YELP DATA CHALLENGE

ANALYZING AND PREDICTING BUSINESS SUCCESS USING BUSINESS ATTRIBUTES AND CUSTOMER REVIEWS

December 2nd, 2018

Sahil Arora, Sanjana Rosario, Riddhi Kanoi, Huy Nguyen

Abstract: Restaurant review platforms have become our default way of searching for restaurants for every occasion. However, when looking for specific attributes of a restaurant or less generic contexts, this search isn't as straightforward. This gives rise to a need to analyze the specific features as well as the customer's perception of restaurants to understand what customers look for in a food business and what distinguishes a successful one from the others. Through this project, we hope to achieve two purposes. First, we want to analyze the reviews and tips given for the businesses to give personalized recommendations for future users about the businesses' ambiance. Second, we want to employ machine learning models and natural language processing to gain insights into the success factors of a business, and predict whether a business is going to succeed based on those factors.

1. DATASET DESCRIPTION

The dataset used for this project is the Yelp Dataset (https://www.yelp.com/dataset). This includes:

- Business.json: includes information directly related to the business such as latitude, longitude, attributes, and categories
- Review.json: includes the reviews received by the businesses above
- User.json: includes the information about the users who wrote the review
- Checkin.json: includes information about check-ins for the businesses
- Tip.json: includes short pieces of suggestion and advice written by the users

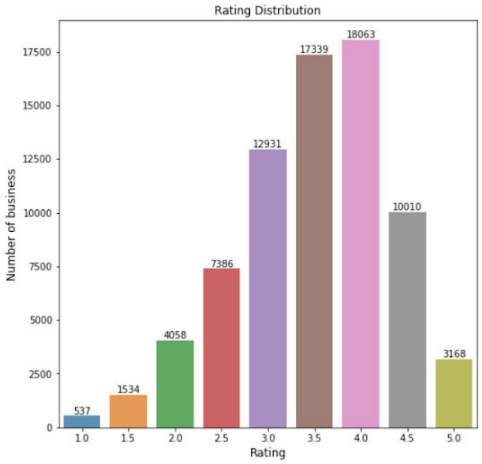
2. DATA PREPROCESSING

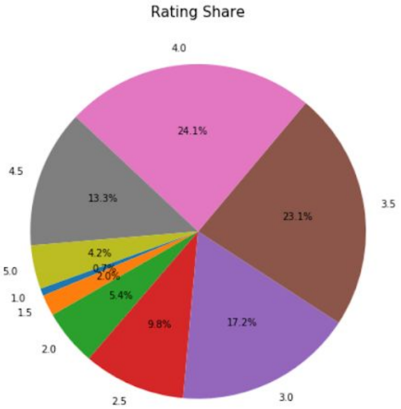
We perform these preprocessing steps:

1. Convert json to csv: For some analyses, we use json_to_csv.py converter provided by Yelp to convert the above files into CSV. For some other analyses, we converted it directly for memory efficiency purpose.
2. Slice the data location-wise and category-wise: Location-wise, for the scope of this project, we will only be looking at business and reviews coming from North American business and users (USA and Canada). In addition, we will only look at the food, drink and restaurant sector.
3. Merge the business, review and user data sets (using left join) to perform exploratory data analysis and building models
4. For machine learning purpose:
    a. Label the attributes with either one-hot encoding (creating dummy variables) or label encoding.
    b. Transform the date features in the reviews dataset to perform datetime operations.

