Lorem Ipsum Dolor

ARE THEY LOVIN' IT?
SENTIMENT ANALYSIS OF YELP! FAST FOOD REVIEWS

Sinead O'Connor
Springboard Science Career Track
April 2020

THE PROBLEM

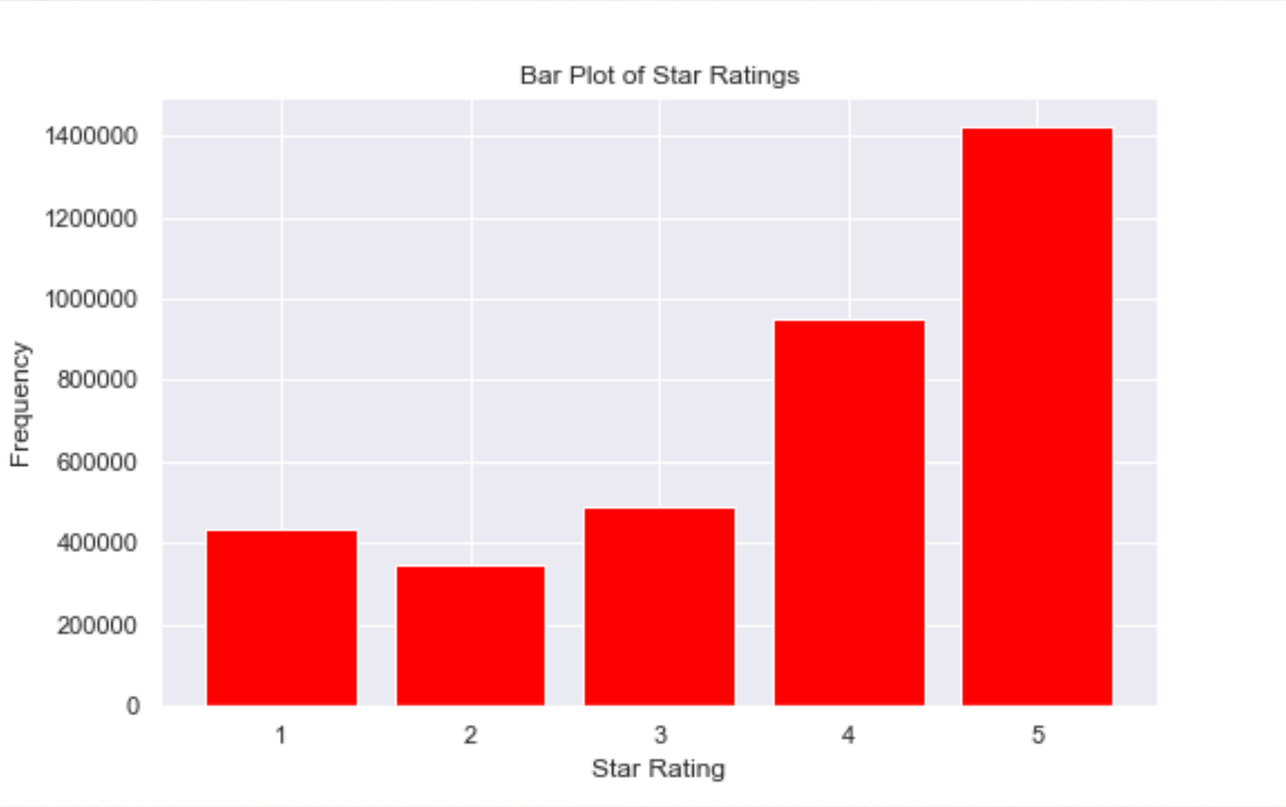
- Goal: Help fast food businesses understand how the public is reacting to their restaurants on social media
- Use supervised learning to predict probability of being associated with a star rating ranging from 1-5

