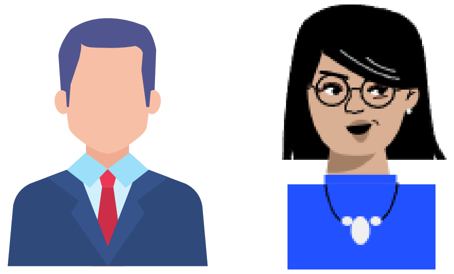
Search Smarter Not Harder:
A Personalized Intranet Recommender System

Capstone Team: Sara Darwish and Shay Kaur
McKinsey Mentor: Suzana Iacob
Faculty Advisor: Professor Alexandre Jacquillat

Developing a Personalized Intranet Recommendation System

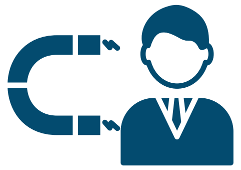


spend considerable time and content is the same for everyone

McKinsey Global Intranet

access to tools, information and expertise including firm benefits, learning portals, etc..

Personalized Recommendations



Increase user engagement

Offer discoverability to less popular intranet webpages

Project Overview

Preprocessing and Exploratory Data Analysis	Cleaned, merged and transformed the three data sources into user-webpages clicks matrix
Modeling	Created and deployed baseline; developed 5 candidate recommender system models
Model Choice and Evaluation	Chose final model and evaluated based on quantitative and qualitative metrics

Data Preprocessing, Matrix Formulation, and Data Limitations

USERS
28 features on employees
(role, location, tenure...)

+

WEBPAGES
focus on subset of well
maintained pages

+

CLICK EVENTS
9 months click analytics

Merged the 3 Databases

Binary User-Webpage Clicks Matrix

$$\begin{array}{cc} \text{40K Users} & \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 1 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \cdots & 1 \end{pmatrix} \\ & \text{448 Webpages} \end{array}$$

0 = User did not click *1 = User click any number of times*

DATA LIMITATIONS

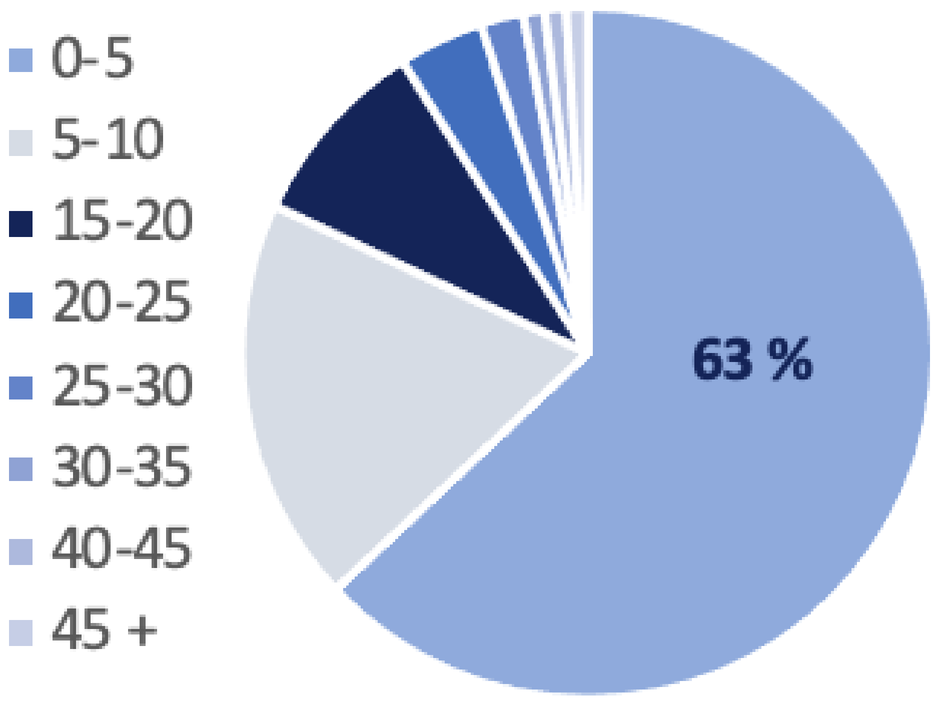
Matrix Sparsity
1.6% of 16M matrix
elements are non-zero