3. DATA EXPLORATORY ANALYSIS

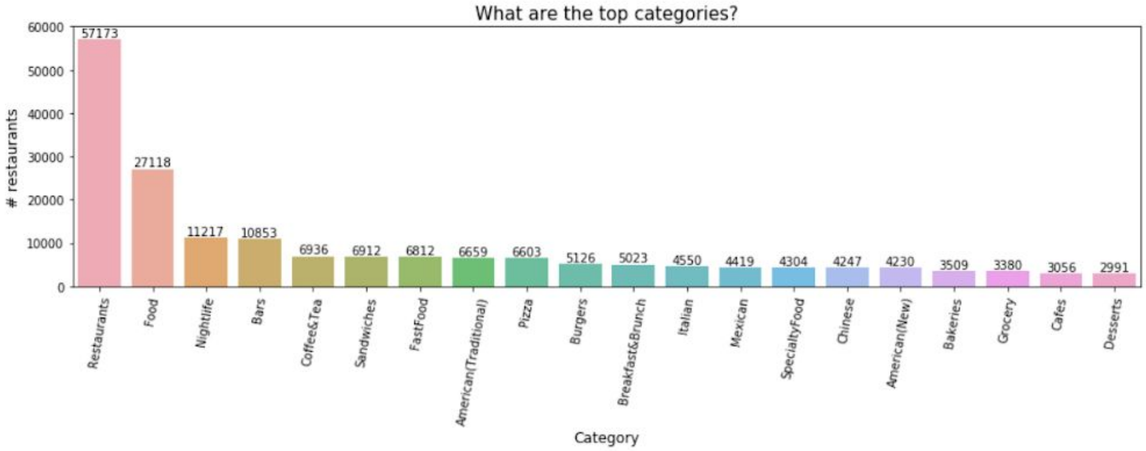
a. Rating share and distribution

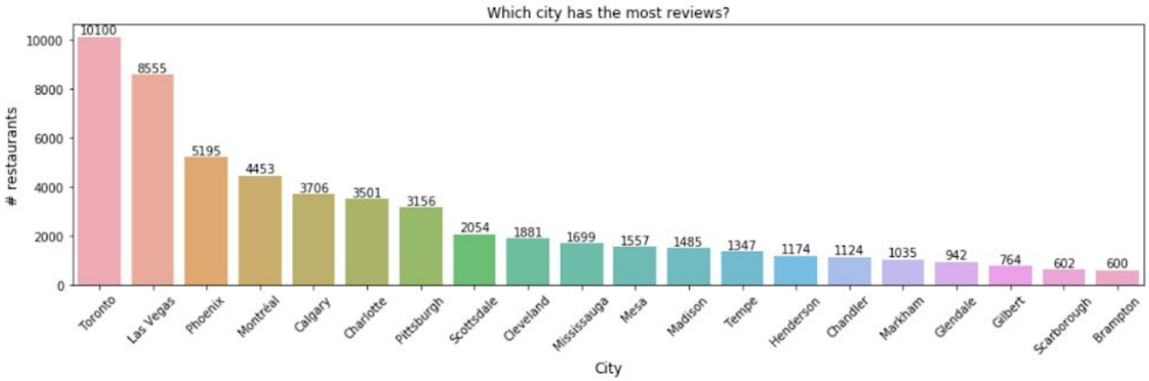




We observe that the ratings are mostly in the range of 3.5 (23.1% of total ratings) to 4 stars (24.1% of total ratings).

