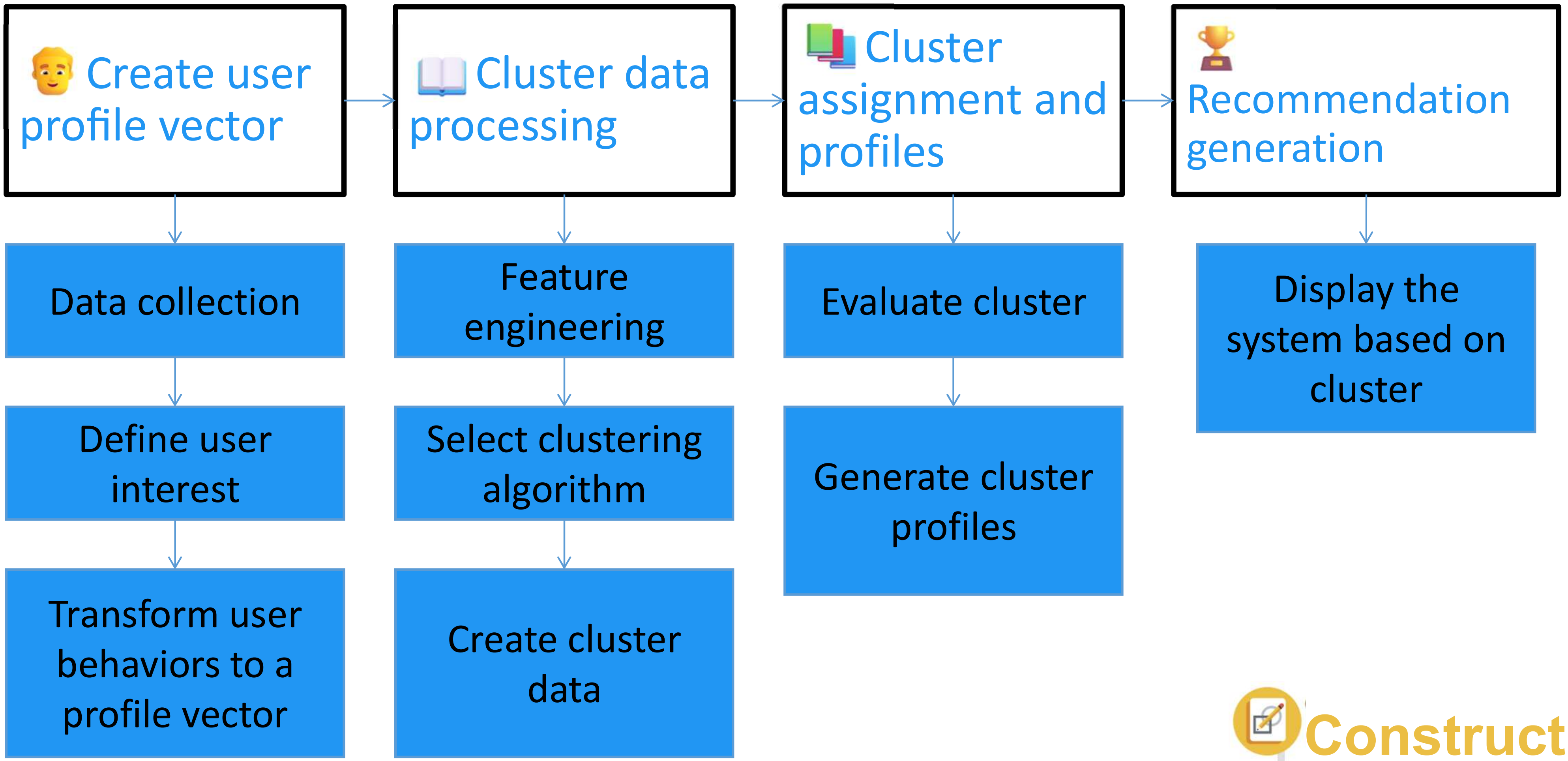
📚 Similarity matrix capture it

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 <u>specific features of courses</u>. allowing users to interpret and understand the reasons behind each recommendation.**

Models and Findings

Clustering-Based Recommender System

Flowchart



Construct

Recommendation Generation

Create user profile vector

Cluster data processing

Cluster assignment and profiles

Recommendation generation

User and Genre

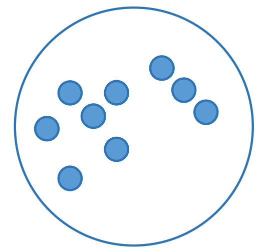
User	Python	Database	...	ML	Blockchain
1		52.00		33.0	6.0
...					

User profile standard scaler

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

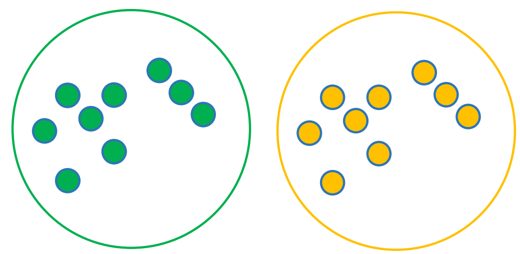
Machine learning (ML) learners:

- ML 101
- ML with python



Database learners:

- SQL 101
- SQL with python

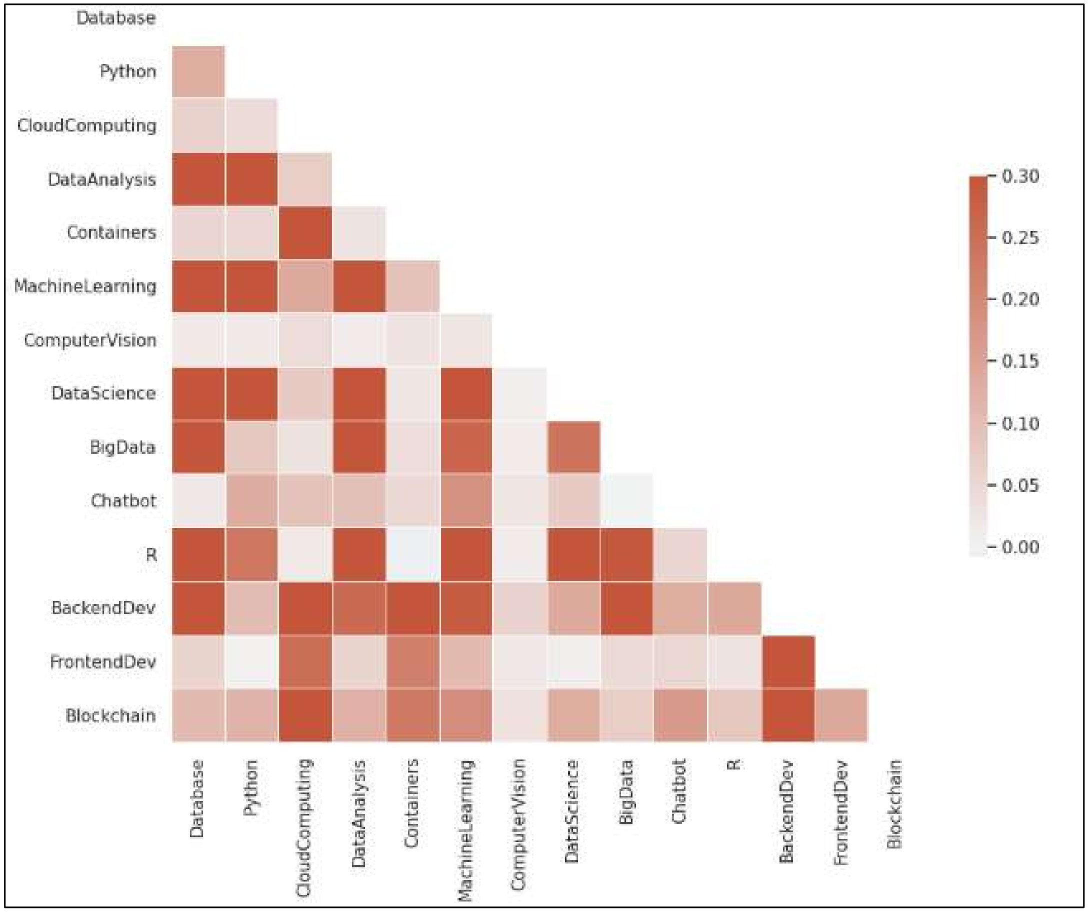


Python:

- Python 101
- Python for analysis

Construct

Covariance matrix of the user profile feature vectors with 14 features

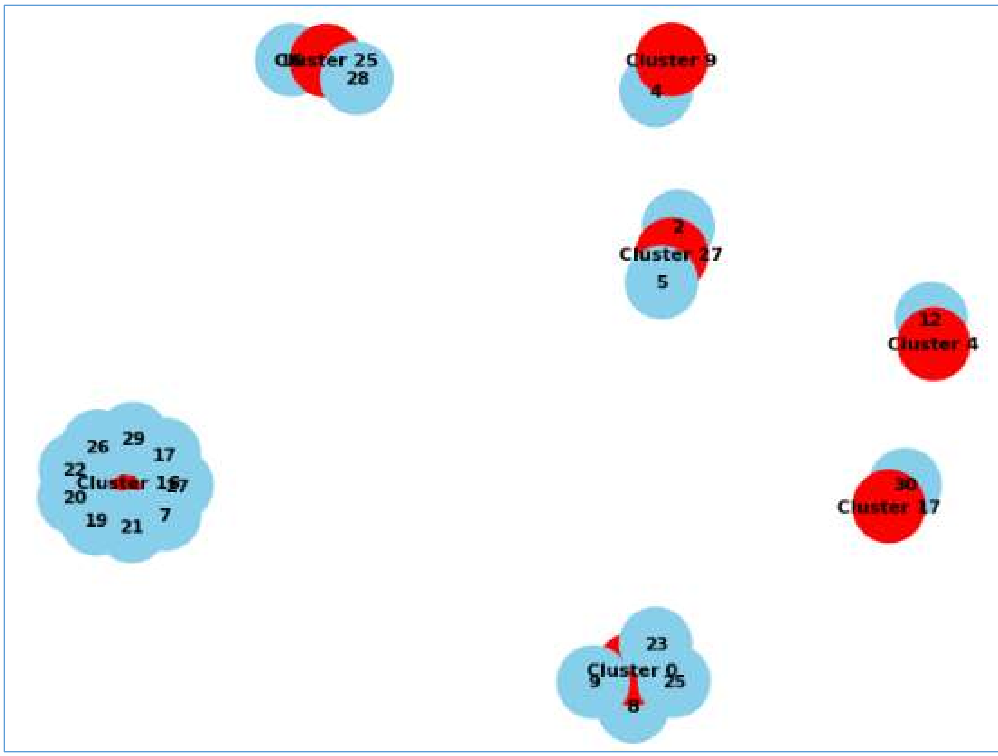


Hot spots shown. Possible to build a recommender system based on cluster.