Implicit Feedback
Frequency of clicks doesn't imply
more usefulness

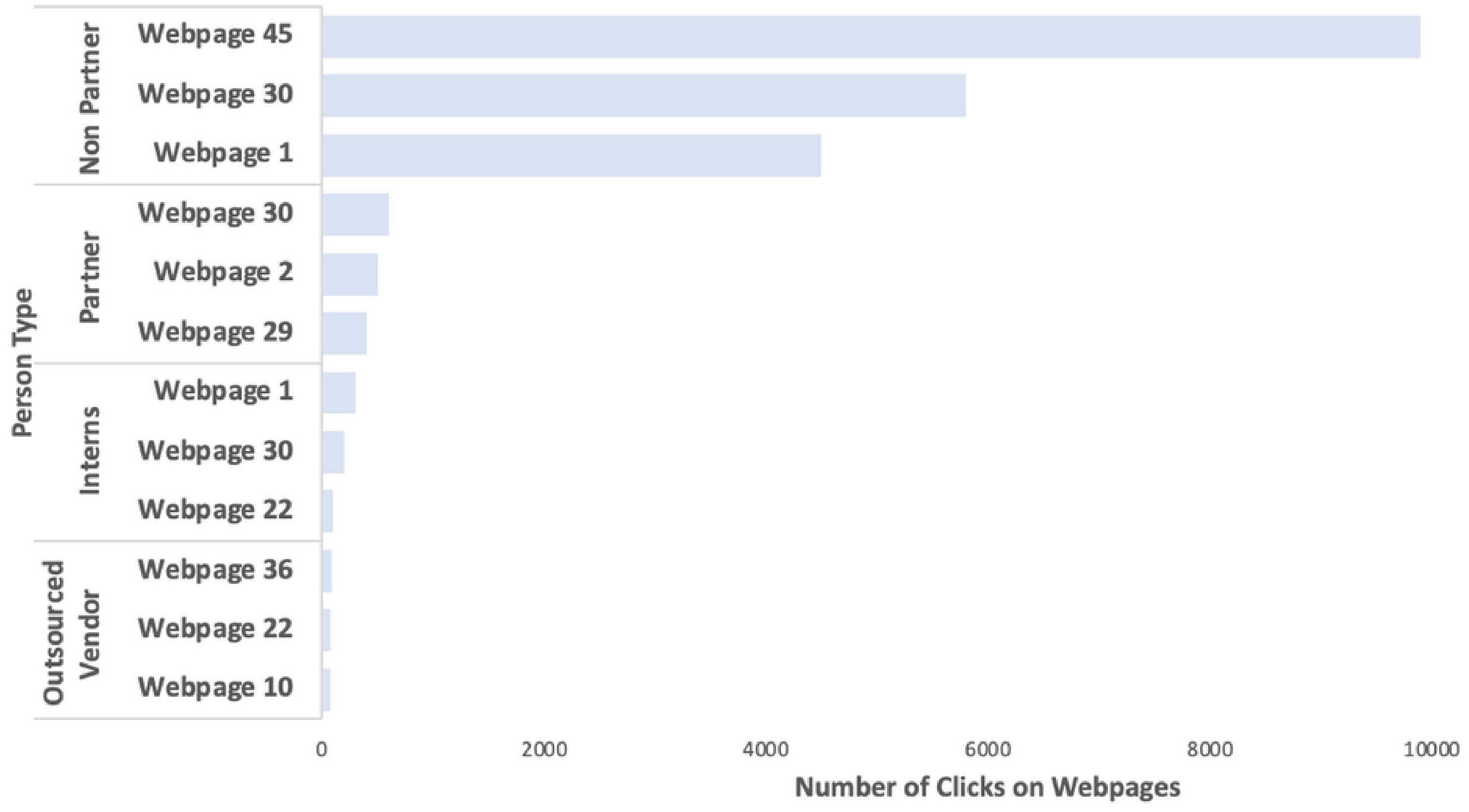
Not Visited (0) ≠ Not useful
(pages were not presented)

