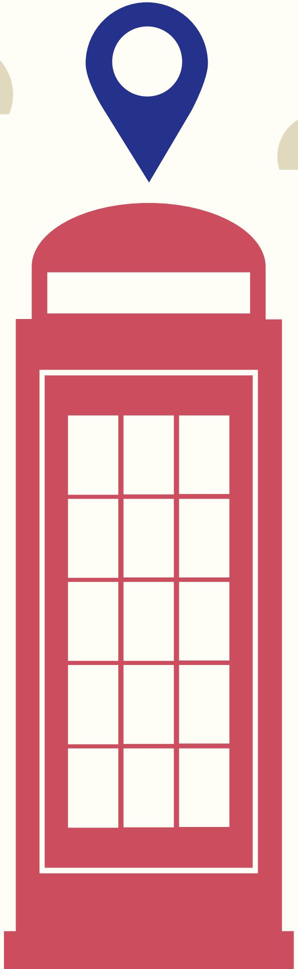
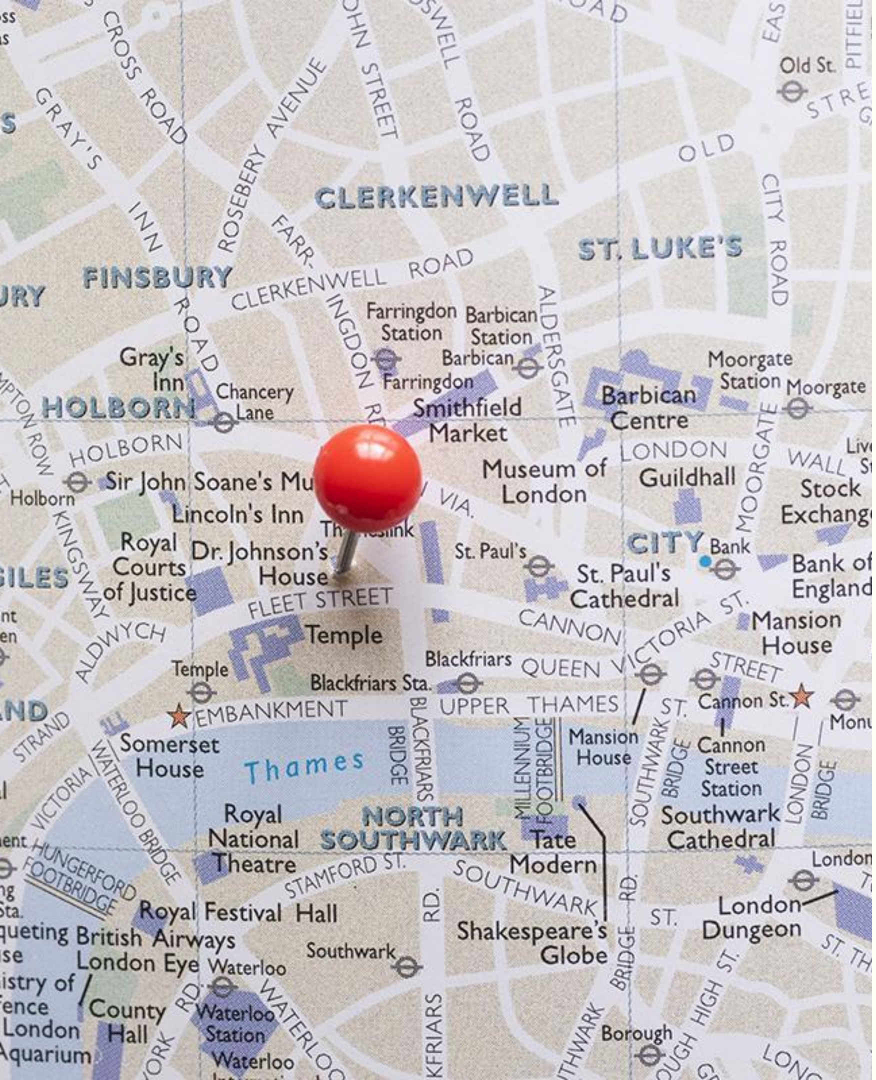


LONDON

TRAVEL GUIDE

Explore the hidden gems!



