Yelp Recommendation System

Capstone Project #1

Jerome Gonzaga

Problem Statement

Current Issue: It can often be very difficult to find a good restaurant, good mechanic, good barber shop, good cafe, etc. People can find a business on yelp and see their associated star rating but it's hard to trust the star rating. Many times I find that people look at reviews and photos to determine if the business fits what they are looking for and even this method can result in mixed conclusions about a business.

Problems to solve

- Is there a way to optimize searching for a business based on other characteristics? If so, what sort of patterns in the dataset can we find in order to base the searches on?

Data Used

https://www.yelp.com/dataset

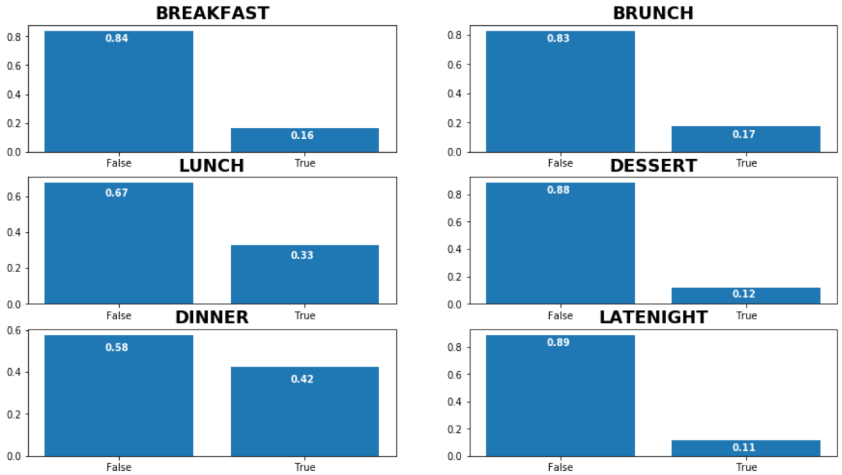
Documentation for the available data:
https://www.yelp.com/dataset/documentation/main

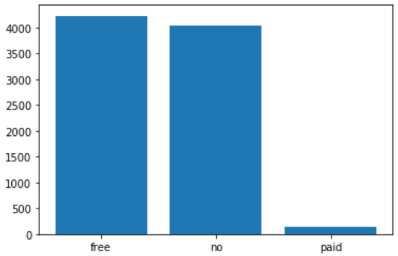
Understanding the Data

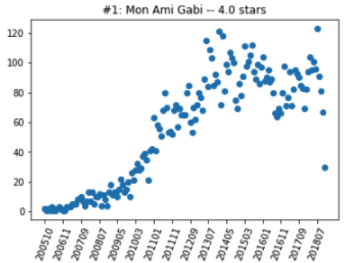
Available Datasets

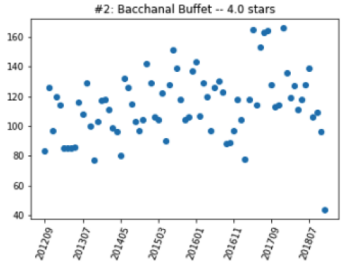
Dataset	Attributes Included
df_business	location data, attributes, and categories
df_photo	Photo id value, associated business id value, caption, category label of photo
df_review	Full review text, user_id of reviewer, review rating, review date
df_tip	Tips and quick suggestions written by users for a business
df_checkin	Business id values, comma-separated list of check-ins
df_user	User's name, number of reviews, votes sent and received, average rating of all reviews, date user joined Yelp, unique user id

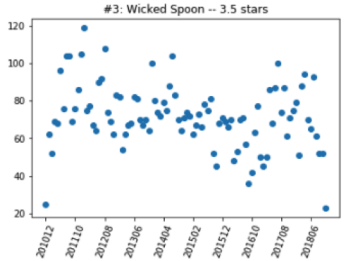
What's in the data?

