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 customer's perception of a 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' ambience. 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

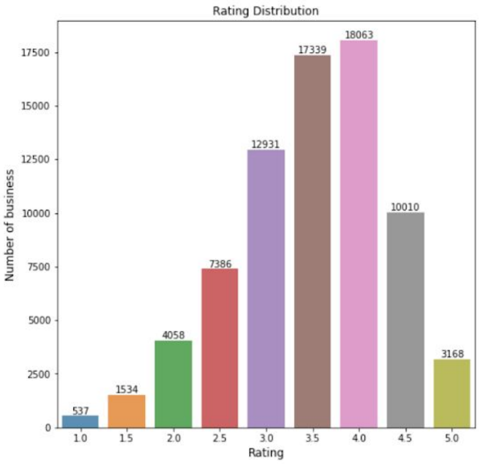
For some analyses, we use json_to_csv.py converter provided by Yelp to convert the above files into CSV. For some other analyses, we converted it directly for memory efficiency purpose.

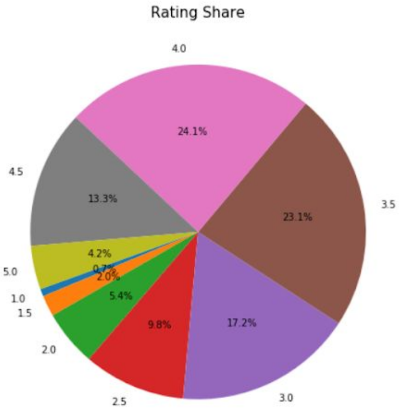
2. DATA PREPROCESSING

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. DATA EXPLORATORY ANALYSIS

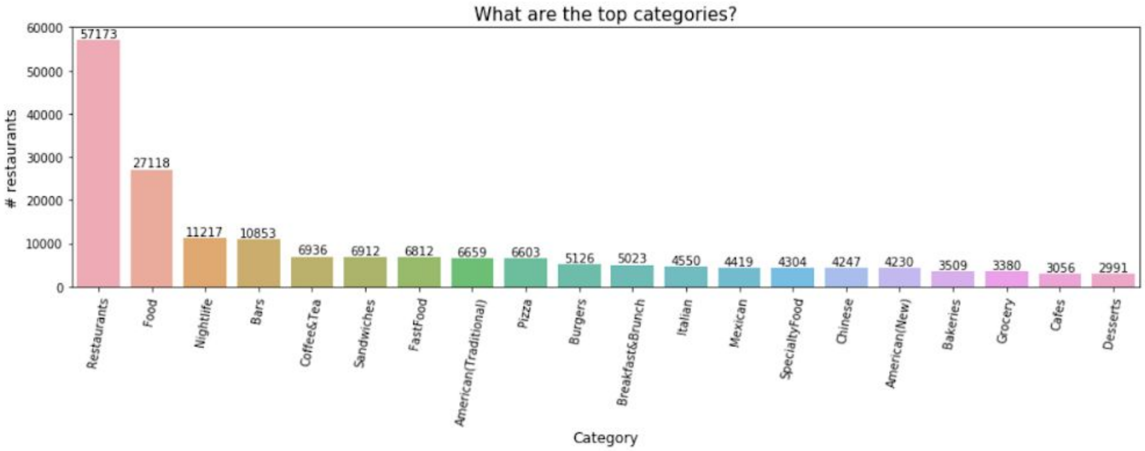
a. Rating share and distribution

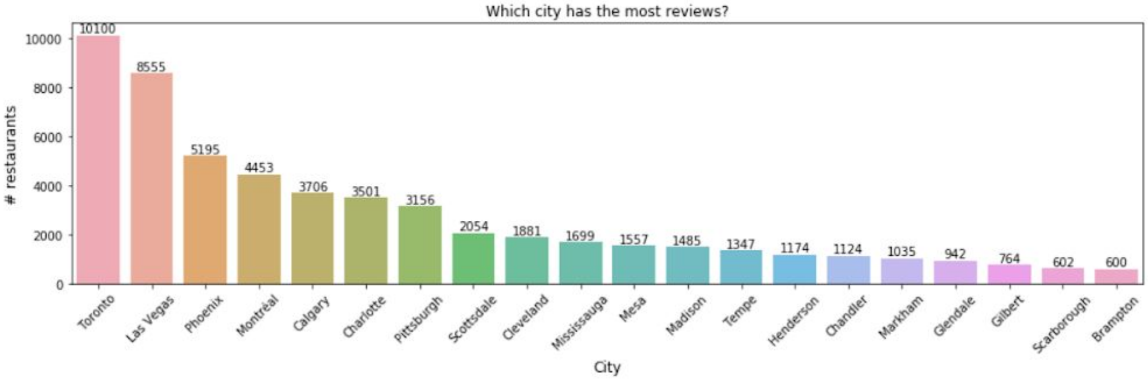




We observe that the ratings are mostly in the range of 3.5 (23.1% of total ratings) to 4 stars (24.1% of total ratings).

b. **Distribution of businesses across categories and cities**





4. ANALYSIS AND METHODS

a. Topic modelling

b. Sentiment Analysis and Word Embeddings

c. Attributes Analysis and Predictions

The machine learning models we use for this project includes 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 perform the best on the set of features that we have as well as the importance of features with respect to each models. 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 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 modelling