Problem Statement

- With so many places of attractions (and strong passport), Singapore travellers are **spoilt for choice**!
- How can we decide **which spots** to visit in our **limited time** (and vacation leave)?
- Some (or most) people just don't want to blindly follow the crowd but want to visit places **where the locals go**.

The Proposal

Travel Recommender

Do the heavy lifting of finding and recommending potential places of interest

Hidden Gems

Exclude the top 50 tourist attractions

Spill the tea!

Why London?

Top travel trends among Singapore travellers in 2023

Written by Arina Sofiah   Category: Mobility   Published: 13 January 2023

Top 10 popular destinations for travel in 2023

1. Bangkok, Thailand
2. Tokyo, Japan
3. Seoul, South Korea
4. Bali, Indonesia
5. Maldives
6. Hokkaido, Japan
7. Phuket, Thailand
8. London, United Kingdom → English - speaking
9. Paris, France
10. Johor, Malaysia

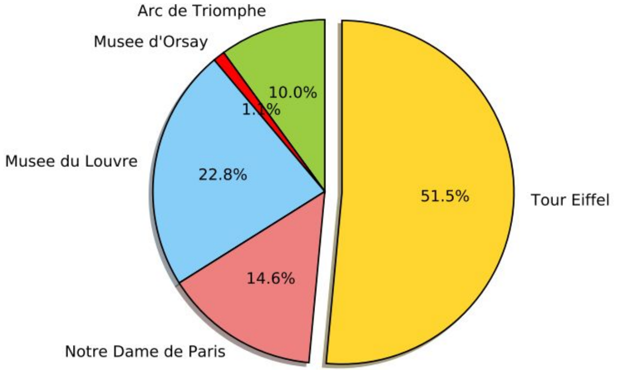
https://www.humanresourcesonline.net/top-travel-trends-among-singapore-travellers-in-2023

Sourcing of Data

#	Touristic Locations	Official Statistic Visitors	TripAdvisor Reviews
1	Notre Dame Cathedral	14,300,000	42,442
2	Musée du Louvre	9,134,000	58,648
3	Eiffel Tower	7,097,302	79,198
4	Musée d'Orsay	3,480,609	40,640
5	Arc de Triomphe	1,200,000	23,689

Table 1: Touristic locations in Paris: annual visitors reported by the official statistic, and number of reviewers in TripAdvisor.

Abstr
search
which
source
in tou
ever h
relatio
visitor
parati
locatic
that s
patter
media



https://ce

Fig. 3: Top tourist locations in Paris based on Instagram's posts.

- Compared official tourism statistics and TripAdvisor, vs Instagram, to find out popularity of locations
- There are differences in the ranking of touristic locations
- More similarity between TripAdvisor and Instagram - both are user-generated
- In conclusion, the study supports social media as a useful data source for touristic marketing and decision making since it can provide real-time insights of tourists' visiting patterns

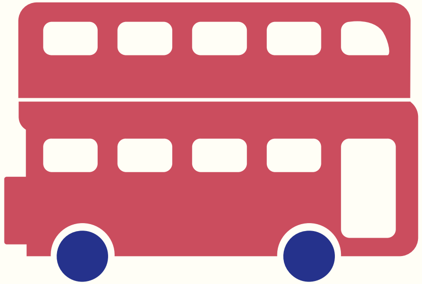
Instagram Datasets

Kaggle

2019 dataset
42M Posts 1.2M Locations
4.5M Profiles

Posts

- ★ User id
- ★ Location id
- ★ Description (caption)
  - ○ Feature engineer: rating