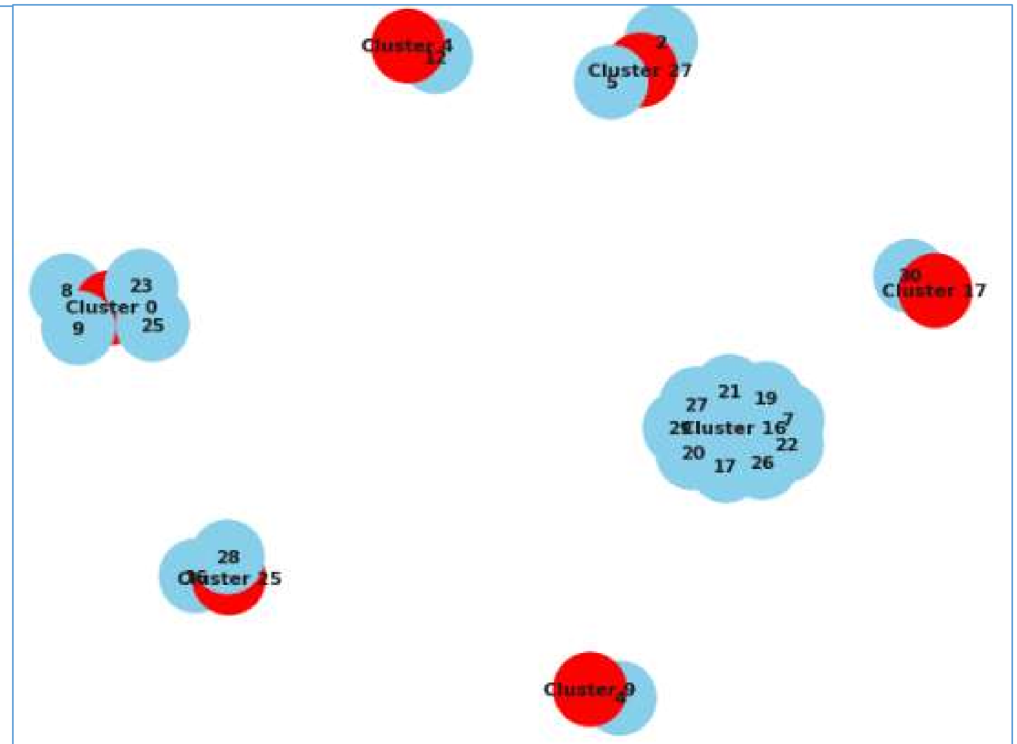
Construct

Evaluation Results

🧑 **Top 20 recomended course based on cluster user profiler feature vectors**



PCA



Construct

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

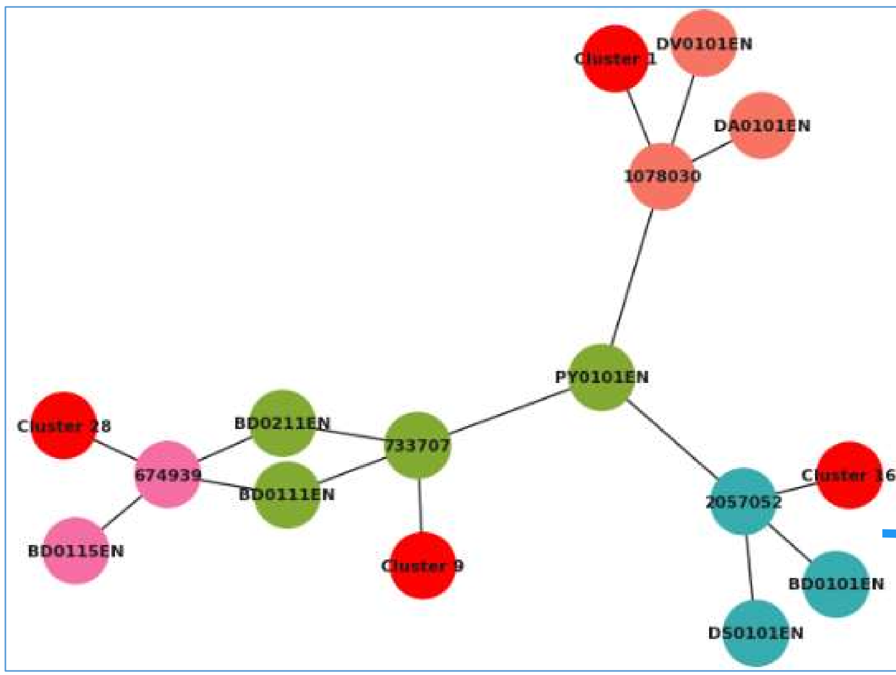
Participate in 15 courses

- Spark
- Hadoop
- Big data

Construct

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

Construct

Recommendation based on cluster

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

3 recommended courses

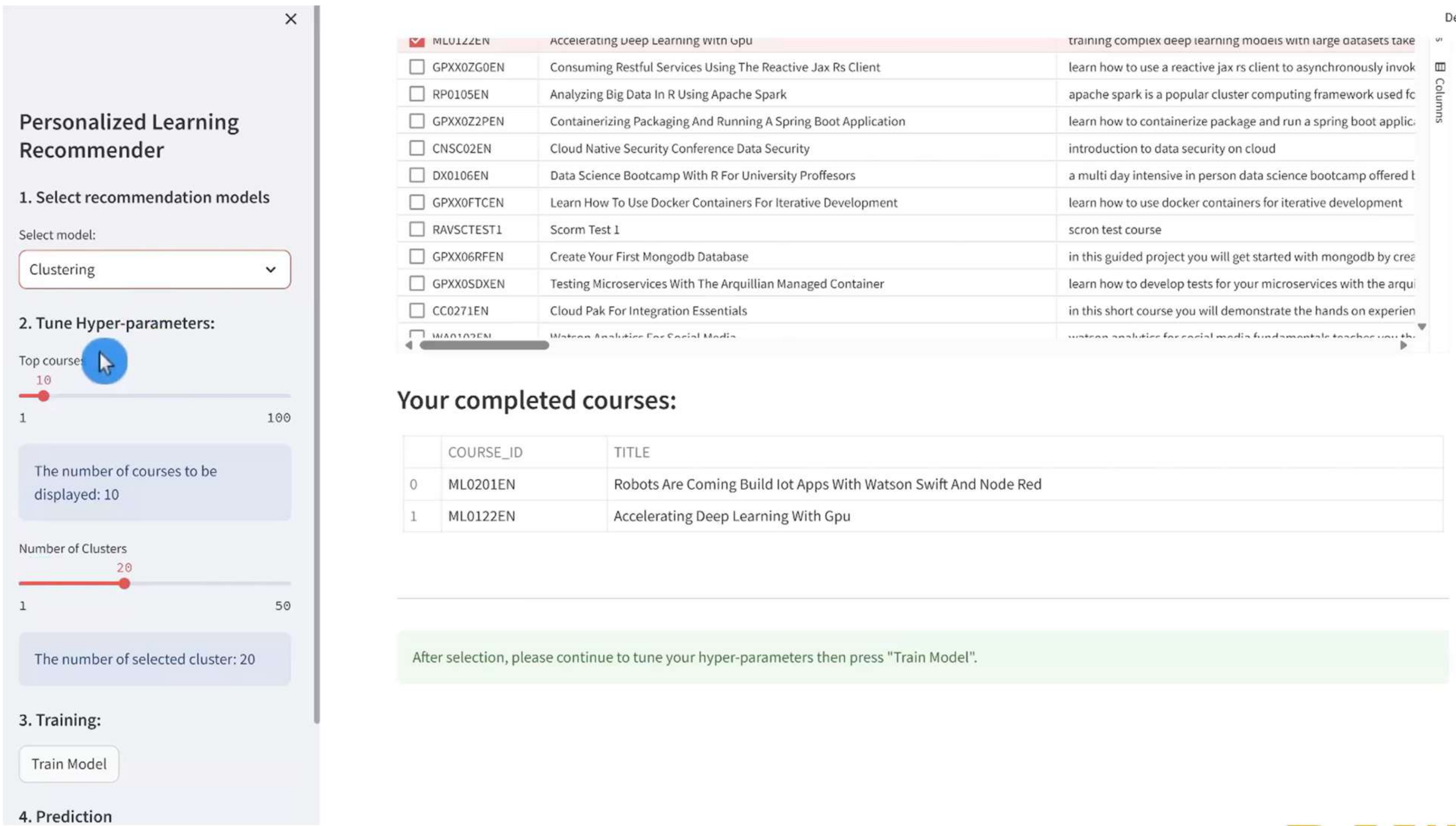
- Python
- Data analysis
- Data science
- Data visualization

- Python
- Data science
- Hadoop
- Spark

- Hadoop
- Maproduce
- Spark

Construct

Recommendation based on cluster and PCA



Summary

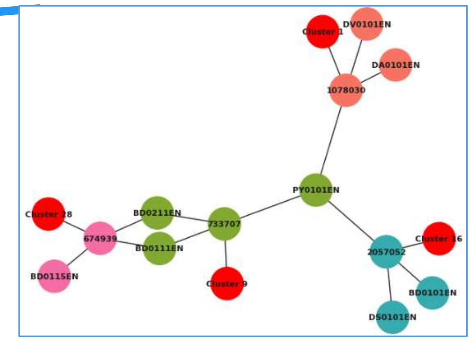
The idea

🧑 **user profile vector**

Cluster data processing

📖 **cluster profile**

📚 Cluster assignment and profile



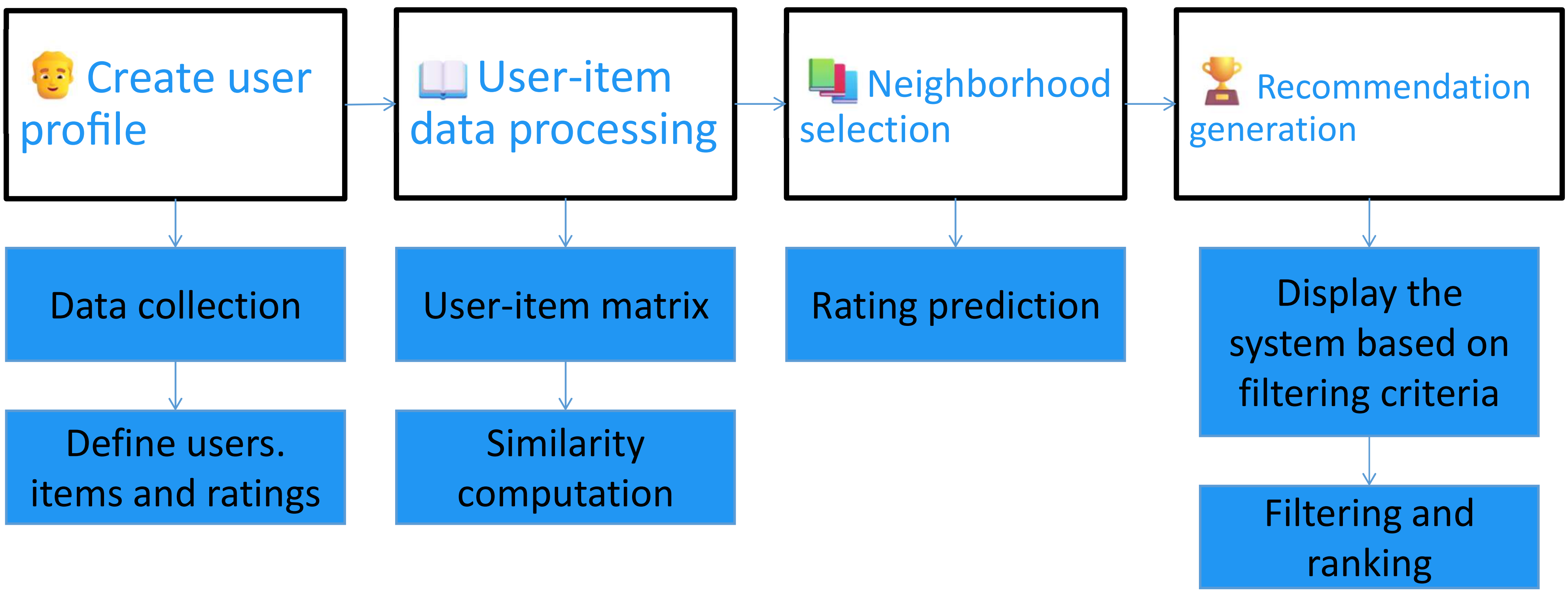
1. **Identify groups of users with similar preferences within the same cluster.**
2. **Users within the same cluster typically share common characteristics or preferences.**
3. **The system focuses on clusters. reducing the complexity from considering every user.**
4. **Recommendations remain relevant from changes.**
5. **PCA helps reduce dimensionality, highlighting key features that influence cluster formation, and improves model performance by simplifying the data while maintaining important patterns**