b. Distribution of businesses across categories and cities

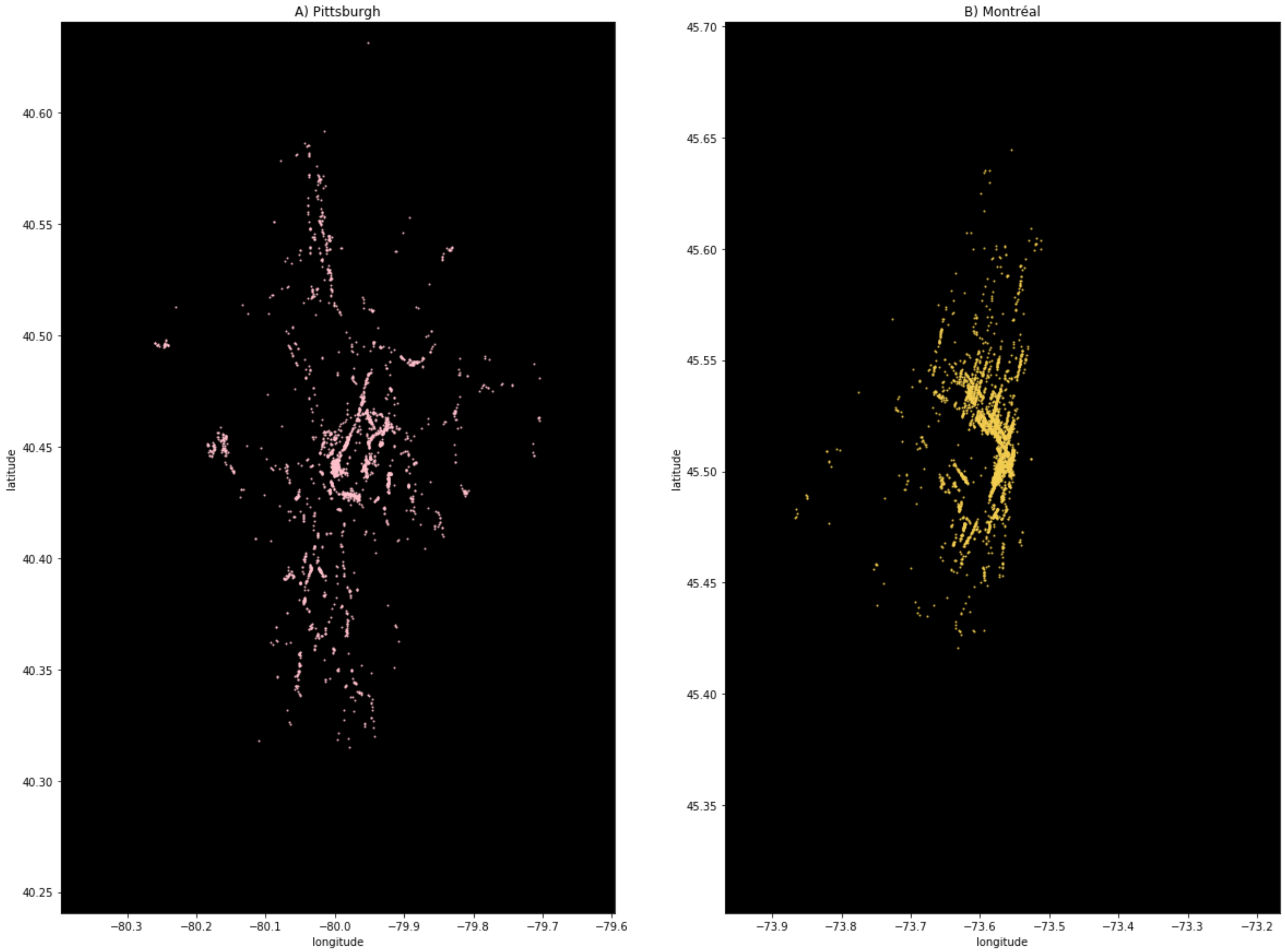




c. Comparison between Montréal and Pittsburgh

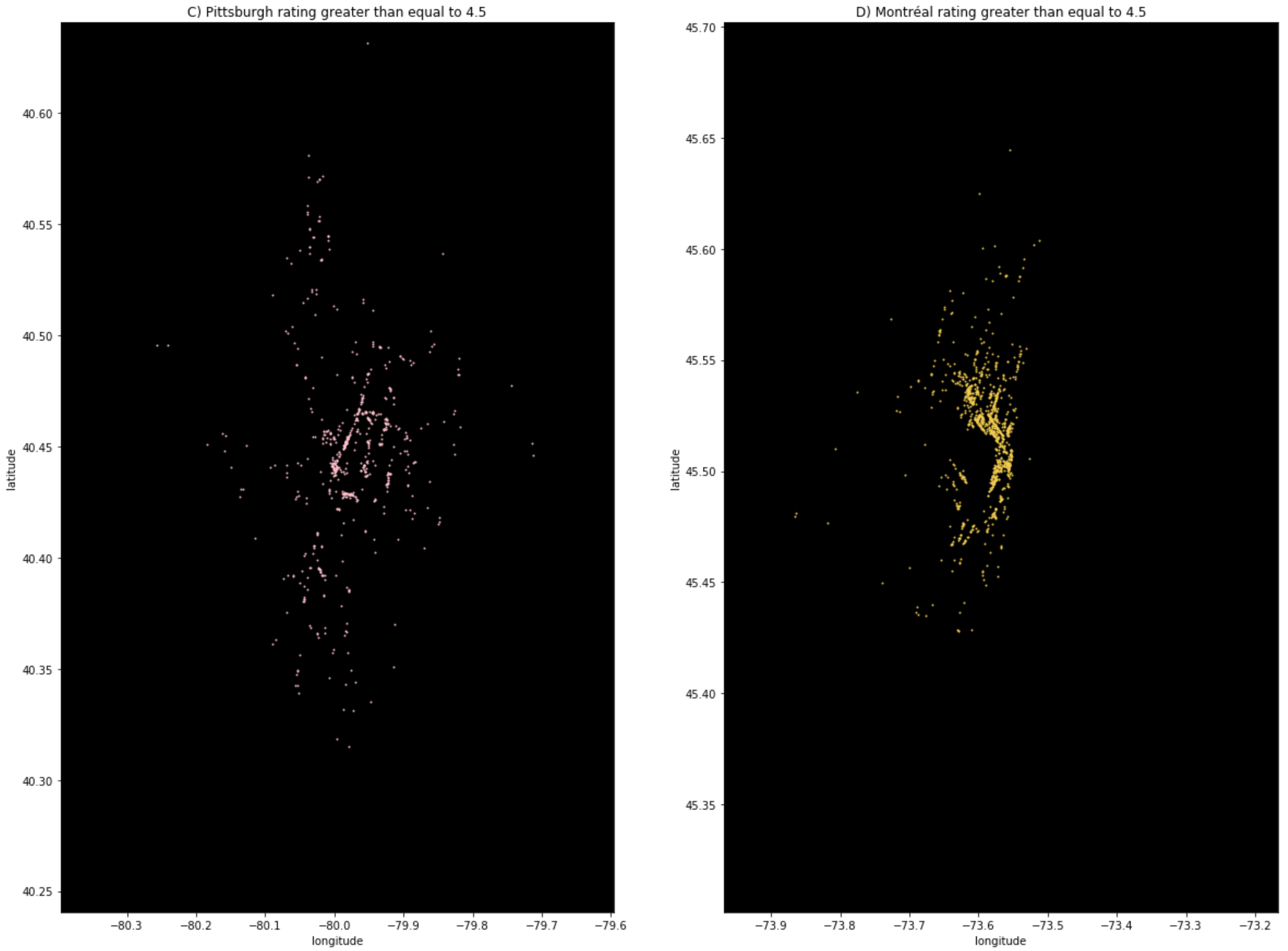
The graph above depicts the number of reviews received by each city. We used this information to consider two cities situated in two different countries to understand the difference in the trends of restaurants and reviews. We considered Montréal (located in Canada) and Pittsburgh (located in America) since the number of reviews received were comparable and they fell within the Top 10 cities with the highest reviews. Thus, keeping these cities in mind we computed our data in the following way:

a) We looked at the ‘latitude’, ‘longitude’, ‘stars’, and ‘review_count’ of the restaurants in both Montréal and Pittsburgh
b) We calculated the max and min values of the latitude and longitude for both cities
c) We imported the library “matplotlib.pyplot” and plotted two graphs in order to visualize the spread out of the restaurants based on the reviews received
d) We selected “Scatter plot” as the graph type and defined the square marker ‘s’, alpha values, and color of the graph.
e) We then printed out both the graphs to get the following result



Looking at the above graph, we instantly notice a difference in the two cities. The data points in Montréal (graph B) is more concentrated when compared to that of Pittsburgh (graph A). Even though Montréal is approximately three times in size (size of Montréal 166.6 mi^2 and size of Pittsburgh 58.35 mi^2) the area covered by its restaurants isn’t as spread out. Additionally, you notice that the data points in both the graphs are more concentrated in the center and dwindles as we move away.