Travel Recommender Needs

- Location id
- Location name
- User id
- User rating

→ **Collaborative filtering model**

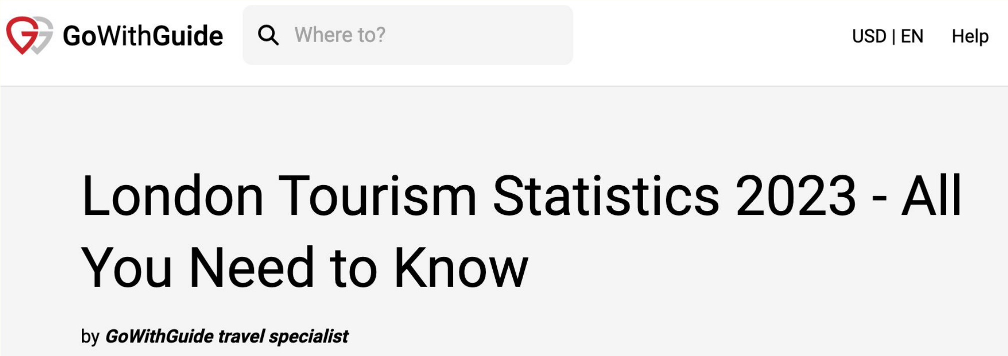
Locations

- ★ Location id
- ★ Location name

TripAdvisor Dataset

- Webscrape for Top 50 tourist spots for exclusion from recommender using BeautifulSoup

Why Top 50?

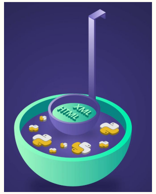


https://gowithguide.com/blog/london-tourism-statistics-2023-all-you-need-to-know-5213

- For tourists, the average length of stay in London equalled 4.6 days
- Assuming a tourist covers 5 spots for 5 days, 25 spots will be covered → recommender will not include an amount double of that

UNDERGROUND

Process



Datasets — Baseline Model — Model Tuning — Alternative Models — Deployment

Datasets - Cleaning & Preprocessing

instagram_posts:

Feature	Type	Description
sid	integer	sequence ID
sid_profile	integer	sequence ID of the profile
post_id	string	Instagram ID of post
profile_id	float	Instagram ID of profile
location_id	float	Instagram ID of location
cts	string	timestamp when post was created
post_type	integer	1 - photo, 2 - video, 3 - multi
description	string	caption of post
numbr_likes	float	number of likes at the moment it was visited
number_comments	float	number of comments at the moment it was visited

Null values

dropped

Datasets - Cleaning & Preprocessing

instagram_locations:

Feature	Type	Description
sid	integer	sequence ID
id	integer	Instagrams ID for that could be used on the website ex: ID=230466055 the url is https://www.instagram.com/explore/locations/230466055
name	string	name of location
street	string	street address
zip	string	zip code
city	string	name of city
region	string	name of region
cd	string	country code

Null values

dropped

City: London

cross check with country code 'GB'

City: London

standardise name

Datasets - Cleaning & Preprocessing

Finding hidden gems of London

1. boat tours and water sports
2. pubs and nightlife
3. sights and landmarks
4. spas and wellness
5. fun and games
6. museums
7. classes and workshops
8. nature and parks
9. markets
10. neighbourhoods

