Amazon Recommendation Systems

Steve Shi, Bhadri Vaidhyanathan,
Vanshika Tibarewalla

Agenda

- Background
- Data Pipeline
- Data Profile
- Exploratory Data Analysis
- Machine Learning Models
- Results and Inference
- Challenges and Future Scope

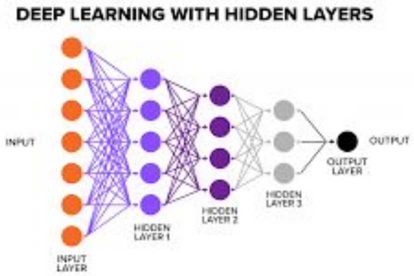
Background: Business Value

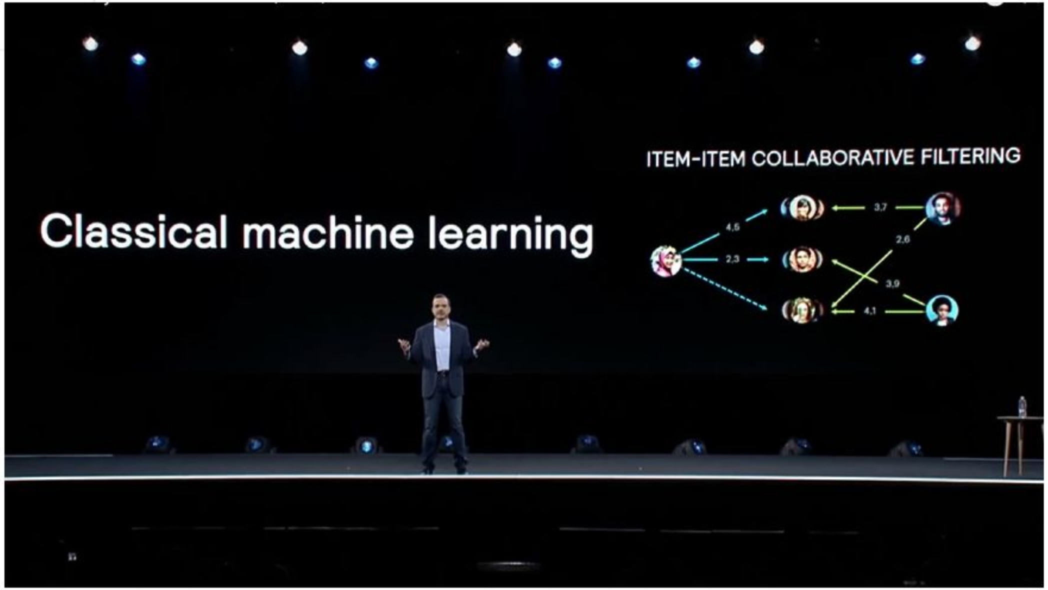
3rd largest company in the world by revenue

200mn customers

12mn products

2003: Amazon.com Recommendations: Item-to-Item Collaborative Filtering





Jeff Wilke, Amazon's consumer worldwide CEO, delivering a keynote presentation at re:MARS 2019

https://www.amazon.science/the-history-of-amazons-recommendation-algorithm

Data Profile

Basic User Rating Data

Used for:
- Memory-based CF
- Model-based CF
- Content-based

Attributes:
- asin (productID)
- reviewerID
- Ratings
- timestamp

Review Text Data

Used for:
- Sentimental Analysis

Attributes:
- asin (productID)
- reviewerID
- reviewText
- Summary

Magazine Description

Used for:
- Content-based models
- Model-based (FM)

Attributes:
- asin (productID)
- categories
- title
- description
- also_buy
- also_view