Exploratory Data Analysis

Users' Clicks Distribution



Visited Content Per Person Type



Low Activity

63% of users have a total of < 5 clicks

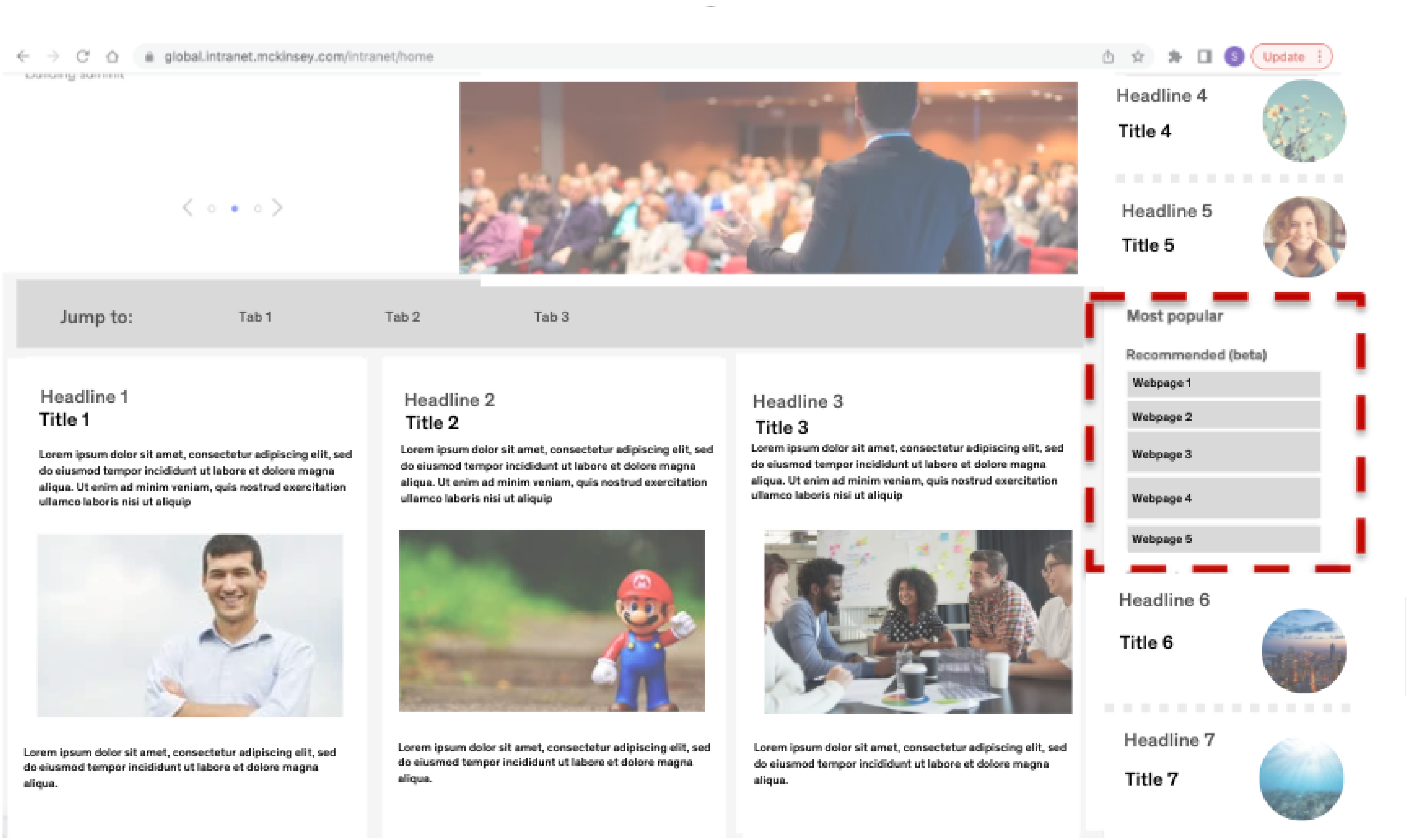
Motivated binary modeling

Assured presence of **signal**

Motivated **baseline** creation

Baseline Creation and Deployment

To act as an initial assessment point to measure the performance of our recommender system models, a non-machine learning baseline was created and deployed



