Amazon Recommender System

Kyla Ronellenfitsch, Rash Pal, Chris Nakamura, Jenny Jiang, Jaclyn Ivanisevic

Agenda

Business Case

Data and Insights

Segment Identification

Modeling

Conclusion

Business Case

Introduction to Recommendation Systems

Recommendation systems identify similarities between users and items and exploit that information to make recommendations

- Benefits:
  - Assists users to find right products
  - For a business, it can increase user engagement
  - Allows producers to meet the right consumers
  - Makes the content more personalized for each user

Types of Recommendation Systems

Content-Based
recommends items based on items similar to user's past interactions

Collaborative Filtering
based on the assumption that people continue to like similar things

Hybrid Approach
combines collaborative and content-based filtering

Classification-Based
understands the features of the user and classifies

Popularity-Based
recommends items viewed and rated by most people

Association Rule Mining
captures relationship between items based on their patterns of co-occurrence across transactions

Business Problem: Collaborative Filtering

Amazon uses a Collaborative Recommendation System. There are two key challenges with this approach:

- **Gray Sheep** are those whose opinions or behaviours are inconsistent and thus do not benefit from collaborative filtering. Similarly, *Black Sheep* are nearly impossible to classify due to their atypical tastes.

- **Cold Start** users are new customers who have had little or no interaction with the system and are not possible to provide personalized recommendations for.

Data & Insights

Data

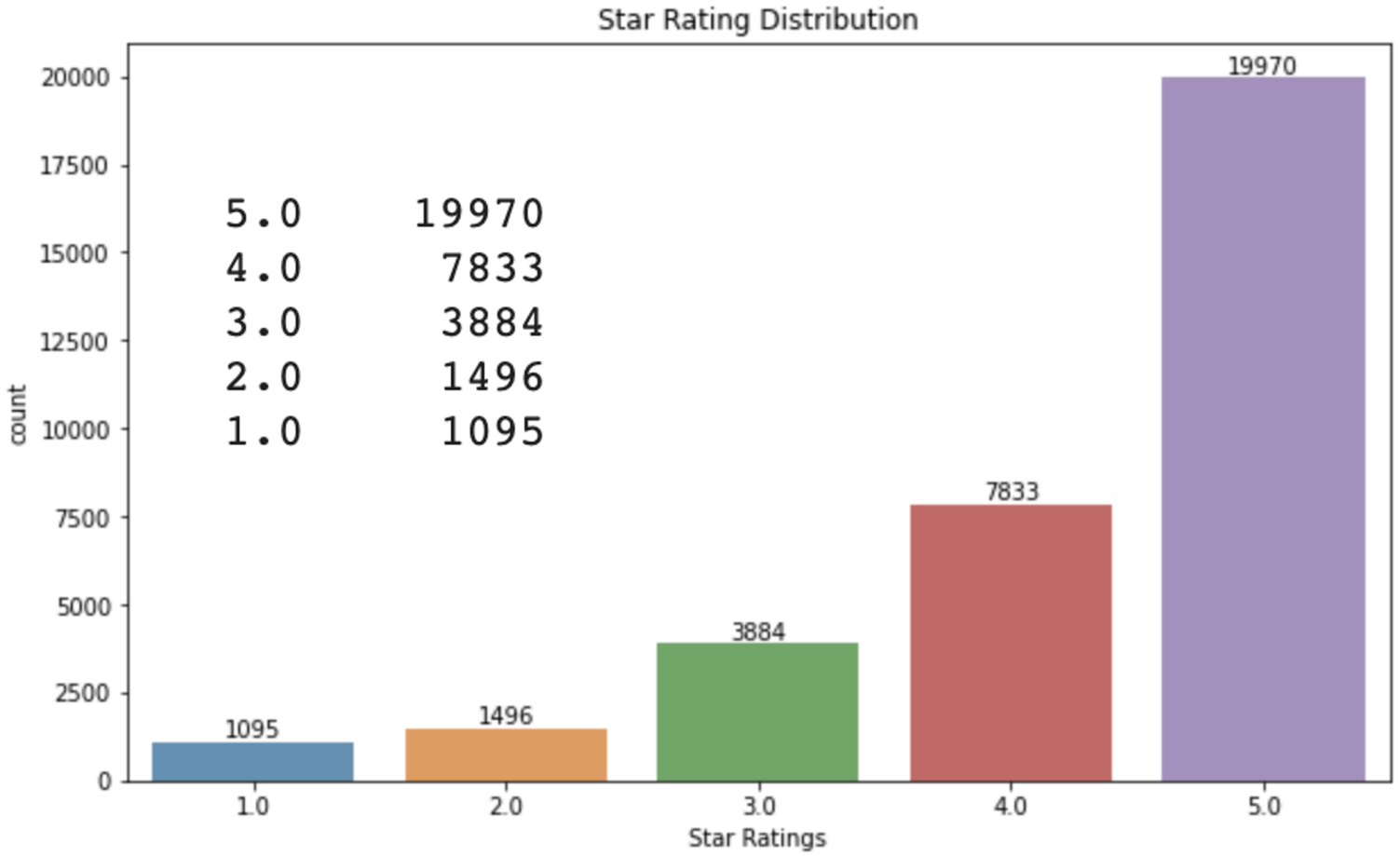
Amazon Luxury Beauty Product Reviews

- Used a subset of data which included only reviewers with at least five reviews
- 34k rows, 12 columns
- Features:
  - Reviewer ID
  - Product ID
  - Overall rating: 1-5 stars
  - Review text (unstructured)
  - Summary: Summary of the review text (unstructured)

Luxury Beauty Product Metadata

- Data set containing information on each product
- Features:
  - Product ID
  - Title (unstructured)
  - Description (unstructured)
  - Price