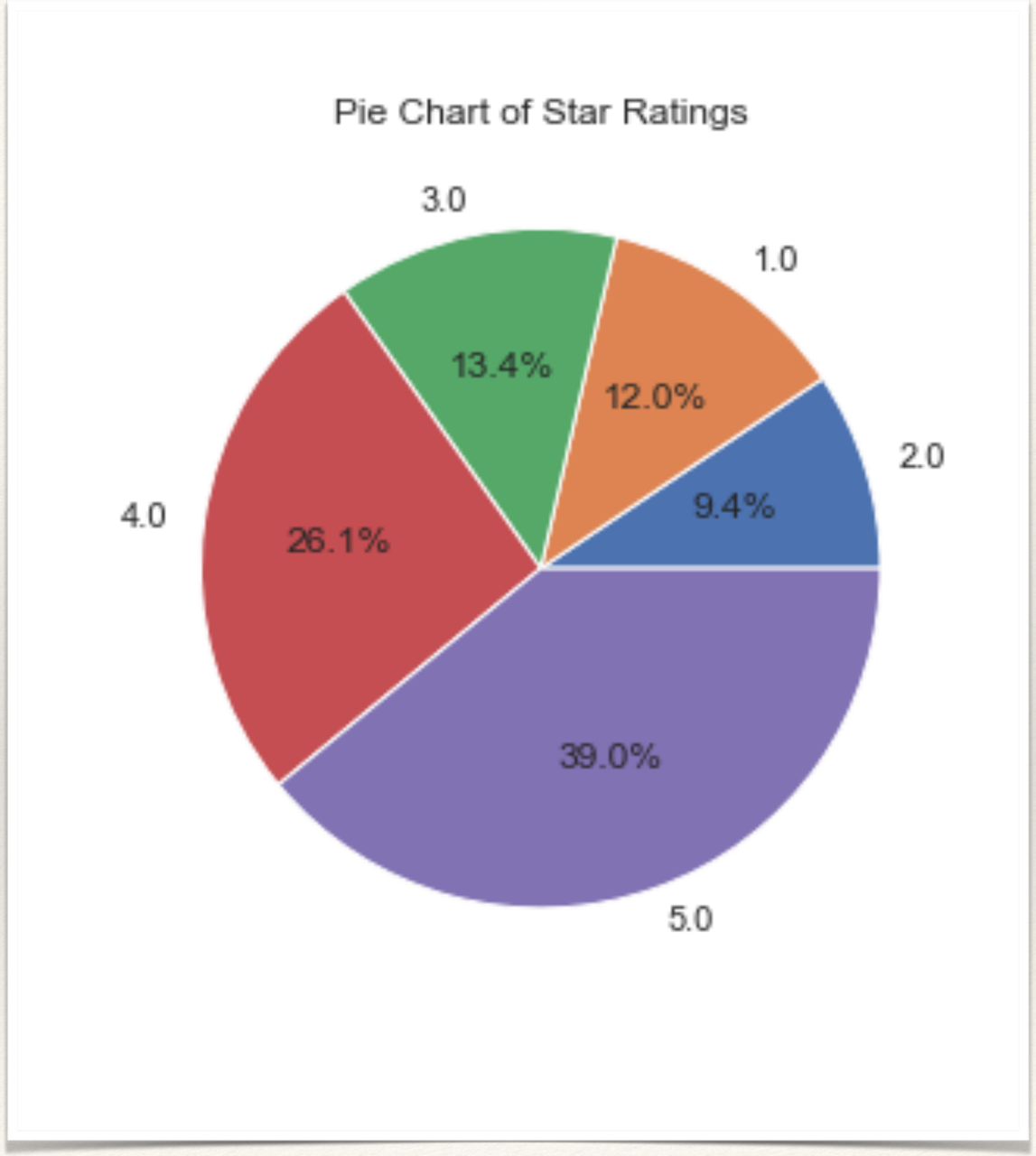
THE DATA

- Data from the Yelp Challenge 2019
- Business file:
  - Includes business name, id, address, number of reviews, and the categories under which the business falls
- Review file:
  - Includes user id of the reviewer, the business id, the rating (out of 5 stars), and the text of the review
- Filtered to obtain just restaurants and reviews of restaurant