Collaborative-Filtering Recommender System using Supervised Learning

Models and Findings

KNN-Based Collaborative Filtering

Flowchart



Construct

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

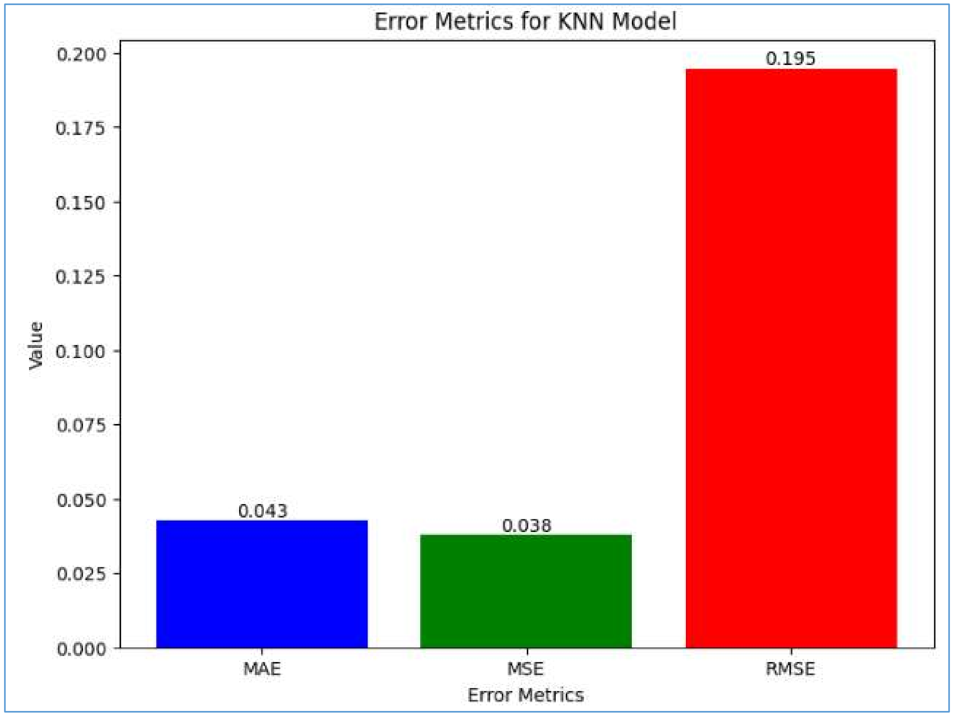
- Spark
- Sql
- Python

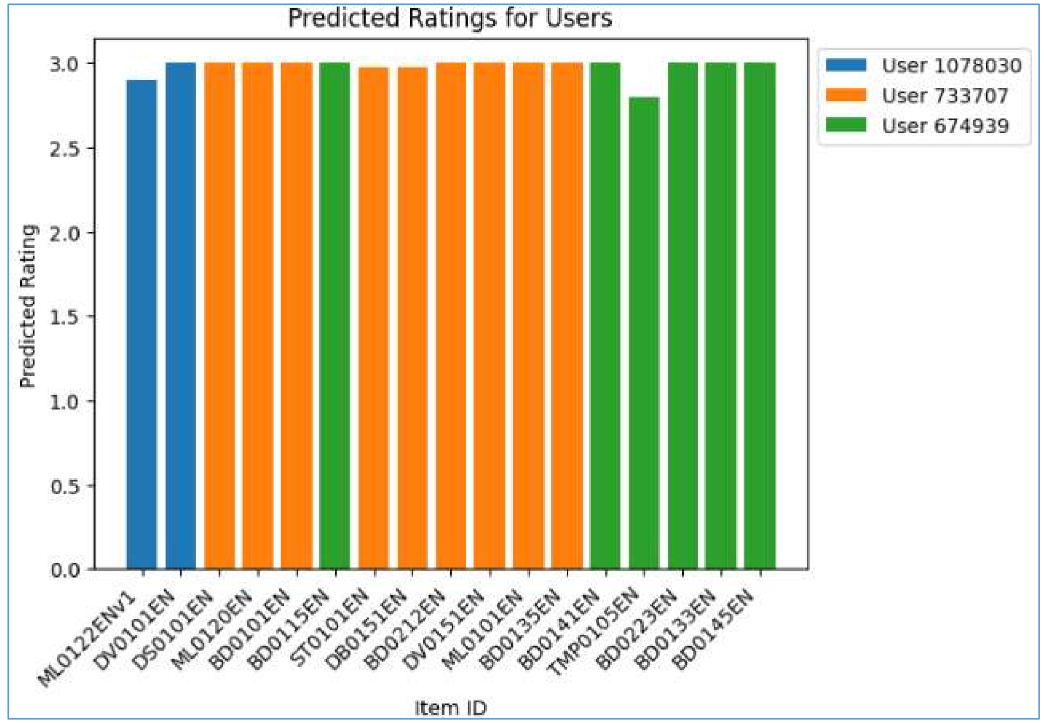
User profile 674939

Participate in 15 courses

- Spark
- Hadoop
- Big data

Evaluation Results





Construct

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

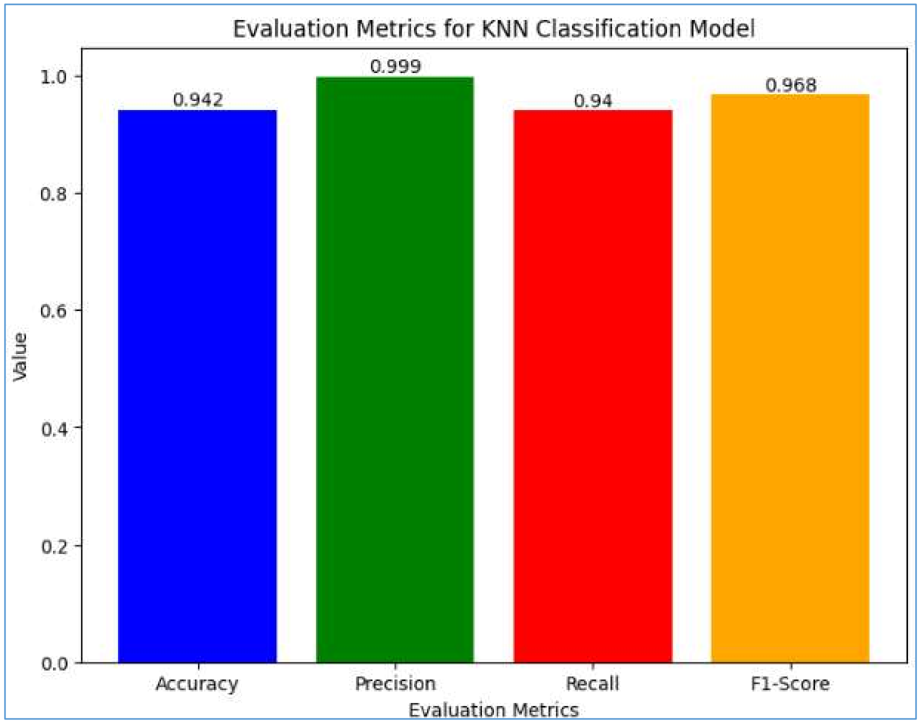
Participate in 15 courses

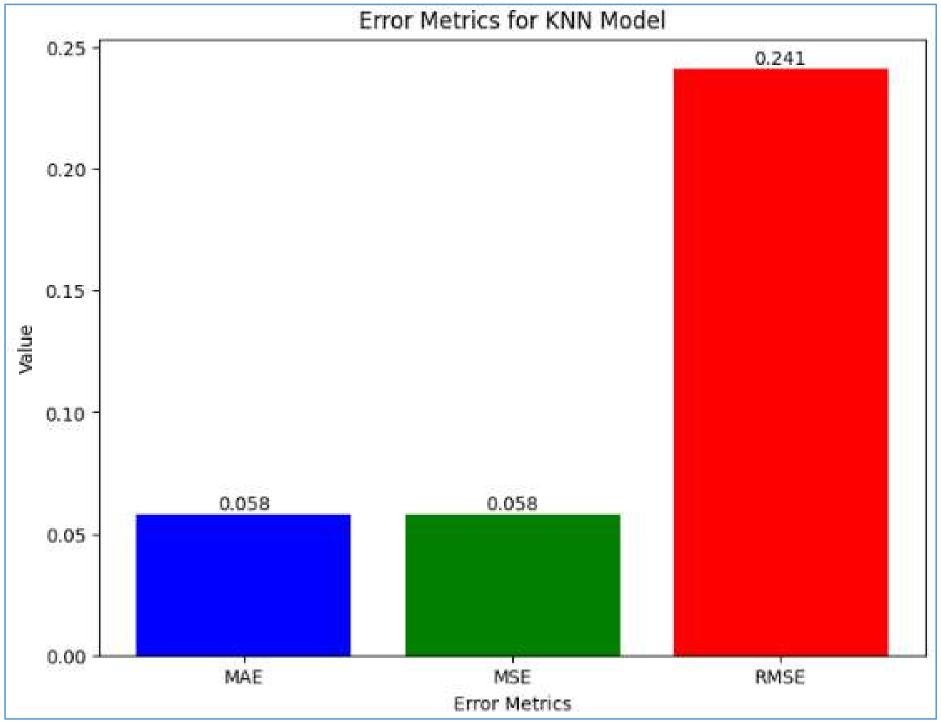
- Spark
- Hadoop
- Big data

- Spark
- Hadoop

Construct

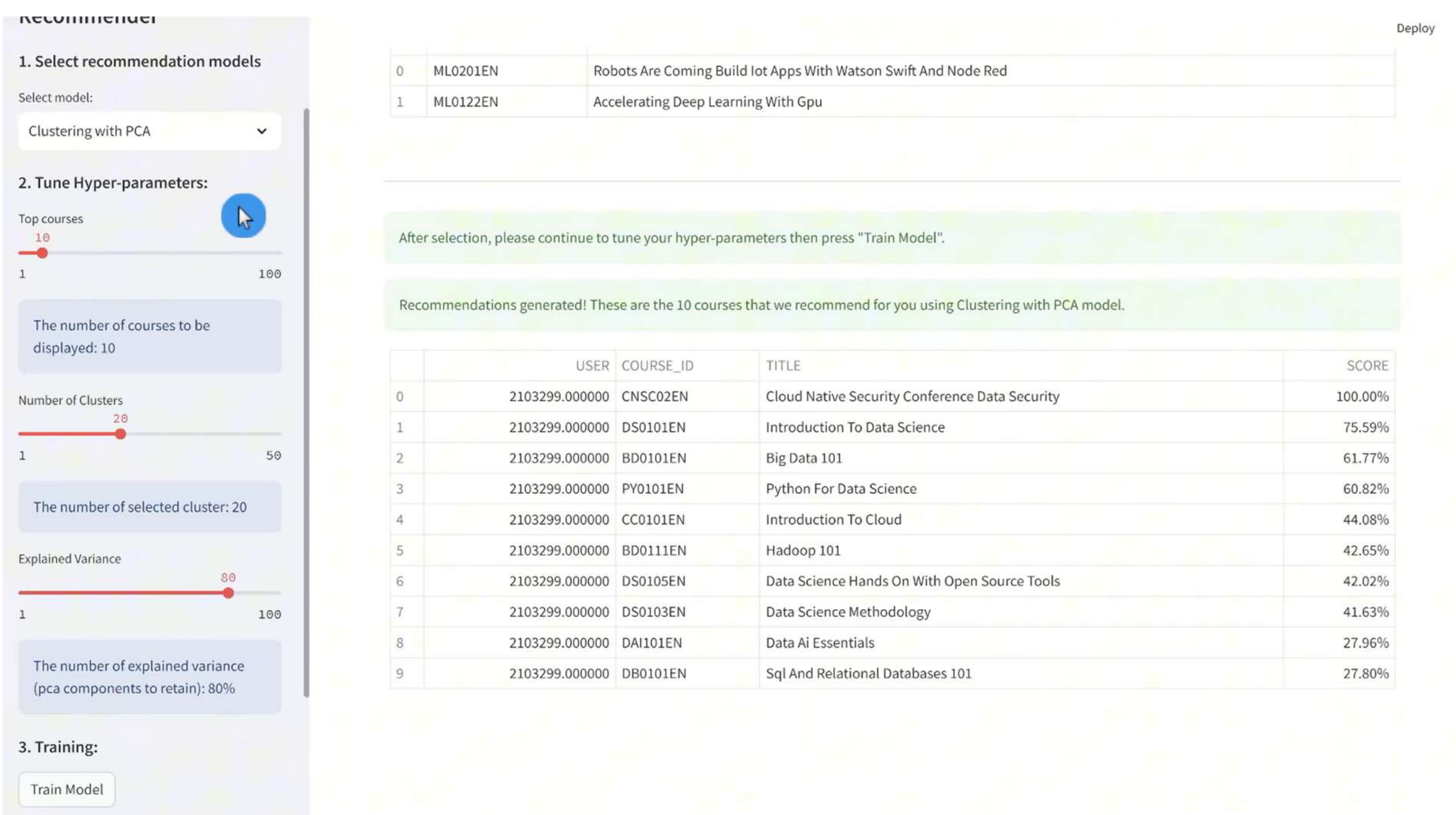
Evaluation Results (binary labels)





Construct

Recommendation based on KNN



0	ML0201EN	Robots Are Coming Build Iot Apps With Watson Swift And Node Red
1	ML0122EN	Accelerating Deep Learning With Gpu

	USER	COURSE_ID	TITLE	SCORE
0	2103299.000000	CNSC02EN	Cloud Native Security Conference Data Security	100.00%
1	2103299.000000	DS0101EN	Introduction To Data Science	75.59%
2	2103299.000000	BD0101EN	Big Data 101	61.77%
3	2103299.000000	PY0101EN	Python For Data Science	60.82%
4	2103299.000000	CC0101EN	Introduction To Cloud	44.08%
5	2103299.000000	BD0111EN	Hadoop 101	42.65%
6	2103299.000000	DS0105EN	Data Science Hands On With Open Source Tools	42.02%
7	2103299.000000	DS0103EN	Data Science Methodology	41.63%
8	2103299.000000	DAI101EN	Data Ai Essentials	27.96%
9	2103299.000000	DB0101EN	Sql And Relational Databases 101	27.80%