We next repeated the same steps in order to filter out restaurants with ratings greater than equal to 4.5 stars. The resultant graphs are as follows:



The above graph depicts restaurants in Montréal and Pittsburgh with ratings greater than equal to 4.5 stars. It is interesting to see how the number of restaurants in Pittsburgh (graph C) reduces far more rapidly than those in Montréal (graph D) when the filter is applied. The surface area covered by restaurants with ratings greater than equal to 4.5 stars is once again more concentrated in Montréal. It isn't as spread out as Pittsburgh. However, the restaurants are still more dense towards the center as compared to the exteriors.

4. ANALYSIS AND METHODS

a. Topic modeling

b. Word Embeddings using Word2Vec

It is essential to develop a sophisticated word representation model to analyze similarities and establish connections among customer reviews and tips. We compute continuous vector representations of words from Yelp's customer reviews using Word2Vec. The Word2Vec model produces a vocabulary, with each word being represented by an n-dimensional numpy array.

These vectors prove to perform much more better on our test set with regards to measuring syntactic and semantic word similarities. Being able to measure similarity means being able to quickly analyze and measure sentiments and topics in customer reviews. The result of this embedding is a spatial space displaying scatter points. The points with the same colors representing words that are highly correlated.

After converting the reviews to vectors we use TSNE, to reduce the dimensionality and visualize the vectors in space. TSNE is pretty useful when it comes to visualizing similarity between objects. It works by taking a group of high-dimensional (100 dimensions via Word2Vec) vocabulary word feature vectors, then compresses them down to 2-dimensional x,y coordinate pairs. The idea is to keep similar words close together on the plane while maximizing the distance between dissimilar words.

c. Attributes Analysis and Predictions

The machine learning models we use for this project include linear regression, logistic regression, decision tree, random forest and bootstrapping. We decided to choose many models and compare the results among them to figure out which model performs the best on the set of features that we have as well as the importance of features with respect to each model. For regression tasks, we use the star ratings (on the continuous scale of 0.0 to 5.0) as the label. For classification tasks, we also 1 and 0 to indicate success (a restaurant achieving 3.5 stars and above on average will receive a 1 and a 0 otherwise).

In both regression and classifications tasks, we ran OLS regression beforehand to narrow down the features that explain the variances in our labels. In addition, for the classification task, we also ran recursive feature elimination to pick the most important features. These features, as well as the features appearing in feature importance as a result of the random forest algorithm, are important to restaurant owners because they can infer future business performance.

5. RESULTS

a. **Topic modeling**

b. **Word Embeddings**

The result for word similarity using Word2Vec is displayed below. Using this tool will help us quickly analyze and understand a customer review and gain insights into the sentiments shared among words.
Here we use Python library Genism's implementation of word2vec model to train our word vectors. In order to better capture the corpus statistics, such as the word-word co-occurrence matrix, we train word vectors using the reviews.

