Cracking the Culinary Code

Using Yelp Reviews to Guide Foodies towards their Perfect Plate

Analee Graig
Zoya Shafique
Rahul Chandani
Syed Faquaruddin Quadri

Agenda

1. Motivations
2. The Yelp Dataset
3. Exploratory Data Analysis
4. Baselines
5. Recommender System
6. Closing Remarks

01.

Motivations

Business Problem and Use Cases

Business Problem & Use Cases

Business Problem:

- Drive ad revenue by encouraging user engagement with the Yelp platform

Proposed Solution:

- Drive engagement by developing an algorithm to make personalized recommendations for new restaurants to users

Recommender Systems

- Can we recommend users restaurants they will like based on their previous ratings and/or the ratings of other users?

Deviations from Conventional ML

- Train and Test sets need some overlap
  - Otherwise run into cold start problem
- How to use only star ratings to model predictions? Can other features be related to star rating?
  - Not necessarily classification or regression
- Focused on personalized recommendations

01.

The Yelp Dataset

Overview

The Yelp Dataset

- Real world data scraped from Yelp
- Contains information on businesses found on Yelp, business reviews, and user information

Quick Statistics

- 6,990,280 reviews
- 1,987,897 users
- 150,346 businesses
- 11 metropolitan areas

Main Challenges

1. Handling JSON Data
2. Data Cleaning
3. Feature selection

Handling JSON Data

- Parsing nested dictionaries not possible with Pandas or JSON package
- Requires custom parsing script to extract restaurant categories and attributes
- Volume of reviews dataset requires pre-processing to be done in Spark


```
{"business_id":"Pns2l4eNsfO8kk83dixA6A","name":"Abby Rappoport, LAC, CMQ","address":"1616 Chapala St, Ste 2","city":"Santa Barbara",
    "state":"CA","postal_code":"93101","latitude":34.4266787,"longitude":-119.7111968,"stars":5.0,"review_count":7,"is_open":0,
    "attributes":{"ByAppointmentOnly":"True"},
    "categories":"Doctors, Traditional Chinese Medicine, Naturopathic\/Holistic, Acupuncture, Health & Medical, Nutritionists","hours":null}
{"business_id":"mpf3x-BjTdTEA3yCZrAYPw","name":"The UPS Store","address":"87 Grasso Plaza Shopping Center","city":"Affton","state":"MO",
    "postal_code":"63123", "latitude":38.551126,"longitude":-90.335695,"stars":3.0,"review_count":15,"is_open":1,
    "attributes":{"BusinessAcceptsCreditCards":"True"},
    "categories":"Shipping Centers, Local Services, Notaries, Mailbox Centers, Printing Services",
    "hours":{"Monday":"0:0-0:0","Tuesday":"8:0-18:30","Wednesday":"8:0-18:30","Thursday":"8:0-18:30","Friday":"8:0-18:30","Saturday":"8:0-14:0"}}
{"business_id":"tUFrWirKiKi_TAnsVWINQQ","name":"Target","address":"5255 E Broadway Blvd","city":"Tucson","state":"AZ","postal_code":"85711",
    "latitude":32.223236,"longitude":-110.880452,"stars":3.5,"review_count":22,"is_open":0,
    "attributes":{"BikeParking":"True","BusinessAcceptsCreditCards":"True","RestaurantsPriceRange2":"2","CoatCheck":"False",
    "RestaurantsTakeOut":"False", "RestaurantsDelivery":"False","Caters":"False","WiFi":"u'no'",
    "BusinessParking":"{'garage': False, 'street': False, 'validated': False, 'lot': True, 'valet': False}",
    "WheelchairAccessible":"True","HappyHour":"False","OutdoorSeating":"False","HasTV":"False","RestaurantsReservations":"False",
    "DogsAllowed":"False","ByAppointmentOnly":"False"},
    "categories":"Department Stores, Shopping, Fashion, Home & Garden, Electronics, Furniture Stores",
    "hours":{"Monday":"8:0-22:0","Tuesday":"8:0-22:0","Wednesday":"8:0-22:0","Thursday":"8:0-22:0","Friday":"8:0-23:0",
    "Saturday":"8:0-23:0","Sunday":"8:0-22:0"}}
```

Handling JSON Data

- Parsing nested dictionaries not possible with Pandas or JSON package
- Requires custom parsing script to extract restaurant categories and attributes
- Volume of reviews dataset requires pre-processing to be done in Spark


```
{"business_id":"Pns2l4eNsfO8kk83dixA6A","name":"Abby Rappoport, LAC, CMQ","address":"1616 Chapala St, Ste 2","city":"Santa Barbara",
    "state":"CA","postal_code":"93101","latitude":34.4266787,"longitude":-119.7111968,"stars":5.0,"review_count":7,"is_open":0,
    "attributes":{"ByAppointmentOnly":"True"},
    "categories":"Doctors, Traditional Chinese Medicine, Naturopathic\/Holistic, Acupuncture, Health & Medical, Nutritionists","hours":null}
{"business_id":"mpf3x-BjTdTEA3yCZrAYPw","name":"The UPS Store","address":"87 Grasso Plaza Shopping Center","city":"Affton","state":"MO",
    "postal_code":"63123", "latitude":38.551126,"longitude":-90.335695,"stars":3.0,"review_count":15,"is_open":1,
    "attributes":{"BusinessAcceptsCreditCards":"True"},
    "categories":"Shipping Centers, Local Services, Notaries, Mailbox Centers, Printing Services",
    "hours":{"Monday":"0:0-0:0","Tuesday":"8:0-18:30","Wednesday":"8:0-18:30","Thursday":"8:0-18:30","Friday":"8:0-18:30","Saturday":"8:0-14:0"}}
{"business_id":"tUFrWirKiKi_TAnsVWINQQ","name":"Target","address":"5255 E Broadway Blvd","city":"Tucson","state":"AZ","postal_code":"85711",
    "attributes":{"BikeParking":"True","BusinessAcceptsCreditCards":"True","RestaurantsPriceRange2":"2","CoatCheck":"False",
    "RestaurantsTakeOut":"False", "RestaurantsDelivery":"False","Caters":"False","WiFi":"u'no'",
    "BusinessParking":"{'garage': False, 'street': False, 'validated': False, 'lot': True, 'valet': False}",
    "WheelchairAccessible":"True","HappyHour":"False","OutdoorSeating":"False","HasTV":"False","RestaurantsReservations":"False",
    "DogsAllowed":"False","ByAppointmentOnly":"False"},
    "categories":"Department Stores, Shopping, Fashion, Home & Garden, Electronics, Furniture Stores",
    "hours":{"Monday":"8:0-22:0","Tuesday":"8:0-22:0","Wednesday":"8:0-22:0","Thursday":"8:0-22:0","Friday":"8:0-23:0",
    "Saturday":"8:0-23:0","Sunday":"8:0-22:0"}}
```

Data Cleaning

- Misspelled city labels
- Duplicate restaurant reviews
- Users with more than one review for the same restaurant
- Sparse and irrelevant restaurant categories
- Non-English language reviews
- Spam reviews

Narrowing Focus to New Orleans

469,337 reviews

201,608 users

2255 businesses

02.