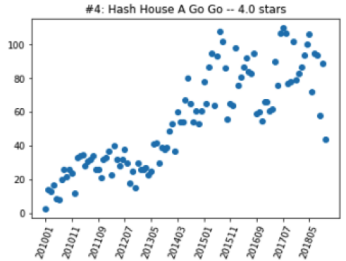




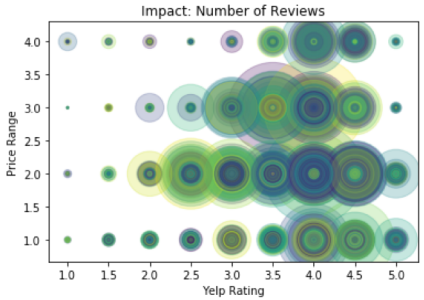


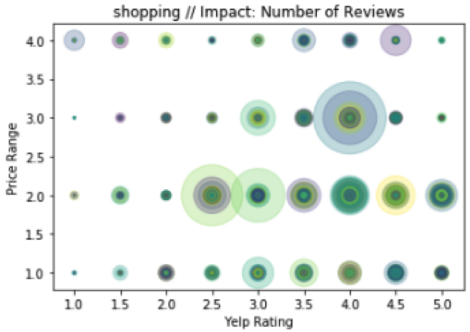


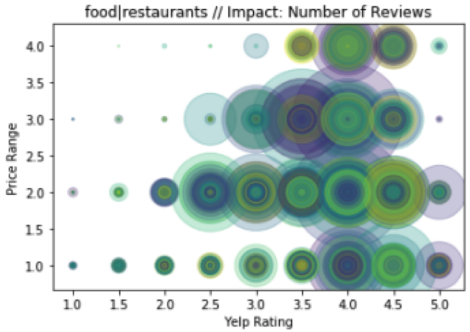


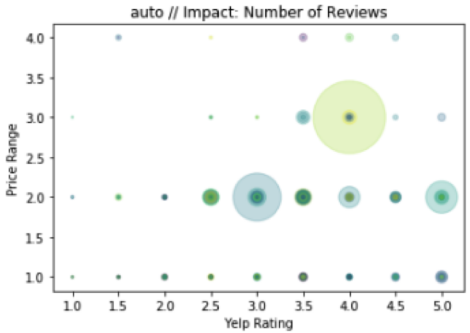


Trend analysis

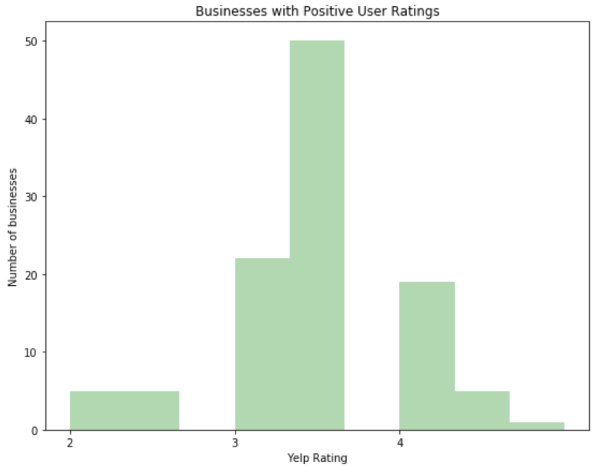




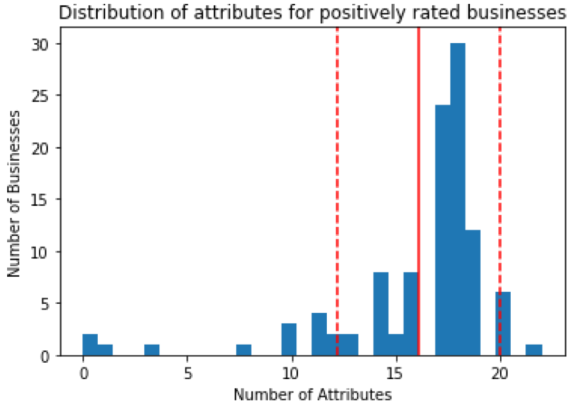


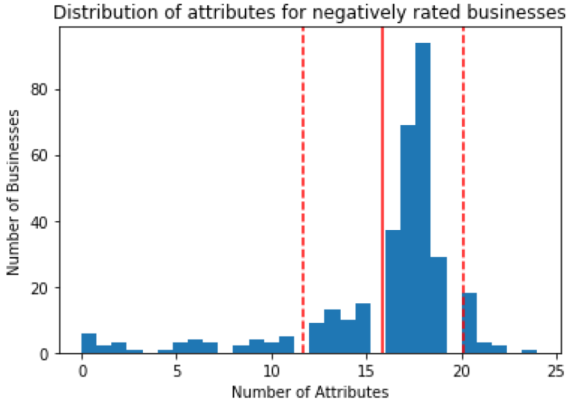


Statistical Analysis



Number of Attributes Posted

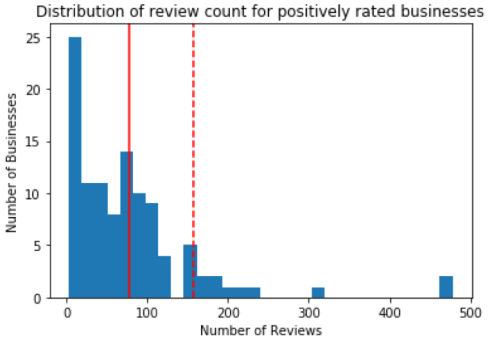


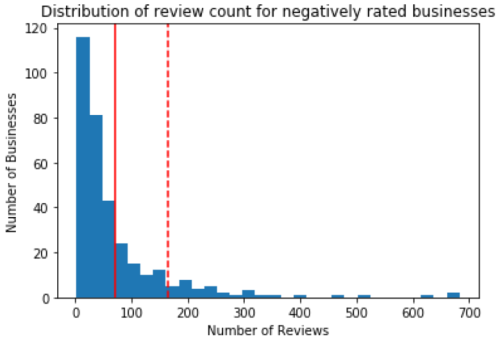


Null Hypothesis: The number of attributes posted on Yelp affects the Yelper's rating of the business
Alternative Hypothesis: The number of attributes posted on Yelp does not affect the Yelper's rating of the business

The p-value is very high at more than 60%. The null hypothesis - that the number of attributes posted affects a Yelper's rating - does not have to be rejected.

Number of Yelp Reviews