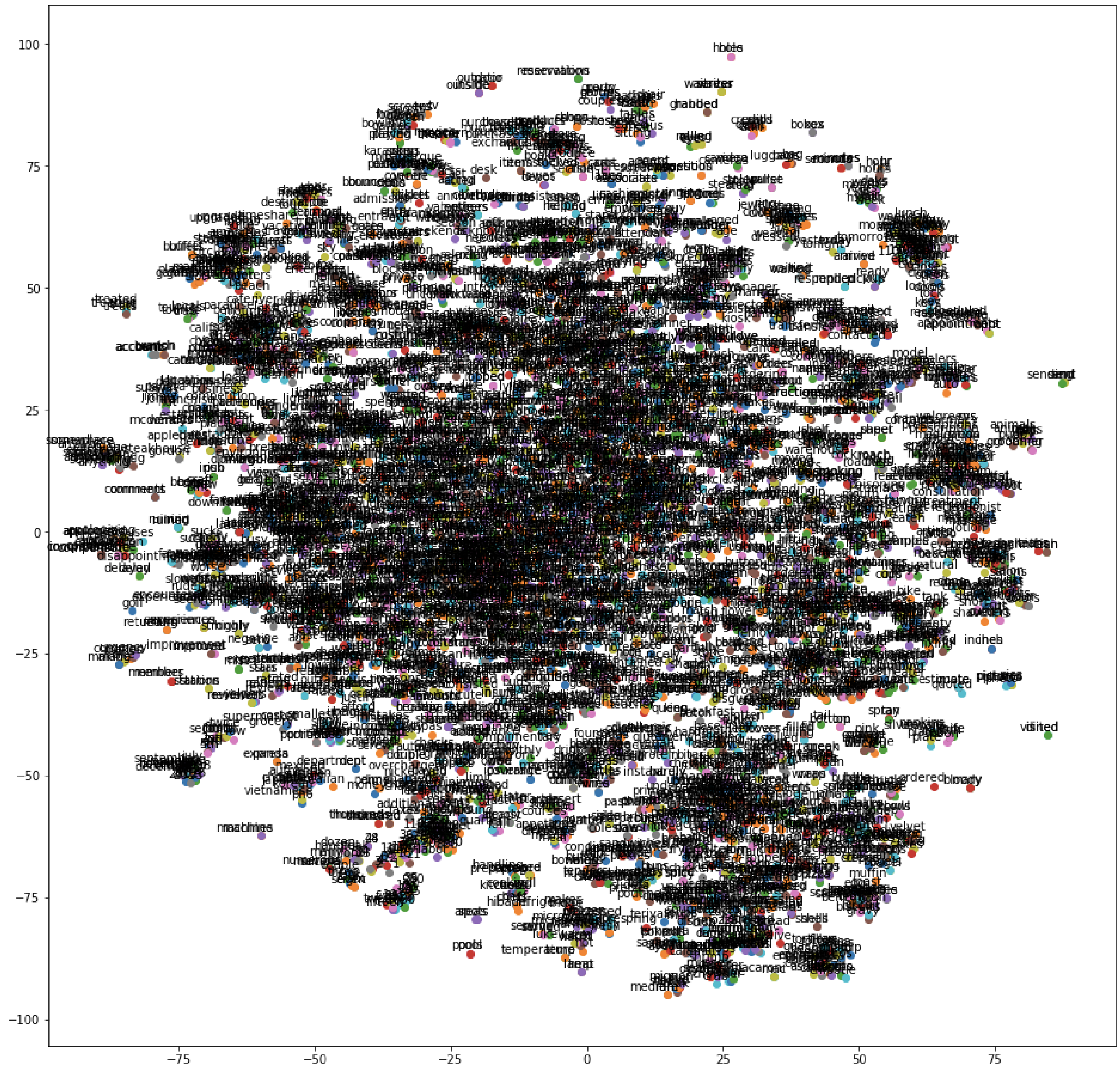
Here Xij is the (i,j) entry of the co-occurrence matrix denoting the number of co-occurrence of word i and word j; w, w̃, b and b̃ are input and output word vector and intercept terms as in Word2Vec. The weight function f(Xij) is defined as,

f(x) = ◎ (x/xmax)α if x < xmax1 otherwise.

<u>Visualization :</u>

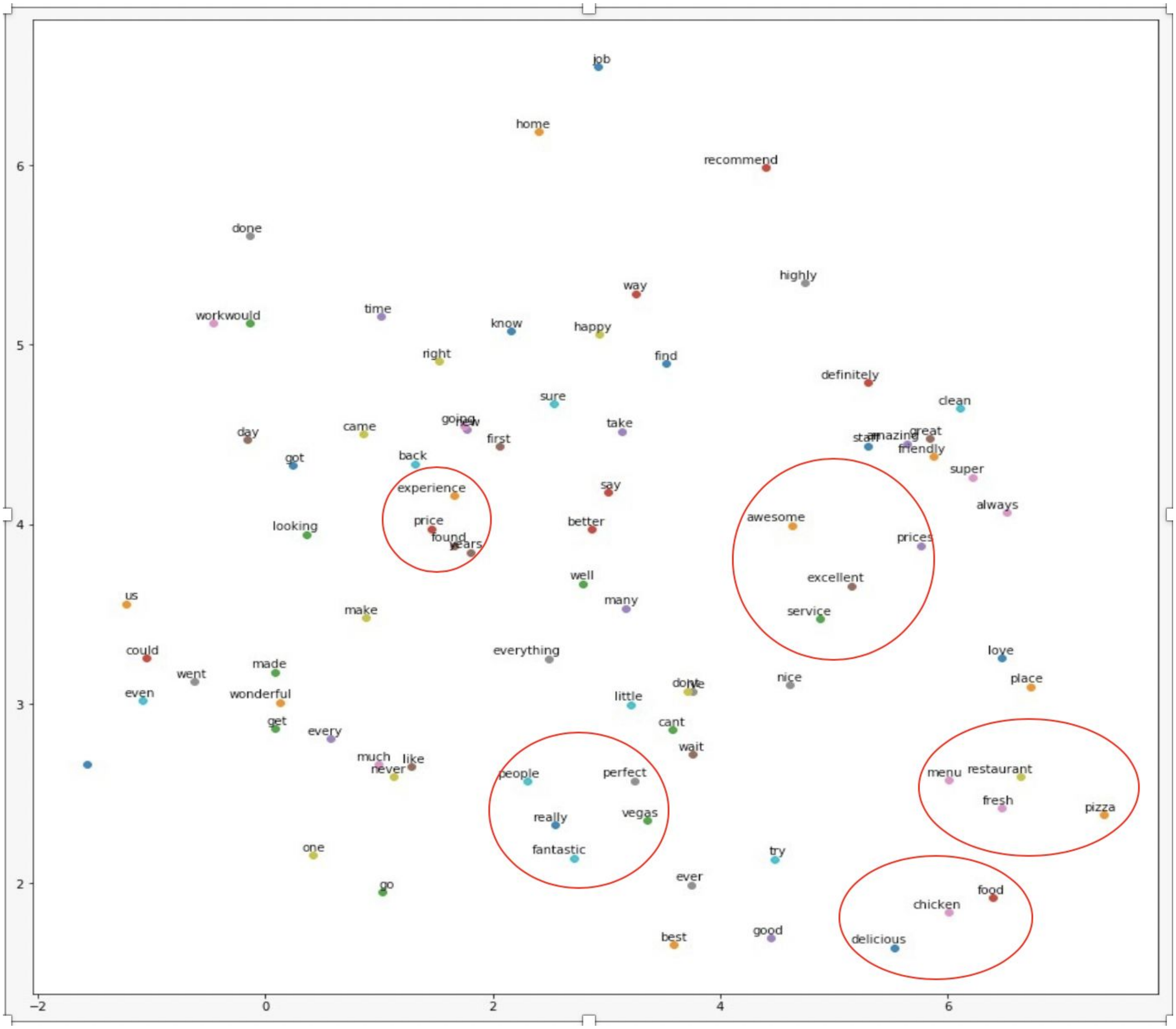
(t-SNE) t-Distributed Stochastic Neighbor Embedding is a non-linear dimensionality reduction algorithm used for exploring high-dimensional data. It maps multi-dimensional data to two or more dimensions suitable for human observation. With help of the t-SNE algorithms, you may have to plot fewer exploratory data analysis plots next time you work with high dimensional data.