EXPLORATORY DATA ANALYSIS

RESTAURANT REVIEWS BY STAR RATING

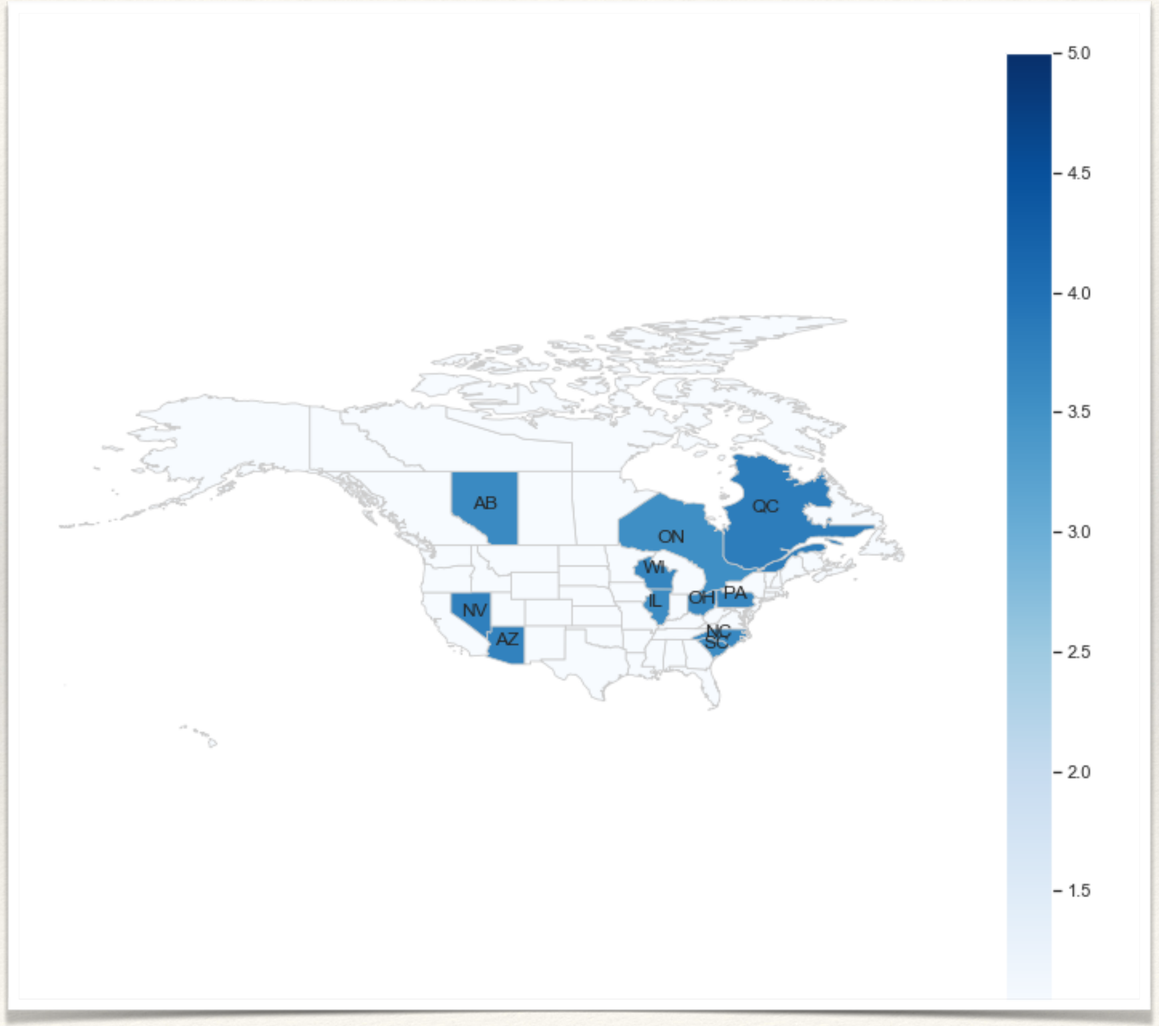




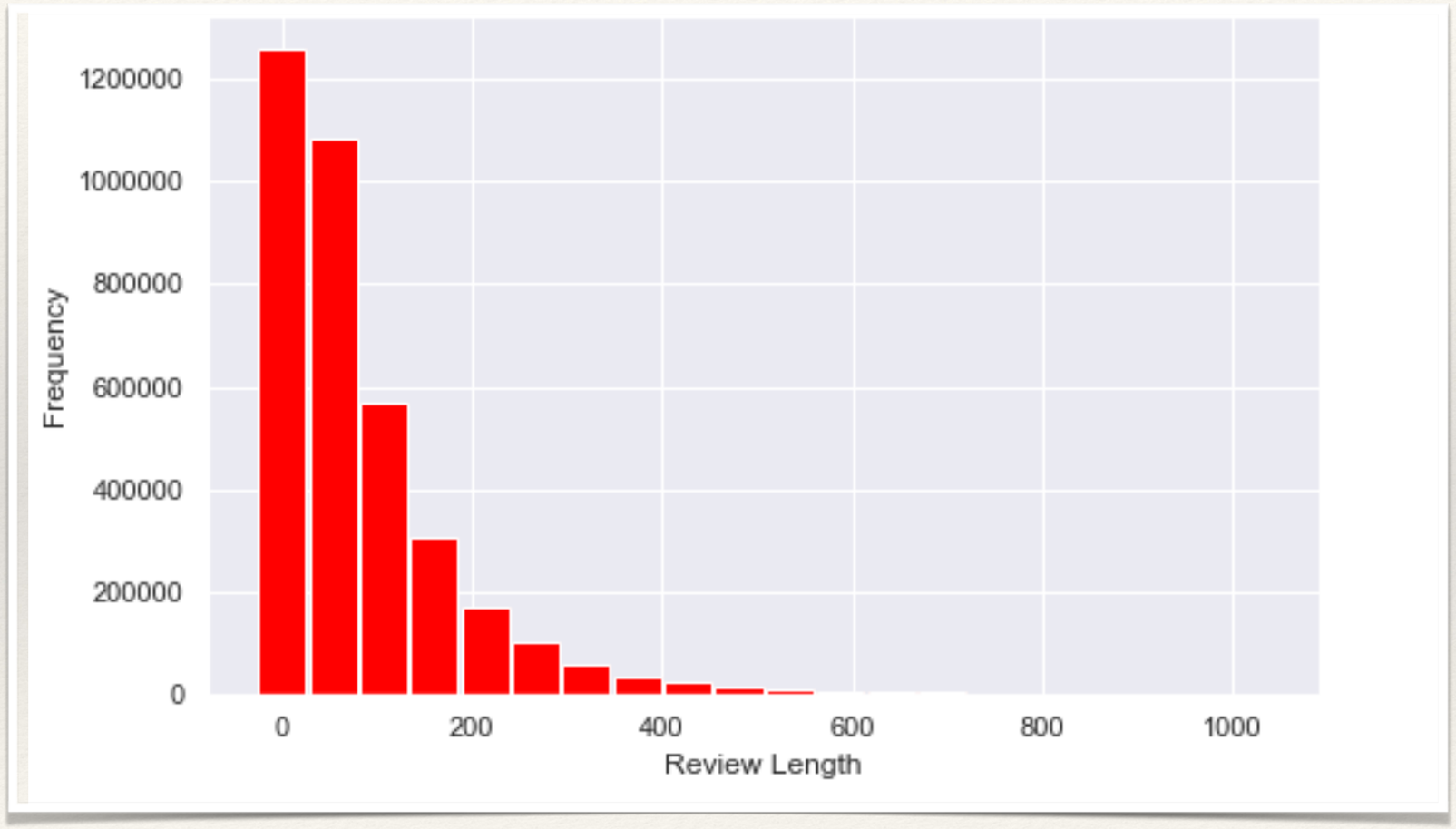
AVERAGE RESTAURANT REVIEW BY STATE/PROVINCE

(for states/provinces with more than 10,000 reviews)

STATE/PROVINCE	AVG REST. STAR RATING	NUMBER OF REVIEWS
ONTARIO, CAN	3.541205	518,338
SOUTH CAROLINA, USA	3.579360	13,905
ILLINOIS, USA	3.588263	25,560
ALBERTA, CAN	3.624103	56,010
NORTH CAROLINA, USA	3.650980	238,187
OHIO, USA	3.666499	202,569
PENNSYLVANIA, USA	3.685806	178,864
WISCONSIN, USA	3.689474	80,573
ARIZONA, USA	3.737472	1,056,142
NEVADA, USA	3.776024	1,154,070
QUEBEC, USA	3.825868	119,105