Summary

The idea

🧑 Create user profile

📖 User-item matrix

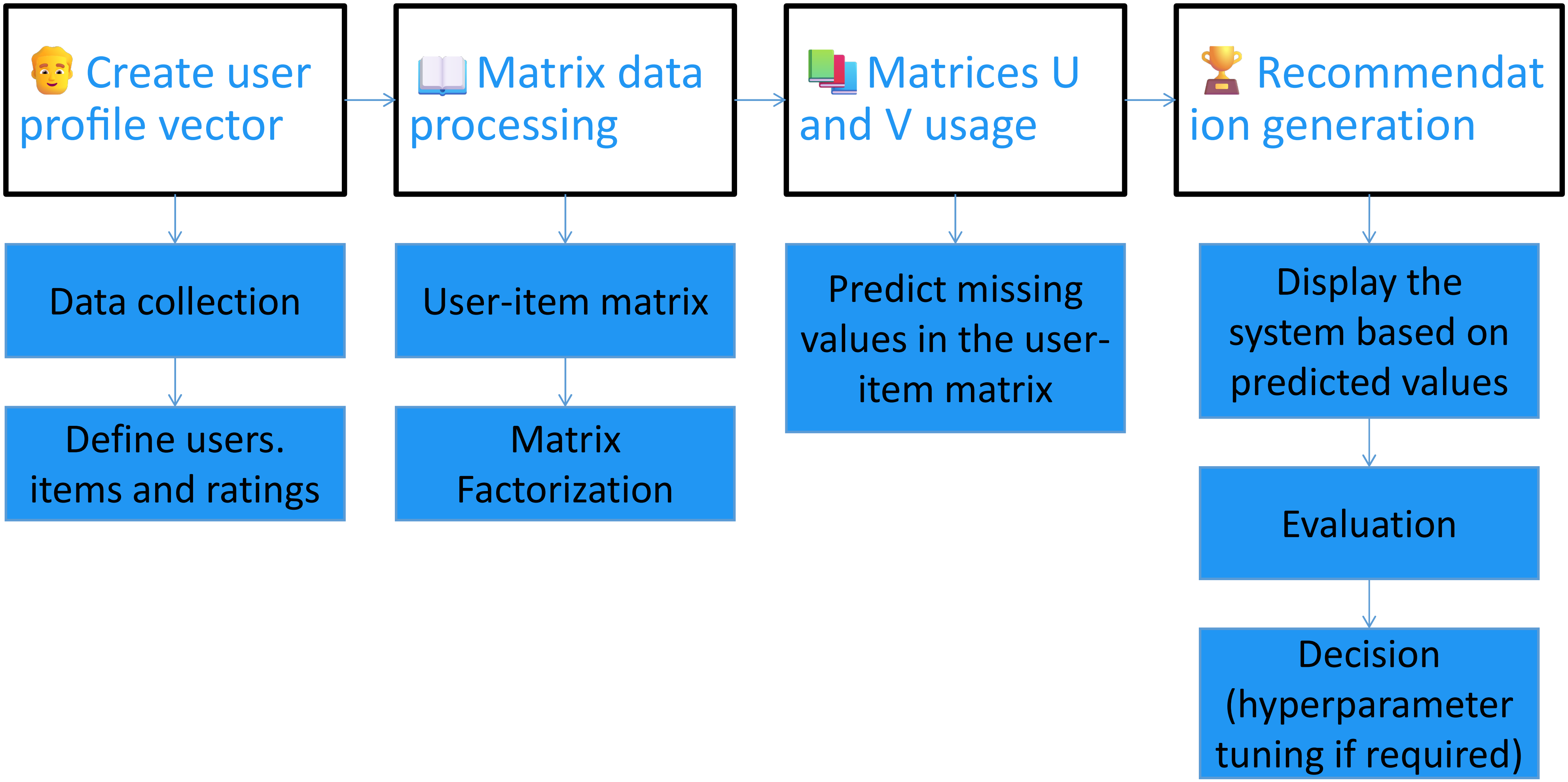
📚 Neighborhood selection

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

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

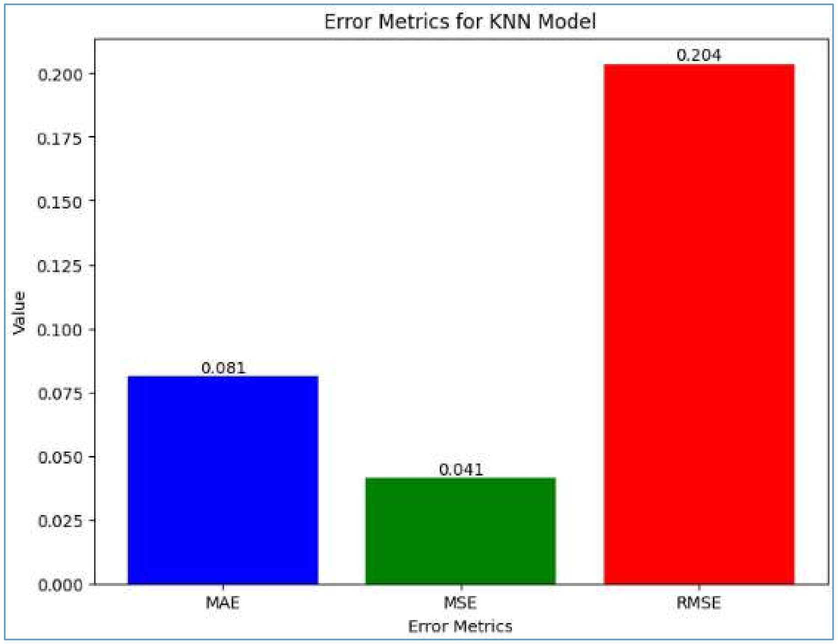
- Spark
- Sql
- Python

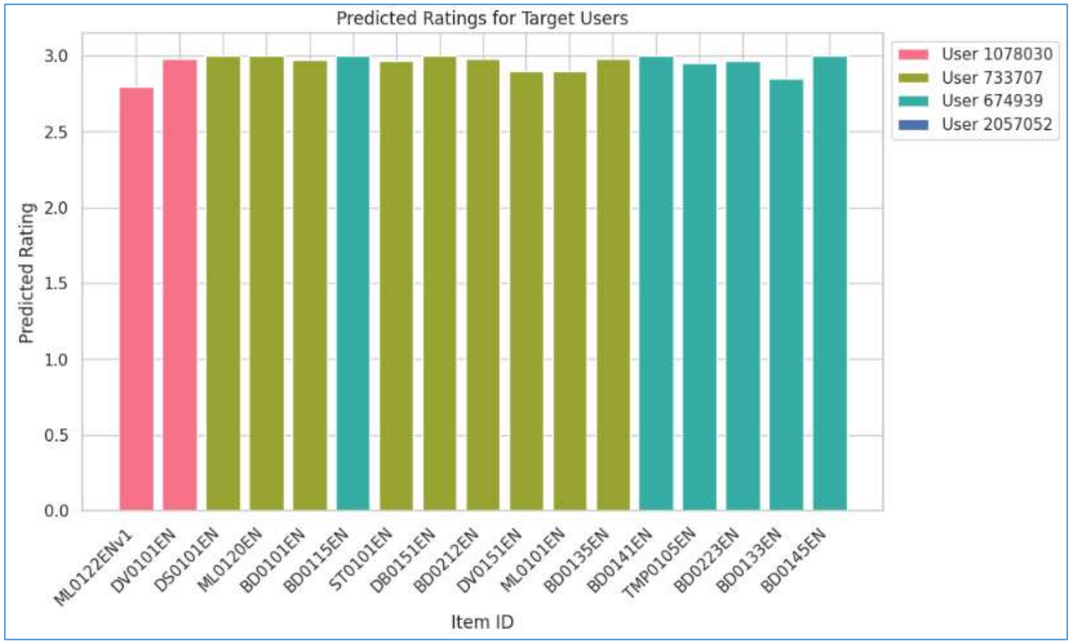
User profile 674939

Participate in 15 courses

- Spark
- Hadoop
- Big data

Evaluation Results





Construct

Recommendation based on NMF

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

Recommended courses

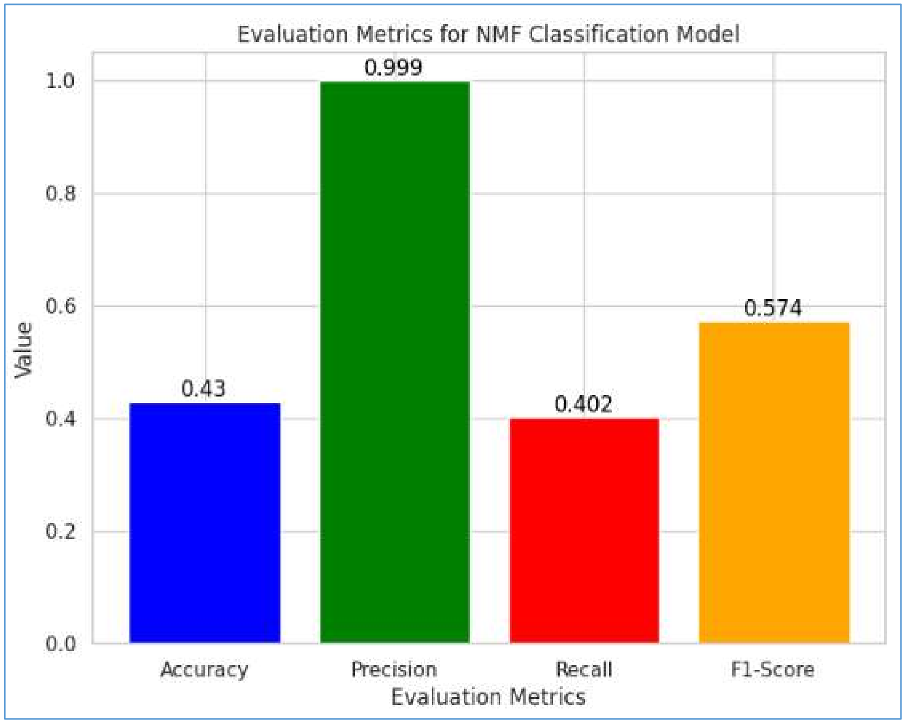
- Data visualization
- Deep learning

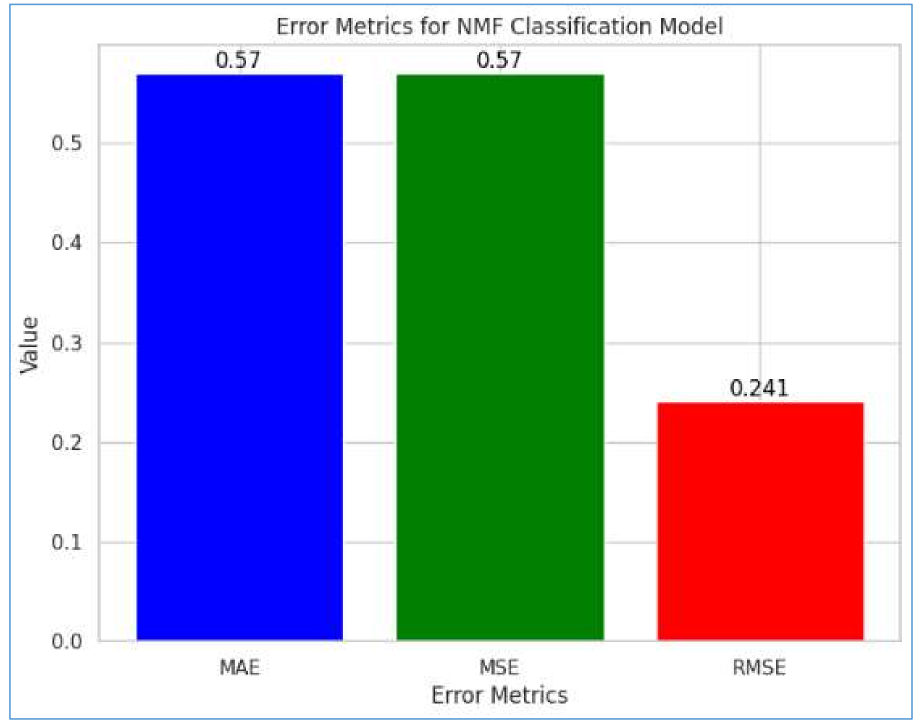
- Data science
- Deep learning
- Big data
- Spark
- Data visualization
- Machine learning
- Statistics
- Nosql

- Spark
- Hadoop
- Apache
- Sql

Construct

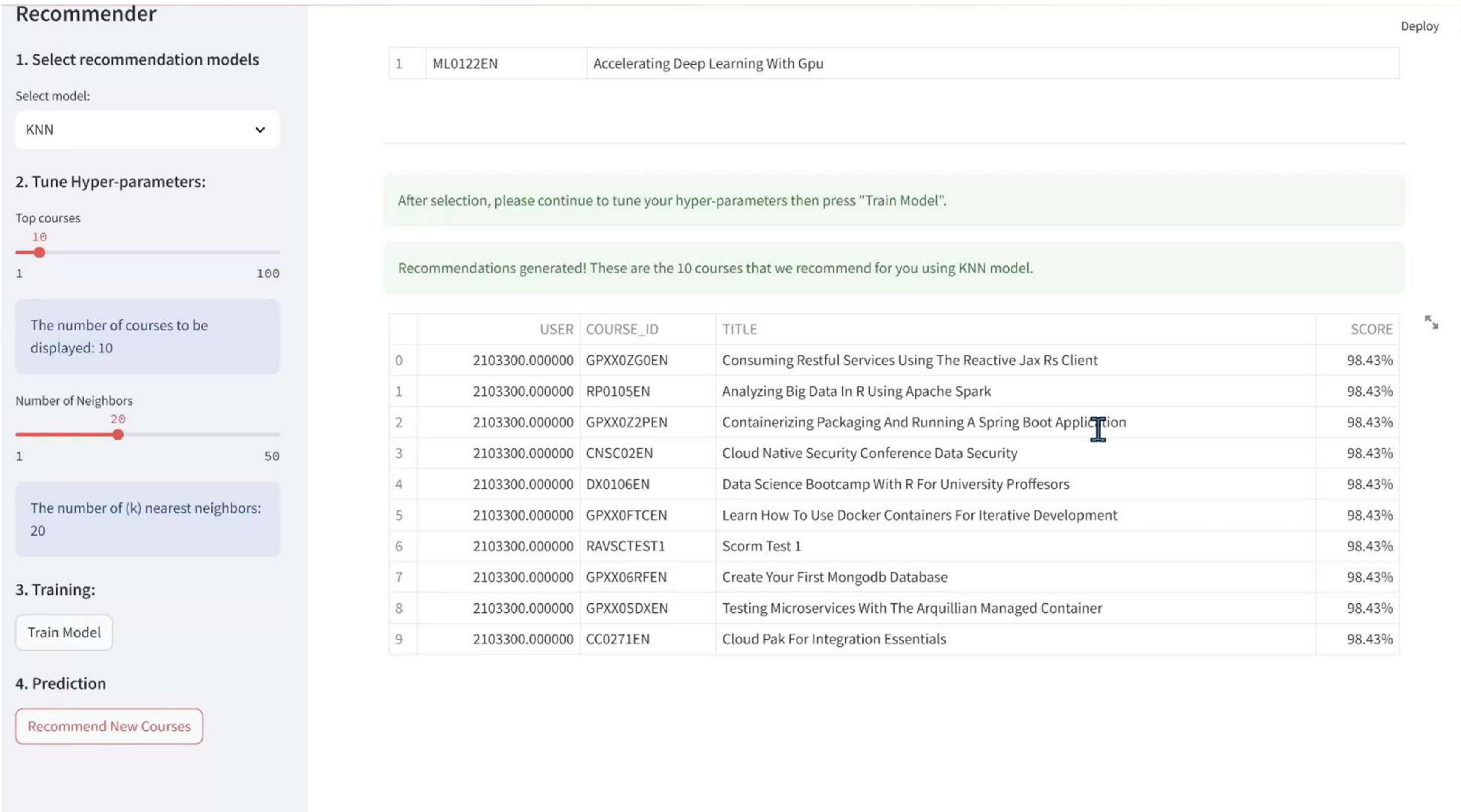
Evaluation Results (binary labels)





Construct

Recommendation based on NMF



	USER	COURSE_ID	TITLE	SCORE
0	2103300.000000	GPXX0ZG0EN	Consuming Restful Services Using The Reactive Jax Rs Client	98.43%
1	2103300.000000	RP0105EN	Analyzing Big Data In R Using Apache Spark	98.43%
2	2103300.000000	GPXX0Z2PEN	Containerizing Packaging And Running A Spring Boot Application	98.43%
3	2103300.000000	CNSC02EN	Cloud Native Security Conference Data Security	98.43%
4	2103300.000000	DX0106EN	Data Science Bootcamp With R For University Proffesors	98.43%
5	2103300.000000	GPXX0FTCEN	Learn How To Use Docker Containers For Iterative Development	98.43%
6	2103300.000000	RAVSCTEST1	Scorm Test 1	98.43%
7	2103300.000000	GPXX06RFEN	Create Your First Mongodb Database	98.43%
8	2103300.000000	GPXX0SDXEN	Testing Microservices With The Arquillian Managed Container	98.43%
9	2103300.000000	CC0271EN	Cloud Pak For Integration Essentials	98.43%

Summary

The idea

Create user profile vector

Matrix Factorization

Matrix data processing

Matrices U and V usage

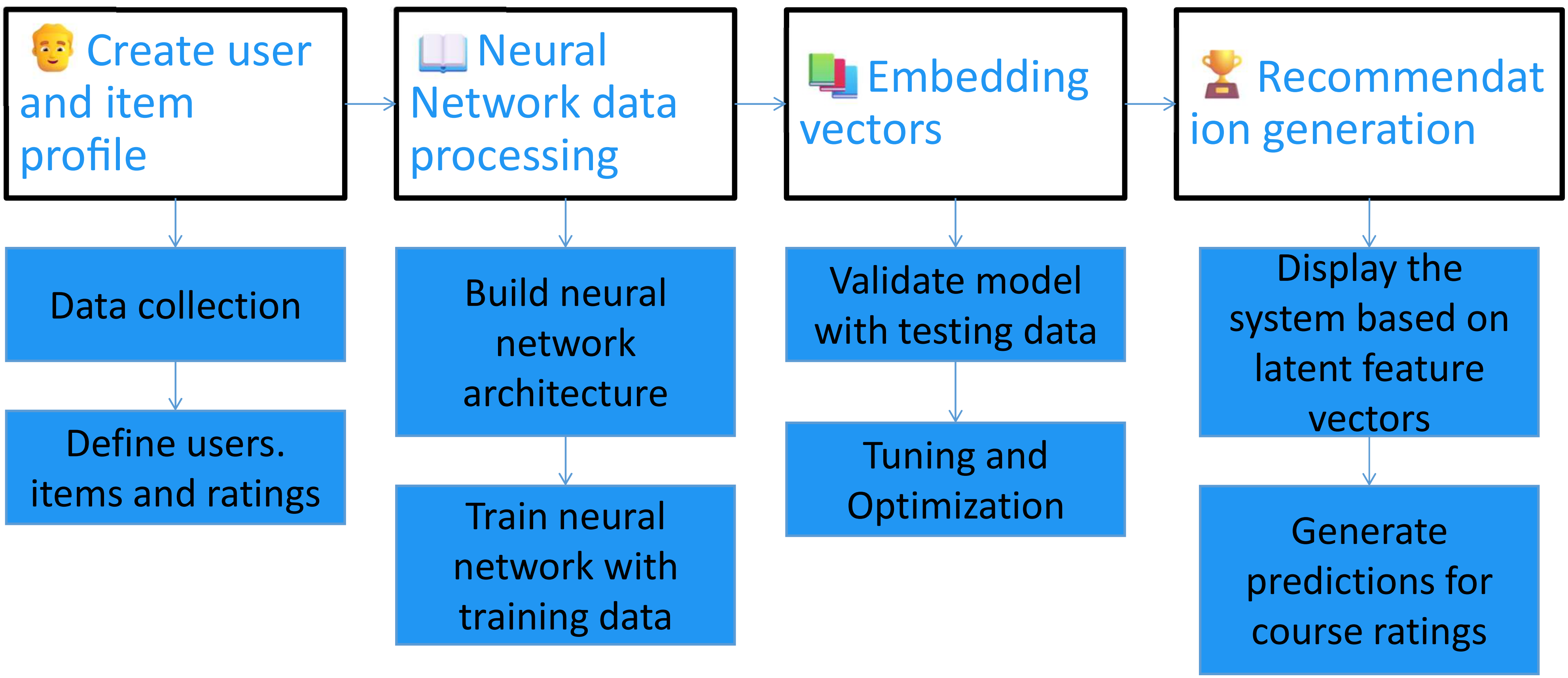
(Predict missing values)

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

Flowchart



Construct

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

User *i*

User Feature Engineering

User Profile feature vector

	genre1	genre2	genre3	...	genreN
user i	1.0	0	1.0	...	0

User Enrollment feature vector

	item1	item2	item3	...	itemN
user i	2.0	0	3.0	...	3.0

Item j

Item Feature Engineering

Item genre feature vector

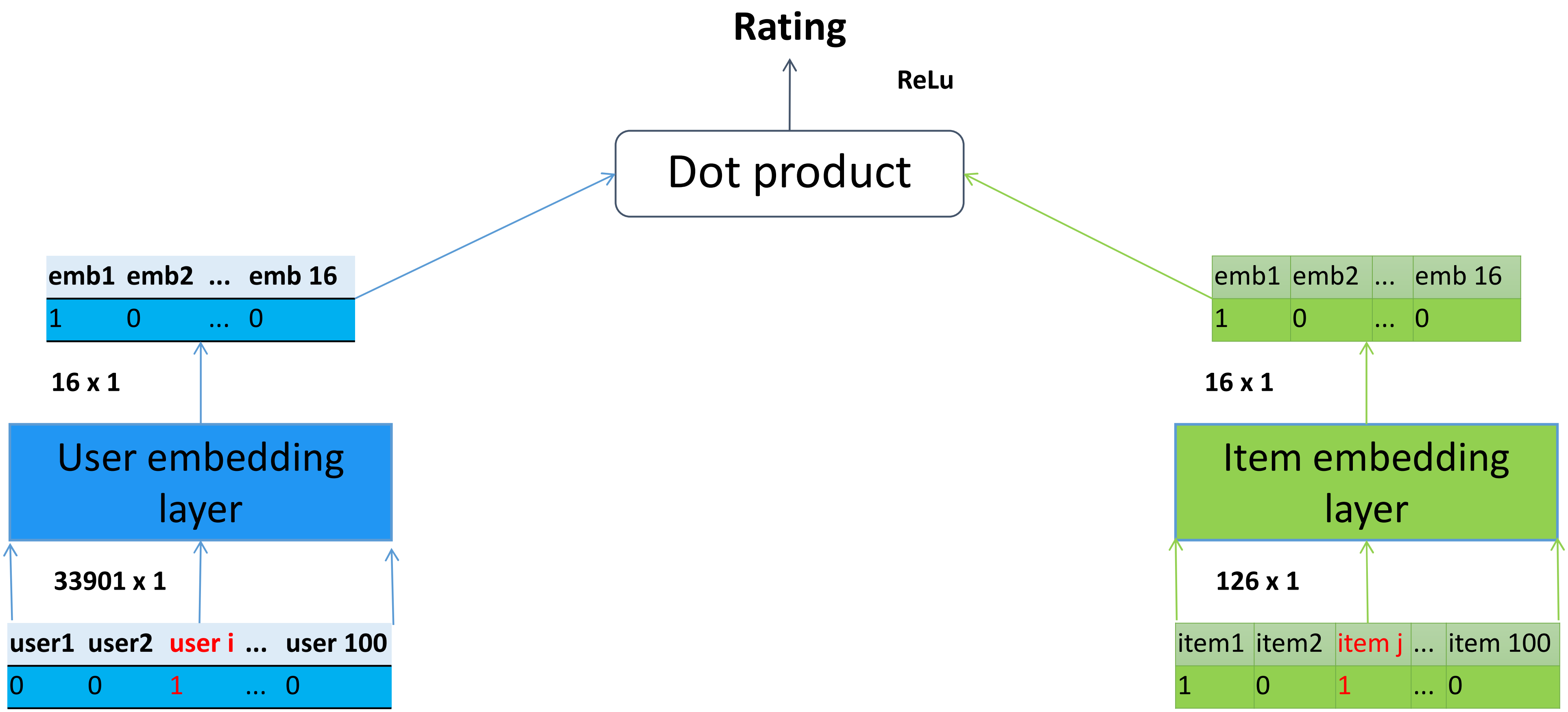
	genre1	genre2	genre3	...	genreN
item j	1.0	0	1.0	...	0

Item Enrollment feature vector

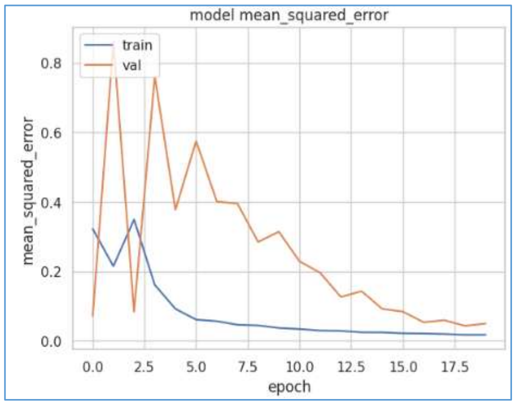
	user1	user2	user3	...	userN
item j	2.0	0	3.0	...	3.0