Data Pipeline

Data Collection

{JSON}

CSV

Data Analysis & Collaboration

GitHub

Packages

NLTK

TextBlob

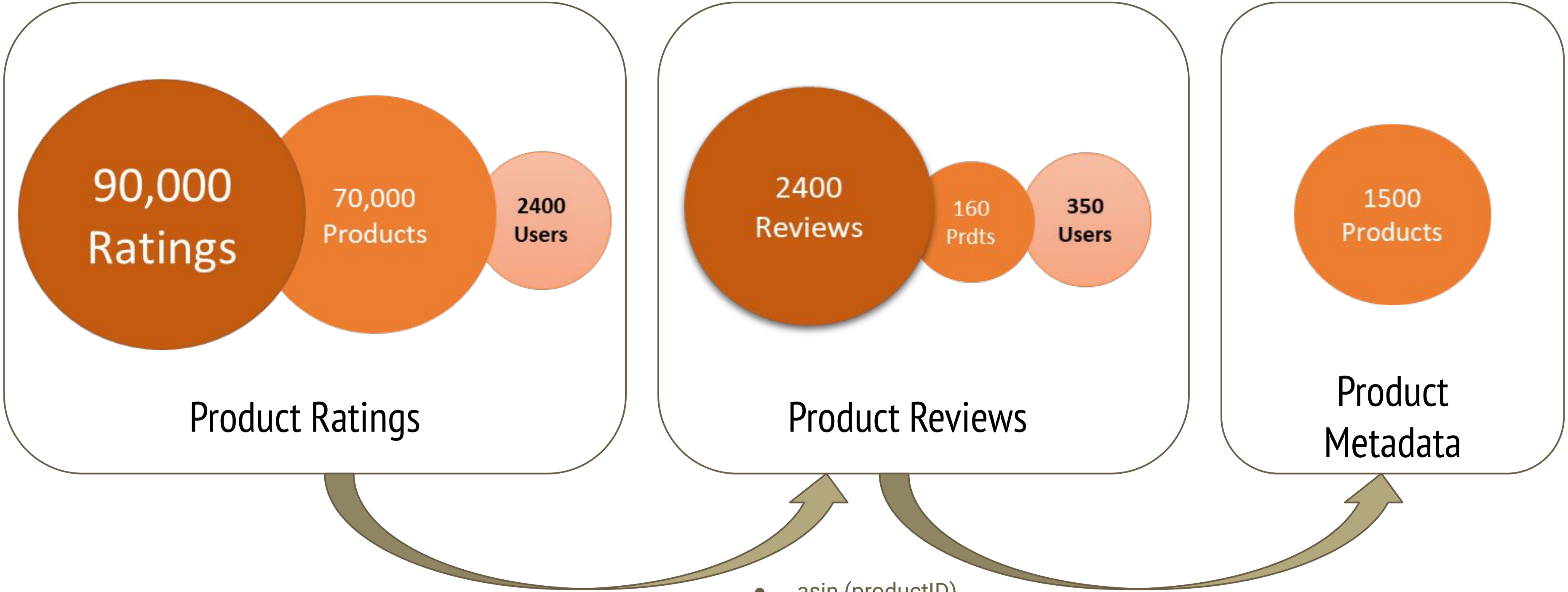
surprise

A Python scikit for recommender systems.

scikit learn

Exploratory Data Analysis (EDA)