Histogram of Word Count Per Restaurant Review

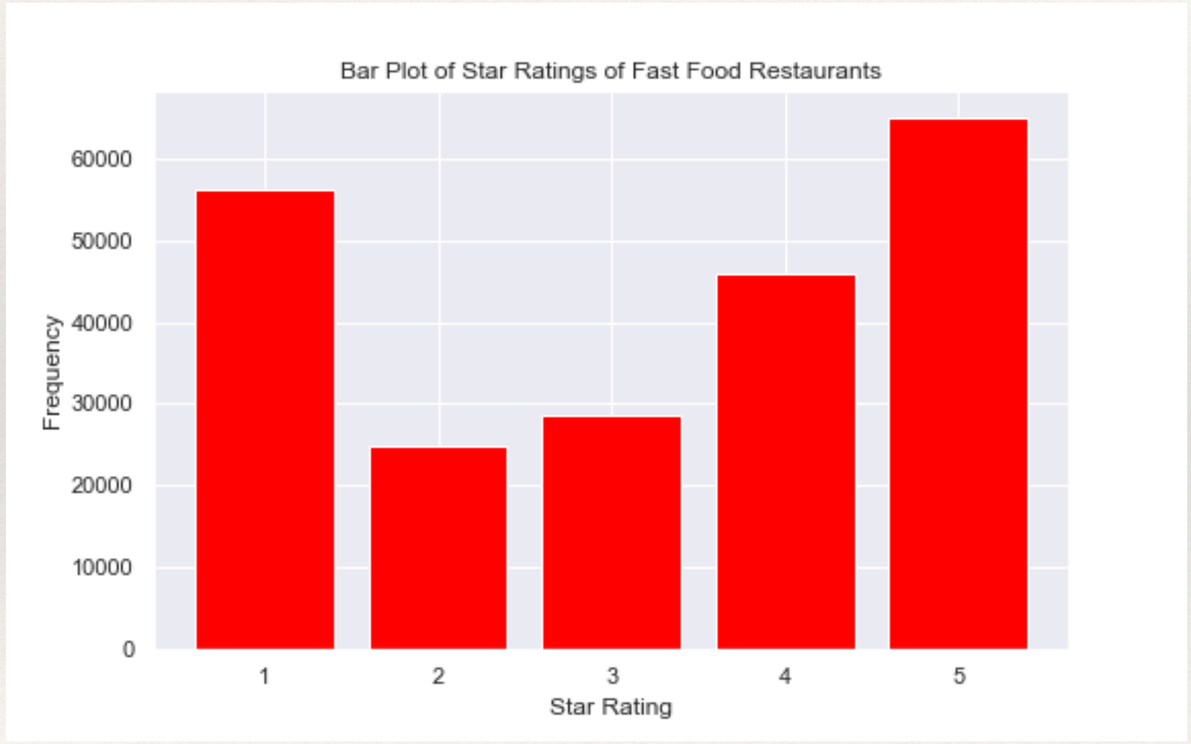


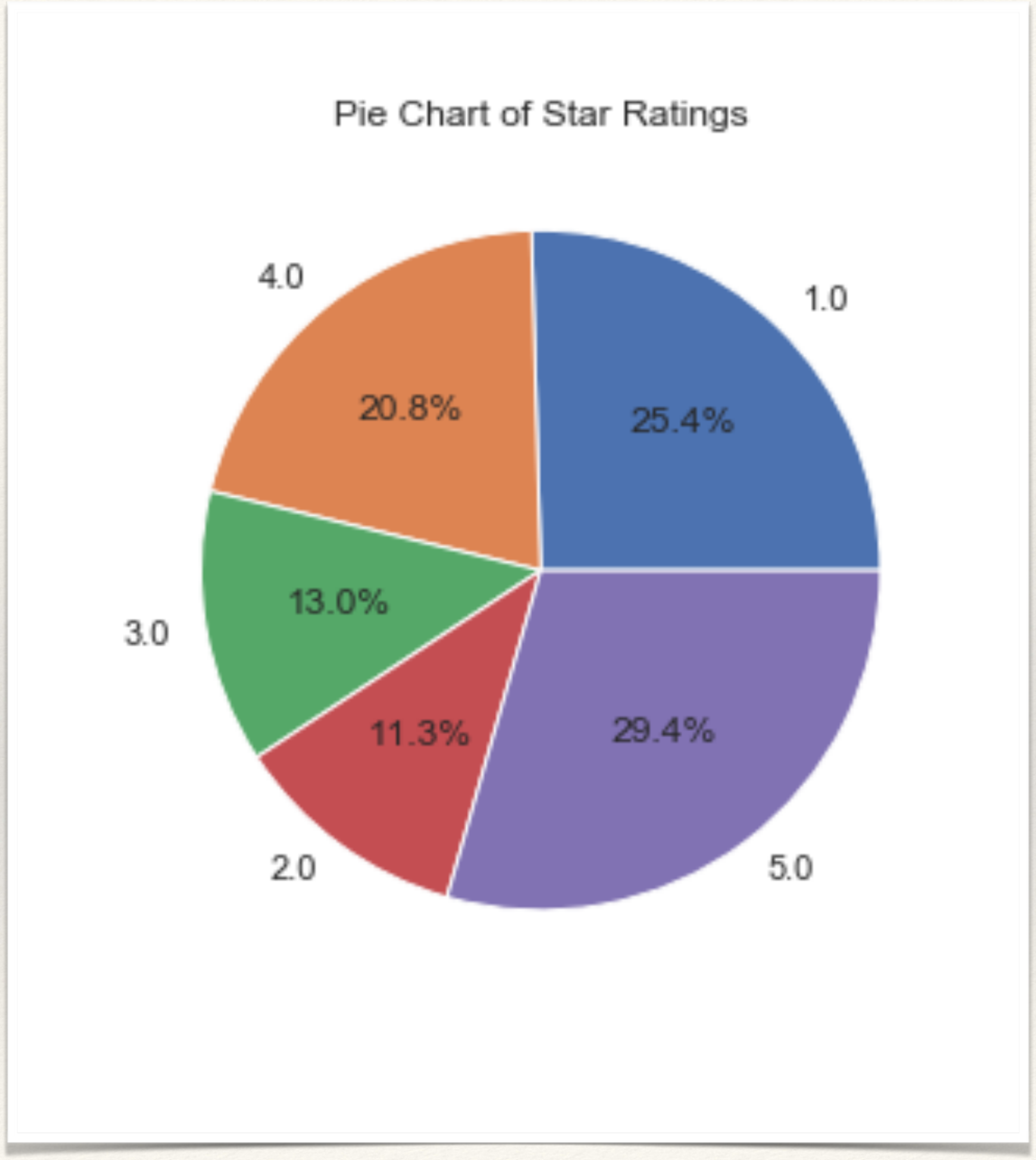
Time Series Plot of Avg Star Rating -

Fast Food Restaurants vs Other Restaurants

- Avg Fast Food Rating: 3.175 starts (220,507 reviews). Fast food ratings seem to have decreased overall from about 3.6 in 2011 to about 2.9 by the end of 2018
- Avg Rating for Other Restaurants: 3.742 stars (3.4 million reviews). Other restaurants ratings experienced overall increase in average reviews per month from about 3.6 in 2013 to about 3.8 at the end of 2018