5

Most Visited Pages

Per Person Type

Per Office Location

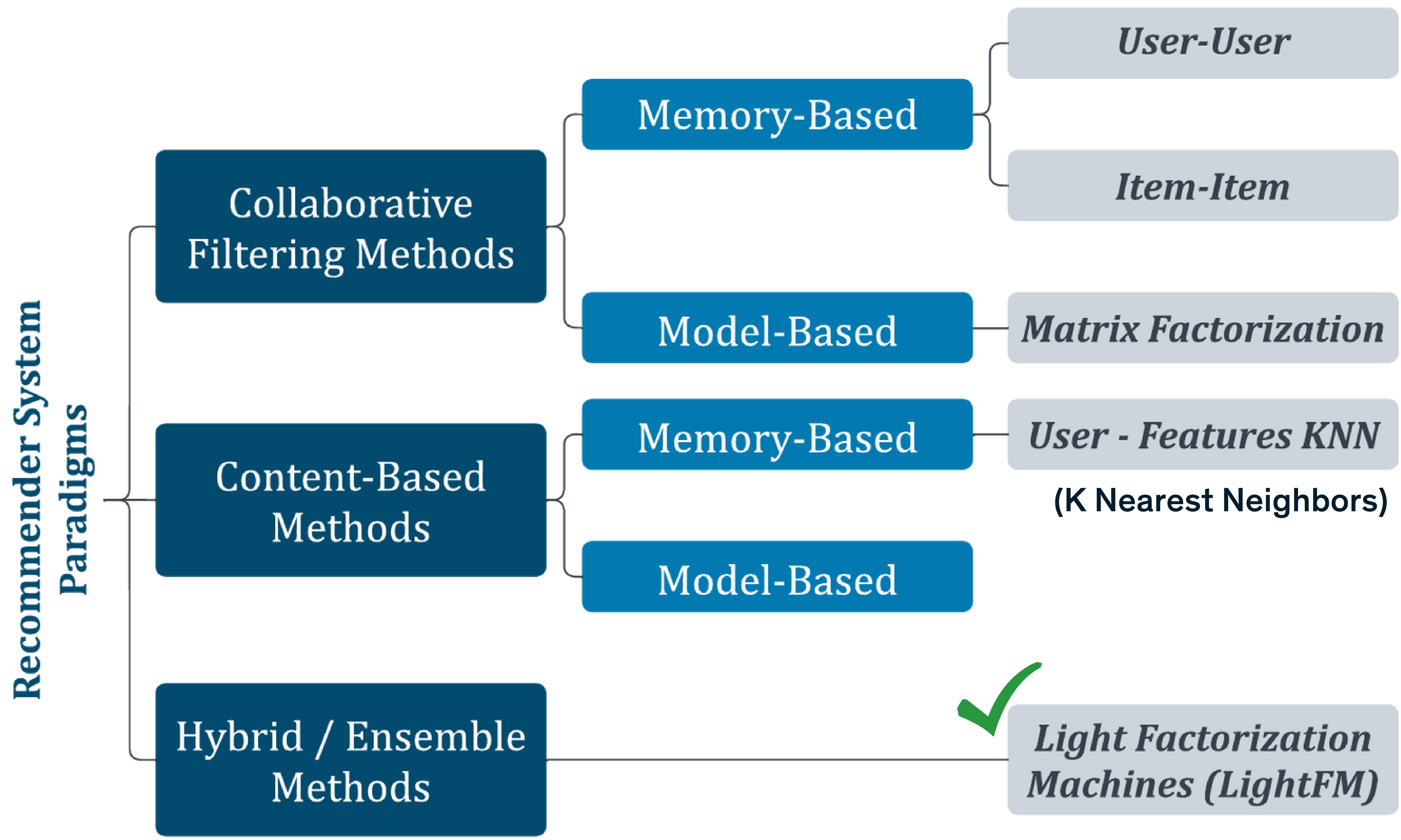
RECALL@K = 0.21

Baseline Productionalized

5 K Clicks Per Week

Explored the Three Paradigms of Recommender Systems

Five Candidate Models



Recall@K - Main Evaluation Metric

Actual web pages that user X has seen:

Website 1
Website 2
Website 3
Website 4
Website 5
Website 6
Website 7
Website 8
Website 9
Website 10

Recall@K - Main Evaluation Metric

Actual web pages that user X has seen:

Website 1
Website 2
Website 3
Website 4
Website 5
Website 6
Website 7
Website 8
Website 9
Website 10

Our Model Output:

Website 1
Website 5
Website 6
Website 11
Website 7

Recall@K - Main Evaluation Metric

Actual web pages that user X has seen:

Website 1
Website 2
Website 3
Website 4
Website 5
Website 6
Website 7
Website 8
Website 9
Website 10

Our Model Output:

Website 1
Website 5
Website 6
Website 11
Website 7

Recall@K (True Positive Rate @K)

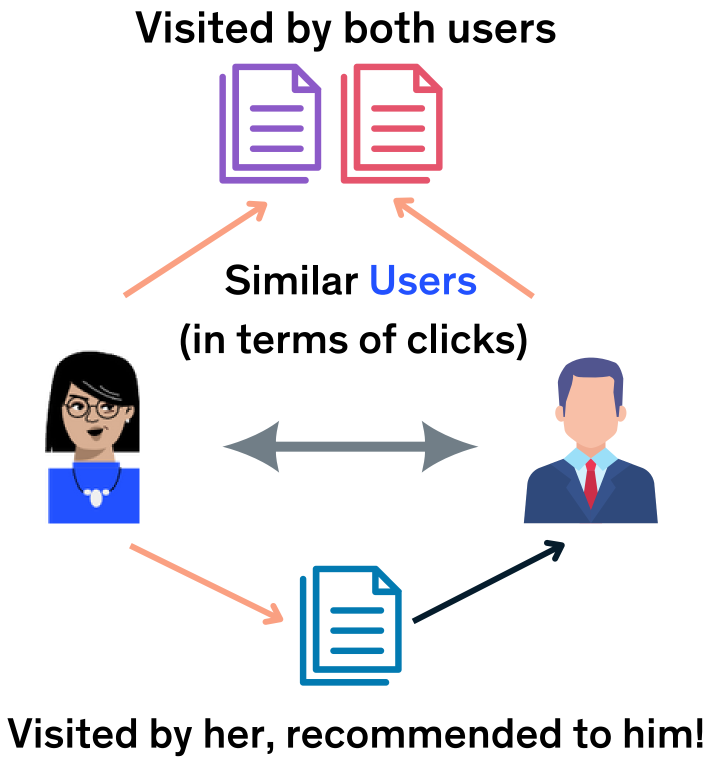
= out of the total # of webpages that the model gave how many has the user visited