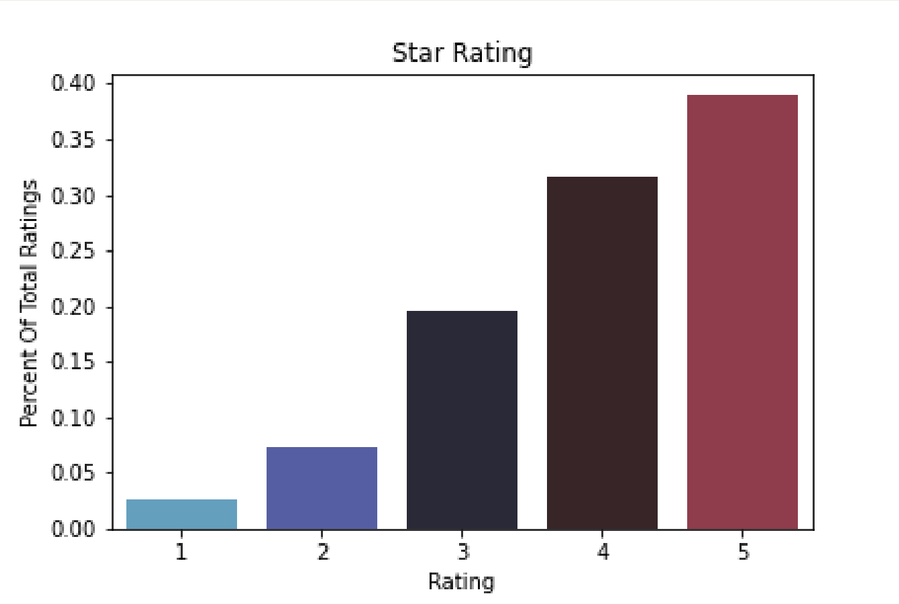
Exploratory Data Analysis

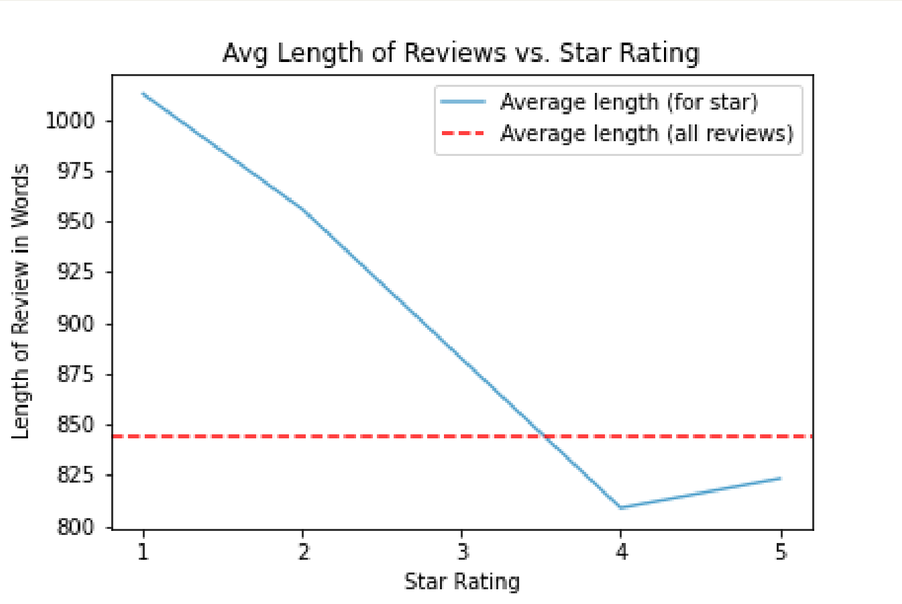
Making Sense of Large-Scale, Real World Data

Goals of Exploratory Data Analysis (EDA)

- Which City Should We Focus On?
- How Much Usable Data Do We Have?
- What Can We Learn From the Restaurant categories and attributes?
- What Can We Learn From Text Reviews?