Latent feature vectors

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

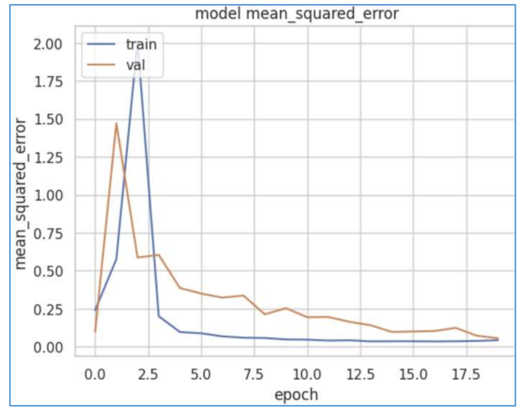


Evaluation Results (Improve performance)




```
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
_________________________________________________________________
```

Before recommender_net




```
Model: "recommender_net_plus"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 user_embedding_layer (Embed  multiple                 1084832
 ding)

 user_bias (Embedding)       multiple                  33901

 item_embedding_layer (Embed  multiple                 4032
 ding)

 item_bias (Embedding)       multiple                  126

=================================================================
Total params: 1,122,891
Trainable params: 1,122,891
Non-trainable params: 0
_________________________________________________________________
```

After Improvement recommender_net_plus

Component	recommender_net	recommender_net_plus	Difference
User Embedding	56,416 params	1,641,022 params	29x larger
Item Embedding	2,016 params	4,052 params	2x larger
Total Params (Trainable Params)	~571K–591K	~1.12M–1.25M	~2x more complex
Validation Error Stability	Fluctuates	More Stable	
Generalization	Overfitting Risk	Improved Generalization	

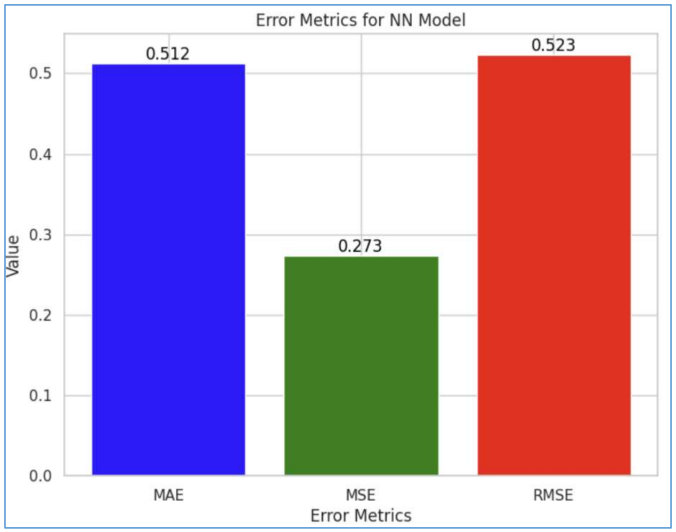
Evaluation Results (Improve performance)

Model	Initial MSE (Start)	Final MSE (End)	Epochs
recommender_net	~0.6	~0.0	17.5
recommender_net_plus	~1.75	~0.25	17.5

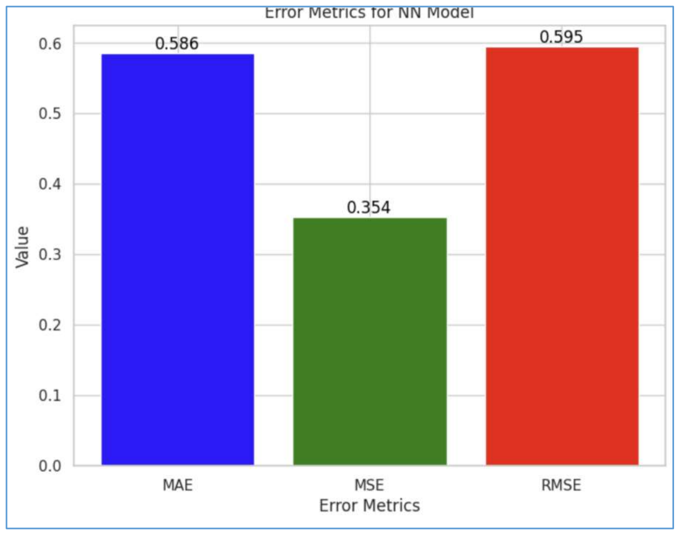
Key takeaways :

- **recommender_net**:
  - Achieves near-zero MSE quickly (by epoch 5), suggesting **overfitting**. The model memorizes training data but fails to generalize (validation MSE likely high, though not shown).
  - Example: Predicts training users' ratings perfectly but fails for new users.
- **recommender_net_plus**:
  - Starts with higher MSE but shows **steady improvement**, indicating better optimization and generalization.
  - Example: Predicts ratings for both training and new users more reliably, though not perfectly
- **What we did**:
  - Increased Embedding Dimensions.
  - Optimized Learning Rate and Training Strategy
  - Added Regularization & Dropout
  - More Training Data or Longer Training Time

Evaluation Results (3rd)



Before recommender_net



After Improvement recommender_net_plus

Key takeaways :

The improved model:

- **MAE (Mean Absolute Error):**
  - After the improvements, the MAE increased slightly to **0.586**.
- **MSE (Mean Squared Error):**
  - MSE increased to **0.354**, suggesting that while the error is slightly higher, the model is becoming more robust to outliers due to better handling of complex relationships between users and items.

Key takeaways :

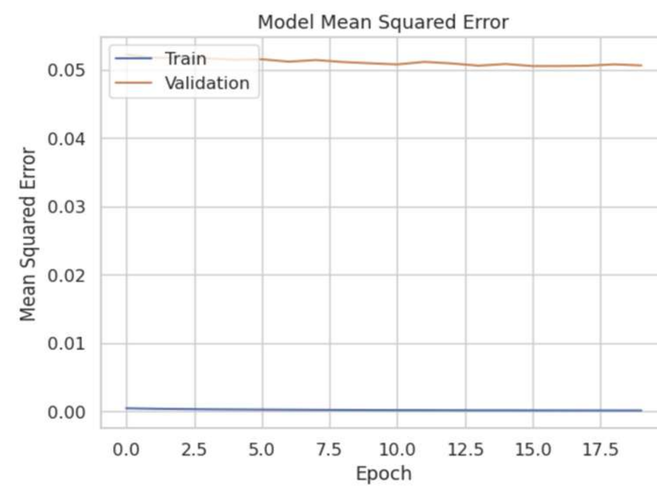
The improved model:

- **RMSE (Root Mean Squared Error):**
  - RMSE shows a slight increase to **0.595**, reflecting that while the error is somewhat higher, the model can handle unseen data better.

2nd Improvement

```
Model: "model"
____________________________________________________________________________________________
 Layer (type)                 Output Shape         Param #     Connected to
============================================================================================
 user_id (InputLayer)         [(None,)]            0           []

 item_id (InputLayer)         [(None,)]            0           []

 mlp_user_embedding (Embedding)  (None, 32)        1084832     ['user_id[0][0]']

 mlp_item_embedding (Embedding)  (None, 32)        4064        ['item_id[0][0]']

 flatten_2 (Flatten)          (None, 32)           0           ['mlp_user_embedding[0][0]']

 flatten_3 (Flatten)          (None, 32)           0           ['mlp_item_embedding[0][0]']

 concatenate (Concatenate)    (None, 64)           0           ['flatten_2[0][0]',
                                                                'flatten_3[0][0]']

 mf_user_embedding (Embedding)  (None, 32)         1084832     ['user_id[0][0]']

 mf_item_embedding (Embedding)  (None, 32)         4064        ['item_id[0][0]']

 layer_0 (Dense)              (None, 16)           1040        ['concatenate[0][0]']

 flatten (Flatten)            (None, 32)           0           ['mf_user_embedding[0][0]']
```

```
Model: "model"
____________________________________________________________________________________________
 Layer (type)                 Output Shape         Param #     Connected to
============================================================================================
 flatten_1 (Flatten)          (None, 32)           0           ['mf_item_embedding[0][0]']

 layer_1 (Dense)              (None, 8)            136         ['layer_0[0][0]']

 multiply (Multiply)          (None, 32)           0           ['flatten[0][0]',
                                                                'flatten_1[0][0]']

 layer_2 (Dense)              (None, 4)            36          ['layer_1[0][0]']

 concatenate_1 (Concatenate)  (None, 36)           0           ['multiply[0][0]',
                                                                'layer_2[0][0]']

 interaction (Dense)          (None, 1)            37          ['concatenate_1[0][0]']

============================================================================================
Total params: 2,179,041
Trainable params: 2,179,041
Non-trainable params: 0
____________________________________________________________________________________________
```



Feature	Recommender Net Plus	Model (from the image)
Architecture Complexity	Simple	Complex (Dual Pathway)
Training Speed	Faster	Slower
Number of Parameters	Lower (Fewer layers)	Higher (Over 2 million)
Computational Cost	Low	High
Interaction Modeling	Linear (Concatenation)	Nonlinear (Concatenation & Multiplication)
Risk of Overfitting	Lower	Higher
Generalization	Good	Dependent on regularization
Recommendation Accuracy	Moderate	Potentially higher
Best for	Small/Moderate datasets	Large/Complex datasets

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.**

Models and Findings

Collaborative Filtering Algorithms Evaluation

Models and Findings

Regression-based Rating Score Prediction using Embedding Features

Flowchart Regression Based

Create user and item profile
- Data collection
- Define users. items and ratings
- Feature engineering

Neural Network data processing
- Build neural network architecture

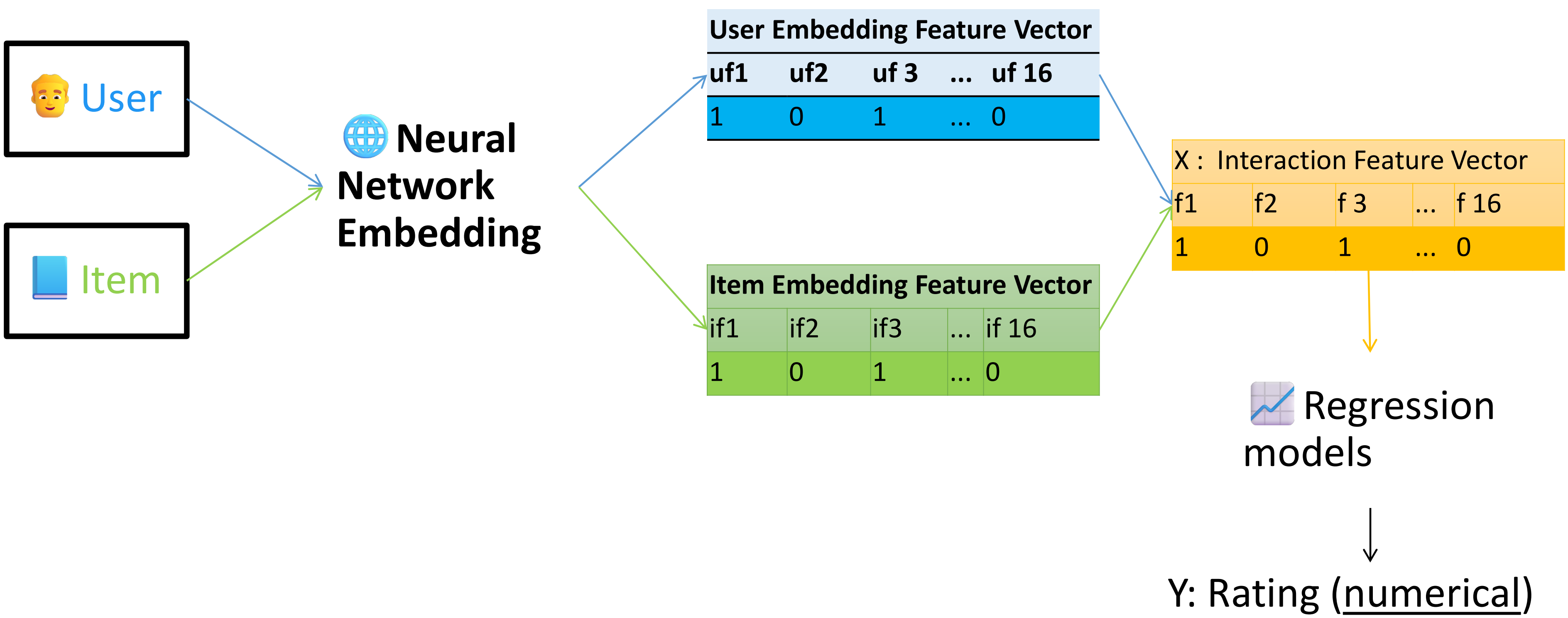
Embedding vectors
- Build regression model with embedding features
- Train neural network with training data
- Validate model with testing data

Recommendation generation
- Display the system based on relevant regression metrics
- Generate predictions for ratings scores

Construct

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**.

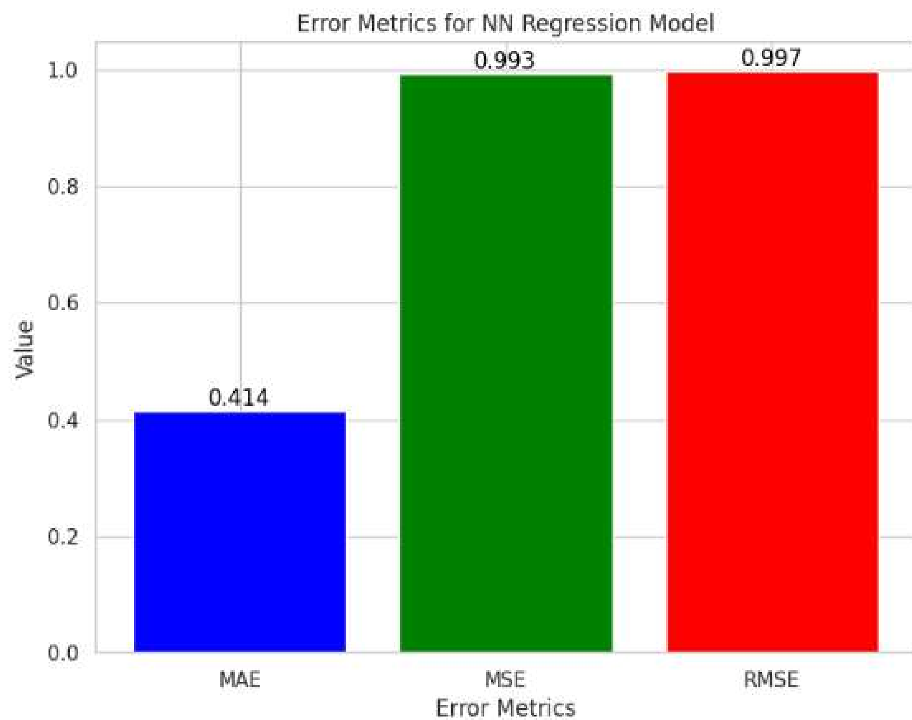


Evaluation Results (3rd)