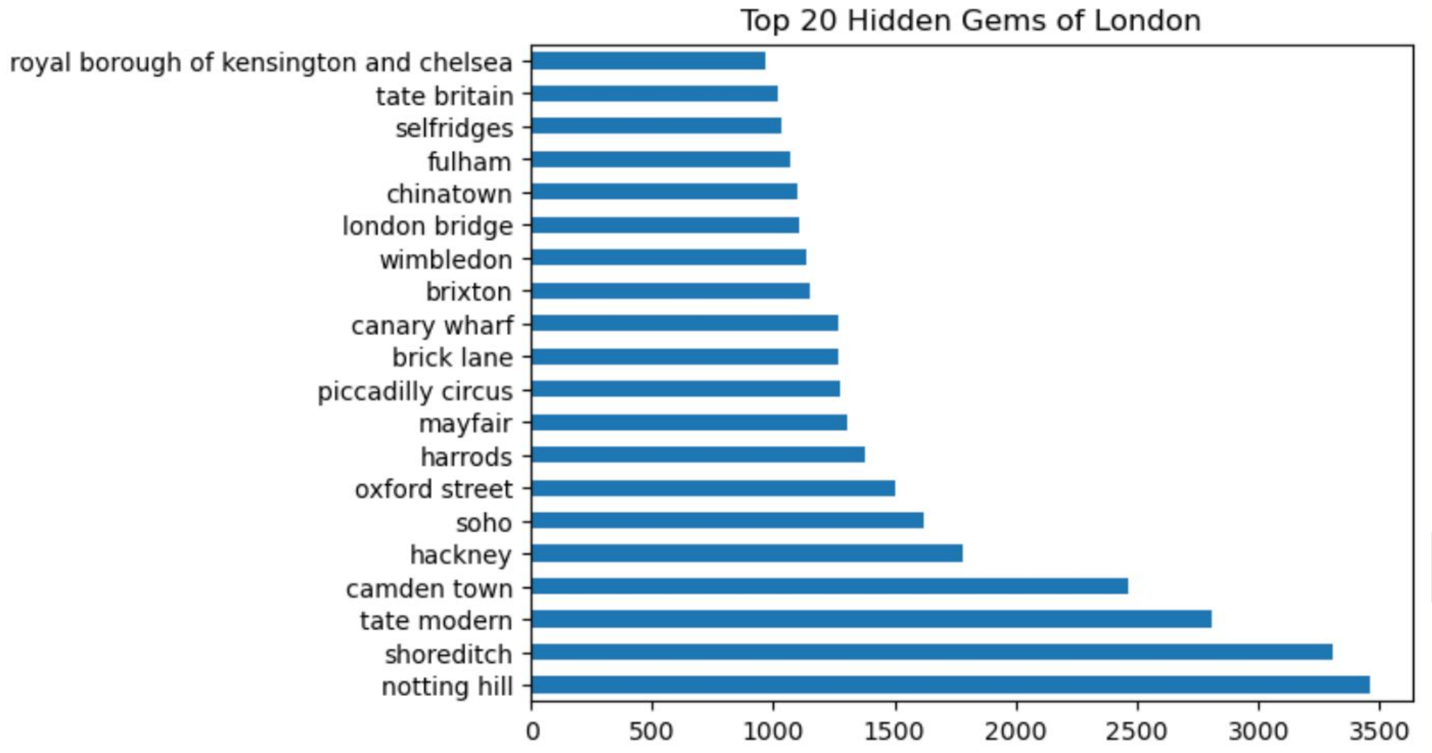
Non-related locations

dropped

Names of locations

standardise name

Hidden Gems



Top 20 Hidden Gems of London
royal borough of kensington and chelsea
tate britain
selfridges
fulham
chinatown
london bridge
wimbledon
brixton
canary wharf
brick lane
piccadilly circus
mayfair
harrods
oxford street
soho
hackney
camden town
tate modern
shoreditch
notting hill
0
500
1000
1500
2000
2500
3000
3500

Feature Engineering: User ratings

Package

VaderSentiment

- ★ Trained on social media data

Ratings

1 - negative
2 - neutral
3 - positive

Evaluation

81% accuracy against hand labelled ratings

Predictions

Accepted for modelling

Modelling

Package: scikit surprise
Algorithm: Matrix Factorisation

surpr!se

Surprise is a Python scikit for building and analyzing recommender systems that deal with explicit rating data.