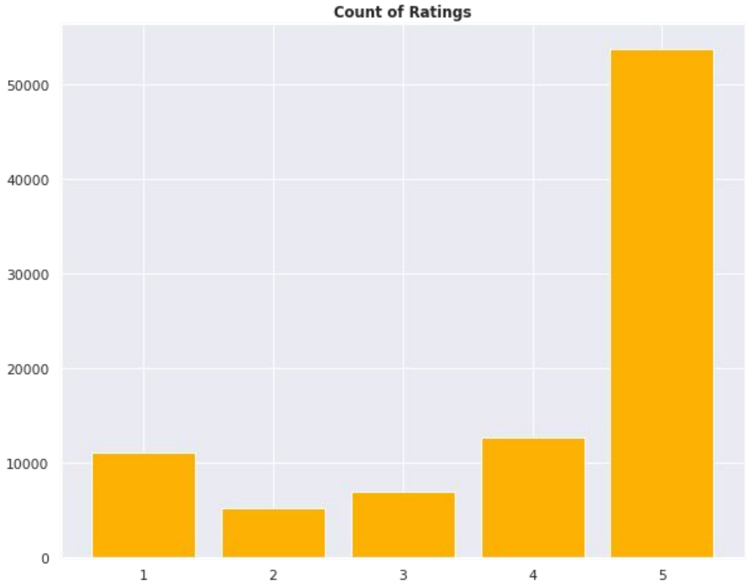
Unique Products and Reviewers



- asin (productID)
- reviewerID

Ratings

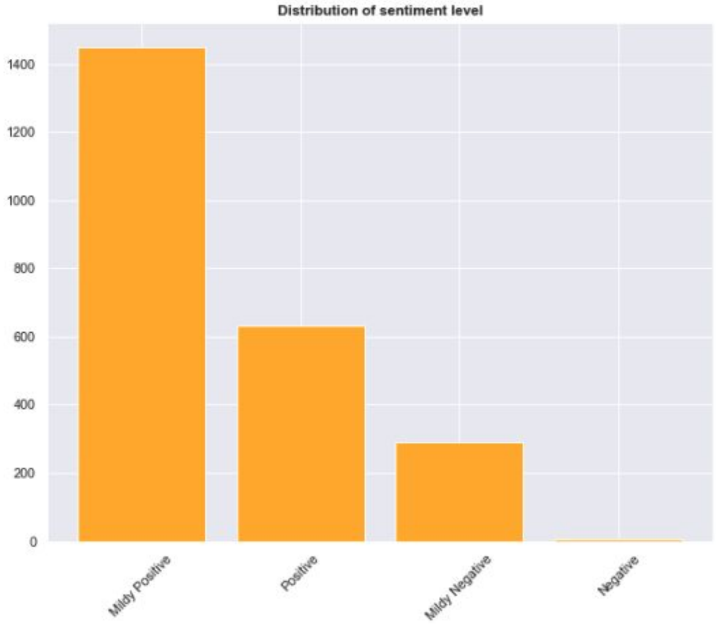
Distribution of ratings



Majority of the products have a rating of 5

Sentiment Level Distribution

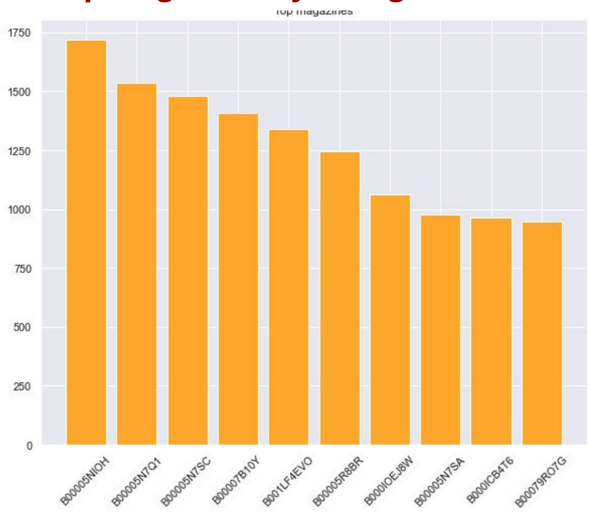
TextBlob

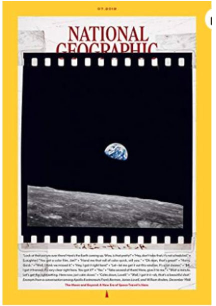


Most products mildly positive

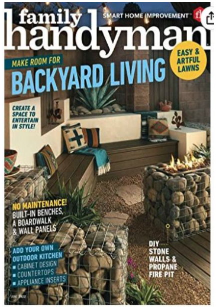