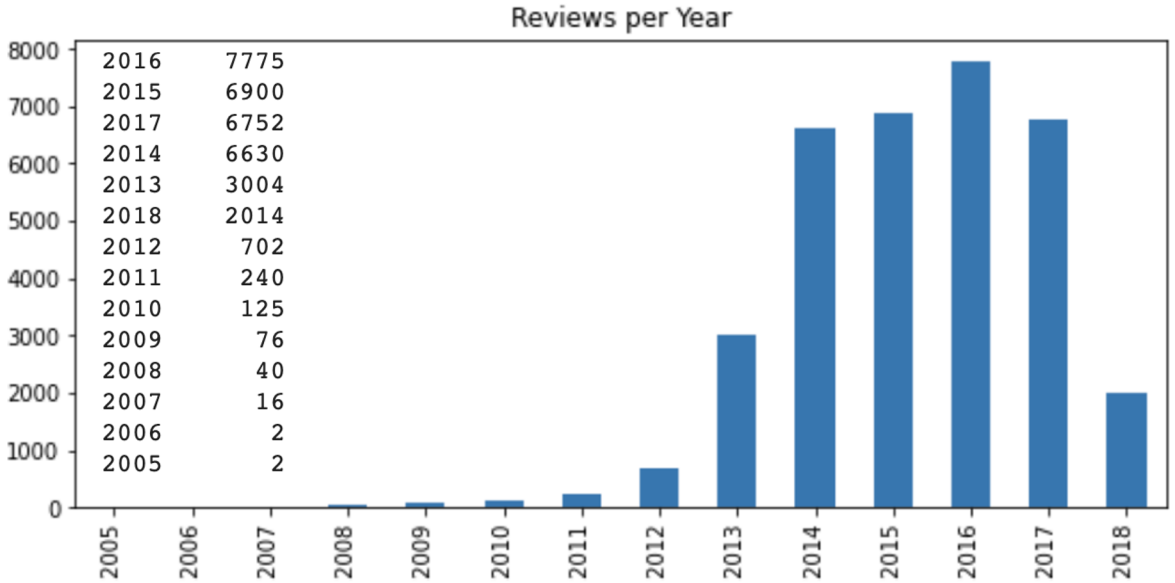
Data Insight - Product Ratings

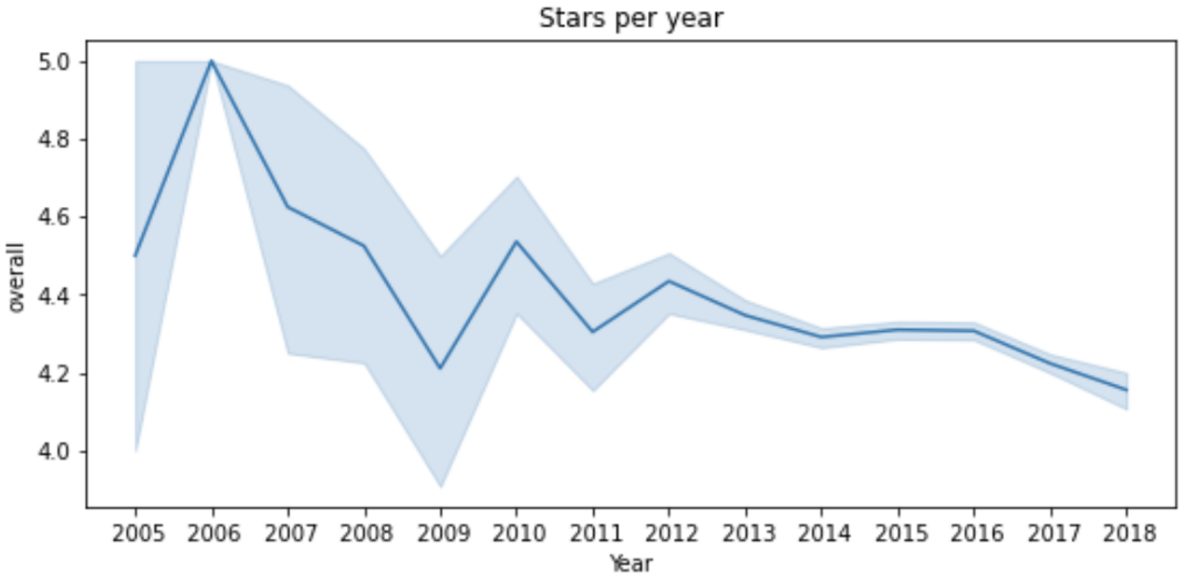


5.0	19970
4.0	7833
3.0	3884
2.0	1496
1.0	1095

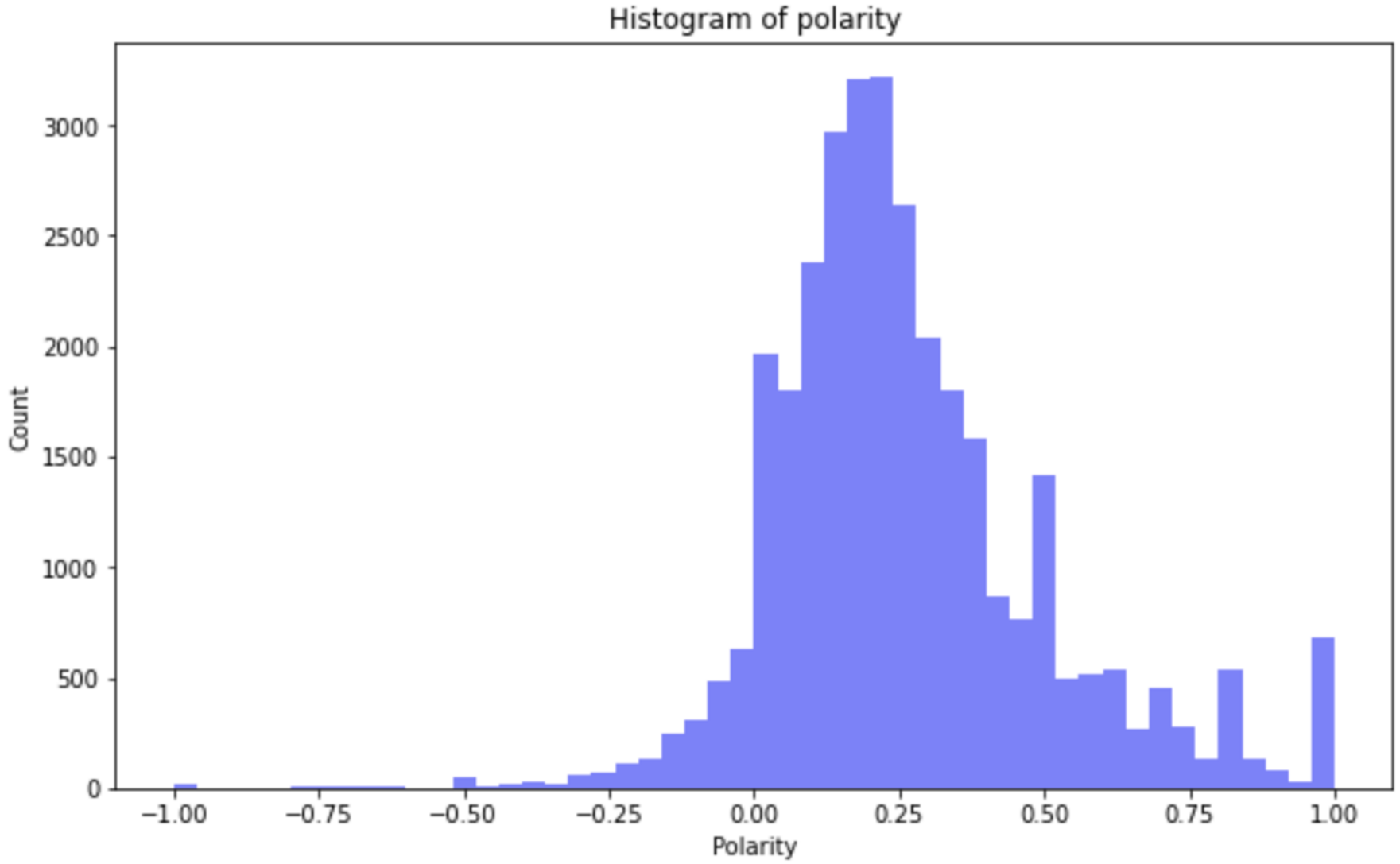
count	34278.00000
mean	4.28616
std	1.03736
min	1.00000
25%	4.00000
50%	5.00000
75%	5.00000
max	5.00000