Matrix Factorisation

Intuition: decomposition of a matrix into product of two or three matrices

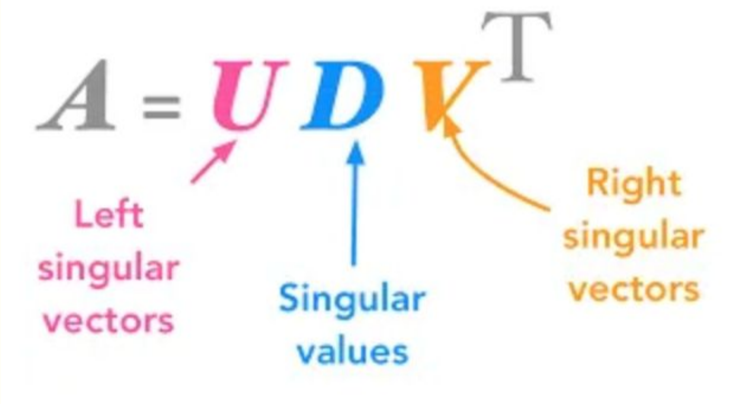
$$\underset{m \times n}{\overset{\hat{X}}{\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & \\ \vdots & \vdots & \ddots & \\ x_{m1} & & & x_{mn} \end{pmatrix}}} \approx \underset{m \times r}{\overset{U}{\begin{pmatrix} u_{11} & \cdots & u_{1r} \\ \vdots & \ddots & \\ u_{m1} & & u_{mr} \end{pmatrix}}} \underset{r \times r}{\overset{S}{\begin{pmatrix} s_{11} & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & s_{rr} \end{pmatrix}}} \underset{r \times n}{\overset{V^{\mathsf{T}}}{\begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \\ v_{r1} & & v_{rn} \end{pmatrix}}}$$

Matrix Decomposition/Factorization into three matrices (SVD)(figure — 1)

Intuition of SVD: matrix X (m,n) can be viewed as a dot product between two or three matrices with each matrix having dimensions of (m,r) and (r,n)

https://medium.com/analytics-vidhya/matrix-factorization-made-easy-recommender-systems-7e4f50504477

Matrix Factorisation (con't)

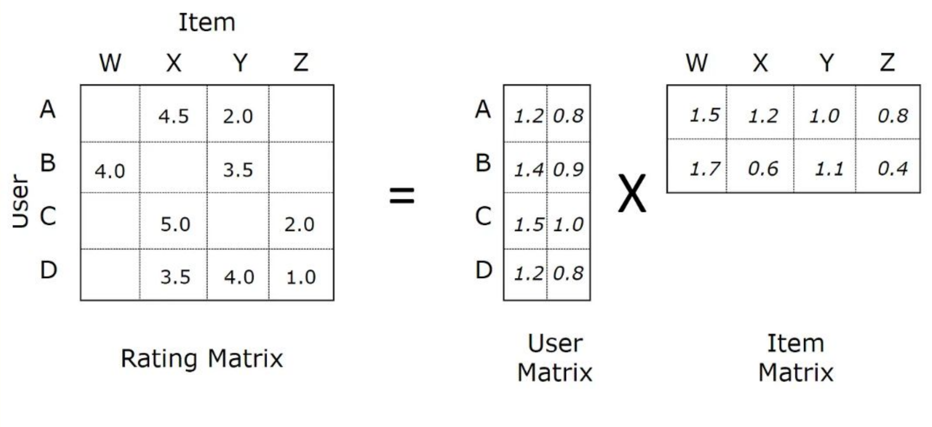


Singular Value Decomposition(figure — 2)

Intuition of SVD: these three matrices are factors of matrix A and if you multiply them, you'll get A

Matrix Factorisation (con't)

Matrix Factorization as Feature Engineering in Recommender Systems



User Item data set decomposed into User and Item Matrices(figure — 3)

After applying Matrix Factorization, we get two matrices, user matrix of shape (nxd) and item matrix of shape (dxm), which are the left and right singular matrices.