$$\frac{4}{5} = 0.8$$

do that per user and get average

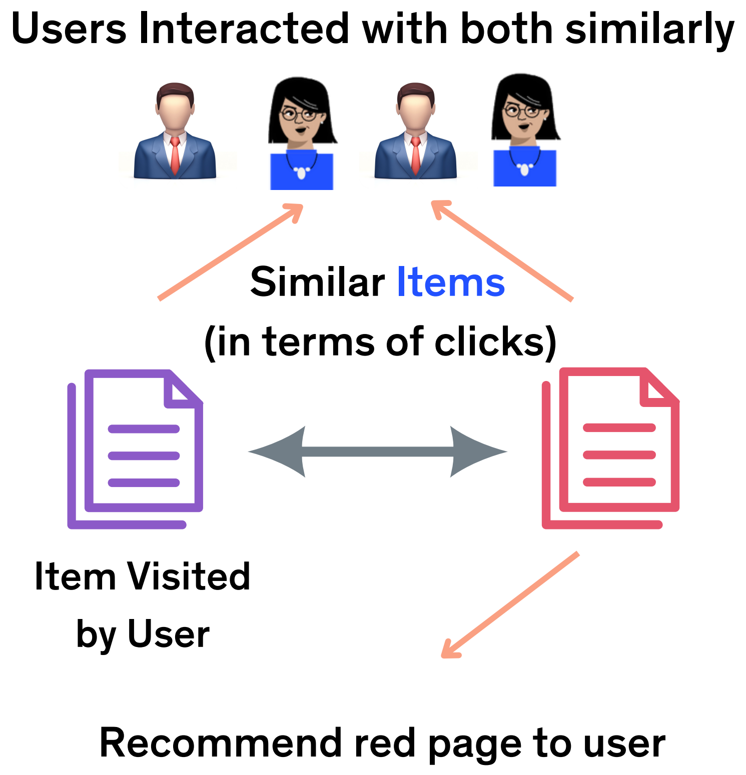
Modeling Approach - Model 1 - 3

1 User-User Collaborative Filtering



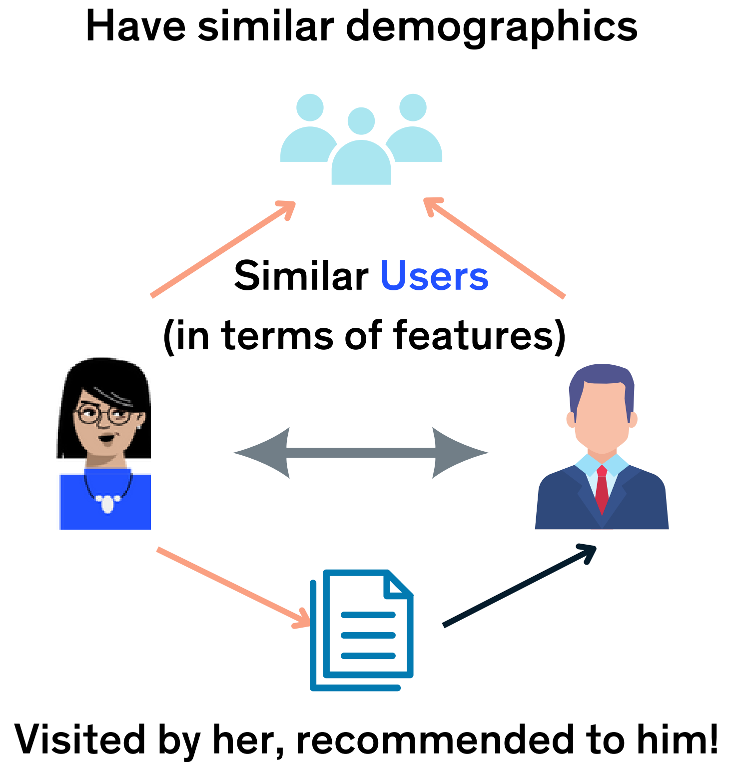
RECALL@K = 0.21

2 Item-Item Collaborative Filtering



RECALL@K = 0.12

3 User- Features KNN (K Nearest Neighbors)



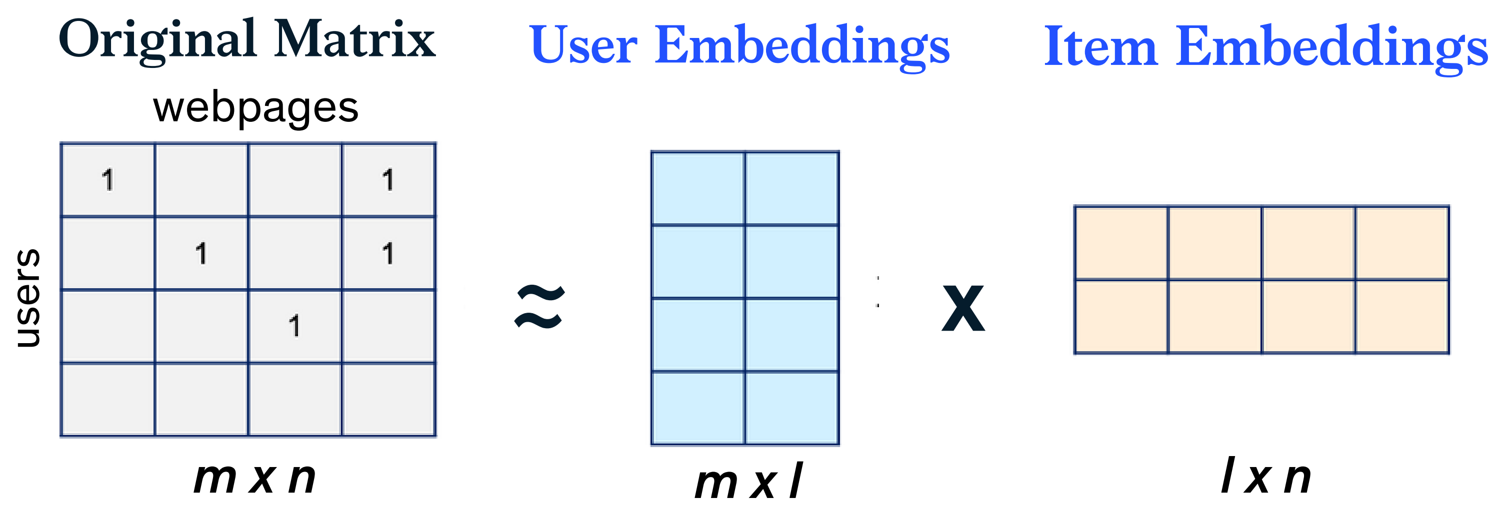
RECALL@K = 0.3

Models 1 - 3 are based on KNN and differ in the similarity metric used

Modeling Approach - Model 4 and 5

4

Matrix Factorization



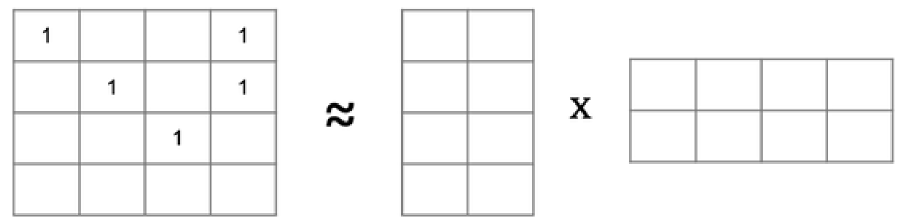
decomposing the sparse user-item binary matrix into a product of two lower dimensional ones representing the user and item embeddings

RECALL@K Test = 0.28

5

Light Factorization Machines (FM)

Collaborative Filtering (Matrix Factorization)



+

Content-based (user-features)

RECALL@K Test = 0.34

80% better than baseline

Deep-dive on Chosen LightFM Model [1]

- ✓ Leverages clicks + features
- ✓ Highest Recall@K
- ✓ Ensemble nature deals well with sparsity and implicit feedback
- ✓ Tackles cold start for new and inactive users