MAE: 0.41428838083033687

MSE: 0.9932500760760065

RMSE: 0.9966193235513781



Construct

Summary

The idea

Create user and item profile

Build regression model

Neural network processing

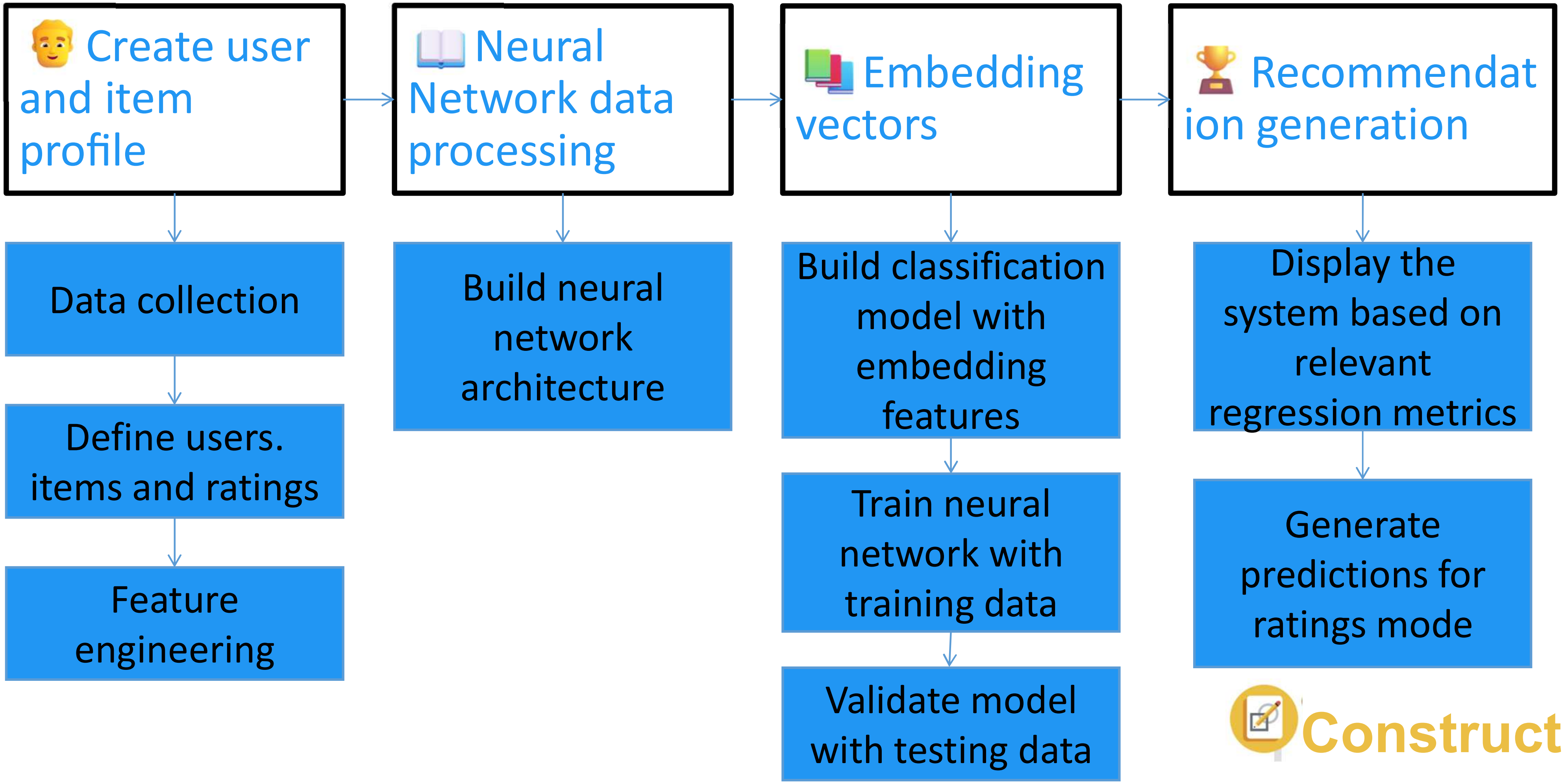
Embedding vectors

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

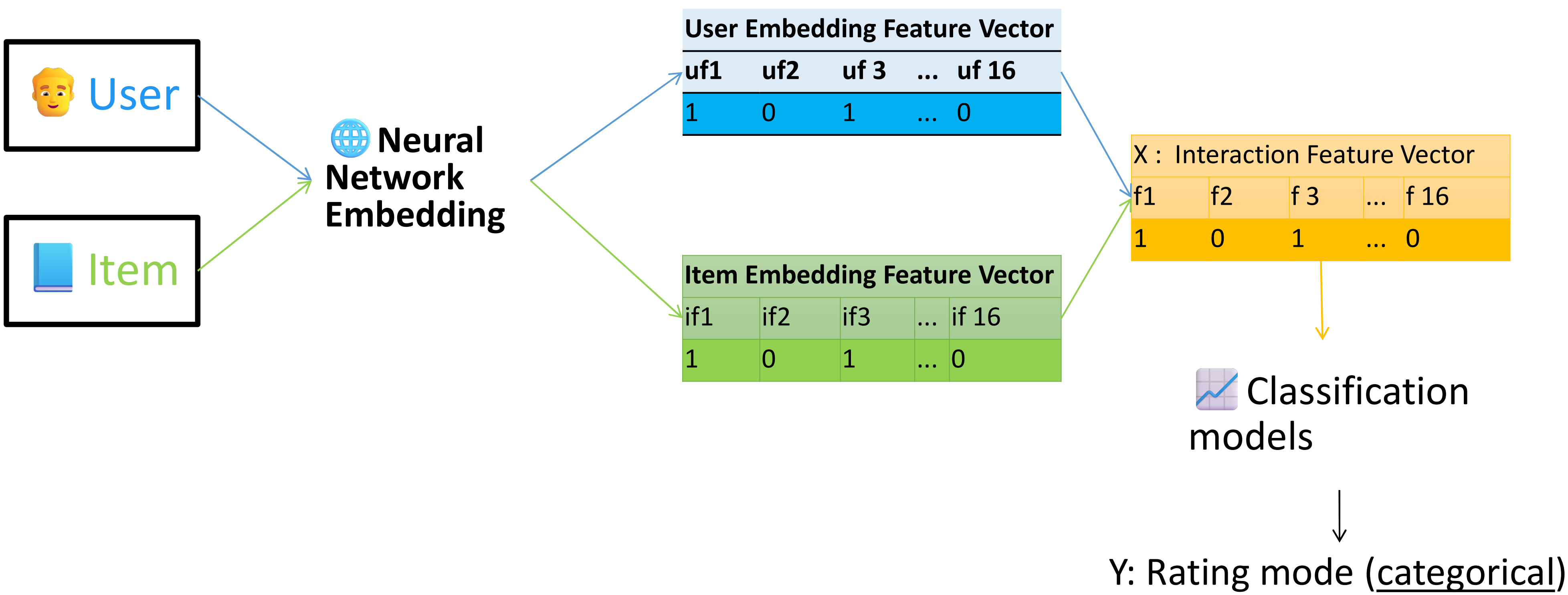
Classification-based Rating Score Prediction using Embedding Features

Flowchart Classification Based

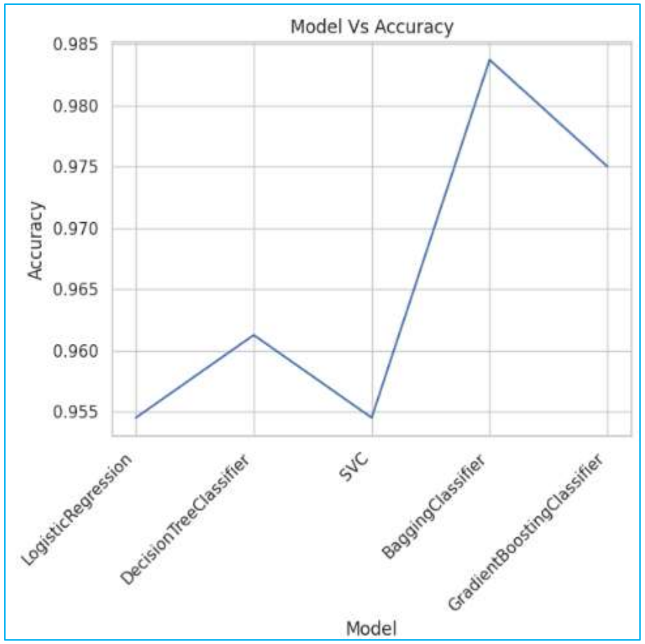


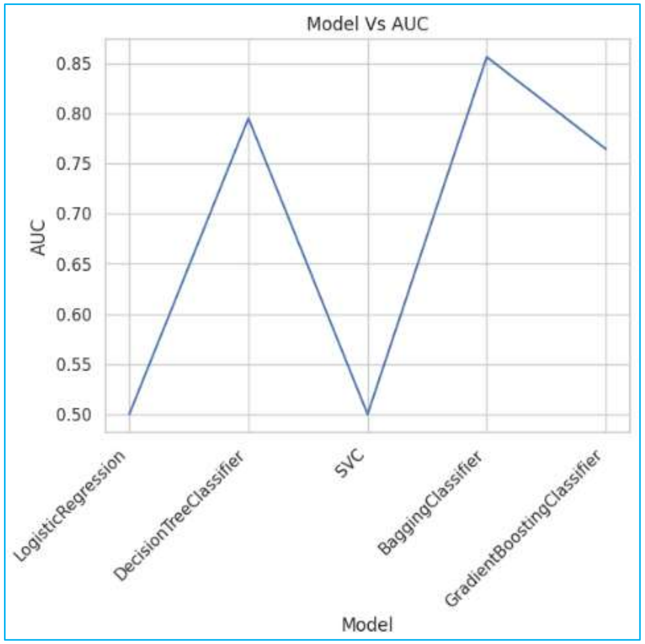
Neural Networks using Embedding Features

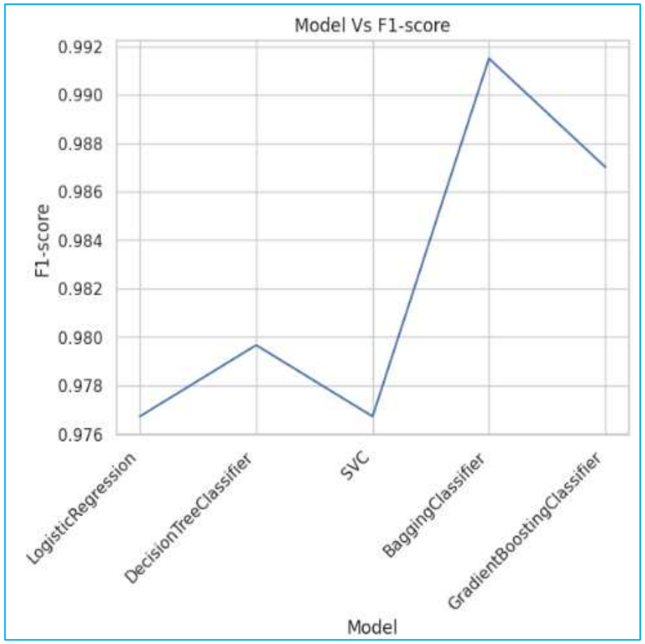
The prediction problem as **a classification problem** as rating only has **two categorical values (Aduit vs. Completion)**



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

Construct

Summary

The idea

🧑 Create user and item profile

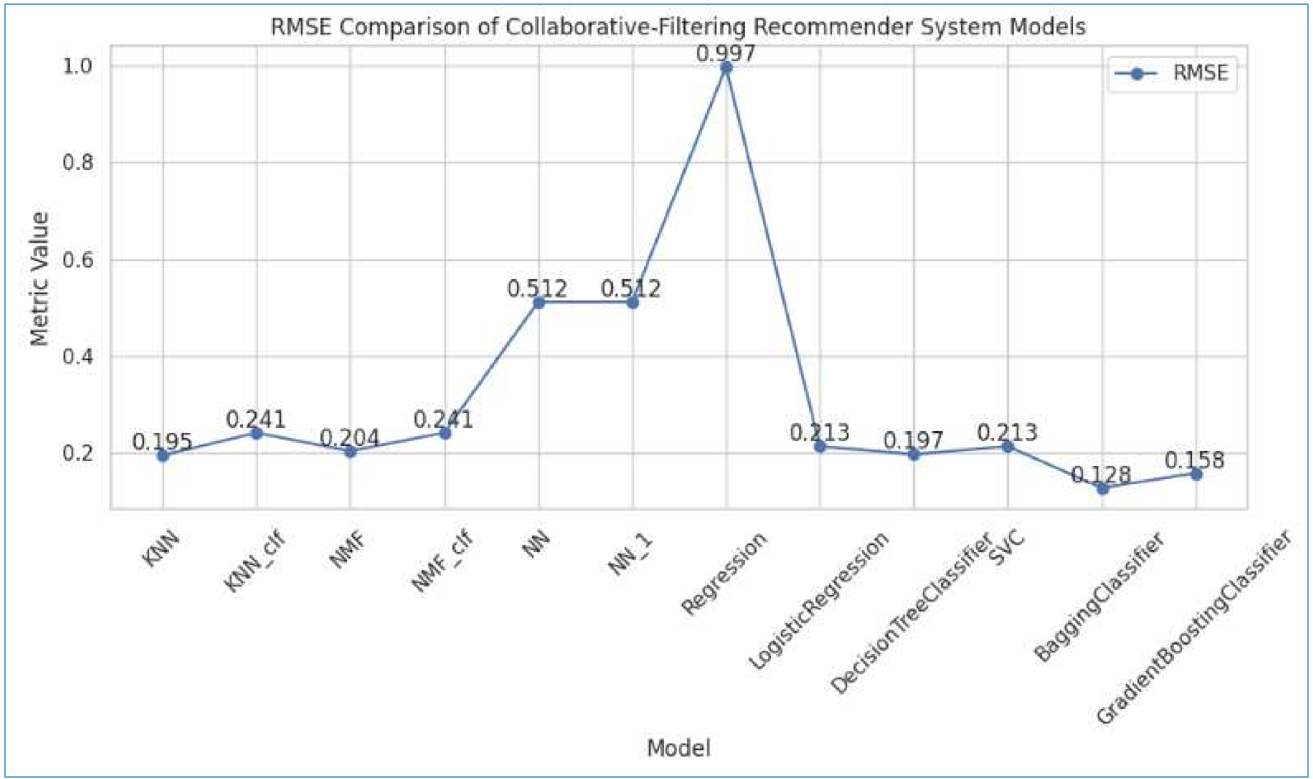
Build classification model

📖 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

Construct

Streamlit

Personal Recommended System

Recommender System in Streamlit

https://youtu.be/7zvritN8VO0

Personalized Learning Recommender System Supervised and Unsupervised Learning Using Streamlit

Unlisted

Wahyu Ardhitama | Subscribe | 0 | Share | Download | Save

7 views Feb 24, 2025

Applying Recommender System in Streamlit with 6 Successful Models:

1. Content-Based Recommender System using Unsupervised Learning
   a. User Similarity
   b. Course Similarity
   c. Clustering
   d. Clustering with PCA
2. Collaborative-Filtering Recommender System using Supervised Learning
   a. KNN-Based Collaborative Filtering
   b. NMF-Based Collaborative Filtering

0:00 - Intro
0:37 - Content-Based Recommender System using Unsupervised Learning
2:10 - User Similarity
3:33 - Course Similarity
5:42 - Clustering
6:45 - Clustering with PCA
9:07 - KNN-Based Collaborative Filtering
10:25 - NMF-Based Collaborative Filtering
11:38 - Closing
12:10 - MIT ADSP (Postpone)

Link:
1. https://www.kaggle.com/code/wahyuardh...
2. https://www.kaggle.com/code/wahyuardh...
3. https://www.kaggle.com/code/wahyuardh...
4. https://github.com/whyzie/Task003-ML-...