Data Insight - Reviews & Rating by Year

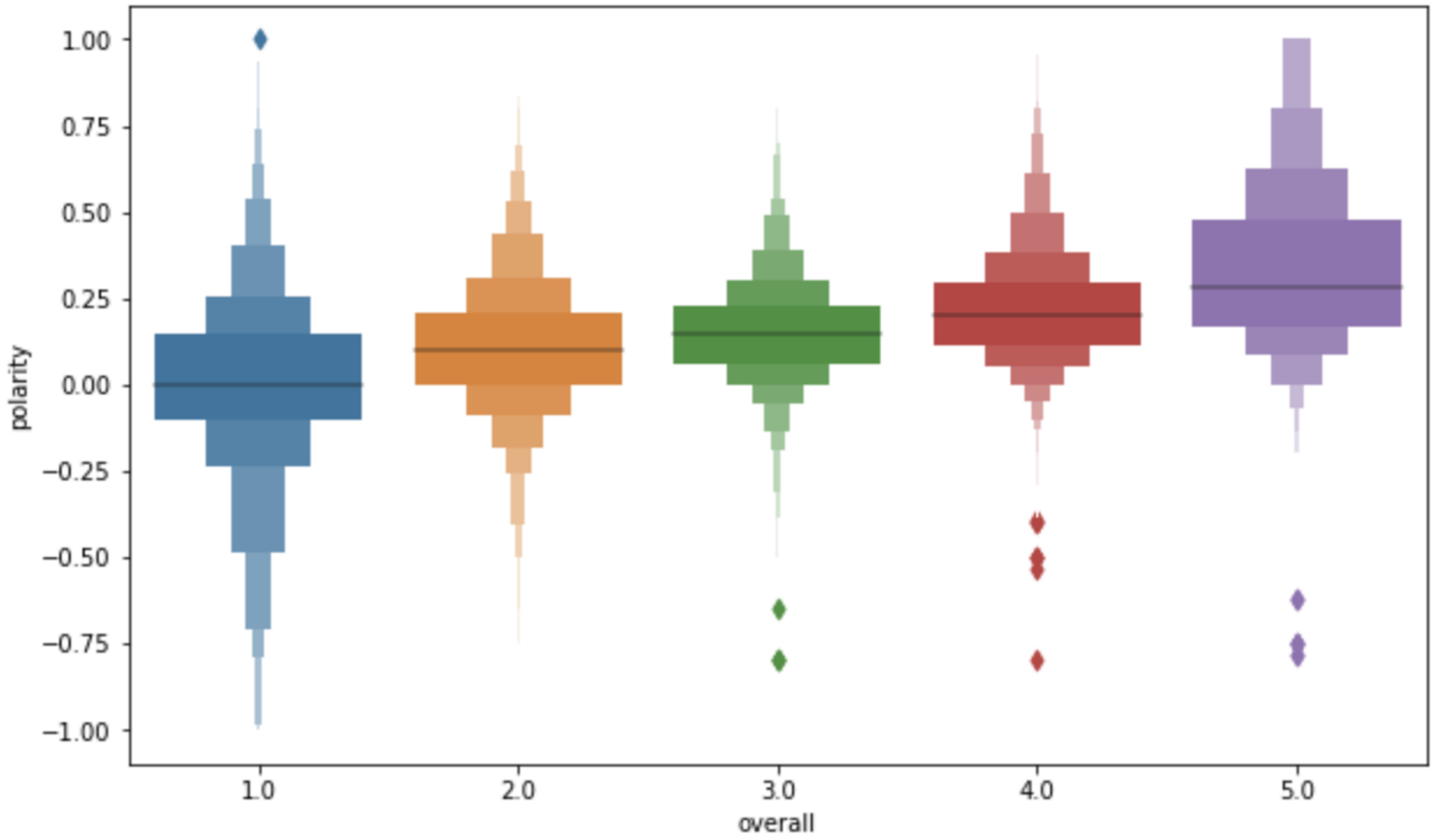




Sentiment Analysis on Product Reviews



Review Sentiment Based on Rating



Review Sentiment - Outliers

Negative Polarity with highest rating

overall	reviewText	polarity
5.0	Very expensive but I am a believer. Helps generate collagen on my face and under eye area. I think it is worth every penny.	-0.175000
5.0	Expensive but worth every penny	-0.100000
5.0	I really like this face wash, because I like to use a facial wash that has suds and isn't drying. This product provides both for me. I'm not sure if my face is "brighter", but I will continue to use it!	-0.037500