NUMBER OF FAST FOOD REVIEWS BY STAR RATING





WORD CLOUDS FOR FAST FOOD REVIEWS BY STAR RATING

1-Star Cloud

2-Star Cloud

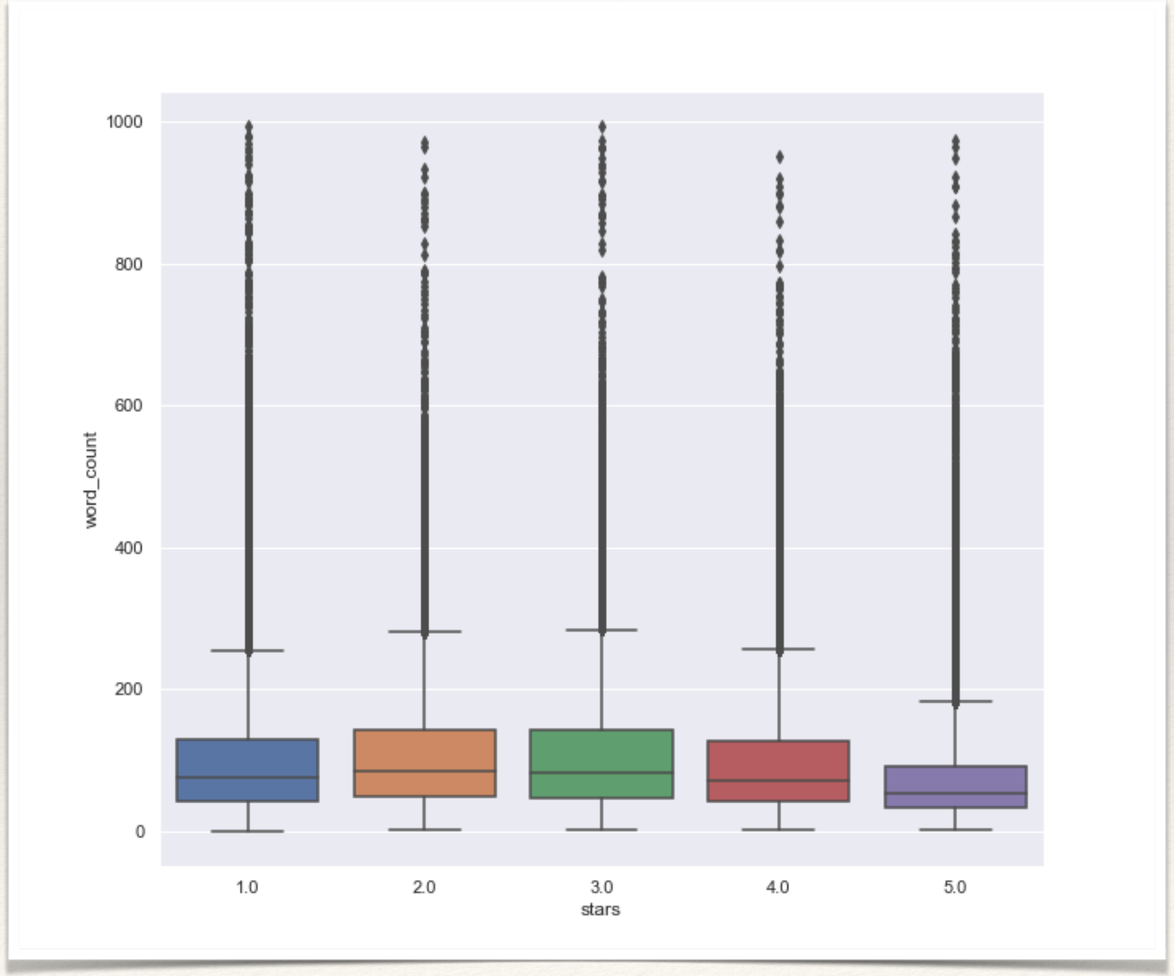
3-Star Cloud

4-Star Cloud

5-Star Cloud

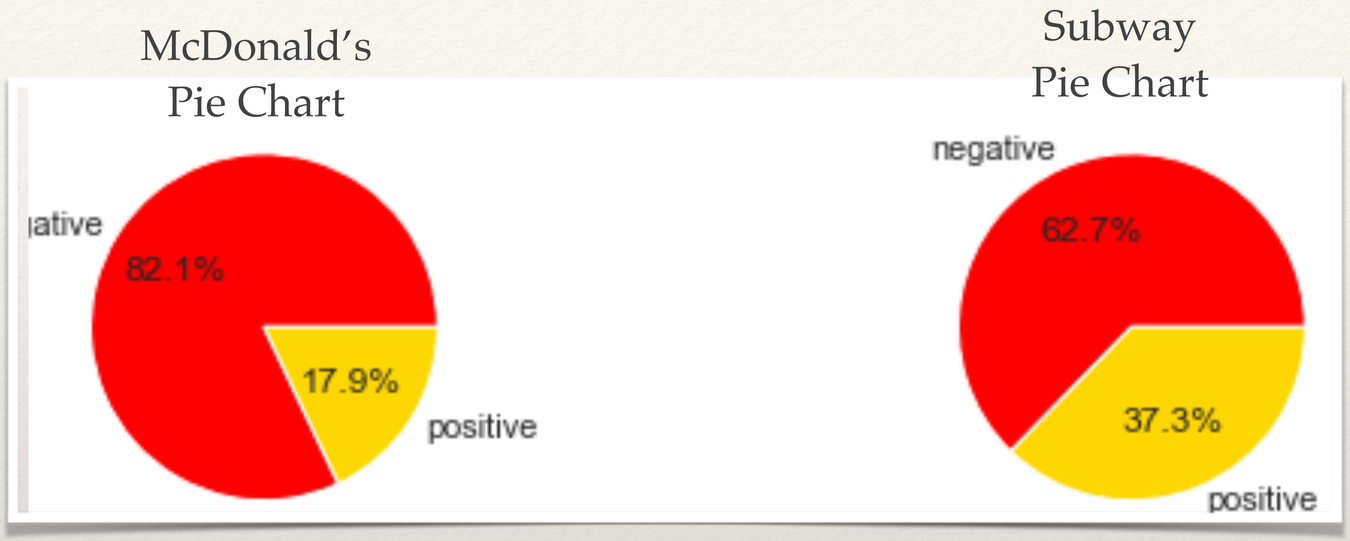
FAST FOOD REVIEW LENGTH BY STAR RATING

- It appears 5-star reviews tend to have the fewest number of words
- Statistically significant difference between at the average review length for at least one of the star ratings