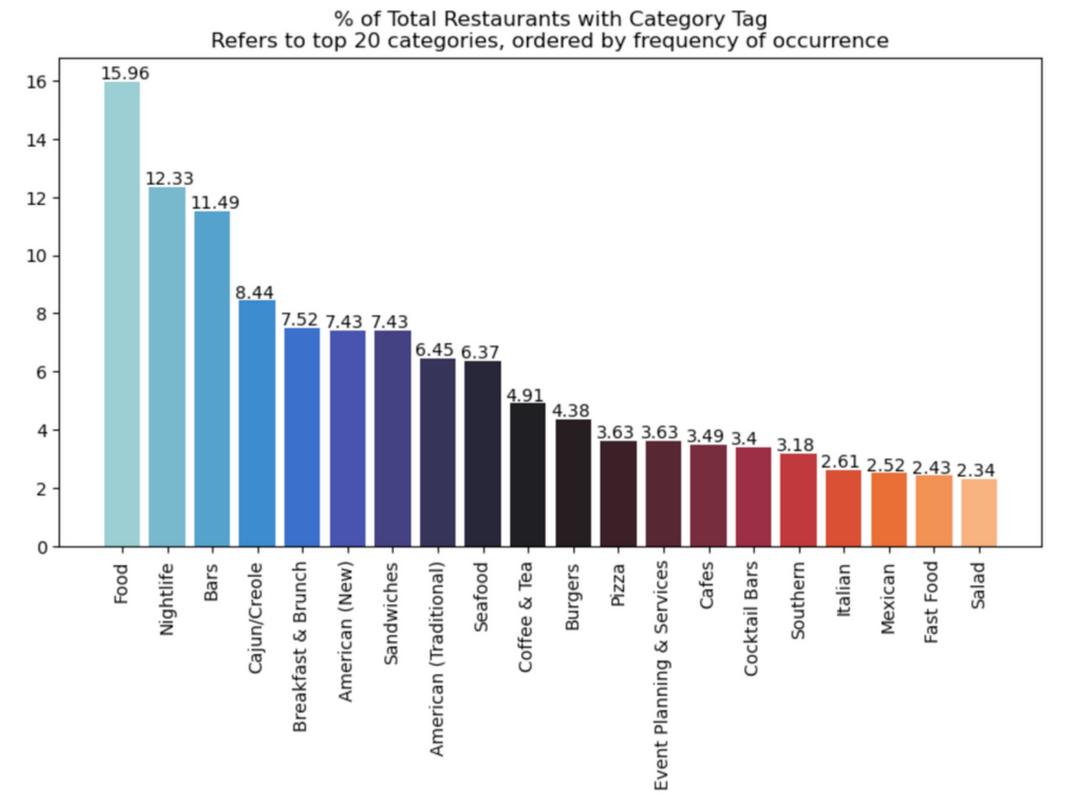
Initial Insights for New Orleans

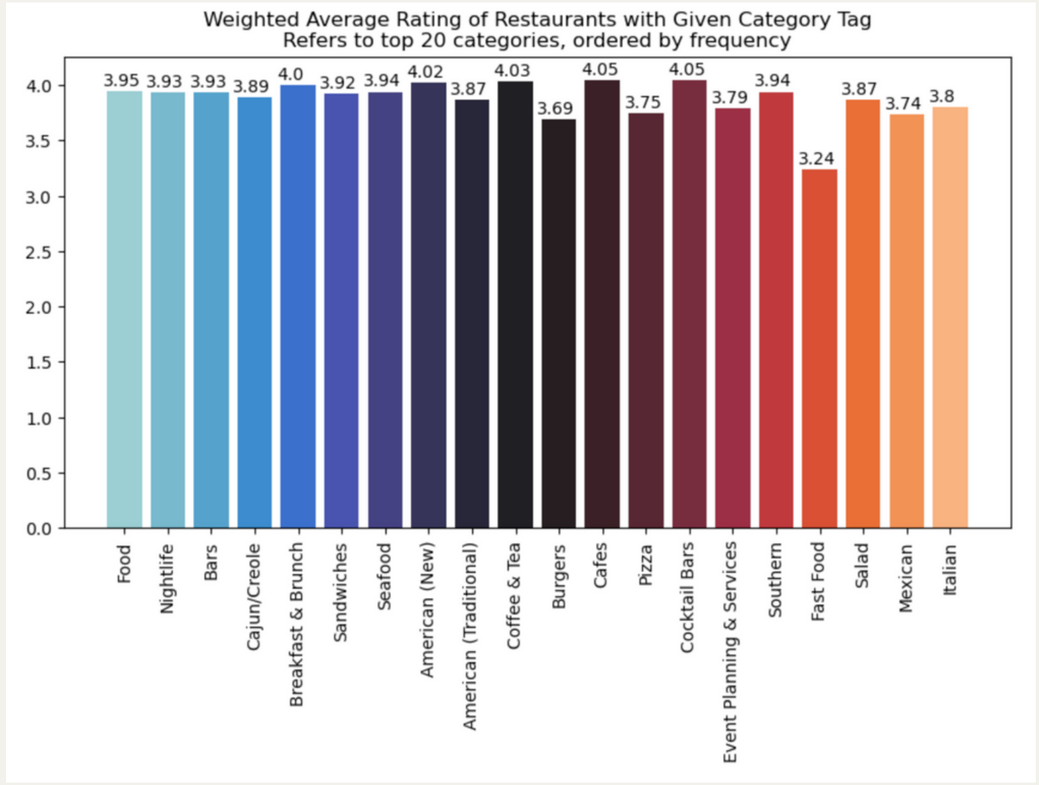




New Orleans Deep Dive - Restaurant Categories

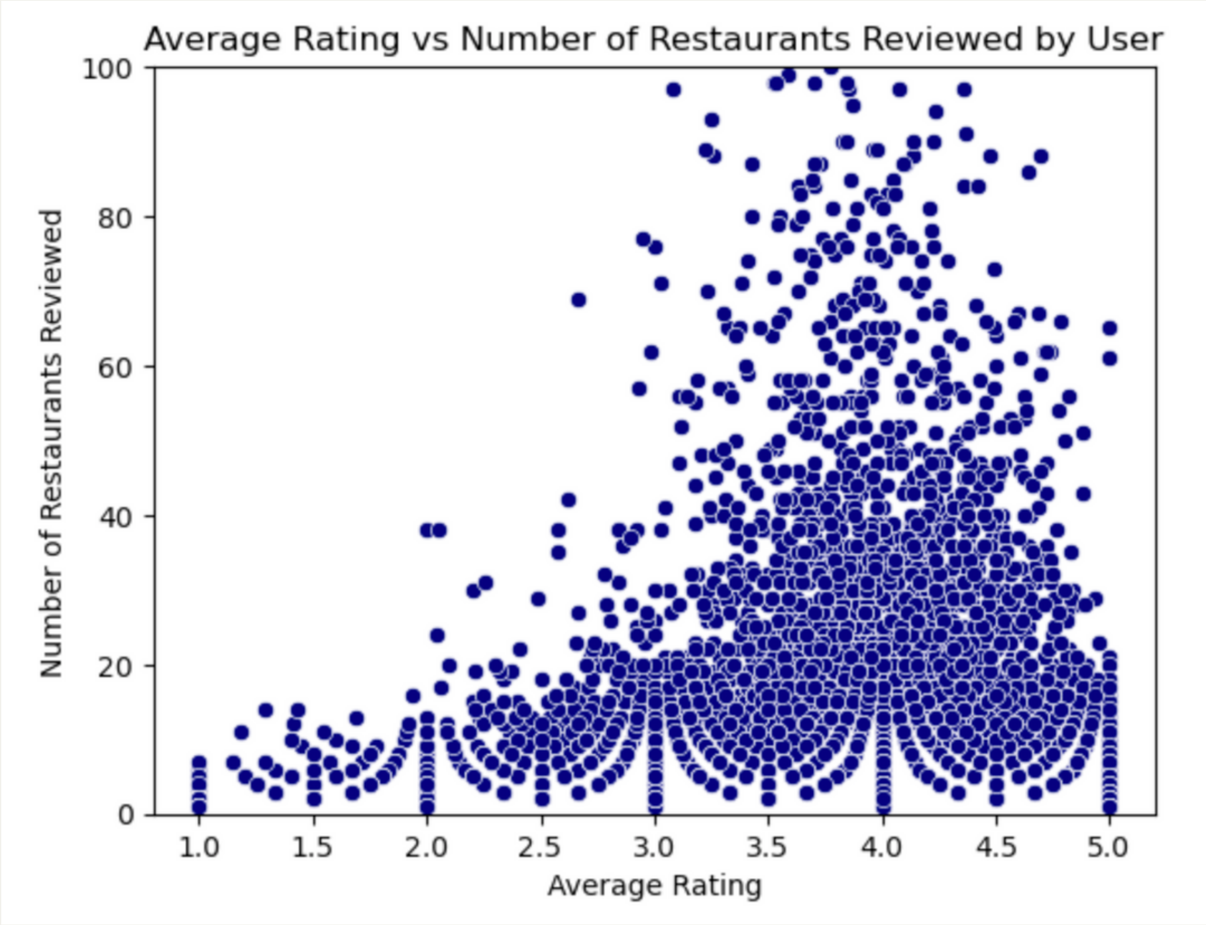
- Majority of tags have a low percentage of total restaurants with that tag. This suggest partially that New Orleans has a diverse restaurant scene
- The second and third highest categories are 'Nightlife' and 'Bars', highlighting New Orleans' prominent nightlife culture