Top magazines/users

Top magazines by rating count





1

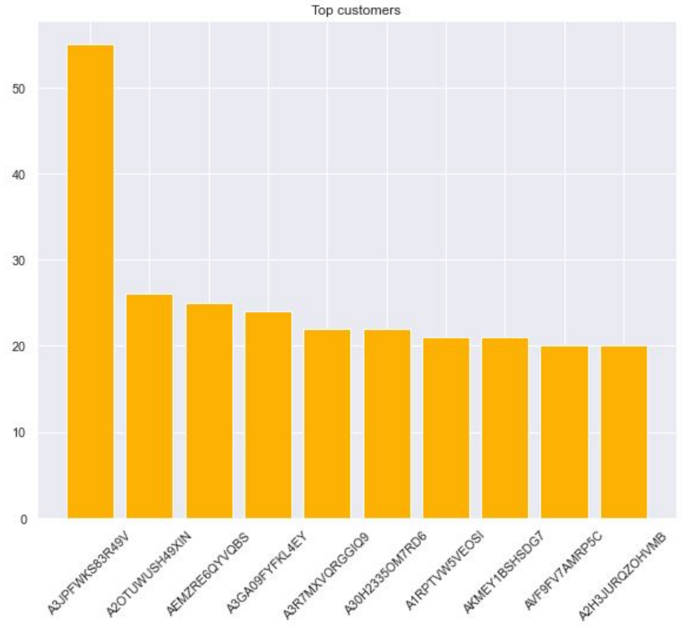


2

3

4

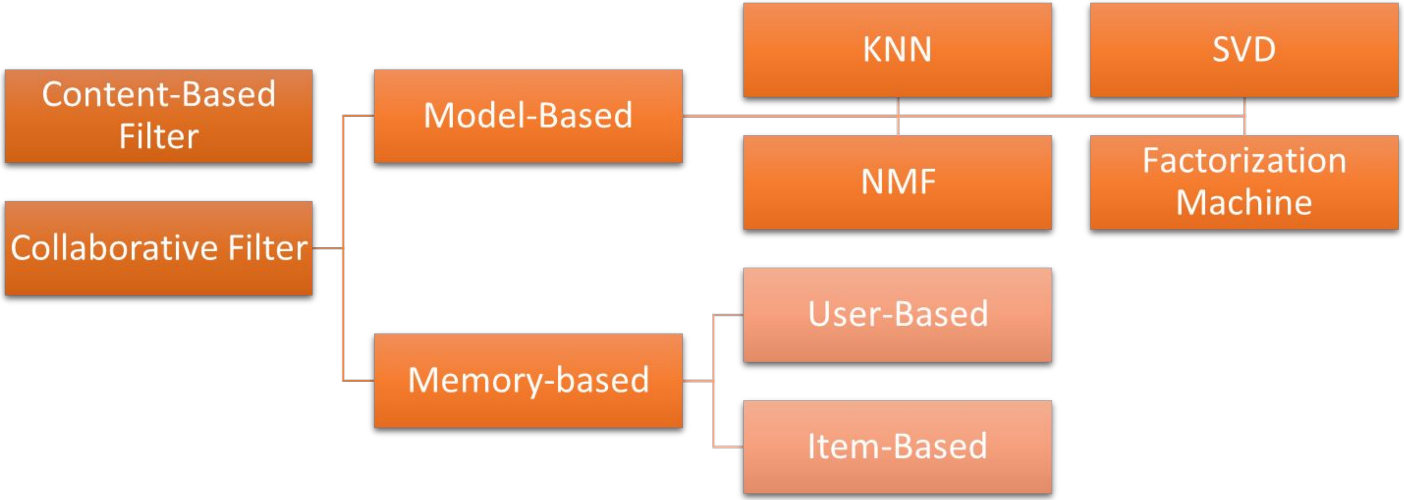
5



Top customers by rating count

Recommendation Systems

Model Approaches



KNN

Feature Engineering:
Active customers

Steps:

1. Count the number of unique product reviews for each unique reviewer
2. Average value found to be 20
3. Drop reviewers with count<20 : 10k reviewers