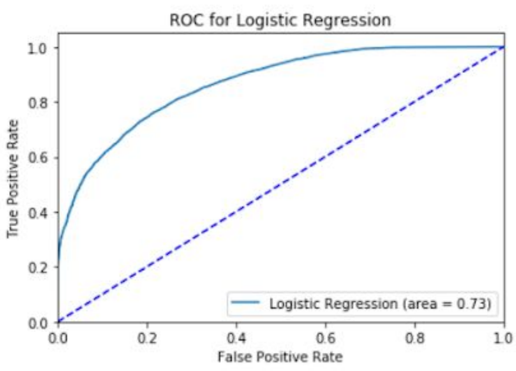
b. Sentiment Analysis and Word Embeddings

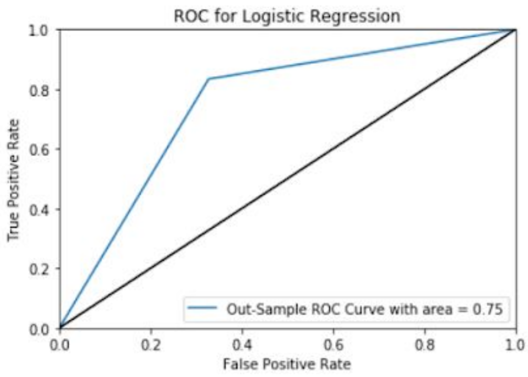
c. Attributes Analysis and Predictions

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

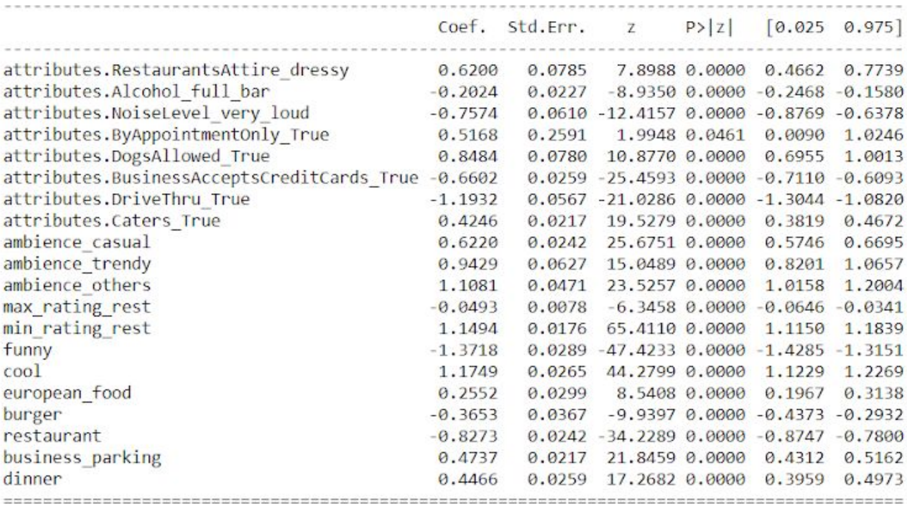
	Mean-squared Error	**Variance Score**
Linear Regression	0.29	57.00%
	Accuracy	**Area under the ROC**
Logistic Regression	79.00%	73.00%
Decision Tree	76.36%	72.00%
Random Forest	77.40%	75.15%
Bootstrapping	76.90%	75.30%

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



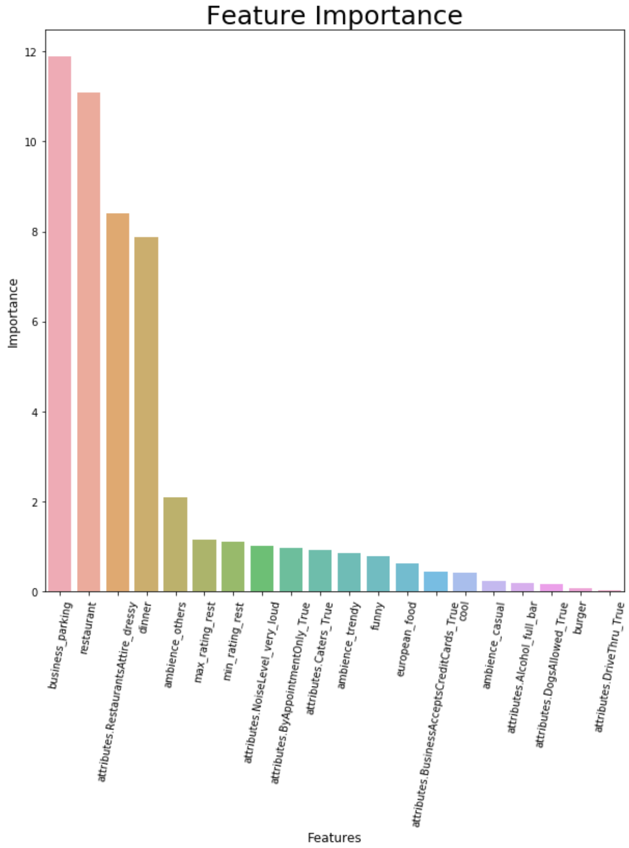


The features that are used as input of the above classification algorithms are:



```
-----------------------------------------------------------------------------------------
                                        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 good facility is essential for a successful restaurants. Customers seem to favor those with good business parking and avoid those with loud noises.
2. Ambience seems to play a decent role in the success of a business. In general, a trendy ambience will help a business to gain interest from customers. It it important to maintain a certain ambience because on average these businesses tend to do better (‘Others’ are agregation of ambiences 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

Businesses can use logistic regression and random forest ensembles to gain insights into how their business will perform based on a number of input features.

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)