Initial Word2Vec embeddings with a low minimum count of occurences:



One of the key challenges was to effectively find the minimum count of occurrences from the entire corpus that plots the t-SNE as well as retains the inherent meaning of the sentences in the words. The picture above depicts one of the initial word2Vec representations that didn't help us draw good inferences.

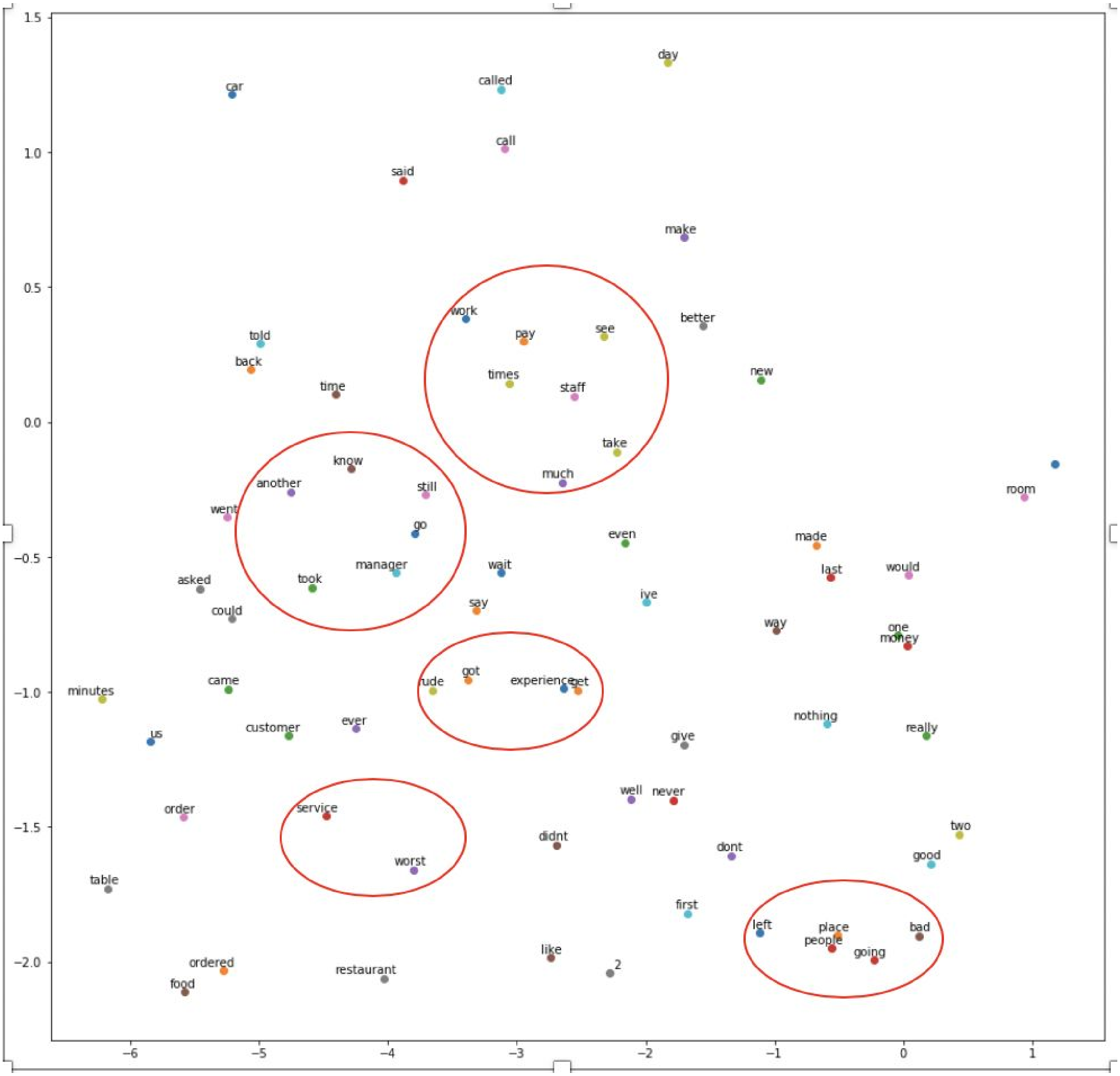
Vector Representation of Entities scoring Well in reviews (> 3.5) :