Top 10 reviewers Sorted by Count

	reviewerID	rating
48609	A3JPFWKS83R49V	55
32315	A2OTUWUSH49XIN	26
60746	AEMZRE6QYVQBS	25
46846	A3GA09FYFKL4EY	24
52524	A3R7MXVQRGGIQ9	22
38444	A30H2335OM7RD6	22
14817	A1RPTVW5VEOSI	21
64002	AKMEY1BSHSDG7	21
69735	AVF9FV7AMRP5C	20
28160	A2H3JURQZOHVMB	20

Steps:

- Used filtered data (users with >20 reviews) to create an item-user matrix
- Use Cosine similarity
- Nearest Neighbour Model
- K = 6

```
Recommendations for B0065MEDRI:

1: B006BFR2U4, with distance of 0.988167194060955:
2: B000INCK4I, with distance of 0.9934467397187514:
3: B000066HVN, with distance of 0.9940040150159131:
4: B00005N7VP, with distance of 0.9941629630396817:
5: B00005NIPE, with distance of 0.9959081886984714:
```

Problem:
Lack of features

Solution:
Use Sentiment Score from Review Summary

Collaborative Filtering - Item-Based

- Use cosine similarity
- Recommend 10 products

Problem:
item-item CF does not have a lot of intersection with Amazon's own 'Also-buy' data

10 recommendations to users who have buy B00005N7Q1

asin	
B00005N7Q1	1.000000
B00006LIR1	0.040265
B00005R8BL	0.036821
B0193CNAIY	0.028030
B00006K1BF	0.028030
B00005N7SA	0.024097
B00005N7SC	0.022225
B007FIR1Z2	0.019820
B007ZUWNA8	0.019072
B00007AZRH	0.018646
B000UMJODW	0.017431

Collaborative Filtering - User-Based

Problem:
Originally Dataset has more than 70000 users, it cost too much time to compute on local laptop

Filtering customers who purchased more than 3 item to reduce computing time

Find most similar user:
A2877WXAPQ7T50 is the most similar user to user A3JPFWKS83R49V

```
reviewerID
A3JPFWKS83R49V    1.000000
A2877WXAPQ7T50    0.304830
```

Run model and test:
- Recommended magazines based on similar user's purchase history
- user A3JPFWKS83R49V already purchased to all of them

```
recommendations to users A3JPFWKS83R49V

asin
B00005N7P0    5.0
B00005N7SC    5.0
B00005N7TL    5.0
B00005NIN8    5.0
B00005NIOC    5.0
```

Content-Based Model

Data Cleaning:

- Removing Null Value
- Removing replicate value
- Clean up description text

	asin	description	category	desc2	clean_desc
0	B00005N7NQ	[REASON is edited for people interested in economic, social, and international issues. Viewpoint...	[Magazine Subscriptions, Professional & Educational Journals, Professional & Trade, Humanities &...	REASON is edited for people interested in economic, social, and international issues. Viewpoint ...	reason is edited for people interested in economic social and international issues viewpoint str...
1	B00005N7OC	[Written by and for musicians. Covers a variety of musical styles and includes transcriptions fr...	[Magazine Subscriptions, Arts, Music & Photography, Music]	Written by and for musicians. Covers a variety of musical styles and includes transcriptions fro...	written by and for musicians covers a variety of musical styles and includes transcriptions from...
2	B00005N7OD	[Allure is the beauty expert. Every issue is full of celebrity tips and insider secrets from the...	[Magazine Subscriptions, Fashion & Style, Women]	Allure is the beauty expert. Every issue is full of celebrity tips and insider secrets from the ...	allure is the beauty expert every issue is full of celebrity tips and insider secrets from the p...

Feature Engineering:

- Keyword extraction with Rake
- Create vector representation
- Create the similarity matrix

key_words	bag_of_words
[reason, edited, people, interested, economic, social, international, issues, viewpoint, stresse...	professionaleducationaljournals professionaltrade humanitiessocialsciences economicseconomictheo...

