Mid-Semester Progress Report

DSA5900 – Spring 2021

Purushotham Vadde

03/24/2021

Introduction

93% of customers will read reviews of local businesses to determine their quality(**BrightLocal**). Social media is filled with a lot of reviews there are thousands of reviews available for each business on social media and we can also see the responses from the business owners for the reviews given by the public. In this project, we will develop a Data science approach to determine what causes business owners to respond to a comment on social media and how likely is a business owner to respond to a comment about their firm.

The outcome of this project will help in providing the key insights about the motivations and actions of a business owner who respond to online reviews for their firms. Developing a machine learning-based model and researching the root causes of engagement by owners will also help to expose other interesting insights into the use of social media by the firms.

Objectives

Technical Project Objectives:

This project aims to build and implement a supervised machine learning model to classify the business reviews on the Yelp website, we will classify the public reviews for which the business owner going to respond.

Individual Learning Objectives:

By this project, we will be developing skills in the field of Natural Language Processing, Machine learning, and Deep learning using packages such as NLTK, sci-kit learn, TensorFlow.

The project will also provide me with skills such as:

- Learning Machine learning and Deep Learning Algorithms for classification of data.
- Hyperparameter Tuning to find the optimal parameter for Machine learning Algorithms.
- Learning the Data Visualization skills.
- Storytelling skills from the insights got from the project.

Data

Ingestion

We collected the reviews from Yelp website through web scraping using tools like Beautiful soup and Selenium. The reviews are collected for multiple cities such as Seattle, Los Angeles, New York, Miami, Palm beach, Oklahoma City in different categories like Plumbing, Painters, Auto Repairs etc. The collected business reviews are stored into a CSV format file.

The data frame consists of 24 features related to the business and customer from the yelp website which includes the customer review and the response for the business owner for the review.

Challenges:

1. By yelp api we can get maximum of 3 reviews only.
2. To scrape the data from website most of the HTML elements will render only after the yelp website loaded in the browser, by using the urllib we will get very minimal data.
3. By searching the business in specific location will results the maximum of 240 business and most of the business will repeated.
4. The HTML element tags are dynamic in nature and keeps changing frequently.