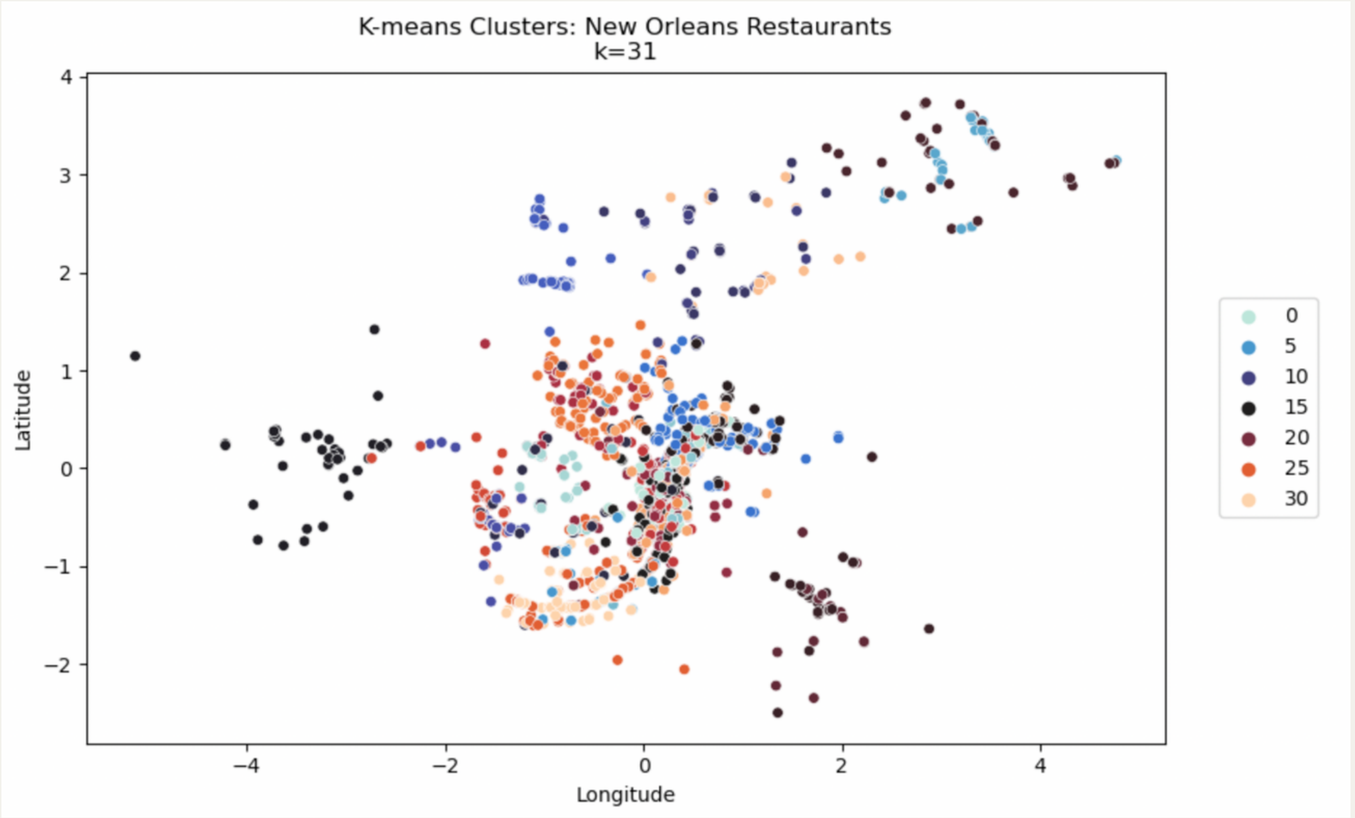
New Orleans Deep Dive - Users

- Majority of users have an average and median rating of about 4.5
- Relationship between the number of restaurants a user has reviewed and their average and median ratings. The more restaurants a user reviewed, the more likely they were to have rated restaurants well
- There is a small population of users who we'd call 'haters', who have fairly low average and median ratings (between 0-2). All of these users have rated less than 40 restaurants