Positive Polarity with lowest rating

overall	reviewText	polarity
1.0	Extremely light!	0.500000
1.0	I love this color (and many of Essie's colors), but the polish gets so many bubbles as it dries that it's unusable.	0.500000
1.0	Great product.	0.800000

Product Segmentation

1

- Predefine 7 **categories** for all beauty products
- Use word2vec to convert words to vectors

2

- Extract nouns from all **product title**
- Use word2vec to convert those nouns to vectors

3

- Calculate similarity score between **categories** and **extracted nouns**
- Pick top 15 extracted nouns that have the highest similarity score in each category

4

- Loop over each product to calculate a **frequency score** for each category
- Choose the category with the highest frequency score as final category for each product

Example

Seven Categories

Makeup

makeup, eyeliner, eyeshadow, lipliner, manicure ...

Skin care

skin, pigmentation, blackhead, redness, moisturizing ...

Hair care

hair, beard, blowdry, frizz, shampoo...

Fragrance

fragrance, perfume, cologne, scent, bodycare ...

Hand care

hand, finger, palm, side, glove ...

Nail care

nail, seal, polish, polishing, scalp ...

Oral care

tooth, lip, cuticle, gum, toothbrush ....

PCA SKIN C-Quench Antioxidant Serum ⇨ SKIN Serum ⇨ Category - Skin care: 1 ⇨ Final category - Skin care

Aspect-based Sentiment Analysis

1

- Predefine **aspects** for one category
- Use word2vec to convert words to vectors

2

- Extract nouns from all **customer reviews** in that category
- Use word2vec to convert those nouns to vectors

3

- Calculate similarity score between **aspects** and **extracted nouns**
- Pick top 15 extracted nouns that have the highest similarity score for each aspects

4

- Loop over reviews of each product to calculate a **sentiment score** for each aspect

Example

Take skin care category as an example

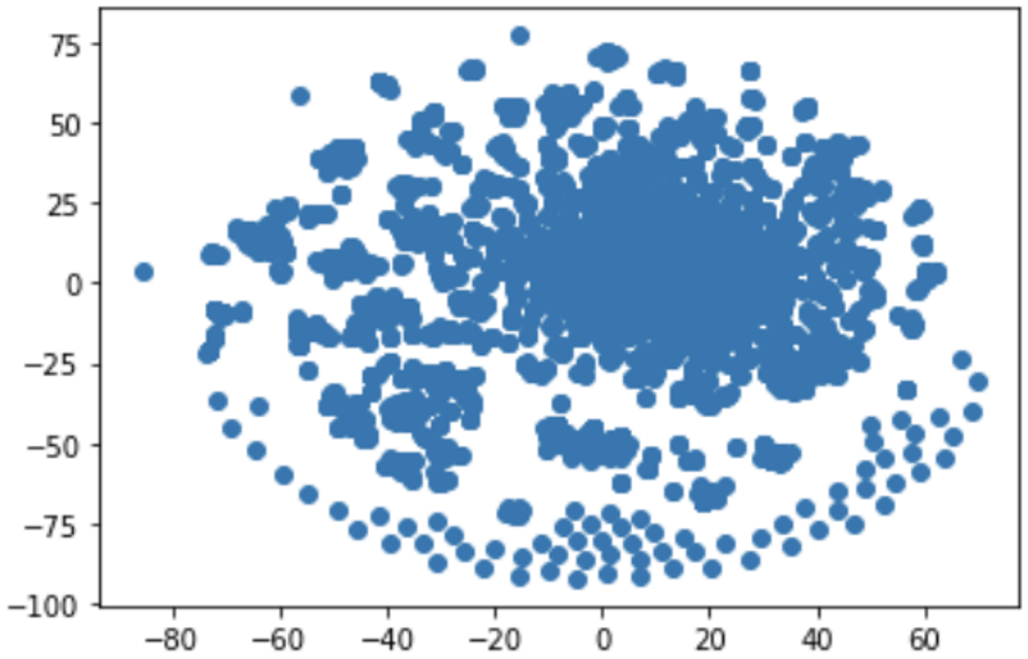
Price	**Moisture**	**Scent**	**Ingredient**
price, pricing, cost, discount, purchase ...	moisture, moist, humidity, wetness, dampness ...	scent, aroma, smell, fragrance, odor ...	ingredient, element, component, additive ....

Love this item. It calms my skin. The smell is good, but I don't like its price ⇨ item, skin, smell, price ⇨ smell in Scent aspect
price in Price aspect ⇨ Aspect - Scent: 1
Aspect - Price: -1

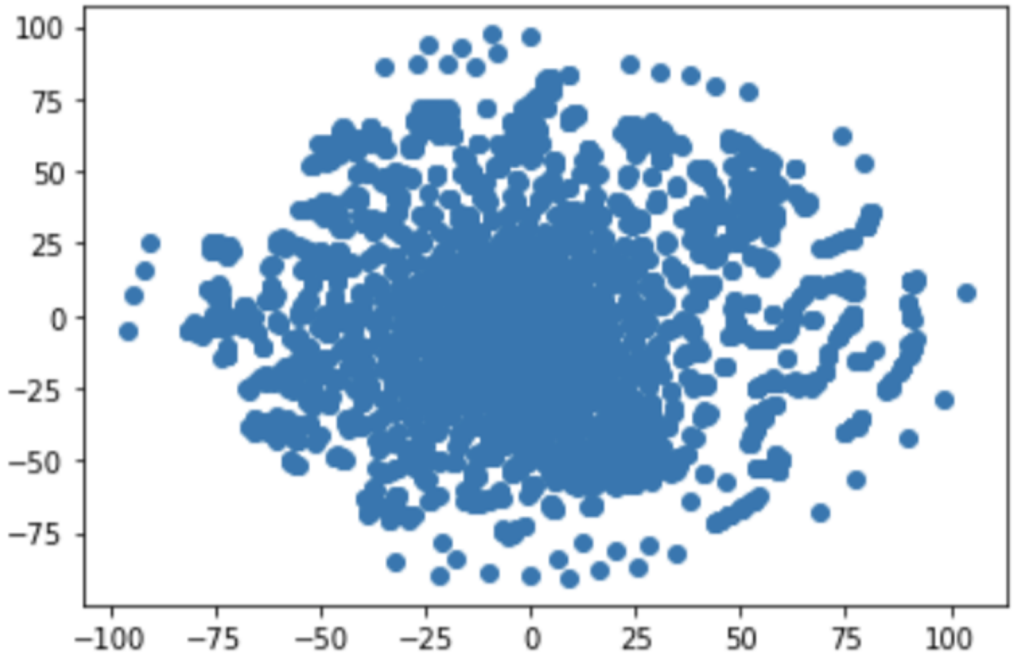
Segment Identification: Black Sheep & Cold Starts

