WHAT ARE THE MOST IMPORTANT ATTRIBUTES TO RESTAURANT BUSINESSES IN AND OUTSIDE THE UNITED STATES?

Reported by: Tam Tran-The
Teammate: Lily Rithichoo
Course: SDS293 Machine Learning
Instructor: Jordan Crouser

I. INTRODUCTION

As customers are becoming more active in writing and reading online reviews to determining the quality of a local business, data floods in from a variety of angles at an increasingly rapid pace. At the same time, as the restaurant industry is becoming more competitive, restaurant enterprises require a detailed analysis of leading factors that customers are looking for in a great restaurant experience. From local businesses' point of view, we are motivated to study the determinants of high-quality restaurants which customers have expressed preference for. By having a holistic view into customers' inclination, businesses are able to make informed decisions that can drive their improvements and ultimately grow their revenues.

II. DATA AND METHOD

We used Yelp Business Dataset for our project. The data set is in JSON format and includes information about local businesses in 444 cities across 4 countries (the United States, Canada, the United Kingdom, and Germany.) Specifically, it consists of 85,901 observations and 15 variables, many of which are nested data frames. These variables cover a wide range of information about the businesses: from name, address, state/city, open hours, categories to all kinds of attributes that describe their characteristics and specialization.

Data wrangling took up the largest chunk of my time spent on this project. Following are the main tasks that I performed in R to tidy the data set and make it ready to be analyzed.

1) Converted the data set in JSON format into a data frame by using the *stream_in* function of package *jsonlite*
    - Since parsing huge JSON strings is challenging and inefficient, this function helps implement line-by-line processing of JSON data over a connection, such as a file or url, and then destroys the connection.
2) Flattened the nested data frame
    - When I worked on this project, I was not aware of the *flatten* function, which is also from the *jsonlite* package. Thus, I only used *dplyr* to "manually" flatten the data frame column by column. Although it was not as efficient as how the function could have worked, I was able to learn a lot more about data wrangling.
3) Recoded NA values
    - NA values takes up a very large part in the original data set. In fact, the number of NA values of most predictors is higher than the number of observations of each level of a predictor. Thus, if I had used *complete.cases()* to keep observations

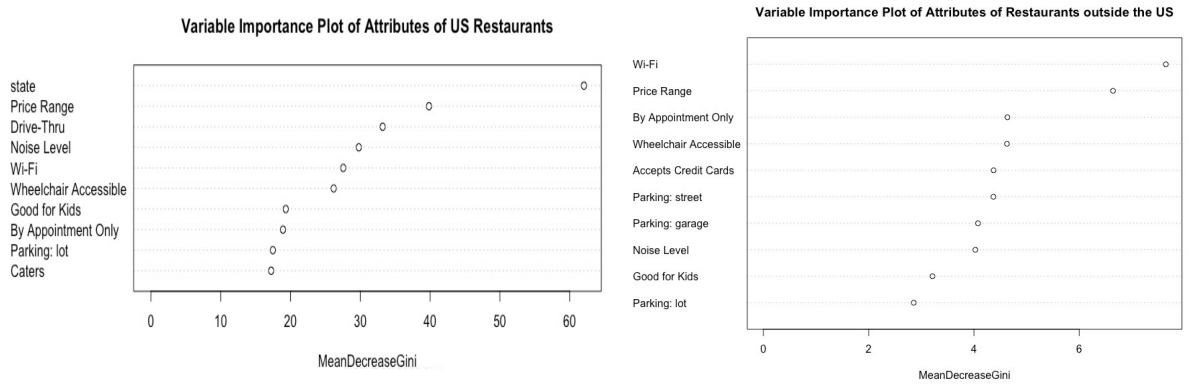
without NA values, there would have been no observation left. That's why I decided to retain these NA values instead of dropping them off, and recoded them into another level of each predictor, named ***MISSING***.

4) Divided the data set into two parts: one that included businesses in the US and one that included restaurants outside the US
    - Since I wanted to compare the model outputs between two kinds of businesses, those located in the US and those located outside, I broke down the whole data set into two parts.
5) Only selected variables of interest, which were ***state***, ***stars***, and ***attributes of restaurant businesses***
    - Here I used ***state***, instead of ***city***, to indicate locations of businesses, since R cannot handle categorical variables with more than 53 levels. On the other hand, my teammate Lily tried using Python to select ***city*** as a categorical variable. Our results in the end turned out to be very similar although we used different variables for location in the model.
6) Transformed ***stars*** (rating of a restaurant, ranging from 1 to 5) into a dummy variable
    - To perform a random forest model, I needed to have a response variable in my data set. I then decided to use ***stars*** to indicate how good a business was and transformed the variable into a binary one. Specifically, I assigned the value "good" if a restaurant was rated above 3 stars, and "bad" otherwise.

I chose to use random forest because with this model, I can obtain an overall graphic of the importance of each predictor, which can be easily interpreted. In addition, random forest can handle high-dimensional data without leading to overfitting. I then fitted a random forest model on both data sets (one includes US businesses and the other includes foreign businesses). For each model, I grew 100 trees and sampled 21 variables, which were 1/3 of all the predictors in the data set, at each split.

I also tried the bagging method where I sampled all 63 predictors in the data set at each split to make sure this model produced similar results to what I obtained from the random forest model.

III. RESULT AND DISCUSSION



State, ***price range*** and ***the availability of drive-through*** are the most important predictors for US restaurants, whereas ***WI-FI***, ***price range*** and ***by appointment only*** prove to be more important than other predictors in determining the quality of restaurants outside the US. These orders of importance remain nearly the same as when I used the bagging method.

On the other hand, the random forest model in Python leads to the result that ***city***, ***price range***, and ***by appointment only*** are the most important factors for US businesses. In terms of foreign businesses, ***city***, ***price range***, and ***noise level*** have higher importance levels than other predictors. The result obtained from bagging methods, again, indicates nearly the same variable importance.

Price range and ***location (states/city)*** are consistently among the most important factors for both U.S. and foreign restaurants to achieve overall reviews greater than or equal to 3 stars. ***Price range*** came up in every model Lily and I fitted while ***location*** appeared in three out of four models. Levels of importance of other attributes besides ***price range*** and ***location*** varied from model to model. This suggests that local businesses should pay special attention to site selection and price strategy.

IV. FUTURE WORK

Going further, one can deploy more models (such as PCA, logistic regression, decision tree, etc.) to find the one with higher prediction accuracy rate as well as continually improve the current random forest model by tuning different parameters to select the optimal ones. If time allows, I would also like to further examine the ***MISSING*** attribute to see if there is any relationship between the lack of mention of a specific attribute and the location and/or other attributes of a restaurant. In addition, I would like to perform text mining on customer reviews for each restaurant to find if there is any factor that is not included in the original data set that might be useful for prediction of good restaurants.