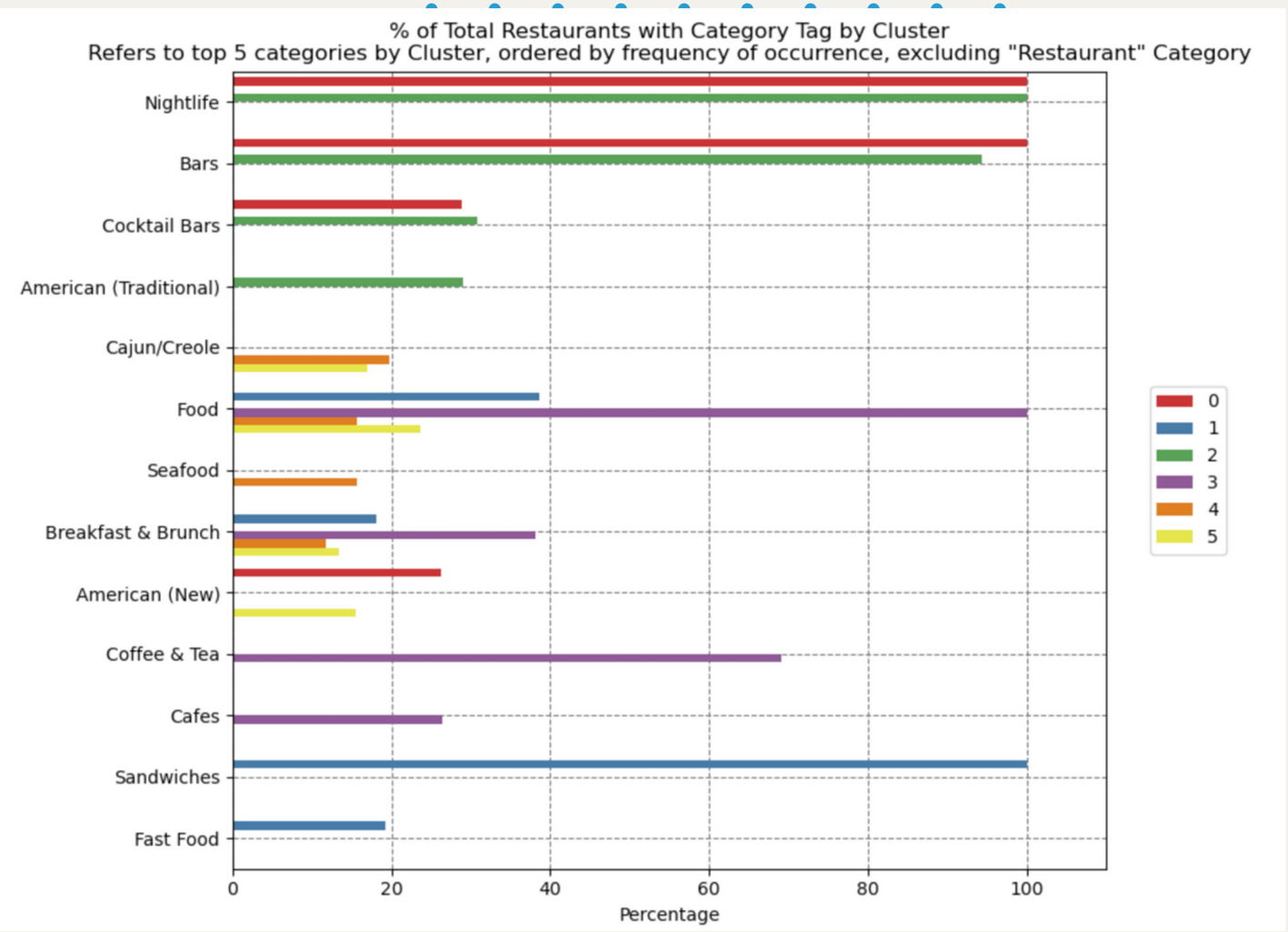
New Orleans Deep Dive
Clustering on Restaurant Attributes

- Clustered on numerical variables (Longitude, Latitude, Number of Reviews per User and Business, Average Star Rating per User and Business
- Not obvious what differentiates different clusters



New Orleans Deep Dive
Clustering on Restaurant Categories

- k = 6
- Divide between traditional restaurants and bars/nightlife venues that also serve food as indicated by Clusters 0 and 2
- Clear delineation for what appears to be bakery/cafe type places, as indicated by Cluster 3
- Cluster 1 appears to contain Fast Food restaurants such as Domino's and McDonald's
- Clusters 4 and 5 map to Cajun/Creole restaurants



03.

Baselines

Using conventional ML and simple statistics

Simple Baselines

Predicting Average Business Rating

Validation Set

RMSE = 0.98
MAE = 0.95

Test Set

RMSE = 0.98
MAE = 0.96

Predicting Average User Rating

Validation Set

RMSE = 0.97
MAE = 0.93

Test Set

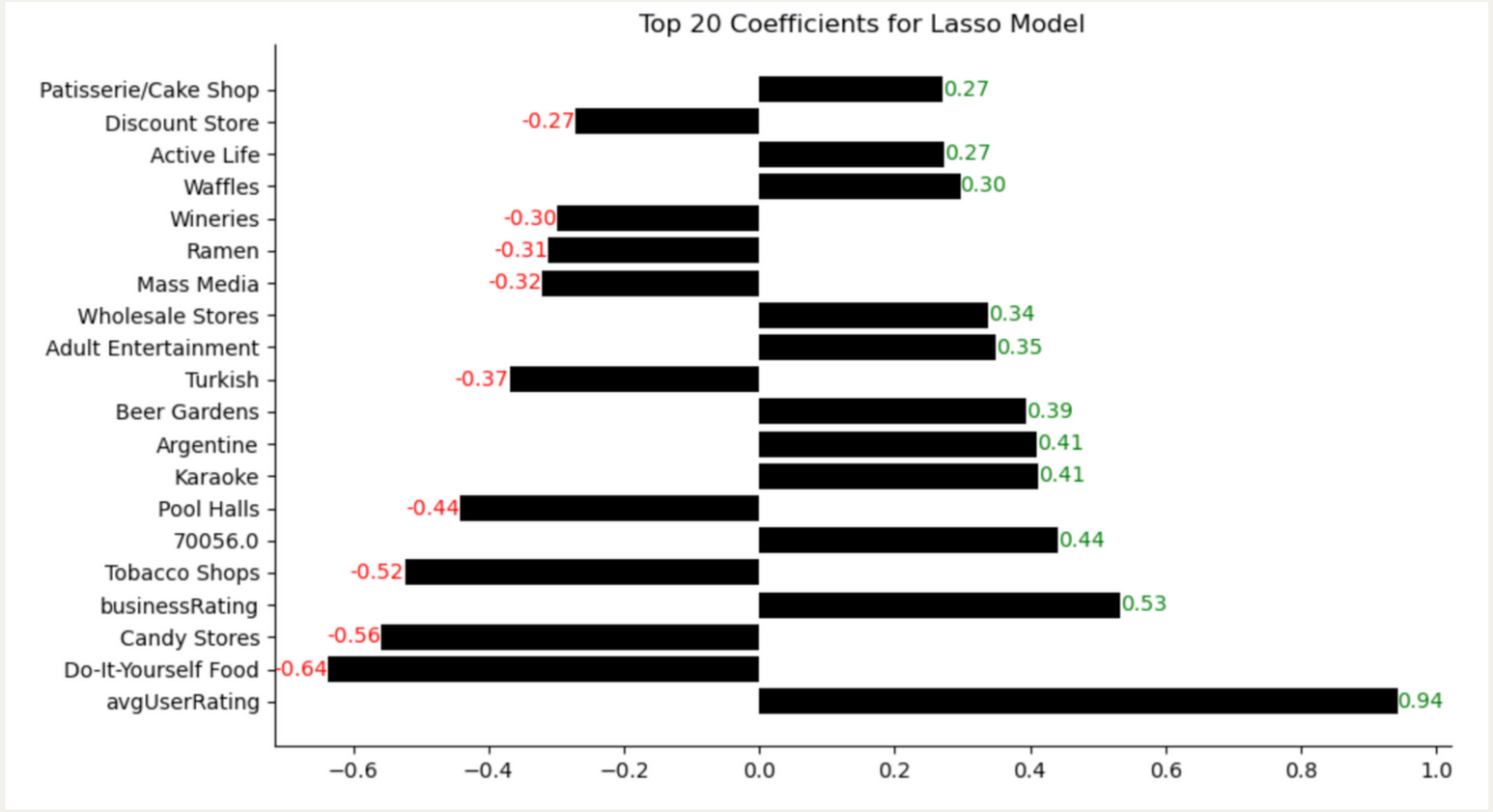
RMSE = 0.93
MAE = 0.87

Lasso Regression

Validation Set:
- CV RMSE: 0.876
- CV R2: 0.258

Test Set:
- CV RMSE: 0.881
- CV R2: 0.250

Takeaways:
- Overall Lasso Regression a poor fit for the data
- Test and Validation evaluation metrics are fairly close
- Alpha = 0.0001 as determined by GridSearchCV
- 113 Coefficients (~33%) of the total dataset sent to 0
- Top Coefficients not surprising given EDA
- Don't have a ton of faith in this model