Inferences for Positive Sentiment reviews:

1. The word "**experience**" of a customer is highly correlated to the word "**back**" and "**bars**". This can help us investigate whether customers are likely to come back to the entity if their experience is good whether be it a bar or a Hair Salon
2. The word "**menu**", "**food**" and "**chicken"** is closely related to the word "**fresh**". This can help us validate whether the restaurants that serve fresh food are more likely to be rated higher.
3. As expected "**excellent**" and "**service**" are correlated due to the fact that people are highly likely to rate an entity higher if the service is good
4. Also, the words"**good**", "**delicious**" and "**chicken"** are spatially correlated leading to the premise that a restaurant with a great chicken dish is more likely to succeed than a restaurant with other specialties

Vector Representation of Entities scoring Poorly in reviews (<3) :



__Inferences for Positive Sentiment reviews__:

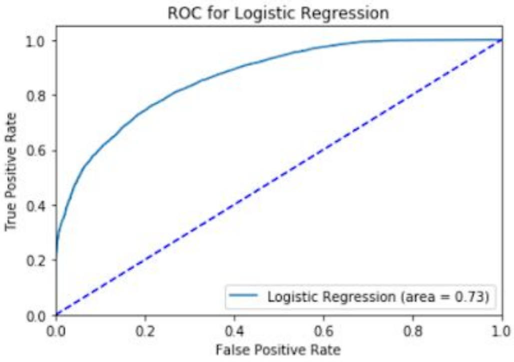
1. The word "**experience**" of a customer is highly correlated to the word "**rude**". This can help us investigate whether an entity is poorly rated solely because of rude behavior of the staff
2. The word "**manager**" and "**service**" are closely related. This can help us validate whether the entities that have unsatisfactory managers are likely to be rated lower
3. As expected "**service**" and "**worst**" are correlated due to the fact that people are likely to rate an entity poorly if the service is bad
4. Also the words "**time**", "**wait**", "**minutes**" and "**took**" are spatially correlated leading to the premise that an entity that takes time irrespective of what the product is highly likely to be rated lower

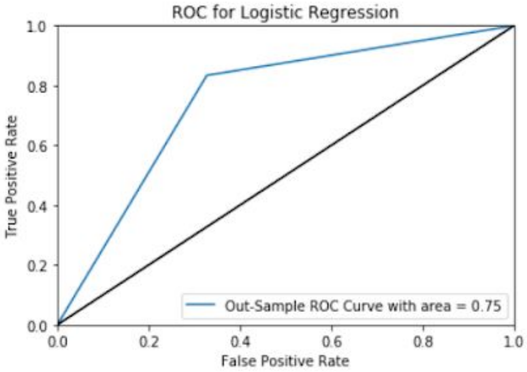
c. Attributes Analysis and Predictions

We achieve a relatively positive result for all of our models. The results are as follows:

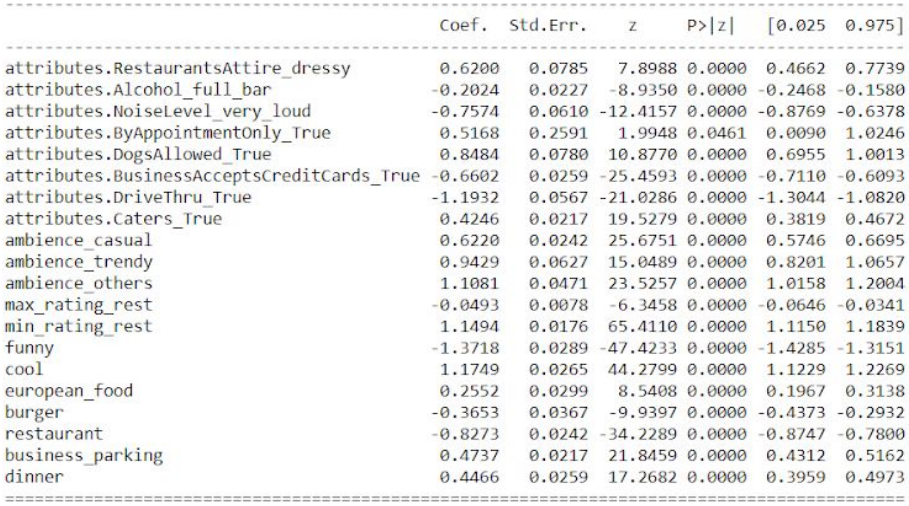
	Mean-squared Error	Variance Score
Linear Regression	0.29	57.00%
	Accuracy	**Area under the ROC**
Logistic Regression	79.00%	73.00%
Decision Tree	76.36%	72.00%
Random Forest	77.40%	75.15%
Bootstrapping	76.90%	75.30%

We suggest using either logistic regression or random forest ensembles because they are relatively simple and yield good results while not computationally expensive. Bootstrapping doesn't perform as well as expected because our dataset is relatively sparse after preprocessing.