Recommender System Streamlit Flowchart

Course Recommendation App

rec_app.py

Preparation and Loading

Import, Basic Web Setup and Load Dataset

Init_recommender_app

Build an Interactive Table

Interact with Training and Prediction

Build an Interactive Table for Training and Prediction

Interact with Hyperparameter

Build an Interactive Table for Hyperparameter

beckend.py

Preparation and Loading

Import and Load Dataset

Interact with File

Add New Ratings and Mapping id and index

Interact with Training

Define Model and Training in accordance to the specified Model

Interact with Prediction

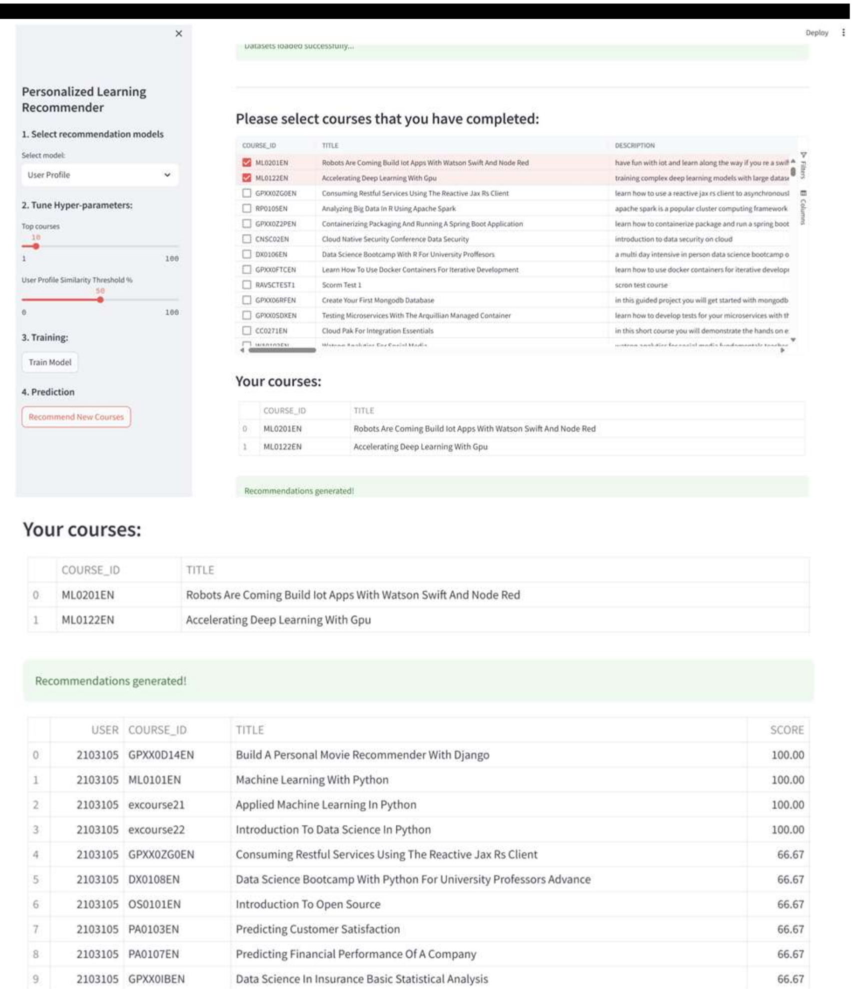
Define Prediction accordance to the specified Model and Parameters

A

A

Interact with Specified Model

Construct model accordance to the specified Model and Parameters



Recommender System Streamlit Flowchart

Course Recommendation App

beckend.py

A

Interact with Specified Model

Construct model accordance to the specified Model and Parameters

Model 0 User Profile Similarity

Set parameter:
enrolled courses = A list of course IDs that the user has enrolled in
course_genres = A DataFrame containing the genre information for all courses.
tres_rec = A threshold value

Return r = the sorted recommendation

Model 1 Course Similarity

Set parameter:
idx_id_dict = A dictionary mapping course indices to course IDs
id_idx_dict = A dictionary mapping course IDs to course indices.
sim_matrix = A similarity matrix representing the similarity between courses
Enrolled courses = A list of course IDs that the user has enrolled in
Tres_rec = A threshold value

Return r = the sorted recommendation

Model 2 Train Cluster

Set parameter:
user_profiles= A matrix of user genre preferences, with users as rows and genre weights as columns.
course_genres = A DataFrame containing the genre information for all courses.
pca= A boolean flag indicating whether to apply Principal Component Analysis (PCA) for dimensionality reduction
exp_var= The amount of variance to preserve if PCA is applied
n_clus = The number of clusters for the KMeans algorithm
init= The method for initializing centroids in the KMeans algorithm.
n_init= The number of times the KMeans algorithm will run with different centroid seeds.
tol = The tolerance for declaring convergence.
max_iter = The maximum number of iterations for the KMeans algorithm.

Return km = Return Kmeans Model

Model 3 With PCA

Set parameter:
data= The input data on which PCA will be performed.
exp_var= The amount of variance to preserve if PCA is applied. An integer with a default value of 80. This means the PCA will aim to capture 80% of the variance in the data by default.

Return df = Return the reduced data

Model 2 Kmean Cluster

Set parameter:
model = A trained clustering model (e.g., KMeans).
user_profiles = A matrix of user genre preferences, with users as rows and genre weights as columns.
rated_courses = DataFrame containing users and the courses they have rated
enrolled courses = A list of course IDs that the user has enrolled in

Return r = the sorted recommendation

Model 4 Train Knn

Set parameter:
k (default 40)= Number of neighbors to consider.
min_k (default 2) = A Minimum number of neighbors necessary to make a prediction.
Metric (default 'msd') = The distance metric to use (e.g., 'msd' for mean squared difference).
user_based = user-based (rather than item based) collaborative filtering
verbose = Whether to print progress to the console.

Return model = Return the trained model

B

Recommender System Streamlit Flowchart

Course Recommendation App

beckend.py

B

Model 4 Knn Predict

Set parameter:
model = A trained KNNBasic model from the surprise library.
enrolled courses = A list of course IDs that the user has enrolled in

Return r = the sorted recommendation

Model 4 Train Embedding Knn Classifier

Set parameter:
ds = A pandas DataFrame containing the dataset with feature columns and the target variable. The last two columns are assumed to be the target variable and an identifier.

Return knn_emb = Return the trained KNN classifier

Model 4 Embedding Knn Classifier Predict

Set parameter:
model = A trained KNNBasic model from the surprise library.
Enrolled_courses = A list of course IDs that the user has enrolled in.
df_regression= A DataFrame containing the user and item embedding. **df_ratings**= A DataFrame containing the ratings data, loaded by default using the load_ratings function.

Return Returns the normalized scores as a dictionary

Model 5 Train NMF

Set parameter:
n_factors (default 10)= The number of latent factors to use in the NMF model
n_epochs (default 20) = The number of iterations of the algorithm.
verbose = progress messages during training .

Return nmf = Return the trained nmf

Model 5 NMF Predict

Set parameter:
model = A trained clustering model (e.g., Non Negative Matrix).
enrolled courses = A list of course IDs that the user has enrolled in.
verbose = A boolean flag to indicate whether to print verbose output during prediction.

Return r = the sorted recommendation

Model 6 Create NCF

Set parameter:
n_users = Number of unique users.
n_items = Number of unique items.
latent_dim_mf = Number of latent dimensions for the matrix factorization (MF) part.
latent_dim_mlp = Number of latent dimensions for the multi-layer perceptron (MLP) part.
reg_mf = Regularization strength for the MF embeddings.
reg_mlp = Regularization strength for the MLP embeddings.
dense_layers = List defining the number of units in each dense layer of the MLP part.
reg_layers = List defining the regularization strength for each dense layer.
activation_dense = Activation function used in the dense layers.

Return model = Return the compiled model

C

Recommender System Streamlit Flowchart

Course Recommendation App

beckend.py

C

Model 6 NCF Data Preparation

Set parameter:
df = A pandas DataFrame containing user IDs, item IDs, and ratings. The shape of the DataFrame is expected to be (n, 3).

Return df_train = the sorted recommendation

Model 6 Build, Train and Validate NCF

Set parameter:
df = A pandas DataFrame containing the preprocessed data with columns 'user_id', 'item_id', and 'interaction'.
val_split = A float representing the fraction of data to be used for validation. The default value is 0.1, meaning 10% of the data will be used for validation.
batch_size = An integer specifying the batch size for the TensorFlow datasets. The default value is 512.
rs = An integer for the random seed to ensure reproducibility when shuffling the data. The default value is 42.

Return ds_train, ds_val

Model 6 Train NCF

Set parameter:
ds_train = A TensorFlow dataset object used for training.
ds_val = A TensorFlow dataset object used for validation.
n_epochs = An integer representing the number of epochs to train the model. The default value is 10.

Return ncf_model, ncf_hist

Model 6 NCF Predict

Set parameter:
model = A trained clustering model (e.g., Neural Network).
enrolled courses = A list of course IDs that the user has enrolled in.

Return r = the sorted recommendation

Model 7 Create Emb

Set parameter:
n_user = The number of unique users.
n_item = The number of unique items (e.g., courses or products).
n_user_latent_dim = The dimension of the latent space for users (default is 16).
n_item_latent_dim = The dimension of the latent space for items (default is 16).
reg_users = Regularization strength for user embeddings (default is 1e-6).
reg_items = Regularization strength for item embeddings (default is 1e-6)

Return model, user_model, item_model
model = The complete embedding model, which includes user and item embeddings combined and additional dense layers.
user_model = A separate model for user embeddings.
item_model = A separate model for item embeddings.

D

Recommender System Streamlit Flowchart

Course Recommendation App

beckend.py

C

Model 7 Emb Data Preparation

Set parameter:
df = A Pandas DataFrame containing user-item interactions. The default value is the result of load_ratings(), a function that presumably loads the ratings data.

Return data, user_id2idx, item_id2idx
Data = A Pandas DataFrame with the user and item IDs replaced by their respective indices.
user_id2idx = A dictionary mapping original user IDs to their corresponding indices.
item_id2idx = A dictionary mapping original item IDs to their corresponding indices.

Model 7 Create Emb Dataset

Set parameter:
df = A Pandas DataFrame containing user-item interactions.
targets: list = A list of target column names to be predicted.
scale: bool = A boolean indicating whether to scale the features. The default value is True.
tt_split: int = The proportion of data to be used for training (train-test split). The default value is 0.8.
val_split: int = The proportion of the training data to be used for validation. The default value is 0.1.
batch_size: int = The batch size to be used for training. The default value is 512.
rs: int = The random state for shuffling the data. The default value is 42.

Return ds_train, ds_val, ds_test

Model 7 Train Emb

Set parameter:
ds = The TensorFlow dataset containing the training data.
df = A Panda DataFrame containing user-item interactions, loaded by default from load_ratings().
epochs: int = The number of epochs to train the model. Default is 10.

Return emb_model_all, user_emb_model_all, item_emb_model_all
emb_model_all: The complete embedding model.
user_emb_model_all: The model for user embeddings.
item_emb_model_all: The model for item embeddings.

Model 7 Emb Regression Dataset

Set parameter:
model = A trained neural collaborative filtering model containing user and item embeddings.
user_id2idx: dict = A dictionary mapping user IDs to indices.
item_id2idx = A dictionary mapping item IDs to indices.
df = A DataFrame containing user-item interactions, loaded by default from load_ratings().

Return regression_dataset

Model 7 Train Regression

Set parameter:
ds = A pandas DataFrame containing the training data. The last two columns are assumed to be the target variable and an identifier (like user ID).

Return lrr = The trained Ridged regression model

Model 7 Regression Predict

Set parameter:
model = A trained clustering model (e.g., Regression).
enrolled courses = A list of course IDs that the user has enrolled in.
ds = A pandas DataFrame containing the dataset with feature columns. The last two columns are assumed to be the target variable and an identifier.

Returning model prediction

Architectural Design

Personal Recommended System

Architectural Design

<table>
<tr><th>App</th><th>Recommender System Data Processing and Model Training</th><th>Data Processing and Model Training</th><th>Data Processing and Model Training Cost</th><th>Components and Services EMR</th><th>EMR Configuration</th><th>EMR Cost</th><th>Storage</th><th>Storage Cost</th><th>Data Transfer</th><th>Data Transfer Cost</th><th>Monitoring & Logging</th><th>Monitoring & Logging Cost</th></tr>
<tr><td>Web</td><td rowspan="2">Apache Spark with Spark Mllib. Employ collaborative filtering or content-based filtering algorithms. Real-time or batch processing to provide recommendations. Integration with the web/app backend to serve recommendations to users.</td><td rowspan="2">Model Training and Inference: Using EC2 instances (p2.xlarge for GPU). Training: 10 hours/month. Cost: $0.90/hour * 10 = $9/month</td><td rowspan="2">$ 9.00</td><td rowspan="2">Apache Spark on Amazon EMR (Elastic MapReduce):

Cluster Configuration:
Master Node: Manages the Spark cluster.
Worker Nodes: Perform data processing and model training.
Instance Types: Choose instances based on processing requirements.
Master Node: m5.large (2 vCPUs. 8 GB RAM)
Worker Nodes: m5.xlarge (4 vCPUs. 16 GB RAM)</td><td rowspan="2">Cluster Configuration:

Master Node: m5.large instance: $0.096 per hour
Worker Nodes: m5.xlarge instances: $0.192 per hour
Assume 4 worker nodes for sufficient processing power.
EMR Costs:

Master Node: $0.096/hour
Monthly cost: $0.096 * 24 * 30 = $69.12
Worker Nodes: 4 * $0.192/hour
Monthly cost: 4 * $0.192 * 24 * 30 = $552.96</td><td rowspan="2">$ 622.08</td><td rowspan="2">Storage for data and model artifacts. Assume 1 TB of storage.
Cost: $0.023 per GB-month
Monthly cost: 1024 GB * $0.023 = $23.55</td><td rowspan="2">$ 23.55</td><td rowspan="2">Assume 1 TB of data transfer per month.
First 1 GB/month is free.
Additional data transfer: $0.09 per GB
Monthly cost: (1024 - 1) * $0.09 = $92.07</td><td rowspan="2">$ 92.07</td><td rowspan="2">AWS CloudWatch:

Basic monitoring is free.
Additional charges for detailed monitoring and logs.
Assume $20/month for detailed monitoring and logs.</td><td rowspan="2">$ 20.00</td></tr>
<tr><td>Mobile</td></tr>
</table>

Project Cost and Benefit Analysis

Personal Recommended System

Project Cost and Benefit Analysis

Metric	Budgeted	Actual (Example)	Variance	Notes
Initial Investment	$(87,833.40)	$(85,000.00)	+$2,833.40	Cost savings in setup
Year 1 Cash Flow	$13,899.60	$14,500.00	+$600.40	Higher early adoption
Year 2 Cash Flow	$20,157.90	$19,800.00	-$357.90	Minor delays in scaling
Year 3 Cash Flow	$27,042.03	$28,200.00	+$1,157.97	Improved customer retention
Year 4 Cash Flow	$34,614.57	$33,000.00	-$1,614.57	Market competition
Year 5 Cash Flow	$42,944.37	$45,000.00	+$2,055.63	Stronger-than-expected growth
Total Cash Flows	$138,658.47	$140,500.00	+$1,841.53	
NPV (10% Discount)	$12,086.39	$14,200.00	+$2,113.61	Higher actual cash flows
IRR	14.36%	15.10%	+0.74%	Outperformed expectations
LTV:CAC Ratio	70,071:1	75,000:1	4,929	Lower CAC achieved
ROI	**15.8%**	**18.2%**	**+2.4%**	Increased profitability

Recommendations:

- Reinvest Year 5 surplus ($2,055) into customer acquisition.
- Monitor Year 4 trends for competitive response strategies.
- Use ROI metrics to secure additional funding or resources.

Key takeaways:

- Actual **NPV (14,200) >** ***Budgeted*** **(12,086)**, validating project ROI.
- **Budgeted ROI (15.8%)** was exceeded by **actual ROI (18.2%)**, indicating higher-than-expected returns.
- **IRR (15.1%)** exceeds the 10% discount rate.
- Lower initial investment (85Kvs.85*Kvs*.87.8K) and **CAC reduction** drove LTV:CAC to **75,000:1**.
- Year 4 dip due to competition; mitigated by Year 5 surge.