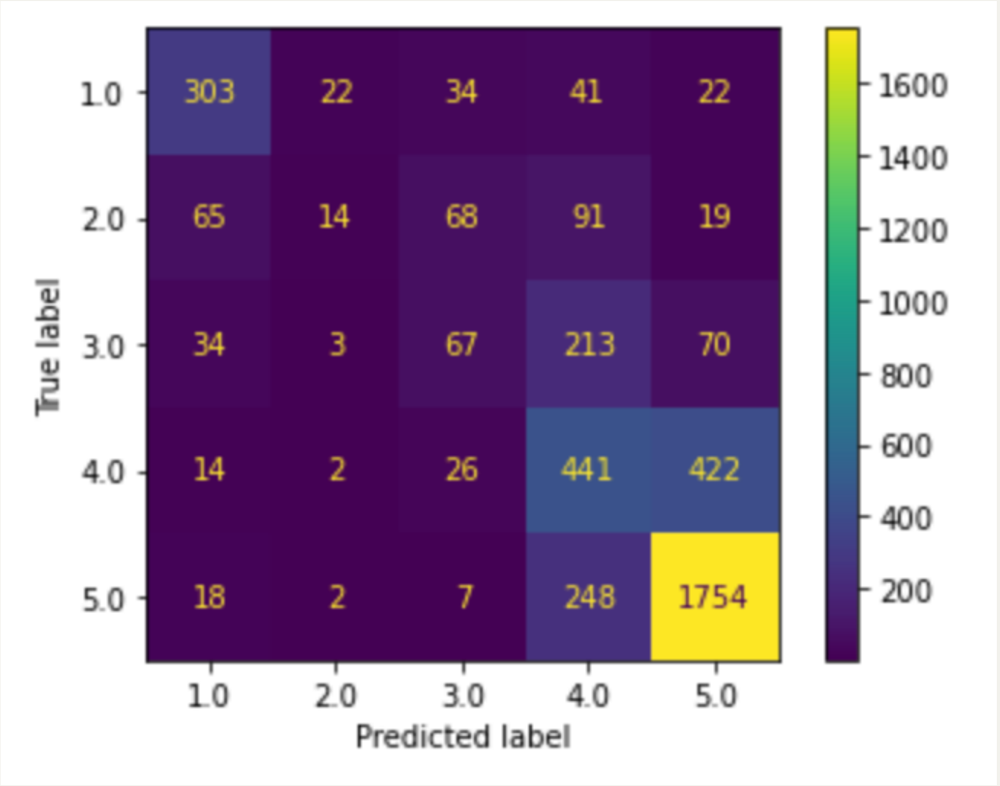


Naive Bayes Classifier

Classification Report

	precision	recall	f1-score	support
1.0	0.70	0.72	0.71	422
2.0	0.33	0.05	0.09	257
3.0	0.33	0.17	0.23	387
4.0	0.43	0.49	0.45	905
5.0	0.77	0.86	0.81	2029
accuracy			0.64	4000
macro avg	0.51	0.46	0.46	4000
weighted avg	0.61	0.64	0.62	4000

Confusion Matrix



03.

Recommender Systems

Customizing Recommendations for Users

1

Collaborative Filtering
Based on star ratings and text reviews
Model Structure: SVD, LDA, TF-IDF, Cosine Similarity

2

Collaborative Filtering
Based on star ratings and restaurant categories
Model Structure: SVD

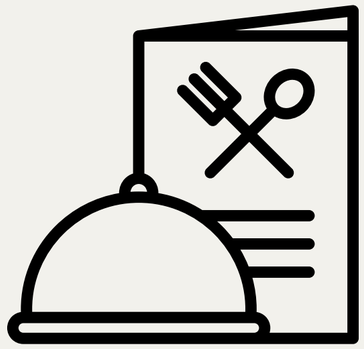
Collaborative Filtering
Based on star ratings
Model Structure: KNN

3

Content Filtering
Based on text reviews
Model Structure: TF-IDF, PCA

4

Final Models



Collaborative Filtering
Based on star ratings and restaurant categories
Model Structure: SVD

Content Filtering
Based on text reviews
Model Structure: TF-IDF, PCA

Collaborative Filtering

Use ratings from similar users to make recommendations

Data preprocessing

- Convert Data to Sparsity Matrix
- Data Standardization
- Train, Test, Split based on Sparse Matrix

Singular Value Decomposition

- Fills in Sparse Matrix
- Base Predictions

Finding Similar Users

- SVD reconstruction is done on a users vs. categories matrices
- Latent vectors represent similarities between users

Recommending Restaurants

- Recommend restaurants with the highest ratings in reconstructed matrix

SVD on Restaurant Categories - Methodology

Sparsity Matrix*

user_id	Acai Bowls	Active Life	Adult Entertainment	African	American (New)
-0LGLx8LP5dq3zcGO4Bebw	NaN	NaN	NaN	NaN	2.666667
-13RX4Gy_F-zoLlenWAo-w	NaN	NaN	NaN	NaN	4.750000
-154QAmLwXOsKqChHSeWJQ	NaN	NaN	NaN	NaN	4.000000
-1MiSauypbVtNnWts4aXpA	NaN	NaN	NaN	NaN	4.000000
-22PBmQh7bBWbNX1irrkPQ	NaN	NaN	NaN	NaN	NaN
-2PlzrbasYWAggcNS2ptGw	NaN	NaN	NaN	NaN	5.000000
-4YEZRxp3TwFKtNlzylx_Q	NaN	NaN	NaN	NaN	3.000000
-6rFcyKGC-B7C89FzLyBeA	NaN	NaN	NaN	NaN	4.000000
-BVK-mFx5n0bPtulAPpXew	NaN	NaN	NaN	NaN	3.000000
-Bo3nX8KSLzUSYVS-6c5Ag	NaN	NaN	NaN	NaN	4.428571

