A Sentiment Analysis of Yelp Reviews

Springboard Data Science Career Track
Capstone Project 1

Introduction

Goals of this Project

- Provide an efficient means of discriminating between positive and negative restaurant reviews.
- Specifically, construct models that reliably classify reviews as positive or negative based on their content.

Motivation for the Project

- Yelp is useful for customers, but could be more useful for business owners.
- A proposal for how to add value for business owners: enable business owners to have greater awareness of customer experiences.

Data Acquisition and Wrangling

The Data

- Approximately 6.5 million reviews of a variety of different businesses
- Acquired directly from Yelp

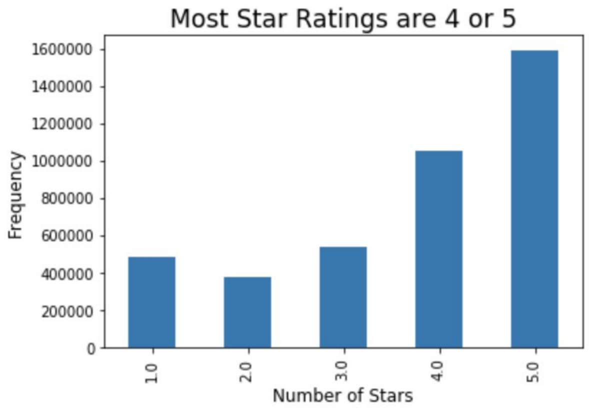
Data Wrangling

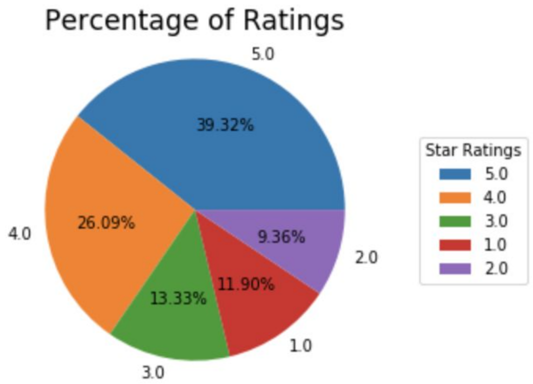
1. Restricted data to reviews of *restaurants*
2. Isolated key variables for analysis: ‘business_id’ | ‘stars’ | ‘text’
3. Eliminated rows with missing or duplicate values
4. Restricted data to reviews in the English language