5 Star Ratings vs. Sentiment

TSNE: 5 Star Ratings



TSNE: Sentiment Score

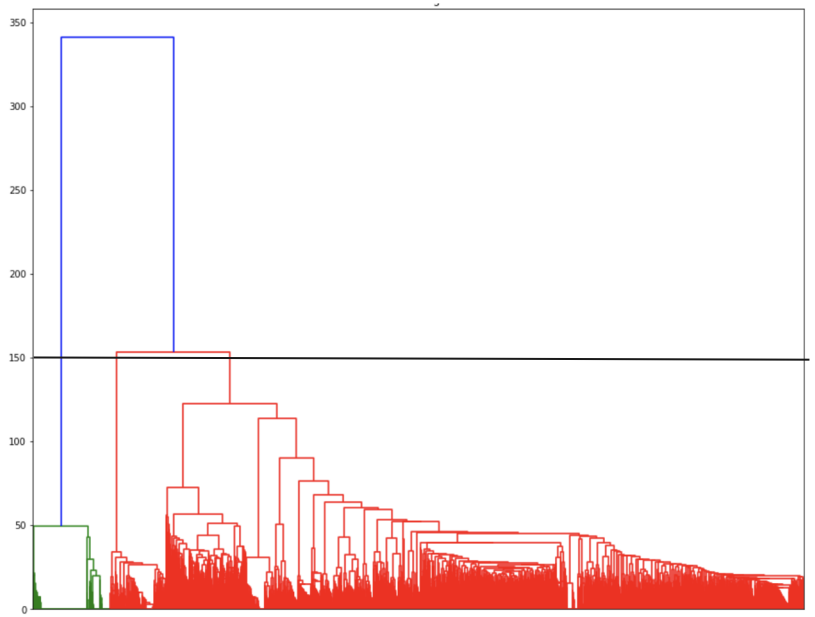


Clustering

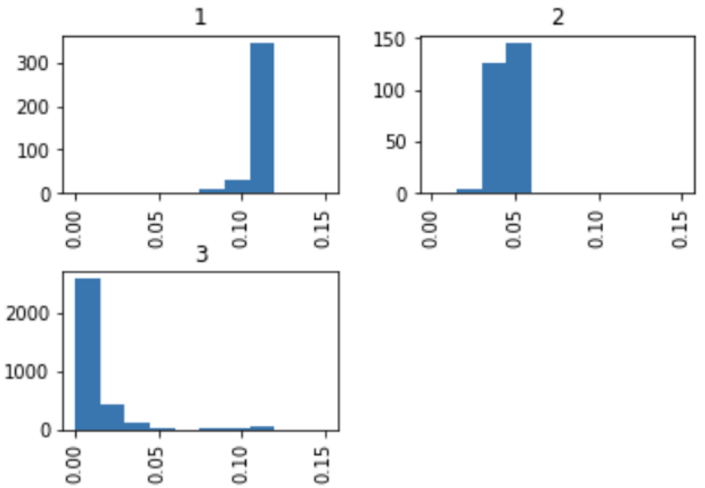
	Cluster: 5 Star	n=	Cluster: Sentiment	n=
K-Means	0	3434	0	3574
	1	385	1	245
DBSCAN	-1	113	-1	229
	0	3706	0	3590
Hierarchical	1	384	1	323
	2	276	2	3496
	3	3159		