Provide recommendation:

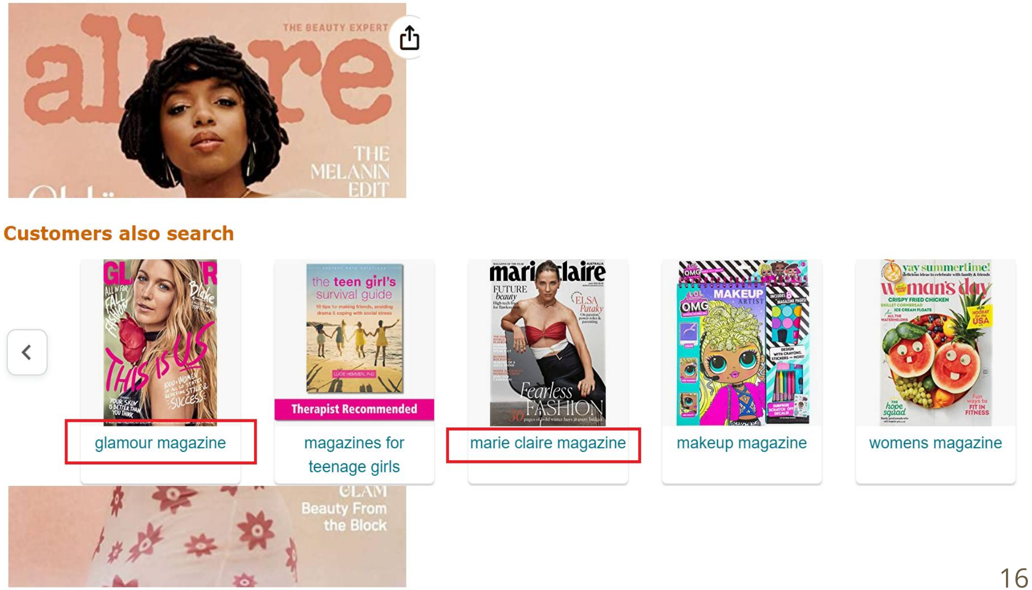
All recommended magazines are related women/fashion magazines

	asin	description	clean_categ
2	B00005N7OD	[Allure is the beauty expe...	[fashionstyle, women]
50	B00005N7QN	[Harper's BAZAAR, the fash...	[fashionstyle, women]
768	B00007AZEO	[Marie Claire Idees focuse...	[fashionstyle, women]
883	B00007M2OH	[Glamour UK is Britain s n...	[fashionstyle, international]
2159	B0007INI2C	[New Beauty is the first p...	[fashionstyle]

Content-Based Model : Validation

Our Recommendations VS Recommendations made in Amazon.com

	asin	description	clean_categ
2	B00005N7OD	[Allure is the beauty expe...	[fashionstyle, women]
50	B00005N7QN	[Harper's BAZAAR, the fash...	[fashionstyle, women]
768	B00007AZEO	[Marie Claire Idees focuse...	[fashionstyle, women]
883	B00007M2OH	[Glamour UK is Britain s n...	[fashionstyle, international]
2159	B0007INI2C	[New Beauty is the first p...	[fashionstyle]



Model Based - Matrix Factorization

NMF and SVD

surpr!se

A Python scikit for recommender systems.