Values pre-Normalization

Strategy:

- For each user, determine average star rating of restaurants with that given category tag

Data Used:

- Restaurants with 50+ reviews
- Users with 10+ ratings
- Users with ratings in 10+ categories
- 4,734 users meet this criteria

Sparsity:

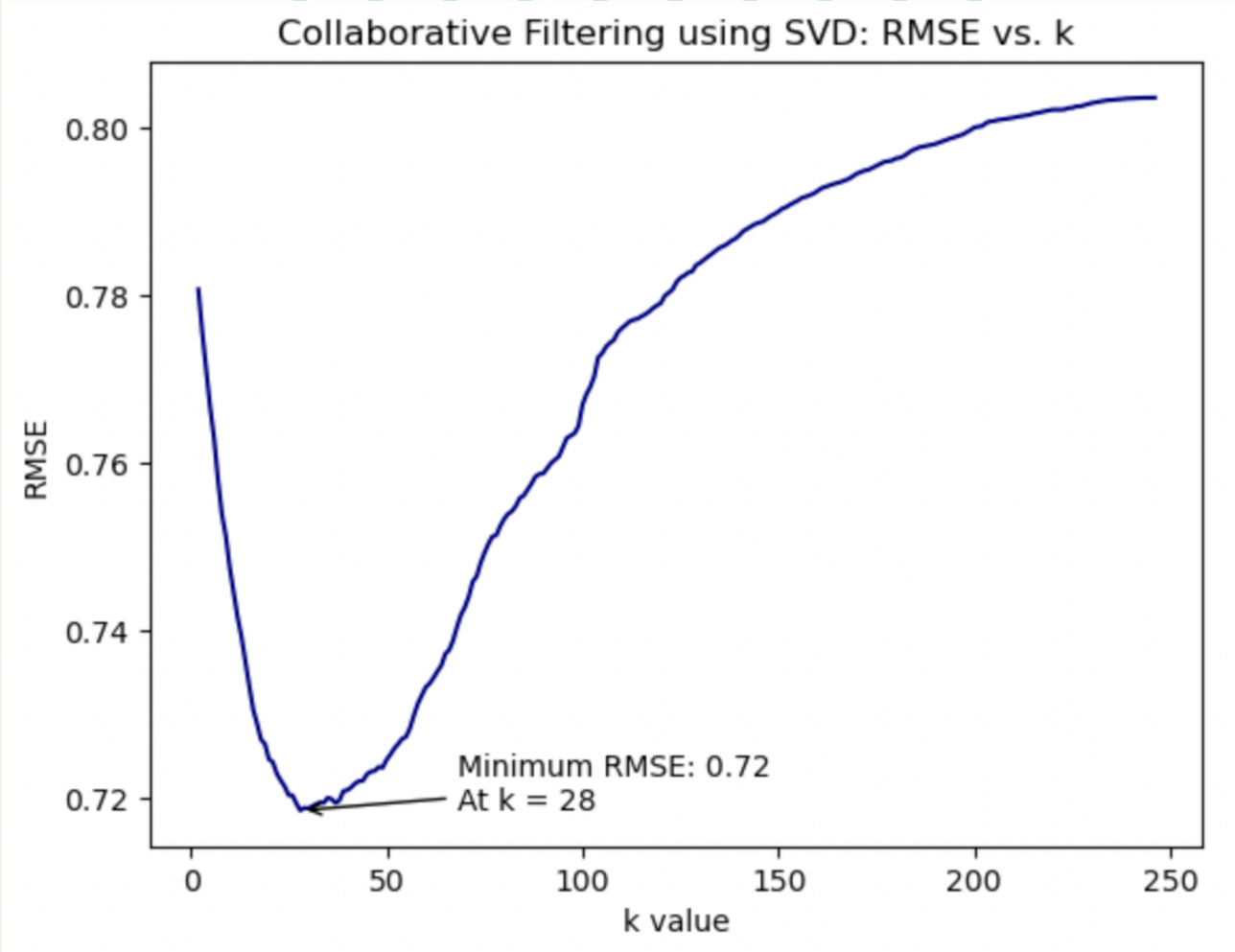
- 83%

SVD on Restaurant Categories - Results

Validation Set:
- RMSE: 0.72

Test Set:
- RMSE: 0.72

Takeaways:
- RMSE on Validation and Test Set roughly equivalent after rounding
- Beats Lasso Regression
- Increased RMSE at higher values of k indicate that model overfits with higher values of k
- We believe this is due to extraneous categories that are sparse and non-predictive as seen by Lasso regression
- RMSE was lower for less-sparse categories



Content Based Filtering

Use user text reviews to make recommendations

Data preprocessing

- Create user and business texts
- Remove punctuations

Feature Extraction

- Tf-idf vectorization
- Apply PCA to reduce to 700 col

Latent Factor Collaborative Filtering

- Generate factorization matrix
- Minimize error in the matrix

Recommending Restaurants

- Based on the matrix
- Filtering based on rating to get top 5 recommendations

Feature Extraction from textual data

Sparse Matrix*

business_id	10	10 minutes	12	15	15 minutes	20	20 minutes	25	30	30 minutes	...	problem	pudding	pulled	quality
--x_BmZbxzK_nx_GHBaRVw	-0.289666	-0.109454	0.038489	-0.225291	0.238462	0.167396	0.016315	-0.114173	0.216055	-0.176475	...	0.001185	-0.008963	0.013055	0.008543
--zb12mw2YK-7j6UaHzm8w	-0.093334	-0.120138	0.029334	-0.139521	0.058224	0.049186	-0.007548	0.044064	0.006012	-0.039793	...	-0.011567	-0.009014	-0.001737	-0.017612
-0__F9fnKt8uioCKztF5Ww	-0.055225	0.414484	-0.023394	0.159493	-0.052941	-0.027068	0.071784	0.019985	0.058636	0.054918	...	-0.004411	0.000052	0.005578	-0.008204
-0ltw8--HLuuIPyOSspqAQ	-0.054615	0.031138	-0.054414	0.015481	0.085246	-0.065940	-0.081257	-0.150943	0.127126	-0.036320	...	-0.004517	-0.001153	0.010079	0.003749
-1XSzguS6XLN-V6MVZMg2A	0.331914	0.097802	-0.030673	0.050489	0.302158	0.026223	0.022331	-0.008237	-0.003277	0.013162	...	-0.001754	-0.008100	-0.011732	-0.007020