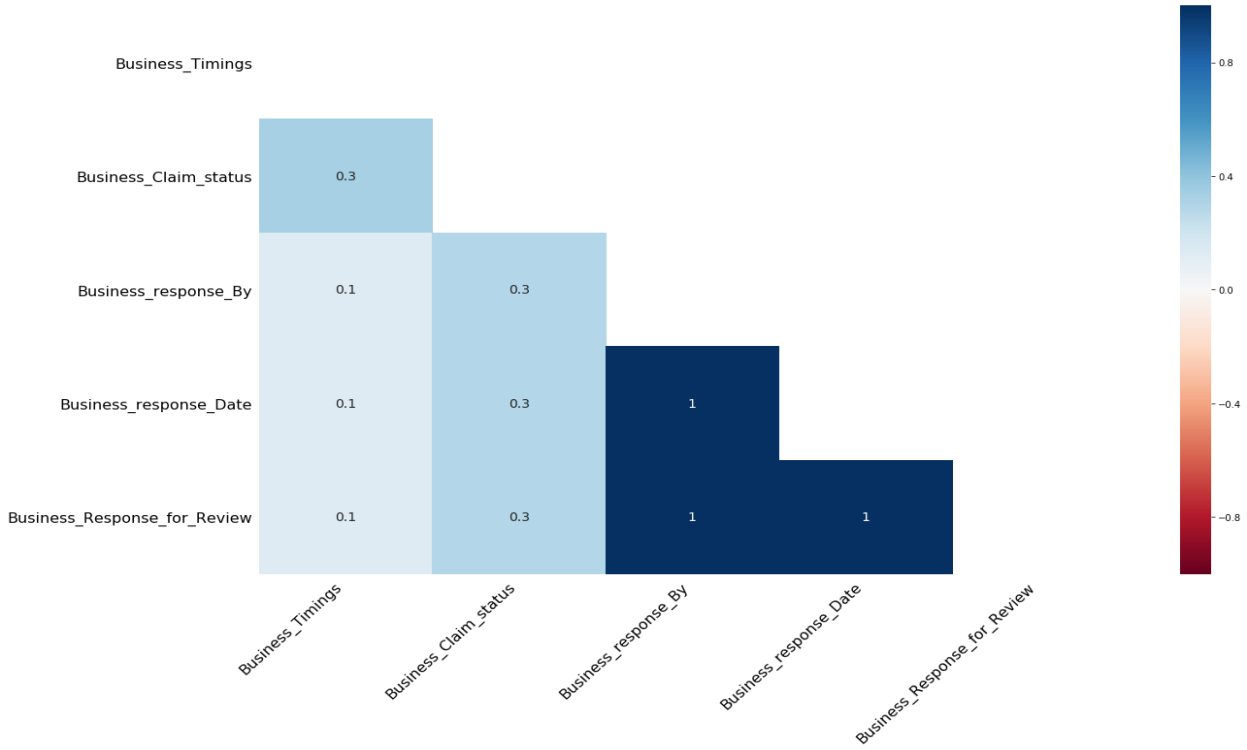
Exploration

In the below figure we can see the 24 features that we extracted from the yelp website through web scraping.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 664 entries, 0 to 663
Data columns (total 25 columns):
Unnamed: 0                          664 non-null int64
Business_Name                       664 non-null object
Business_Address                    664 non-null object
Business_ReviewCount                664 non-null int64
Business_Rating                     664 non-null float64
Business_Photos_Count               0 non-null float64
Business_Timings                    550 non-null object
Business_Claim_status               282 non-null object
Customer_Name                       664 non-null object
Customer_Friends_count              664 non-null int64
Customer_Reviews_count              664 non-null int64
Customer_Photos_count               664 non-null int64
Customer_Elite                      664 non-null object
Customer_Elite_Year                 664 non-null object
Customer_Rating                     664 non-null int64
Customer_Review                     664 non-null object
Customer_Review_Date                664 non-null object
Customer_Review_Uploaded_Photos     664 non-null object
Customer_Review_Useful              664 non-null int64
Customer_Review_Funny               664 non-null int64
Customer_Review_Cool                664 non-null int64
Business_response_By                51 non-null object
Business_response_Date              51 non-null object
Business_Response_for_Review        51 non-null object
Business_Response                   664 non-null int64
dtypes: float64(2), int64(10), object(13)
memory usage: 129.8+ KB
```

Missing Values in DataFrame:

We can see the missing data in data frame in the below plot we can see that most of the data is missing in Business_response_By, Business_response_Date, Business_Response_for_Review features.



Word Cloud:

The below plot represents the word cloud for the customer reviews for business with most used words in the customer review.

Preparation:

Each feature in the data frame is handled separately and applied feature extraction techniques to prepare data for modeling.

Handling the Missing values:

1. The below features are dropped from the data frame as the features has more than 90% of data is missing.
    a. Business_response_By
    b. Business_response_Date
    c. Business_Response_for_Review
    d. Business_Photos_Count

Categorical Features:

The below categorical features are encoded to numerical features using the label encode.

- Business_Name
- Business_Claim_status
- Customer_Elite

Date Feature:

The below date feature is converted into separate features such as day, month, and year.

- Customer_Review_Date

Address Feature:

The Business_Address feature is converted into separate features such as State and Zip code and the derived state and zip code features are encoded using label encoder.

Customer Name Feature:

The Gender of each customer is extracted based on the name of the customer using the gender guesser package.

Customer_Review:

The below feature extraction techniques are applied on the Customer_Review feature and new features are created.

- Text Cleaning(Removing the unwanted text from the review)
- Lemmatization of reviews
- Extract the word count feature by finding the count of words in reviews.
- Extract the character count feature by finding the count of characters in reviews.
- Extract the word length feature by finding the average word length in reviews.
- Performing the Sentiment analysis on the review and find the sentiment of the customer review.
- The Sentiment score 0 represent neutral, 1 represent negative and 2 represent positive

Features Extracted from the Customer Review:

	review_word_count	review_char_count	review_avg_word_length	polarity	subjectivity	Sentiment_Score
0	116	623	4.379310	0.155556	0.477778	2
1	47	247	4.276596	0.253939	0.484242	2
2	66	390	4.863636	0.456250	0.656250	2
3	39	173	3.435897	0.750000	0.400000	2
4	33	189	4.757576	0.337500	0.612500	2

Methodology

Build a baseline classification model with the above data frame using the Random Forest classifier and achieved an accuracy of 90%.

Techniques

Need to build the Classification models with below Algorithms.

- Random Forest
- XGBoost
- ANN

Process Validation

Results and Analysis

Deliverables

References

Note:

The above model is built by collecting data for plumbing category in Oklahoma city only.

The data collection for other categories such as HVAC, Painters, Auto Repair in Seattle, New York, Palm Beach, Los Angeles, Miami, and Oklahoma city has to collect to build the final model.