Storytelling and Inferential Statistics

Data Exploration

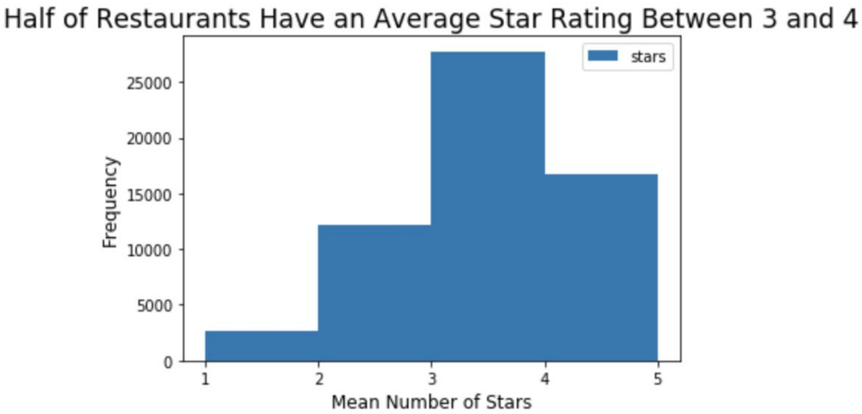
Star Rating Data





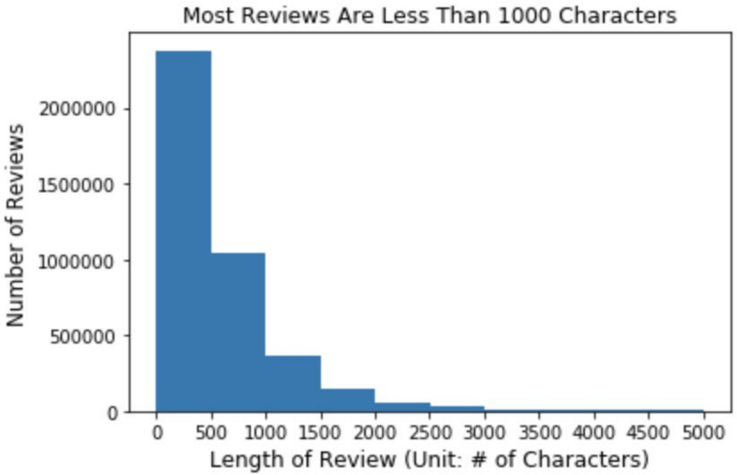
Data Exploration

Star Rating Data cont'd



Data Exploration

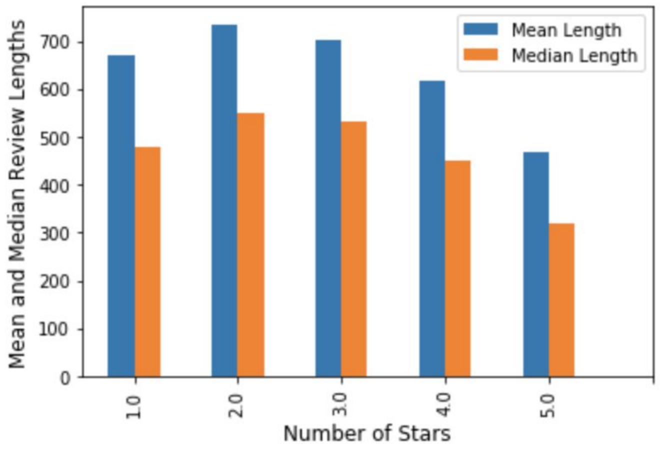
Review Data



Data Exploration

Interactions Between Star Rating Data & Review Data

Star Rating	Mean Length
1.0	669
2.0	734
3.0	702
4.0	616
5.0	469



Statistical Analysis

- Question: Are the differences in mean review length between star ratings statistically significant?
- One-way analysis of variance (ANOVA)
- Null Hypothesis: The mean review lengths are the same for each star rating.
- Alternative Hypothesis: At least one of the mean review lengths differs from another.
- Alpha = 0.05
- Results: | F-statistic: 36.22 | p-value ≅ 0 |
- Conclusion: at least one of the mean review lengths differs from another in the population.

Statistical Analysis cont'd

- Question: Which differences in mean review length between star ratings are statistically significant?
- T-tests (using Bonferroni correction to control for familywise error) of the differences in mean review length between six pairs of star ratings: 1&4, 1&5, 2&4, 2&5, 3&4, 3&5
- Conclusions: the differences between each of these pairs is statistically significant (except between 1& 4 stars)

Baseline Modeling

Text Preprocessing and Representation

- Make all words lowercase
- Remove punctuation
- Remove stopwords

- Represented with a frequency vectorization (CountVectorizer)

Modeling Variations

- Logistic Regression
- Tuned and tested four variations:

<u>Version 1</u> • No Stratification in train-test-split • L2 Regularization for fitting the model	<u>Version 2</u> • No Stratification in train-test-split • L1 Regularization for fitting the model
<u>Version 3</u> • Stratification in train-test-split • L2 Regularization for fitting the model	<u>Version 4</u> • Stratification in train-test-split • L1 Regularization for fitting the model

- Version 4 performed the best

Extended Modeling

Modeling Variations

- Random Forest
- Varied four modeling parameters:
    1. Text Representation (used both frequency vectorization–CountVectorizer–and weighted frequency vectorization–TfidfVectorizer)
    2. Number of Estimators/Decision Trees (from 50 to 150)
    3. Number of Features (all, log2, sqrt)
    4. Measure of Node Purity (gini, entropy)

- Best performing variation:
    1. Frequency vectorization (CountVectorizer)
    2. 100 estimators/decision trees
    3. Sqrt number of features
    4. Gini index as measure of node purity

Summary of Findings

Model Performance

- Test Data: 300,000 observations (103,656 in negative class, 196,344 in positive class)

Performance Metrics:	Logistic Regression	Random Forest
Accuracy	0.90	0.87
Precision	Positive Class: 0.90 Negative Class: 0.89	Positive Class: 0.85 Negative Class: 0.90
Recall	Positive Class: 0.95 Negative Class: 0.81	Positive Class: 0.96 Negative Class: 0.68
F1-Score	Positive Class: 0.92 Negative Class: 0.84	Positive Class: 0.90 Negative Class: 0.78

- Key Finding: Logistic Regression outperformed Random Forest with respect to almost every performance metric.

Conclusions and Future Work

Future Directions

- Using stemming or lemmatization in the text preprocessing
- Using n-grams (for different values of n) as the features in the vector representation of the text
- Adding additional features to influence classification (e.g. length of review, location of business, demographics of reviewer, etc.)
- Using different values for minimum or maximum document frequency for a word to be included as a feature
- Performing train-test-split *before* creating vector representation of text
- Using other algorithms for classification (e.g. Naive Bayes, Support Vector Machines, Neural Nets, etc.)
- Develop an ensemble of different models.

Recommendations for Client

Recommendations

- Continue refining the Logistic Regression model developed in this project.
- Explore some of the future directions mentioned previously.