Hierarchical Clusters

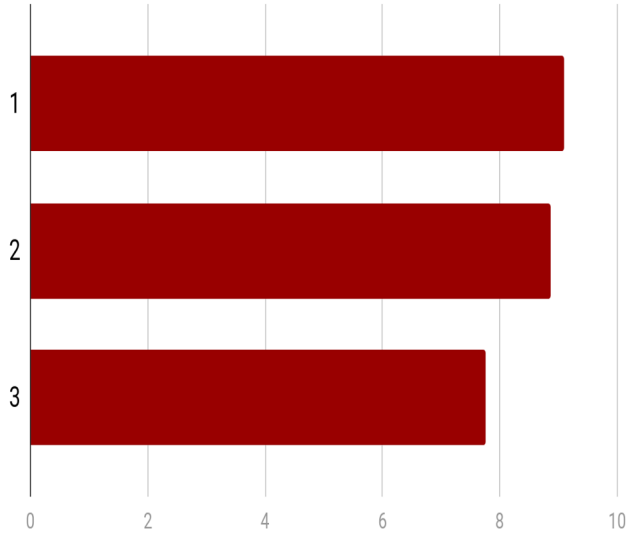
Dendrogram



Average Cosine Similarity

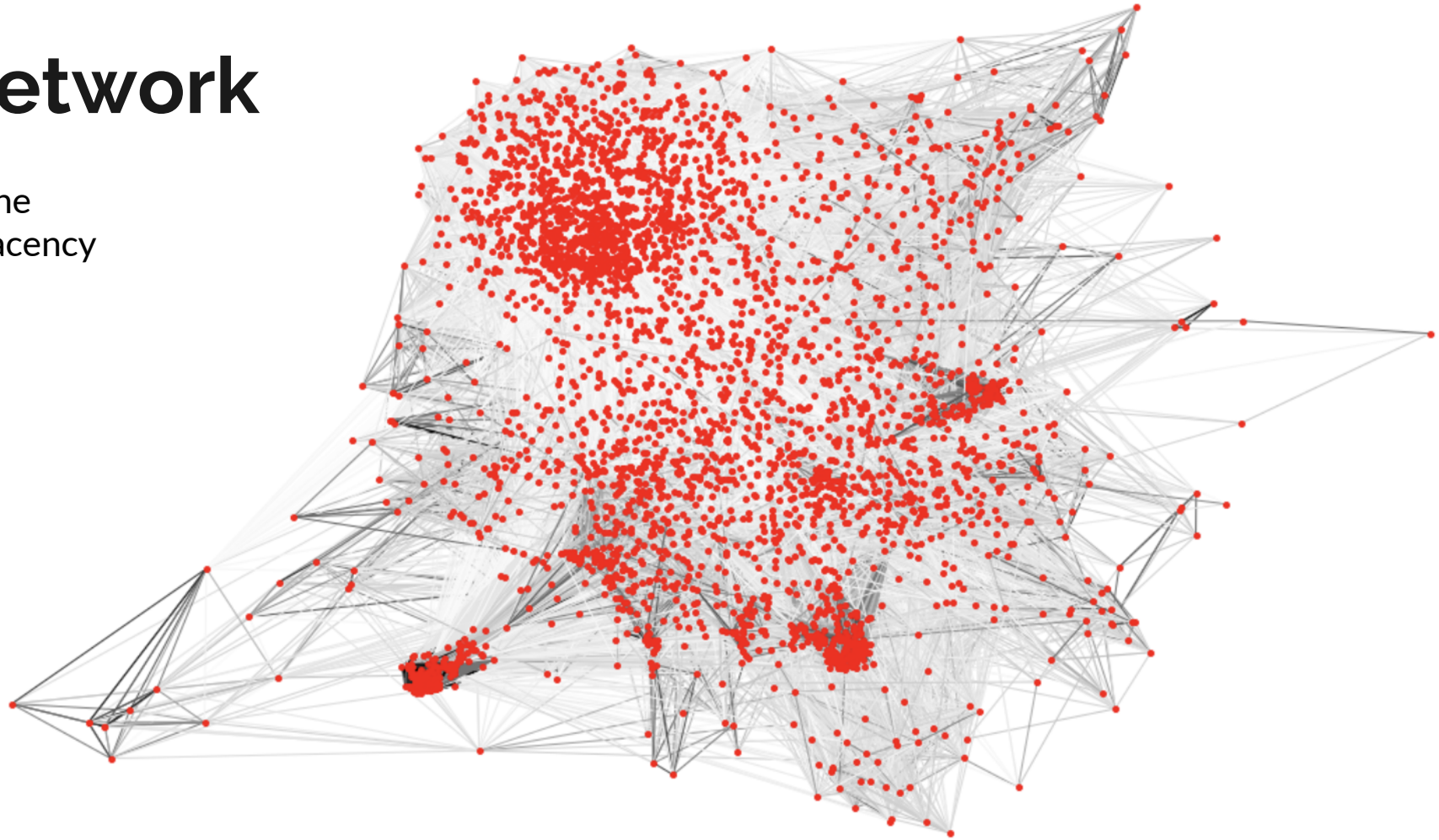


Average Products Purchased

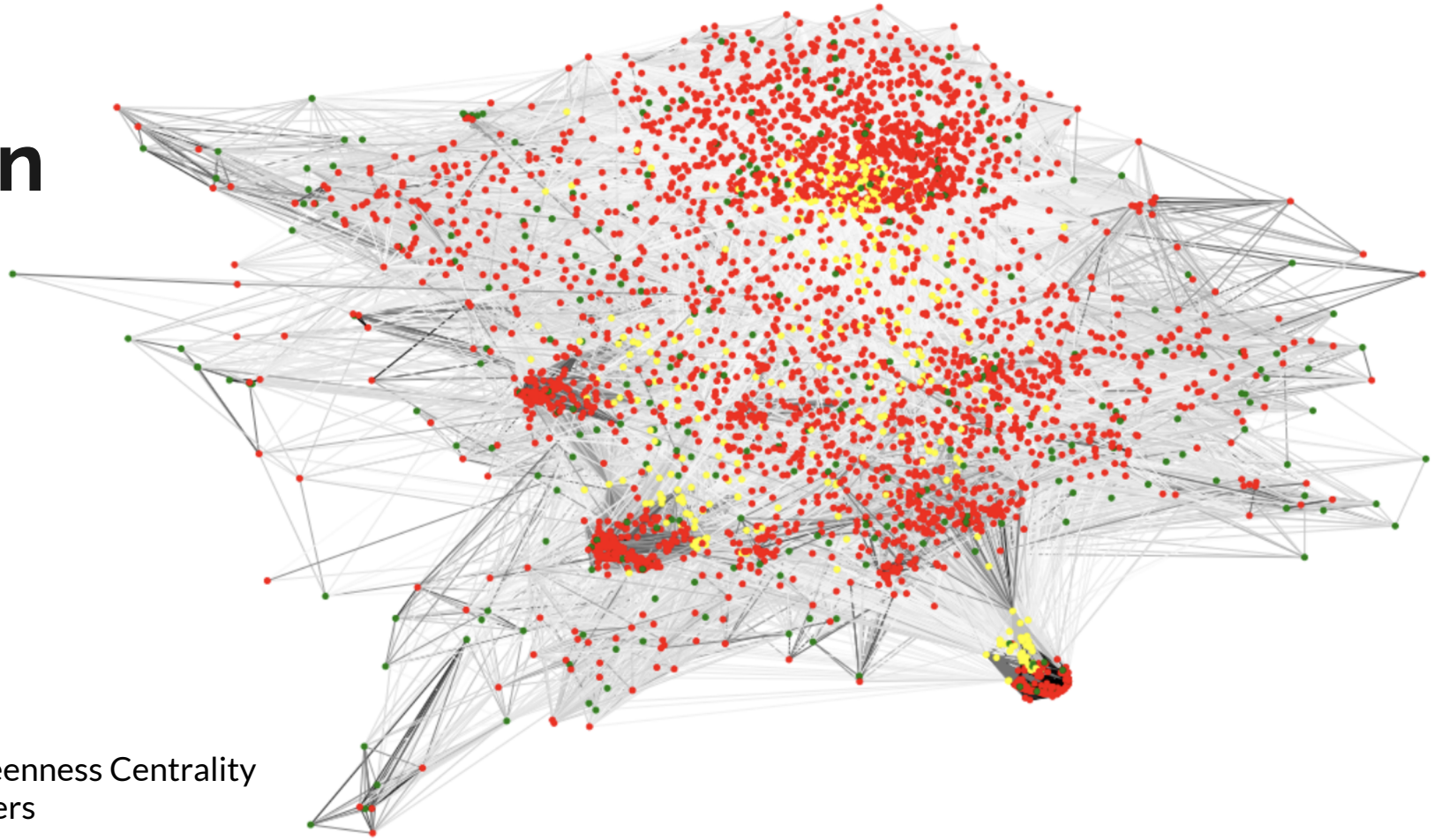


Customer Network

Network analysis using cosine similarity as weights for adjacency matrix



Segment Identification



Cold Start: <= 40 degrees
Gray Sheep: >= 90th percentile Betweenness Centrality
Typical Customer: Remaining customers

Modeling

Content Based Recommendation

Model Construction & Results

- Recommendations based solely on item information
  - Construct a similarity matrix of items based on cosine similarity of the vectorized words in the metadata title and description
  - For an instance of a purchase by a customer, compute the most similar items to the item purchased
  - Precision at 5 is a calculation the portion of the top 5 recommendations that are purchased
- The precision at 5 result for our dataset: 19.5%

Example:

- Customer purchased OPI Nail Lacquer, Not So Bora-Bora-ing Pink
- Top 5 most Similar items per cosine similarity matrix:
  1. OPI Nail Lacquer, She's a Bad Muffuletta!, 0.5...
  2. OPI Nail Lacquer, Cajun Shrimp, 0.5 fl. oz.
  3. OPI Nail Lacquer, Pale to the Chief, 0.5 Fl Oz
  4. OPI Nail Lacquer, That's Hula-rious!, 0.5 Fl Oz
  5. OPI Nail Lacquer, My Very First Knockwurst, 0....
- 2nd most similar item was purchased by the customer