Conclusion

Conclussion Part 1

- **Project Management**:
  - **Clear roadmap** and **defined milestones** ensured that project objectives were met on time and within scope.
  - **Resource allocation** and **risk management** enhanced decision-making, leading to optimized resource use.
- **Agile Scrum**:
  - **Iterative development** allowed for continuous improvements and faster delivery of working features.
  - **Customer feedback loops** were integrated, ensuring that user-centric features were prioritized, boosting system adoption.
- **Lean Six Sigma**:
  - **Process optimization** through Lean Six Sigma reduced waste and increased efficiency, particularly in data processing and model training phases.
  - **Quality improvements** minimized errors and bottlenecks, leading to smoother deployment and enhanced system performance.
- **Budgeted ROI (15.8%)** was exceeded by **actual ROI (18.2%)**, indicating higher-than-expected returns.

Conclussion Part 2

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.

Points learn from Netflix’s Story (Credit)

Points from Netflix's Story

Business Case:

- In 2022, after a decade of remarkable growth, Netflix seemed to have reached a plateau.
  - **Competition** from new streaming services was intensifying, particularly in the US.
  - **Geopolitical conflicts** urged **the company's exit from certain regions**, where it had an expanding customer base.
  - **Rising inflation** was making **users more price-sensitive**, restricting Netflix's capacity to raise its subscription fees.

Problem Statements:

- **Netflix went from adding over 165,000 new customers daily** in the first quarter of 2020, to reaching a plateau in 2021, when **out of 60 million subscribers, approximately 30 million more people were using shared accounts**.
- In the first half of 2022, **Netflix experienced the worst six-month period in its history**, losing customers and prompting the company to **fire hundreds of people** and **scale back its programming**.
- This downturn led to a significant drop in its share price, **wiping out approximately $200 billion in market value**.

Points from Netflix's Story

Root Cause Analysis:

- Some users shared their accounts with partners or children they lived with, which was generally considered acceptable.
- Others shared with friends or relatives in different locations, a more problematic and common scenario.
- Additionally, there were instances of individuals sharing passwords with dozens of people, often reselling accounts to those unwilling or unable to pay through traditional means.

- This was Greg Peters approach, when the numbers **revealed untapped potential in how many subscribers would share their accounts - approximately 30 million more people were using shared accounts**..

Experimenting and Testing Possible Solutions:

- The company's management initiated two measures:
  - **Blocking Password Sharing:** aiming to curb the loss of potential revenue by ensuring that only paying subscribers could access Netflix's content.
  - **Introducing an Ad-Supported Version:** a new subscription tier designed to attract cost-sensitive customers who might be willing to endure advertisements in exchange for a lower subscription fee.
- Netflix developed a model to identify users who are traveling and differentiate them from those using someone else's password - **Blocking Password Sharing – The Experiment**

Points from Netflix's Story

- After identifying account sharers, **the next step was to decide how to make these users pay**.
  - On one side there was the belief that **Netflix should charge by residence**, similar to cable TV. This belief came from Reed Hastings, Netflix co-founder.
  - This would require users to pay per home and get another account for different locations.
  - The argument opposing this view was that **the residence model contradicted a core principle of Netflix: the ability to take the service anywhere**.
  - The alternative was **an individual user model**, allowing customers to access Netflix wherever they went, **with an additional fee for adding new users to their accounts**.

The Experimentation Roll Out:

- In 2022, Netflix introduced the **user model in Chile, Costa Rica, and Peru**, while **the residence model was deployed in five other Latin American countries**.
- This region, with its high incidence of password sharing, **served as an ideal testing ground due to common language (Spanish) and similar payment challenges**, as many residents lacked bank accounts.
- The results were clear-cut: **the subscriber-centric model was more successful.**

The Outcomes:

- Last year, **Netflix added 30 million new subscribers**, and in the first quarter of this year, **another 9.3 million.**
- According to Netflix, its **ad-supported plan now hosts 40 million monthly active users worldwide**, a significant **increase from 23 million in January**.
- This model not only increased the number of subscribers but also reduced churn rates (the rate at which customers unsubscribe) and minimized negative feedback on social media platforms.

Points from Netflix's Story

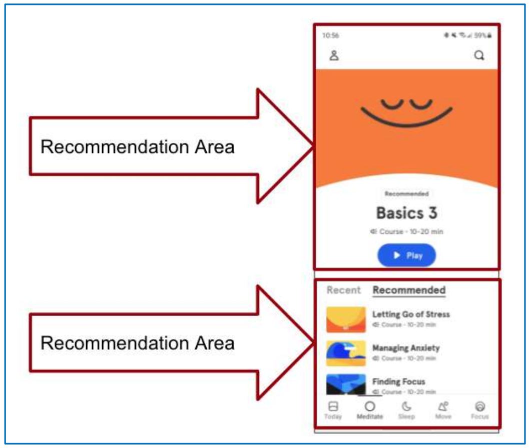
The subscriber-centric model's success can be attributed to several factors:

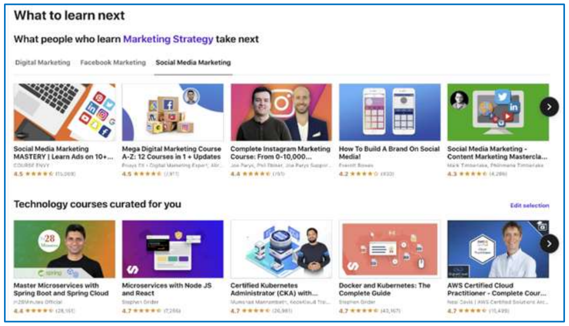
- **Flexibility**: users appreciated the ability to access Netflix from anywhere without being tied to a single household.
- **Cost-Effectiveness**: while there was an additional fee for adding new users, this was often seen as a better alternative than forcing each household member to have a separate subscription.
- **Reduced Friction**: the model minimised disruptions to the user experience, which could have led to dissatisfaction and cancellations.

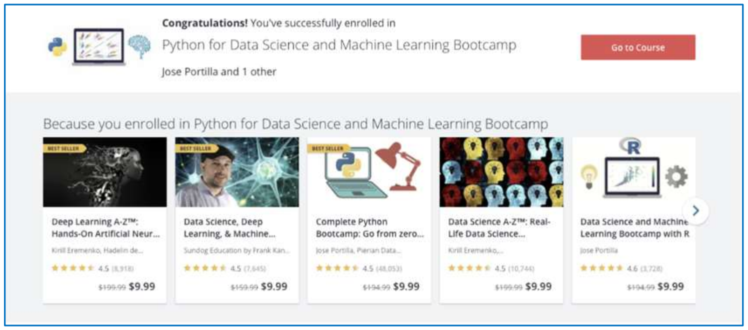
Project Related – AB Testing, Marketing Campaign and TV Ads and Process Capability (Credit)

Recommender System

How many click from the control group and the experimental group?

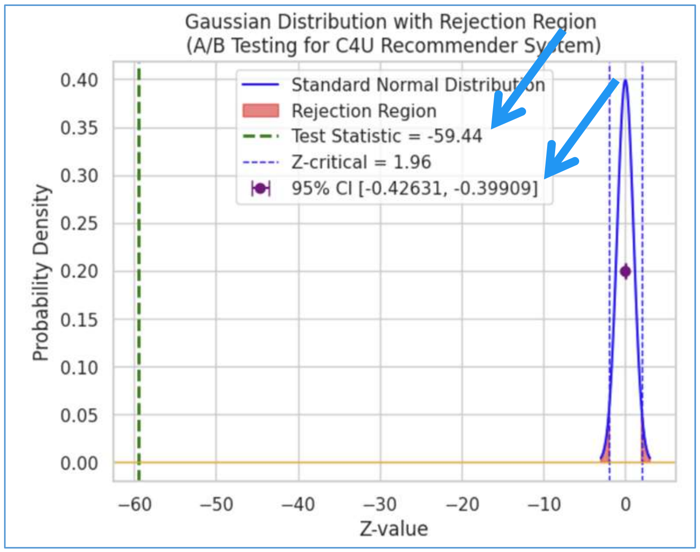






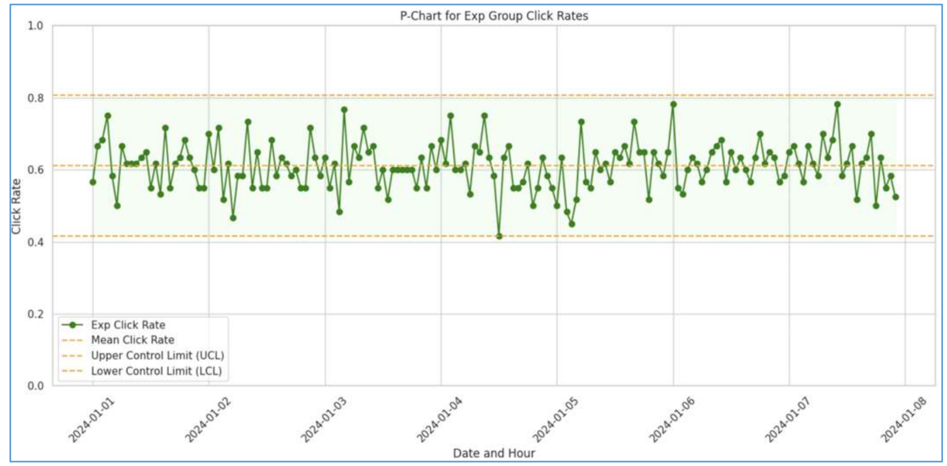
Recommender System

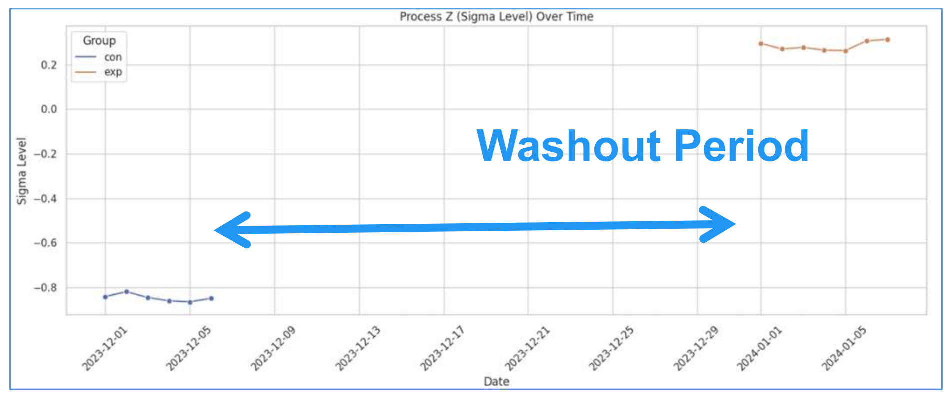
Probability



Z-Value

Key takeaways : Will applying Control Chart and Process Capability add value to the Experimentation





Washout Period

Con (Baseline)

Exp

Recommender System

How many converted with our ads vs product service announcement?

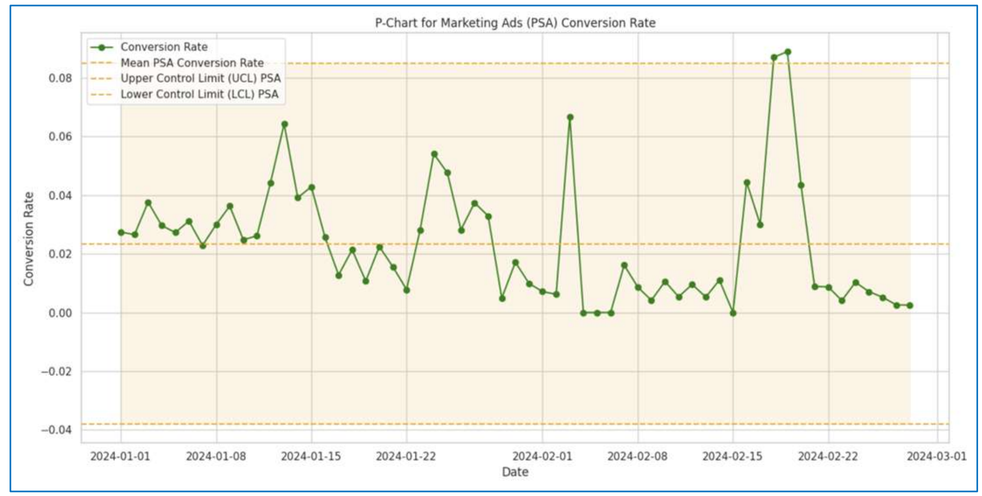
Ads

Products Service Announcement (PSA)

Plan

Ads and PSA

How many converted with our ads vs product service announcement?



Key takeaways : Should we compare to different Ads and determine which one is the most effective one?

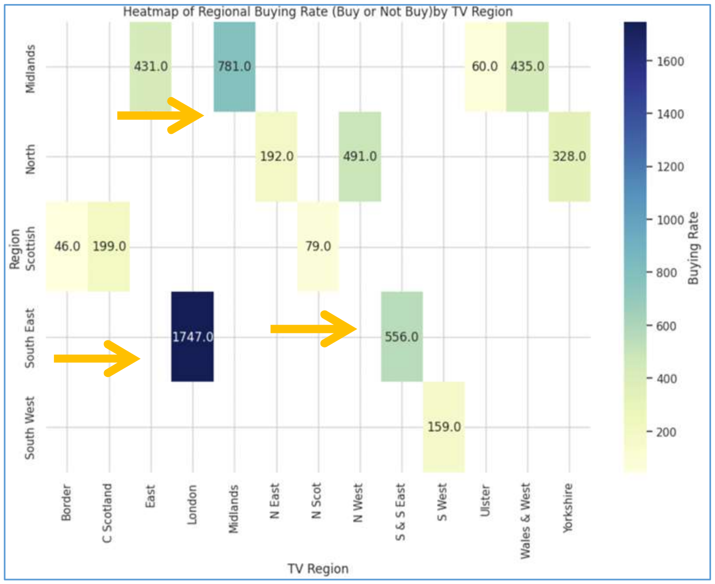
Plan

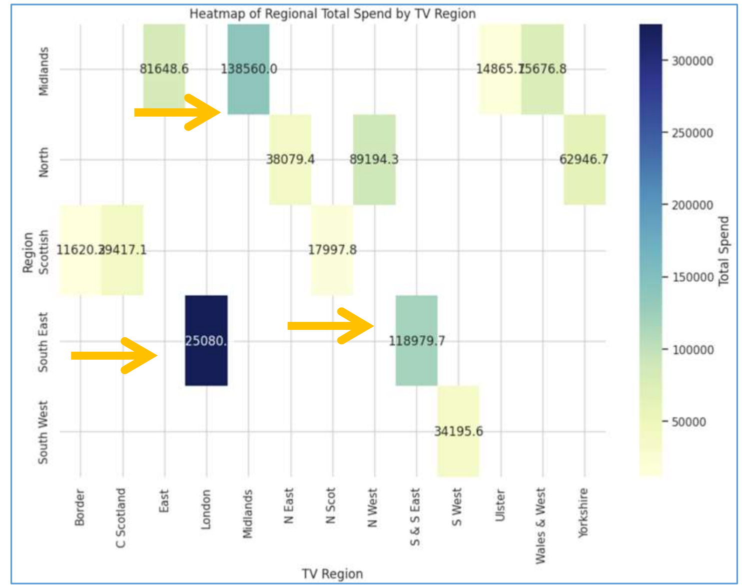
Recommender System

How many purchase with our tv campaign?

TV Campaign

Region and TV Region





- The highest buyers and total history spending are from London, South & South East, and Midlands.
- The middle range for both are from East, Wales & West, and Yorkshire.

Key takeaways : Is TV Ads still relevant? Which location is significant to our sales

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