Proportion of Positive Review-

McDonald's vs Subway

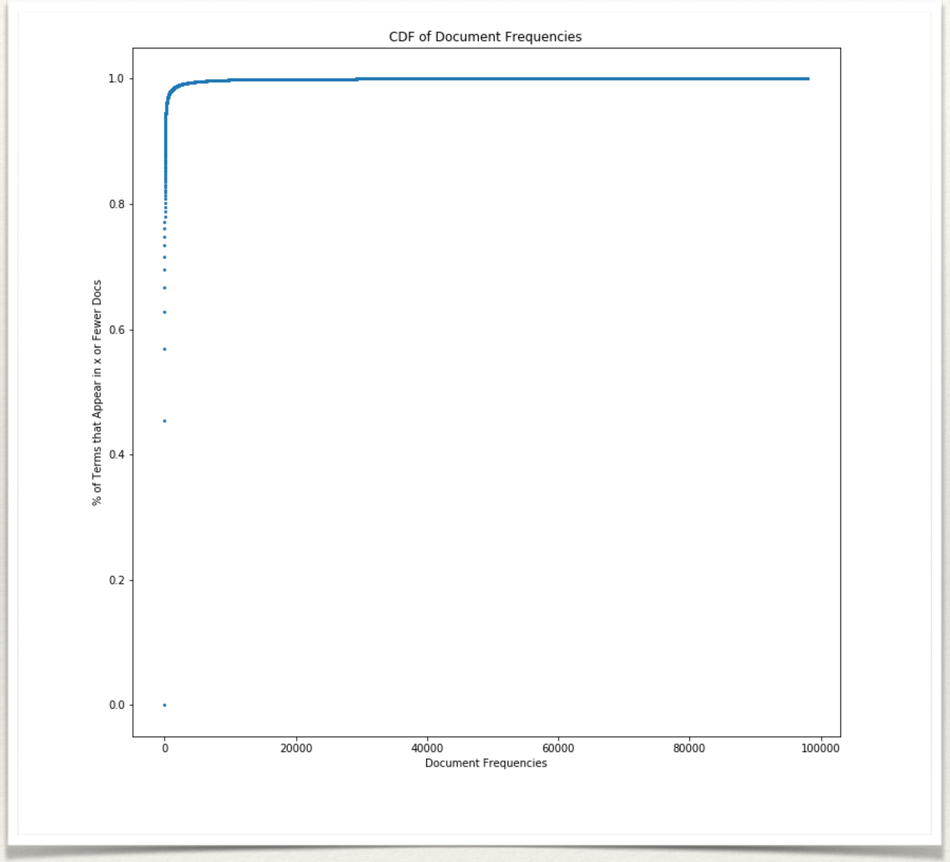


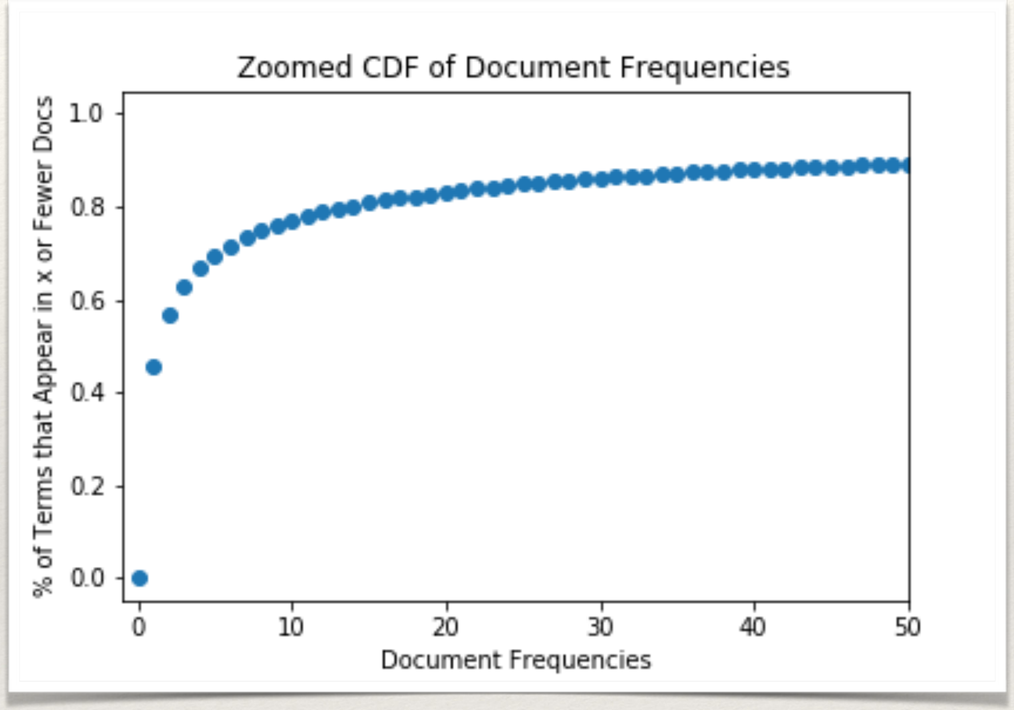
- Positive Review: 4 or 5 star review
- Negative Review: 1, 2, or 3 star review
- There is statistically significant difference between the proportion of positive reviews for McDonald's and for Subway

MACHINE LEARNING

5-Class Classification-
Naive Bayes & Logistic Regression
Feature Engineering using CountVectorizer on Unigrams from the Reviews