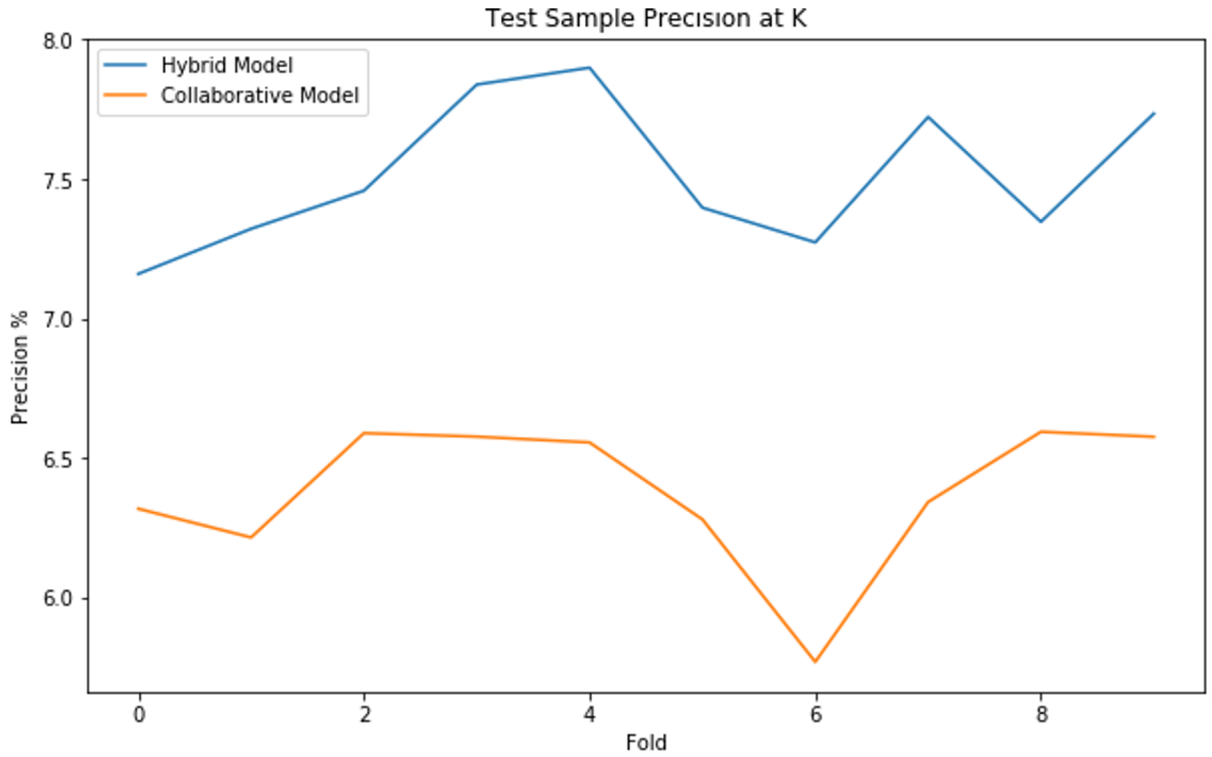
Light FM: Hybrid Approach

Model Construction & Results

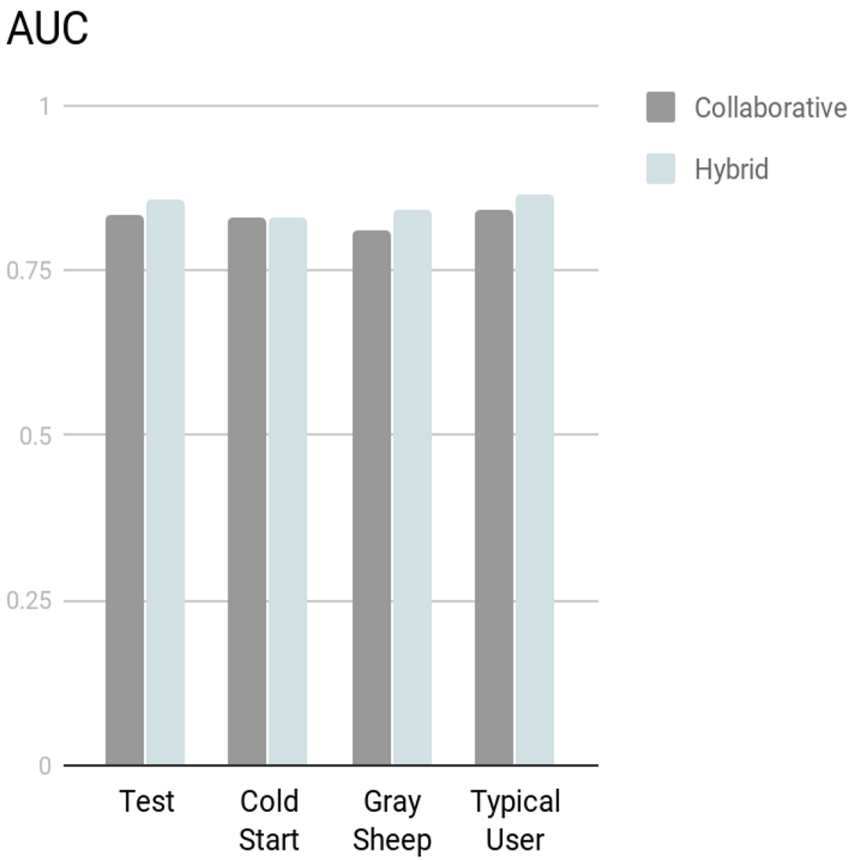
- Hybrid model operates on user item interactions with the ability to also take in item and user features.
- Offers a solution to the “cold-start” problem as item and user features allow for more accurate recommendations than a collaborative model before any interactions have taken place
- Operates using implicit feedback of ratings, as well as item description from product metadata
- Item descriptions were parsed using rake to extract keywords
- This bag-of-words for each item was vectorized into a matrix of item and word features
- For each item, we used cosine similarity to determine the 10 most similar products
- Using this similarity we performed clustering on the items
- We passed the cluster label into the model as an item feature

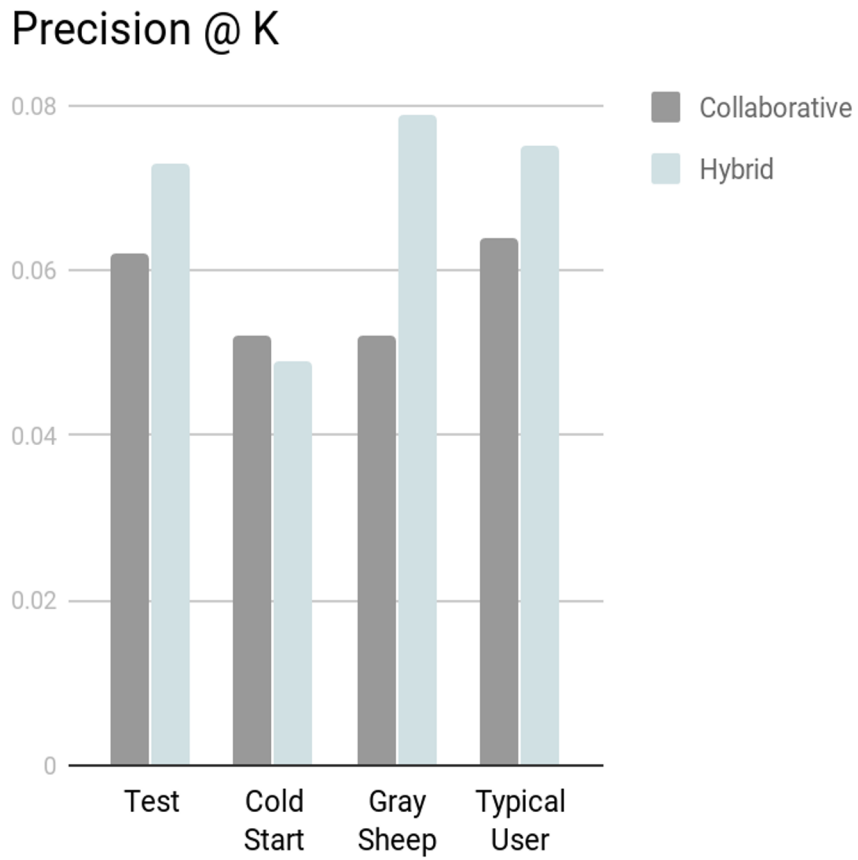
Light FM: Hybrid Approach

Incorporating item features boosted Precision at K from ~6.2% to ~7.5%



Performance on Segments





Light FM: Hybrid Approach

Example

Previous Purchases

title
La Roche-Posay Serozinc Face Toner for Oily Sk...
ampton Sun SPF 15 Super Hydrating Face Cream,...
L'Occitane Best of Collection
L'Occitane Luxurious Divine Star
SKIN&CO Roma Truffle Therapy Serum, 1.0 fl...