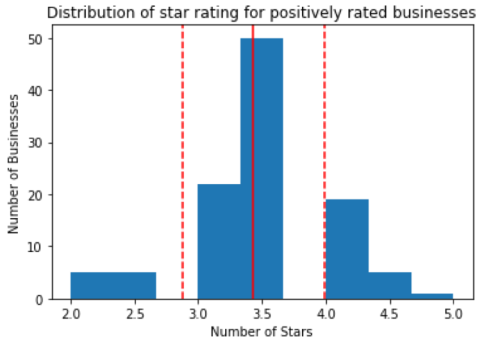


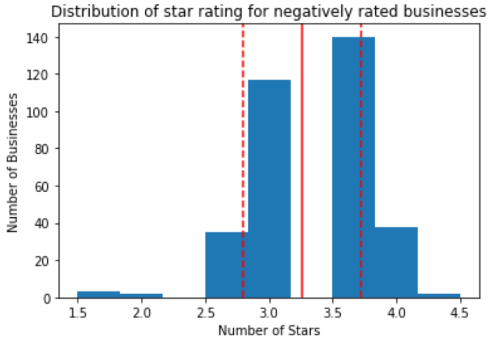


Null Hypothesis: The number of reviews posted on Yelp affects the Yelper's rating of the business
Alternative Hypothesis: The number of reviews posted on Yelp does not affect the Yelper's rating of the business

The p-value is very high at more than 50%. The null hypothesis - that the number of Yelp reviews online affects a Yelper's rating - does not have to be rejected.

Yelp Business Rating

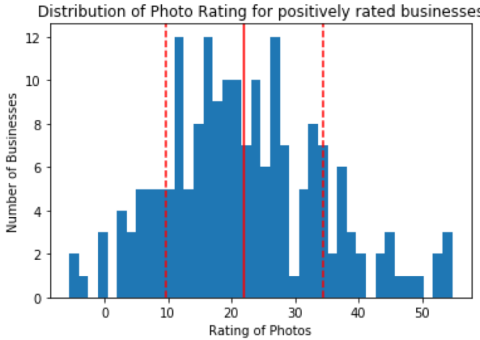


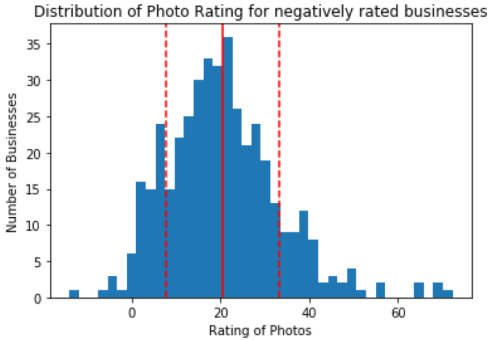


Null Hypothesis: The star rating posted on Yelp affects the Yelper's rating of the business
Alternative Hypothesis: The star rating posted on Yelp does not affect the Yelper's rating of the business

The p-value is less than 0.5%. The null hypothesis - that the business' posted Yelp rating affects a Yelper's rating - can be rejected.