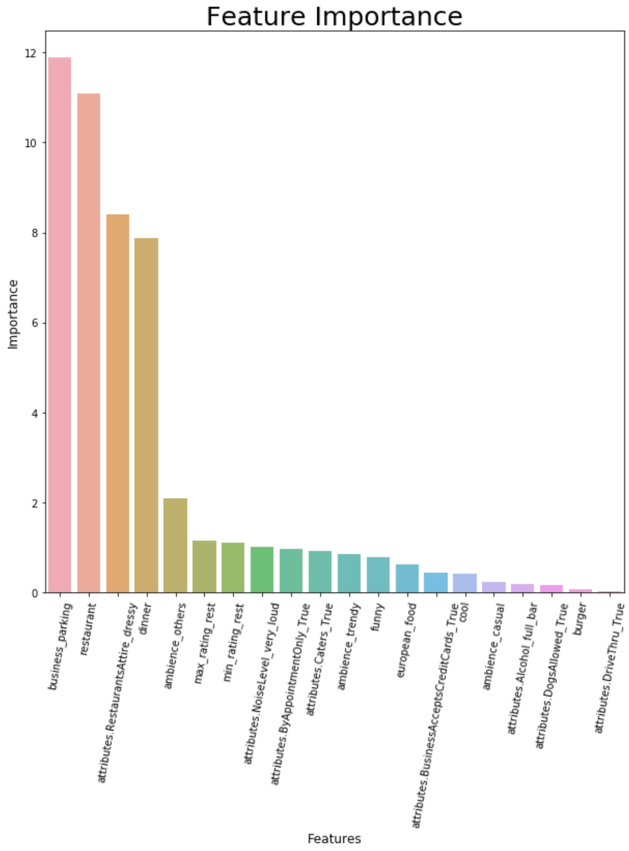


The features that are used as input of the above classification algorithms are:



```
--------------------------------------------------------------------------------------------
                                             Coef.  Std.Err.     z     P>|z|   [0.025  0.975]
--------------------------------------------------------------------------------------------
attributes.RestaurantsAttire_dressy         0.6200    0.0785   7.8988 0.0000  0.4662  0.7739
attributes.Alcohol_full_bar                -0.2024    0.0227  -8.9350 0.0000 -0.2468 -0.1580
attributes.NoiseLevel_very_loud            -0.7574    0.0610 -12.4157 0.0000 -0.8769 -0.6378
attributes.ByAppointmentOnly_True           0.5168    0.2591   1.9948 0.0461  0.0090  1.0246
attributes.DogsAllowed_True                 0.8484    0.0780  10.8770 0.0000  0.6955  1.0013
attributes.BusinessAcceptsCreditCards_True -0.6602    0.0259 -25.4593 0.0000 -0.7110 -0.6093
attributes.DriveThru_True                  -1.1932    0.0567 -21.0286 0.0000 -1.3044 -1.0820
attributes.Caters_True                      0.4246    0.0217  19.5279 0.0000  0.3819  0.4672
ambience_casual                             0.6220    0.0242  25.6751 0.0000  0.5746  0.6695
ambience_trendy                             0.9429    0.0627  15.0489 0.0000  0.8201  1.0657
ambience_others                             1.1081    0.0471  23.5257 0.0000  1.0158  1.2004
max_rating_rest                            -0.0493    0.0078  -6.3458 0.0000 -0.0646 -0.0341
min_rating_rest                             1.1494    0.0176  65.4110 0.0000  1.1150  1.1839
funny                                      -1.3718    0.0289 -47.4233 0.0000 -1.4285 -1.3151
cool                                        1.1749    0.0265  44.2799 0.0000  1.1229  1.2269
european_food                               0.2552    0.0299   8.5408 0.0000  0.1967  0.3138
burger                                     -0.3653    0.0367  -9.9397 0.0000 -0.4373 -0.2932
restaurant                                 -0.8273    0.0242 -34.2289 0.0000 -0.8747 -0.7800
business_parking                            0.4737    0.0217  21.8459 0.0000  0.4312  0.5162
dinner                                      0.4466    0.0259  17.2682 0.0000  0.3959  0.4973
============================================================================================
```

The important feature suggested by our Random Forest Ensembles are as follows:



From the above feature, we have the following suggestions for the restaurant owners:

1. Maintaining a good facility is essential for a successful restaurant. Customers seem to favor those with good business parking and avoid those with loud noises.
2. The ambiance seems to play a decent role in the success of a business. In general, a trendy ambiance will help a business to gain interest from customers. It is important to maintain a certain ambiance because on average these businesses tend to do better ('Others' are aggregations of ambiances that are not trendy or classy).
3. Other features are harder to interpret and built into a strategy. However, they should be kept in consideration for further analysis.

6. CONCLUSIONS

In this project, we employ many techniques to analyze users' reviews as well as machine learning algorithms to classify business to predict their successes. Logistic regression and random forest ensembles are shown to perform well enough compared to other models and can be used to gain insights into how their business will perform based on a number of input features.

In future works, we plan to extend the Word2Vec model to infer hidden messages in the user reviews and use more sophisticated machine learning algorithms such as neural networks and XGBoost to improve our success predictions even more.

REFERENCES:

1. Yelp Dataset and Data Challenge Website: https://www.yelp.com/dataset/challenge
2. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space" (2013)
3. Hood et al., "Inferring Future Business Attention" (2014)