Obliphica Professional Seaberry Moisture Cream...
Mustela Baby Wipes, Unscented

Recommendations

title	recommendation_scores
L'Occitane 20% Shea Butter Hand Cream, 5....	2.171483
L'Occitane 15% Shea Butter Foot Cream Enriched...	1.417230
Mustela Baby Wipes, Unscented	1.068356
DERMASURI Rice Milk Brightening Face Exfoliato...	0.873441
Dermablend Smooth Liquid Foundation with SPF 2...	0.849193
Mustela 1.2.3. Diaper Rash Cream, Baby Skin Pr...	0.844605
La Roche-Posay Thermal Spring Water for Sensit...	0.777537
L'Occitane Cleansing & Softening Almo...	0.759642
Meaningful Beauty Vitality Oil	0.672057
Obagi Professional-C Serum, 1 fl. oz.	0.667480

Conclusion

By using a Hybrid model with collaborative and content based components, we saw a 25% improvement in precision at 5 compared to collaborative only model

We were able to gain insights from the data:

- Sentiment polarity for text reviews
- Aspect based sentiment analysis
- Ideally we would continue to explore incorporating this into recommendations

Limitations:

- Clustering customers based on text data from reviews became challenging due to sparsity of the matrix
- Would benefit from more information like customer demographics
- Data is limited to explicit feedback provided by reviews. Given more implicit data (repeat purchases, clicks, etc.) we could improve the efficacy of our model