HYPERPARAMETER TUNING-
CHOOSING A MIN AND MAX DOCUMENT FREQUENCY





- Based on the CDF of Document Frequencies, I chose min_df = 3, max_df = 70,000

Results

Naive Bayes

```
[Training Classification Report:]
              precision    recall  f1-score   support

         1.0       0.73      0.82      0.77     42004
         2.0       0.42      0.35      0.38     18620
         3.0       0.50      0.44      0.47     21441
         4.0       0.55      0.52      0.54     34079
         5.0       0.73      0.78      0.75     48604

   micro avg       0.64      0.64      0.64    164748
   macro avg       0.59      0.58      0.58    164748
weighted avg       0.63      0.64      0.63    164748

Training Accuracy:  0.6413188627479545

[Test Classification Report:]
              precision    recall  f1-score   support

         1.0       0.72      0.80      0.76     13981
         2.0       0.34      0.27      0.30      6207
         3.0       0.40      0.35      0.37      7093
         4.0       0.48      0.46      0.47     11552
         5.0       0.70      0.74      0.72     16083

   micro avg       0.59      0.59      0.59     54916
   macro avg       0.53      0.53      0.53     54916
weighted avg       0.58      0.59      0.59     54916

Test Accuracy:  0.5944169276713526
```

Logistic Regression

```
[Training Classification Report:]
              precision    recall  f1-score   support

         1.0       0.77      0.91      0.84     42004
         2.0       0.61      0.38      0.47     18620
         3.0       0.60      0.46      0.52     21441
         4.0       0.60      0.52      0.56     34079
         5.0       0.72      0.85      0.78     48604

   micro avg       0.69      0.69      0.69    164748
   macro avg       0.66      0.62      0.63    164748
weighted avg       0.68      0.69      0.68    164748

Training Accuracy:  0.6946184475684075

[Test Classification Report:]
              precision    recall  f1-score   support

         1.0       0.75      0.88      0.81     13981
         2.0       0.43      0.27      0.33      6207
         3.0       0.44      0.35      0.39      7093
         4.0       0.51      0.43      0.46     11552
         5.0       0.68      0.82      0.74     16083

   micro avg       0.63      0.63      0.63     54916
   macro avg       0.56      0.55      0.55     54916
weighted avg       0.60      0.63      0.61     54916

Test Accuracy:  0.627831597348678
```

- Logistic Regression performed better with respect to each test metric
- In both models, we observe imbalance in the classes of the dataset
- Overfitting for some of the classes

Logistic Regression Confusion Matrix

		PREDICTED RATING				
		1 Star	**2 Star**	**3 Star**	**4 Star**	**5 Star**
TRUE RATING	**1 Star**	12307	949	292	143	290
	2 Star	2736	1660	1114	366	331
	3 Star	869	945	2479	1895	905
	4 Star	345	219	1365	4914	4709
	5 Star	260	100	333	2272	13118

REDEFINING THE PROBLEM

- To address the problem of imbalance we will label 4- and 5-star reviews as “positive” and 1-, 2-, and 3-star reviews as “negative”
- BINARY CLASSIFICATION

SUMMARY OF MODELS USED & RESULTS

Note: The same train/test split was used for each model

MODEL	CLASS	PRECISION	RECALL	F1	SUPPORT	LOG LOSS	ACCURACY
NaiveBayes-CountVec, Unigrams, min_df = 3, max_df = 70K	Negative	0.88	0.79	0.83	27281	0.9544	0.8439
	Positive	0.81	0.89	0.85	27635		
LogReg-CountVec, Unigrams, min_df = 3, max_df = 70K	Negative	0.89	0.88	0.88	27281	0.3008	0.8847
	Positive	0.88	0.89	0.89	27635		
RandForest-CountVec, Unigrams, min_df = 3, max_df = 70K	Negative	0.87	0.87	0.87	27281	0.3818	0.8667
	Positive	0.87	0.87	0.87	27635		
LogReg-TfidfVec, Unigrams, min_df = 3, max_df = 70K	Negative	0.89	0.88	0.89	27281	0.277	0.8864
	Positive	0.88	0.89	0.89	27635		
LogReg-TfidfVec, Uni & bigrams, min_df = 35, max_df = 45K	Negative	0.91	0.90	0.91	27281	0.2419	0.9067
	Positive	0.91	0.91	0.91	27635		
RandForest-TfidfVec, Uni & bigrams, min_df = 35, max_df = 45K	Negative	0.88	0.88	0.88	27281	0.3690	0.8772
	Positive	0.88	0.88	0.88	27635		

BEST MODEL WITH RESPECT TO ALL TEST METRICS:

Logistic Regression, TfidfVectorizer, Unigrams & Bigrams

```
[Training Classification Report:]
              precision    recall  f1-score   support

           0       0.93      0.92      0.93     82065
           1       0.92      0.93      0.93     82683

   micro avg       0.93      0.93      0.93    164748
   macro avg       0.93      0.93      0.93    164748
weighted avg       0.93      0.93      0.93    164748

Training Accuracy:  0.9269611770704348

[Test Classification Report:]
              precision    recall  f1-score   support

           0       0.91      0.90      0.91     27281
           1       0.91      0.91      0.91     27635

   micro avg       0.91      0.91      0.91     54916
   macro avg       0.91      0.91      0.91     54916
weighted avg       0.91      0.91      0.91     54916

Test Accuracy:  0.906784907859276

AUC:  0.9664413624942066

Log loss:  0.24192484745324933
```