Quality of Posted Photos

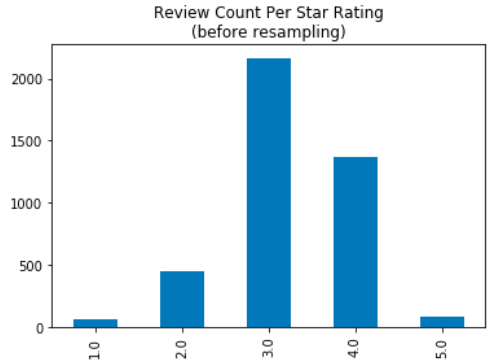


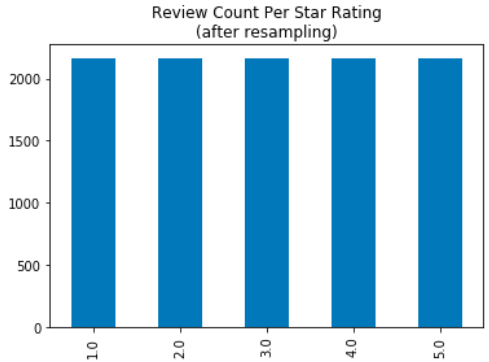


Null Hypothesis: The quality of the photos posted on the business' Yelp page affects the Yelper's rating of the business
Alternative Hypothesis: The quality of the photos posted on the business' Yelp page does not affect the Yelper's rating of the business

The p-value is somewhat high at about 17%. The null hypothesis - that the quality of the business photos affects a Yelper's rating - does not have to be rejected; however, further analysis needed...

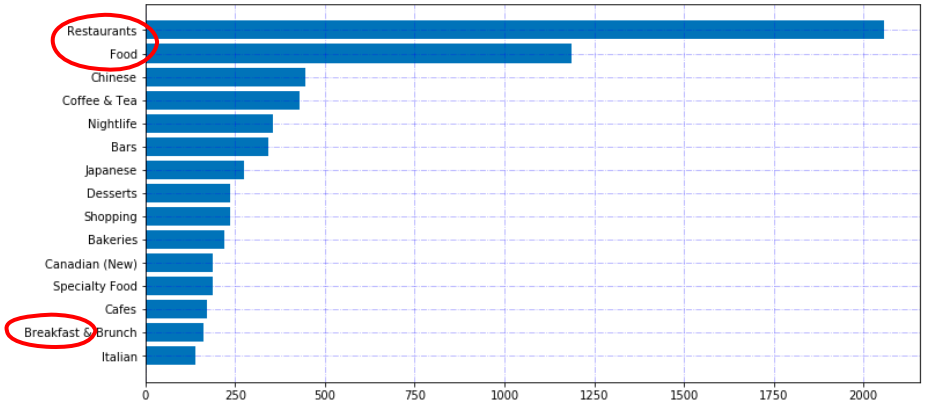
In-Depth Analysis (with Review Texts)





- Analysis of review texts from one Yelp user
- Yelp ratings based primarily on sentiment of review text
- Issue: disproportionate amount of reviews for each Yelp rating
- Potential Solutions: resampling, down sampling or combination of both
- Solution used: resampling to get same number of review texts as max review count of any Yelp rating score

Building a training set




```
# Instantiate Pipeline object: pl
pl = Pipeline([

        ('vec', CountVectorizer(ngram_range=(1,2),
                                analyzer='word',
                                token_pattern=TOKENS_ALPHANUMERIC,
                                stop_words=stopwords,#'english',
                                lowercase=True,
                                tokenizer=textblob_tokenizer
                               )),
        ('standard', StandardScaler(with_mean=False)),
        ('trunc', TruncatedSVD(n_components=500, n_iter=10)),
        ('clf', OneVsRestClassifier(LogisticRegression(solver='lbfgs', max_iter=500)))
    ])
```