Goal: Predict rating for products that user has not tried

Python Package - Surprise
Algorithms:

- Non-negative Matrix Factorization(NMF)
- Singular Value Decomposition (SVD)

Methodology:

- **KFold Cross Validation**

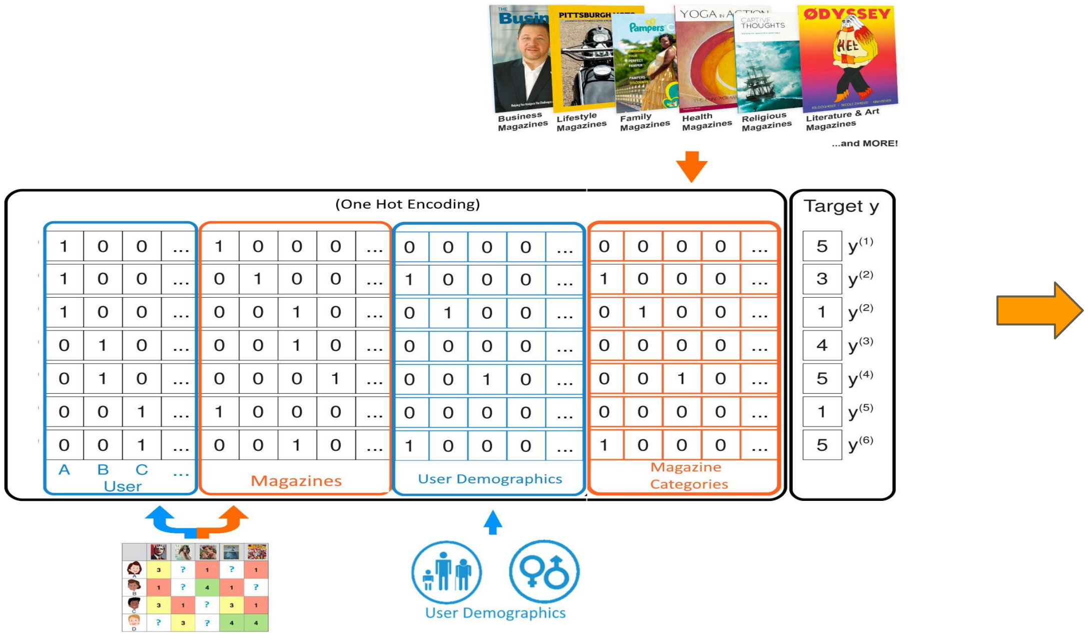
Parameter Modelling		Mean RMSE	
No. of rating/product or user	**Kfold**	**SVD**	**NMF**
All	5	1.37	1.39
>1	5	1.23	1.18
>1	10	1.23	1.16
>2	10	1.13	1.1

Model Error Output:
'User and/or item is unknown.'

Factorization Machines

Goal: Include more features and detect more latent factors

Data Augmentation: Age Group and Gender data sampled from MovieLens dataset



	Actual	Prediction	delta
0	3	2.601987	0.398013
1	2	3.914806	-1.914806
2	5	4.912905	0.087095
3	5	4.660848	0.339152
4	5	4.959595	0.040405
...	...	...	...

RMSE = 1.006

Challenges & Future Scope

Challenges:

- **Collaborative filtering:** Compute power not sufficient to run all recommendation smoothly.
- **Content-based model:** do not have expertise and time to design and assign attributes.
- **Model based Matrix Factorization(NMF and SVD):** Cold Start problem
- **Factorization Machine:** Heavy feature engineering, very large feature space

Future Scope:

- Using Sentiment and Emotion Analysis to run a classification model.
- Field Aware Factorization machines for inclusion of multiple latent factors

Thank you!

Questions?

Github

https://github.com/battery-code/evaluation_RecommendationSystems

References

- Data Source: Amazon *https://jmcauley.ucsd.edu/data/amazon/*
- https://www.kdnuggets.com/2017/02/natural-language-processing-key-terms-explained.html
- John Snow Labs: https://nlp.johnsnowlabs.com/
- NLTK Pre Trained pipeline for Sentiment Analysis: https://www.analyticsvidhya.com/blog/2021/06/rule-based-sentiment-analysis-in-python/
- Matrix Factorization: https://www.youtube.com/watch?v=ZspR5PZemcs
- Model Based: https://towardsdatascience.com/how-you-can-build-simple-recommender-systems-with-surprise-b0d32a8e4802
- Factorization Machines: https://www.analyticsvidhya.com/blog/2018/01/factorization-machines/