Matrix Factorisation (con't)

MF is a cutting edge technique which is hidden in other methods as well like, PCA(dimensionality reduction), clustering etc

$$\hat{r}_{ui} = q_i^T p_u.$$

(figure — 5)

p is the user matrix and ***q*** is the item matrix

$$\min_{q^*, p^*} \sum_{(u,i) \in \kappa} (r_{ui} - q_i^T p_u)^2 + \lambda(||q_i||^2 + ||p_u||^2)$$

objective function(figure — 6)

Goal: Find matrices ***q*** and ***p*** by minimising the objective function wrt ***q*** and ***p***
Method: Gradient descent

Note: *the first half of the equation is nothing but* ***Squared loss*** *and second half is the* ***L2 Regularization***

Modelling & Performance Tracking

mlflow

	train mae	test mae
Baseline: SVD	0.4433	0.5286
SVD GridSearch	0.4061	0.5193
NMF	0.1041	0.5244
NMF GridSearch	0.2117	0.5050
MF NN	0.5108	0.5492

Considerations:

- Metrics
- Overfit

Metrics

In the context of recommendation systems we are interested in recommending top-N items to the user

k = 100, threshold = 2.5

1

Precision@k

the proportion of recommended items in the top-k set that are relevant

$$\text{Precision@k} = \frac{|\{\text{Recommended items that are relevant}\}|}{|\{\text{Recommended items}\}|} = 0.8032$$

2

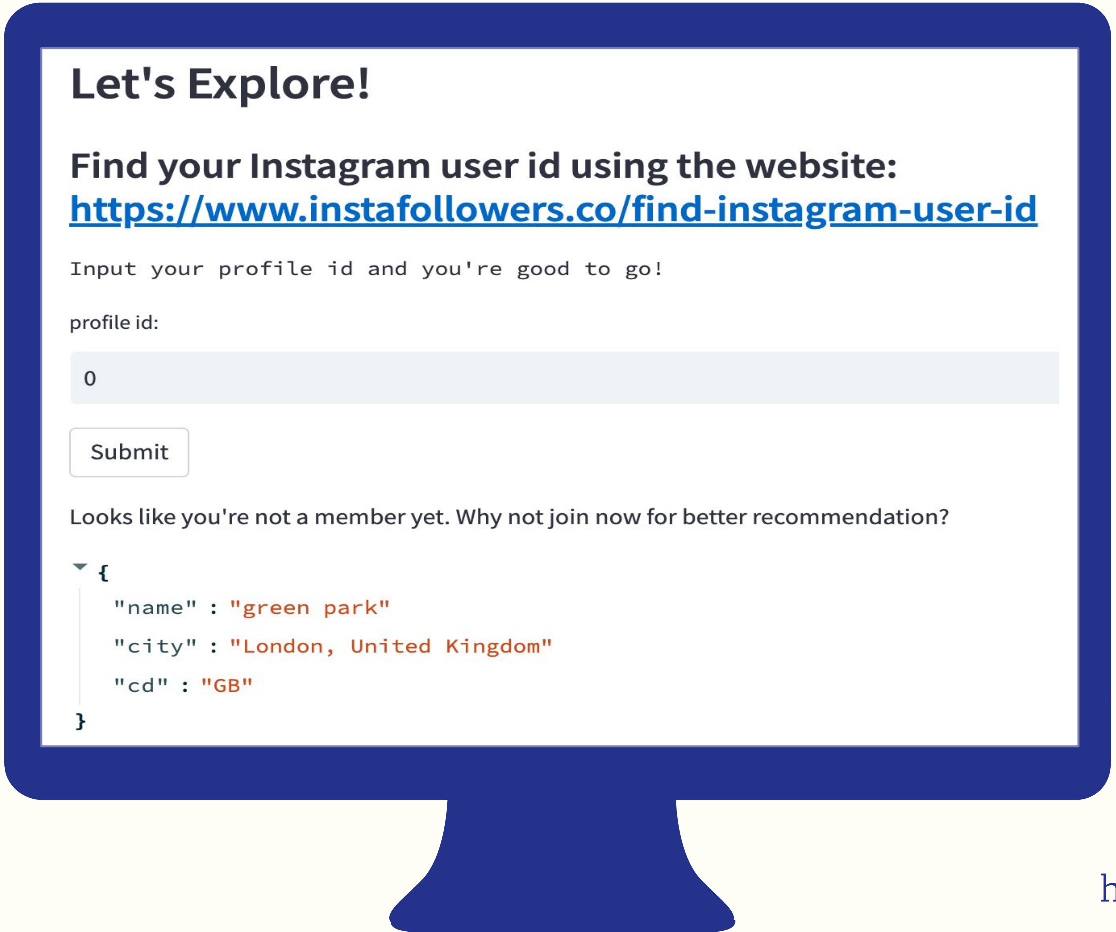
Recall@k

the proportion of relevant items found in the top-k recommendations

$$\text{Recall@k} = \frac{|\{\text{Recommended items that are relevant}\}|}{|\{\text{Relevant items}\}|} = 0.7748$$