5 rows × 700 columns

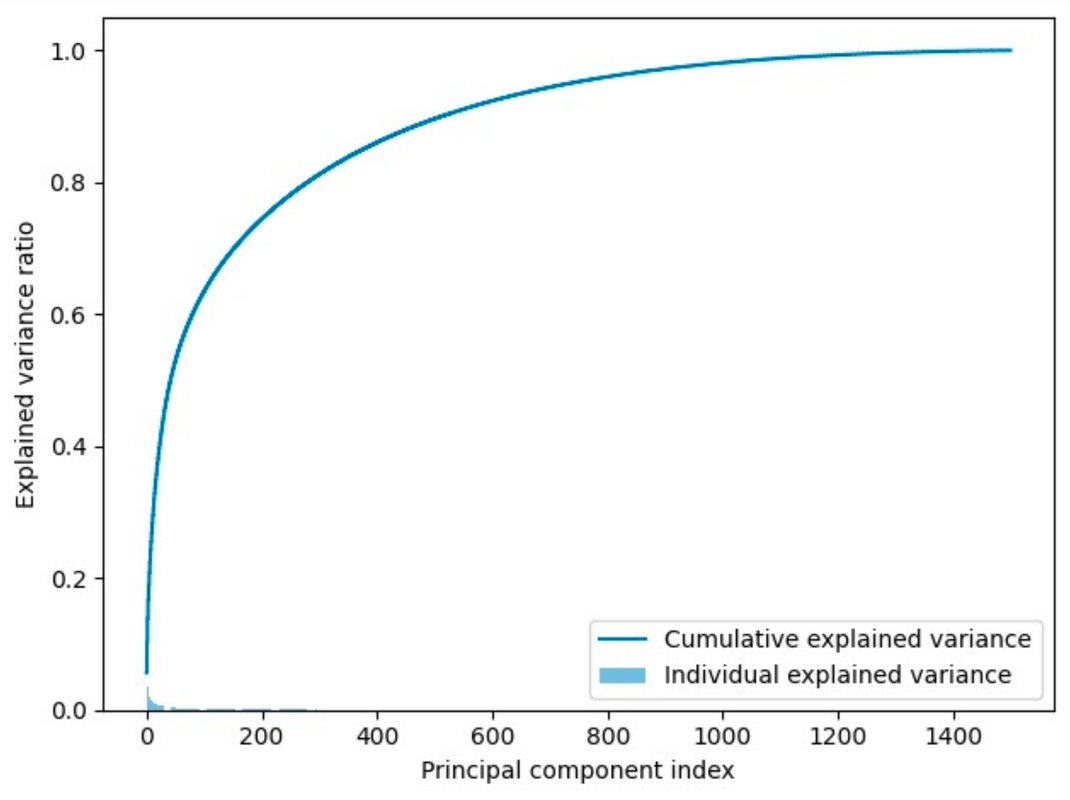
Data Used:

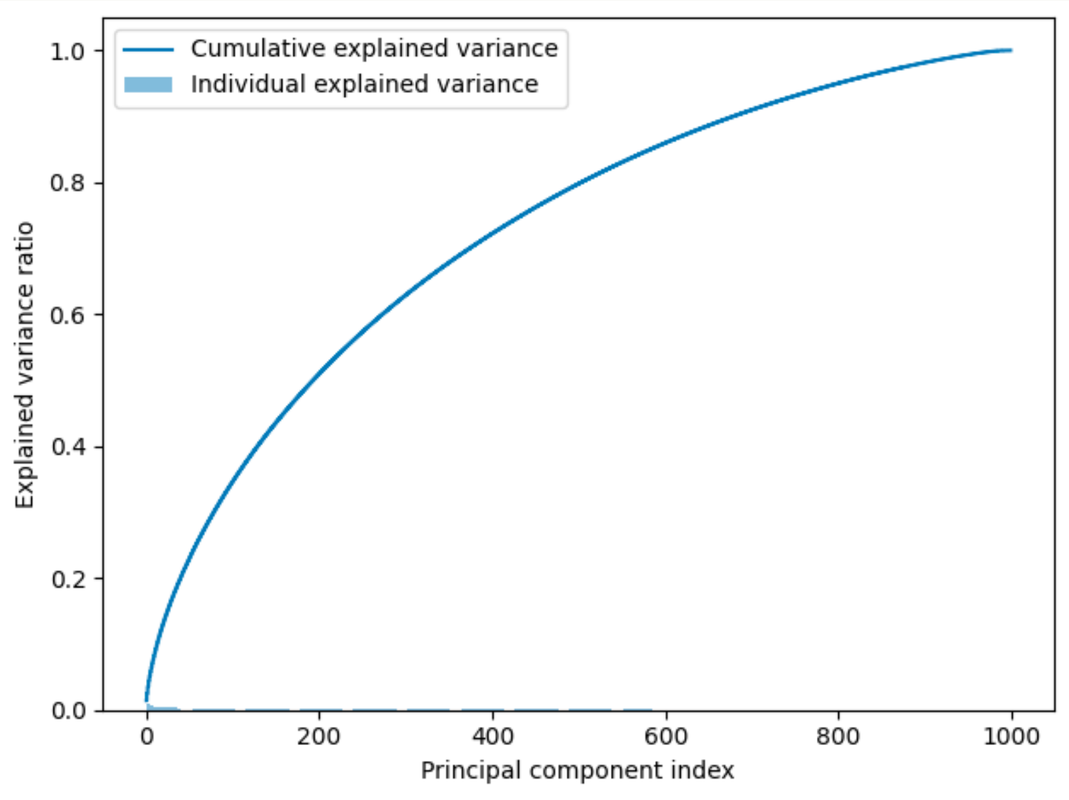
- User_id, business_id
- Review text combined

Strategy:

- Generating Vectors by Tfldf vectorization
- Reduce features using PCA

PCA reduction on text data





Results for a certain user

```
recommender('-13RX4Gy_F-zoLIenWAo-w')

Rock Bottom Lounge
4.0

Howl At the Moon
4.0

Starbucks
2.5

Bourbon Saloon
4.0

China Lights
4.5
```

User's reviews are taken into account to get features

03.

Closing Remarks

Customizing Recommendations for Users

Closing Remarks

Wrangling Real World Data	• Difficult to parse • Gathering insights and cleaning was a big bottleneck in terms of time • Many interesting features that we could not use
Conventional ML Algorithms in the Face of Recommender Systems	• Conventional ML algorithms have relatively high errors/low accuracies for prediction based on available features • Can run into problems with very sparse data and high dimensionality
Building Collaborative and Content Based Filters	• SVD helps combat sparsity problems by finding latent representations of data • Each filtering method has its own pros and cons but outperform conventional ML methods

Thank you!

GitHub Stats

Syed has commits under two names: SyedFaquar and =