STEP 1: Incorporating Features in Embeddings

	CSP	ESP	FSP
User 1	1	0	0
User 2	0	0	1
User 3	0	1	0

User Feature Matrix

×

	Consulting	Operations
CSP	0.9	0.1
ESP	0.2	0.8
FSP	0.1	0.9

User Features in Terms of User Latent Features

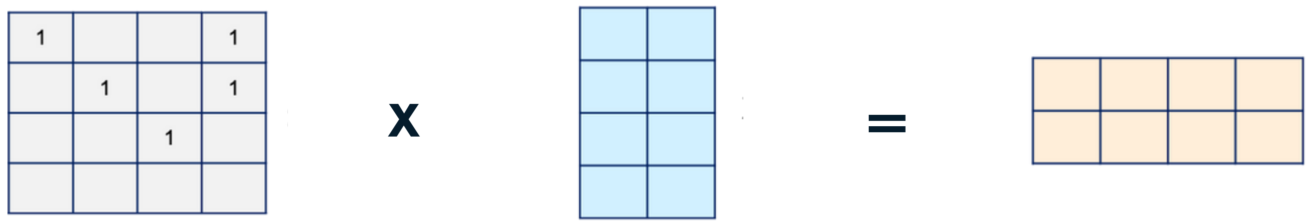
=

	Consulting	Operations
User 1	0.9	0.1
User 2	0.1	0.9
User 3	0.2	0.8

User Embeddings

Illustration on subset of user features - the same is done for item features

STEP 2: Matrix Factorization



User Embeddings Item Embeddings Predictions

[1] Kula, Maciej, Metadata Embeddings for User and Item Cold-start Recommendations, 07 2015.

Thank You!

Model 1 - User- User Collaborative Filtering

	McKinsey Translator	Rydoo	My Benefits (US)	Self Serve	Growth, Marketing
Jennifer	?	1	1	1	1
Suzana	1	1	0	1	1
Matt *most similar to Jennifer	0	1	1	1	1
Andrej	1	0	0	0	1

Model 1 - User- User Collaborative Filtering

	McKinsey Translator	Rydoo	My Benefits (US)	Self Serve	Growth, Marketing
Jennifer	**0.9**	1	1	1	1
Suzana	1	1	0	1	1
Matt	0	1	1	1	1
Andrej	1	0	0	0	1

Model 2 - Item-Item Collaborative Filtering

	McKinsey Translator	Rydoo	My Benefits (US)	Self Serve	Growth, Marketing
Jennifer	?	1	1	1	1
Suzana	1	1	0	1	1
Andrej *most similar to Jennifer	0	1	1	1	0
Matt	1	0	0	1	1

Model 2 - Item-Item Collaborative Filtering

	McKinsey Translator	Rydoo	My Benefits (US)	Self Serve	Growth, Marketing
Jennifer	**0.7**	1	1	1	1
Suzana	1	1	0	1	1
Andrej *most similar to Jennifer	0	1	1	1	0
Matt	1	0	0	1	1

Model 3 - User- Features KNN

	Person Type	Job Category Code	Department	Office	Country
Jennifer	Non Partner	FSP	T&D Internal Engagement	New York	United States
Suzana *most similar to Jennifer	Non Partner	FSP	T&D Internal Engagement	Waltham	United States
Andrej	Non Partner	FSP	T&D Internal Engagement	Prague	Czech Republic
Matt	Partner	CSP	Consulting	Cairo	Egypt

McKinsey Translator
?
1
0
1

Model 3 - User- Features KNN

	Person Type	Job Category Code	Department	Office	Country
Jennifer	Non Partner	FSP	T&D Internal Engagement	New York	United States
Suzana	Non Partner	FSP	T&D Internal Engagement	Waltham	United States
Andrej	Non Partner	FSP	T&D Internal Engagement	Prague	Czech Republic
Matt	Partner	CSP	Consulting	Cairo	Egypt

McKinsey Translator
0.8
1
0
1

KNN Model Calculation

Nearest Neighbors	1	2	3	4	5
Value for Webpage 1 (x)	0	0	1	1	0
Cosine Similarity (y)	0.2	0.4	0.8	0.4	0.2
x * y	0	0	0.8	0.4	0