CONFUSION MATRIX		PREDICTED RATING	
		NEGATIVE	POSITIVE
TRUE RATING	NEGATIVE	24,686	2,595
	POSITIVE	2,524	25,111

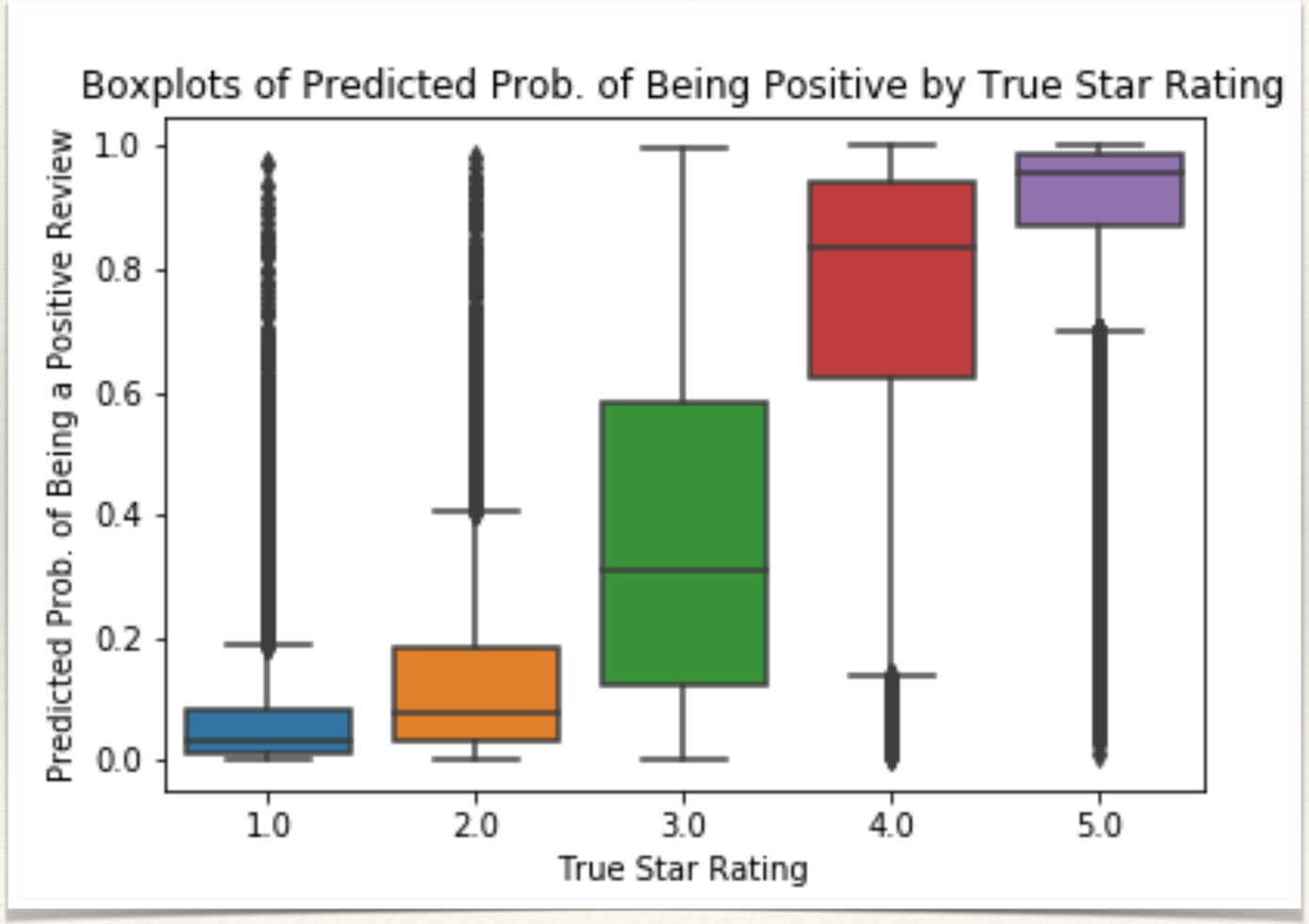
Business Interpretation of Choosing Model Based on Certain Test Metrics

Test Metric	Current Best Value	Business Interpretation
Accuracy	0.9068	90.68% of total predictions were correct. It may be reasonable to choose the model with the highest accuracy since the classes are pretty balanced.
Precision	0.91	Both the positive and negative class had precision of 0.91 meaning when the model predicted the review was positive it was correct 91% of the time, and when it predicted the review was negative, it was also correct 91% of the time.
Recall	0.91	The model has 0.90 recall on the negative class (class 0) and 0.91 on the positive class (class 1), meaning when the review was truly negative, the model correctly predicted it was negative 90% of the time, and when the review was truly positive, the model was correct 91% of the time.
F1 Score	0.91	This is a combination of precision and recall.

Most Positive & Most Negative Words

15 Most Positive Tokens			15 Most Negative Tokens	
Coef.	**Pos. token**	**Ranking**	**Neg. token**	**Coef.**
12.707526	**delicious**	**1**	**worst**	-12.022870
12.185547	**great**	**2**	**rude**	-9.116245
10.390878	**best**	**3**	**horrible**	-8.199477
10.221421	**amazing**	**4**	**bland**	-7.820095
9.521001	**awesome**	**5**	**terrible**	-7.429462
8.990877	**excellent**	**6**	**slow**	-6.952828
8.589614	**love**	**7**	**ok**	-6.472946
7.219532	**so good**	**8**	**not worth**	-6.406678
6.728885	**perfect**	**9**	**disappointing**	-6.367086
6.477385	**fantastic**	**10**	**mediocre**	-6.361814
6.351732	**really good**	**11**	**at best**	-6.109693
6.268007	**fresh**	**12**	**disgusting**	-6.077620
6.205882	**friendly**	**13**	**slowest**	-5.999895
5.306896	**loved**	**14**	**average**	-5.961712
5.113817	**yummy**	**15**	**dry**	-5.825962

Predicted Probability of Review Being Positive By Underlying True Star Rating



- Clearly observe that overall reviews that were given 5-star ratings tended to have the highest probability of being positive, 4-star the second highest, 3-star the 3rd highest, 2-star the 4th highest, and 1-star the lowest.
- 3-star reviews had largest spread, which is expected

RECOMMENDATIONS TO THE CLIENT

- For now, use the Logistic Regression Logistic Regression model fit using term frequency-inverse document frequency vectorization with unigrams and bigrams, as it performs best with respect to every test metric measured.
- I also suggest filtering comments with very low predicted probability of the comments being positive will allow clients to quickly identify very negative feedback that may require rapid responses from the business's social media department. It will also allow clients to identify reviews with high probabilities of being positive, so they can acknowledge the commenter's kind words.