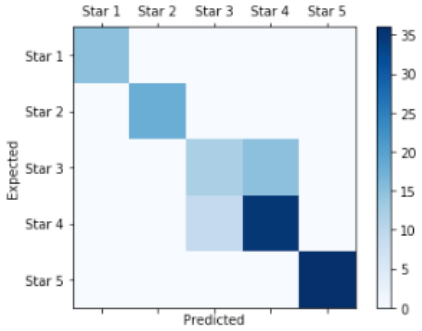
- Review texts taken from one, or more, categories from specific Yelp user of interest
- Create a pipeline to process and tokenize review texts

Validate model with test set

Classification Report

	precision	recall	f1-score	support
1.0	1.00	1.00	1.00	15
2.0	1.00	1.00	1.00	18
3.0	0.57	0.44	0.50	27
4.0	0.70	0.80	0.74	44
5.0	1.00	1.00	1.00	36
accuracy			0.83	140
macro avg	0.85	0.85	0.85	140
weighted avg	0.82	0.83	0.82	140

Confusion Matrix

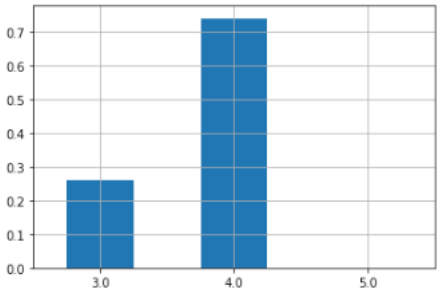


Top Features From Model

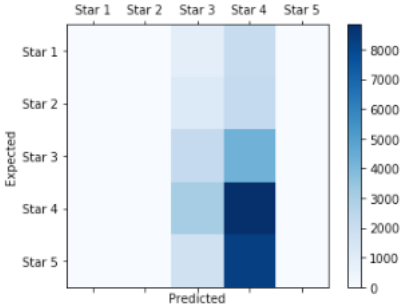
1. 1.5
2. guest
3. 2.5 star
4. 3 rel
5. pan
6. ethiopian food
7. pan
8. 2 8.99
9. 16 element
10. 10 great
11. 10.45 omelett
12. 12 hour
13. 28
14. 9.49
15. chicken nice
16. focaccia
17. 13 mac
18. 12 roast
19. landwer
20. 2016 appet

Testing Recommendation System

Predicted Yelp Ratings From Model

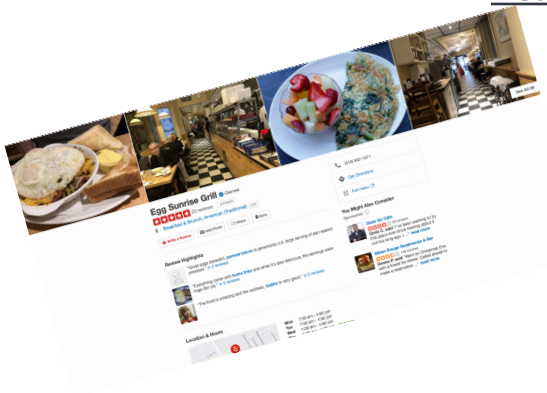


Confusion Matrix



- Compiled review texts from all other Yelp users within Yelp category specified for model

Recommended Businesses

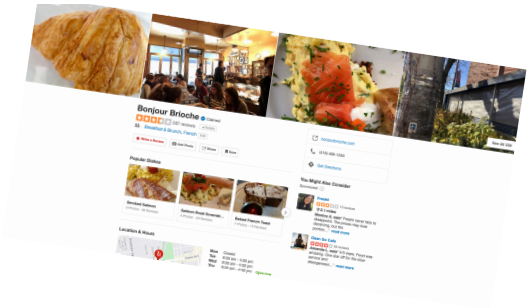


4-Star (Predicted) Rated Businesses

name	stars_business	review_count
Egg Sunrise Grill	5.0	15
Nord Bistro	5.0	8
Trius + Aim	5.0	5
Bailey's Cafe	5.0	5
Eggsmart	5.0	4
Café Polonez	4.5	225
Sunny Morning	4.5	186
Huevos Gourmet	4.5	163
Rashers	4.5	122
Takht-e Tavoos	4.5	114

5-Star (Predicted) Rated Business

name	stars_business	review_count
Bonjour Brioche	3.5	273



Recommendations of businesses predicted to have 4-star or 5-star rating based on other Yelper's review text

For Future Modifications:

- Analyse if other features (number of reviews, hours of operation, amenities offered, etc.) are better indicators to predict a Yelp rating score
- Add more stopwords to filter out 'non-essential' words

Thank you.