https://medium.com/@m_n_malaeb/recall-and-precision-at-k-for-recommender-systems-618483226c54

Streamlit Deployment



Included top hidden gems as default for new profiles
→ no cold start problem

https://london-recsys.streamlit.app/

Conclusions

✓ **Collaborative-filtering recommender system**

Matrix factorization algorithm (SVD)

✓ **Metrics**

Precision@k = 0.8031
recall@k = 0.7748
k=100, threshold=2.5 / 3

✓ **Streamlit deployment**

Shuffled recommendation

Limitations & Further Works

Locations

There are still locations which are not part of the intended ten categories of attractions present in the data

Personalisation

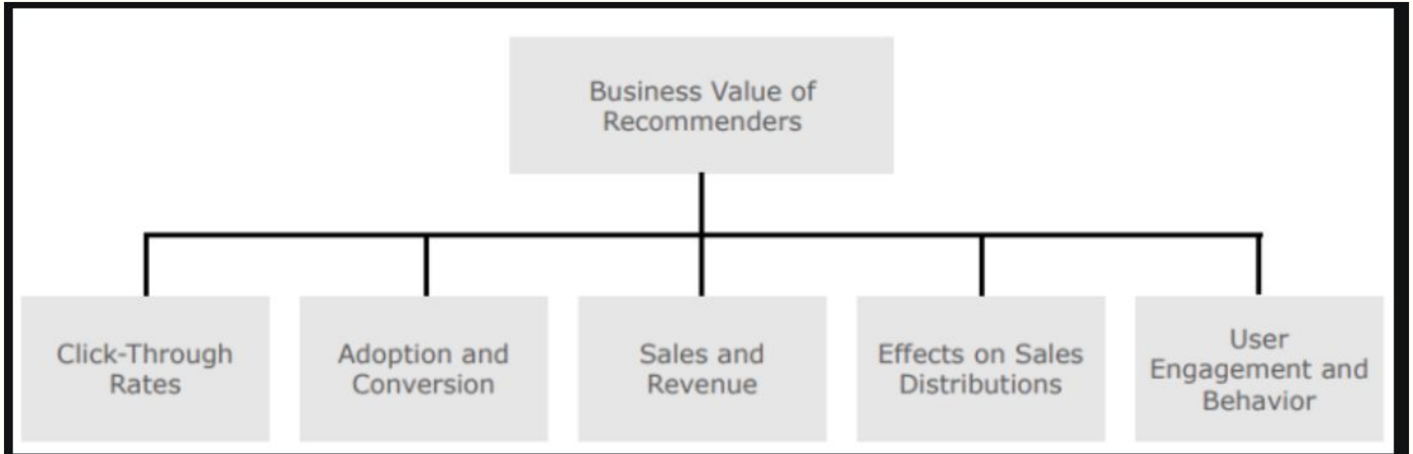
Recommendations for users whose Instagram ID is not part of the data are generic

- ➢ Zero shot classification was attempted in classifying locations based on the intended categories, but did not perform well
- ➢ Further works to explore at how classification can be done efficiently and accurately to create a **hybrid** recommender system

Metrics

Other metrics such as recommendation-centric metrics and business metrics have to be considered to determine how real customers react to the produced recommendations in terms of the company's business strategy through A/B testing.

Limitations & Further Works (Con't)



Business metrics for recommender systems | Source

https://neptune.ai/blog/recommender-systems-metrics

Thank YOU
& have fun in
London!