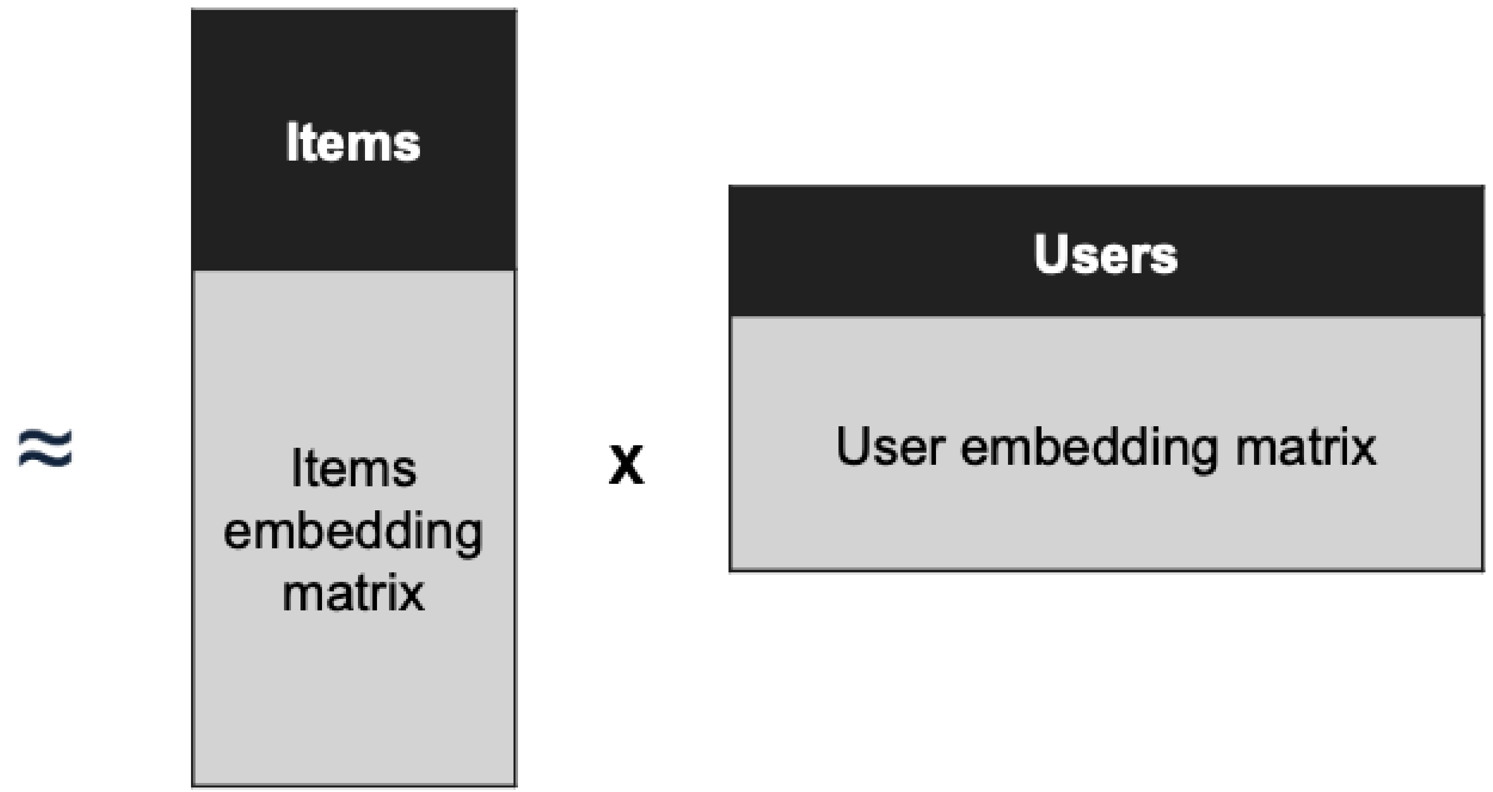
→ Take mean of this

0+0+0.8+0.4+0 / 5 = 0.24 → **Prediction for User 1 Webpage 1 Interaction**

Deep dive on Matrix Factorization

- Quick Recap: Model 2 &3 predicts based on interaction of users and items independently and matrix factorization does this concurrently
- The user x webpage matrix approximated by a combination of two matrices of lower dimension
- The preferences of a user and item can be represented by a small number of hidden factors --> embeddings

	Concur	Rydoo	GHD	...
Suzana	1	1	1	0
Jennifer	1	1	0	0
Andrej	0	0	1	1
...	0	1	0	0



Deep dive on Matrix Factorization

Items

Users	Concur	Rydoo	GHD
Suzana	1	1	1
Jennifer	1	1	0
Andrej	0	1	1

User Embeddings

Users	*"Travelling & Expense"*	*"User Support"*
Suzana	0.6	0.4
Akshata	0.8	0.2
Andrej	0.5	0.5

- Say we have k hidden factors
- Then for each user those hidden factors represent characteristics about the user (e.g Suzana may have 60 % liking towards traveling and expense and 40% user support website)
- Similarly, the hidden factors for webpages may be how much the webpage - Concur relates to the category "Traveling and Expense"