What are the most important attributes to restaurant businesses in and outside the United States?

Tam Tran The and Lily Rithichoo

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.

Data & Methods

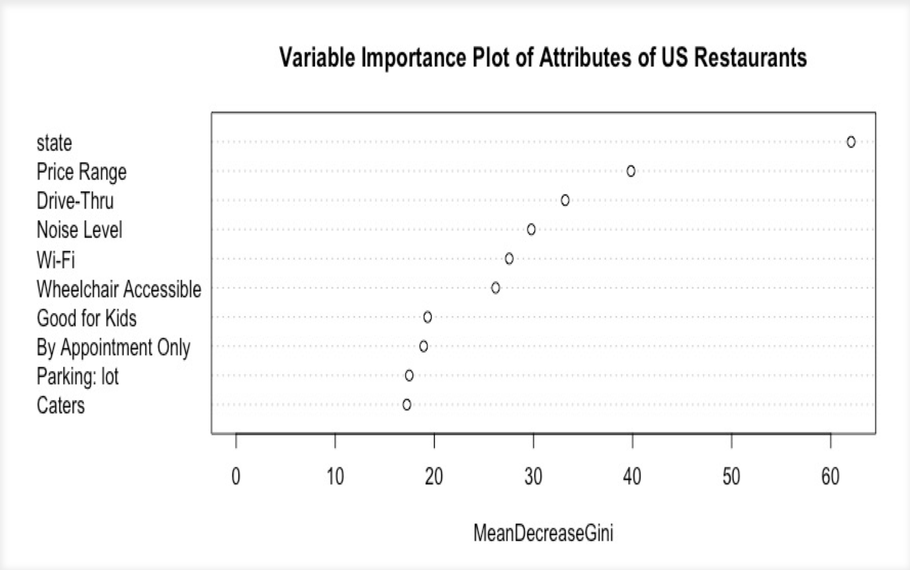
We used Yelp Business Dataset for our project. The dataset includes information about local businesses in 10 cities across 4 countries. There were four main data wrangling tasks that we performed in R:

1) Converted the dataset which was originally in json format into a data frame; flattened the nested data frame
2) Recoded NA values into another level of each predictor, named ***MISSING***
3) Divided the data set into two parts: one that included businesses in the US and one that included restaurants outside the US; only selected variables of interest
4) Transformed ***stars*** (rating of a restaurant, ranging from 1 to 5) into a dummy variable; assigned the value "good" if a restaurant is rated above 3 stars, and "bad" otherwise

We then fitted a random forest model on both data sets. For each model, we grew 100 trees and sampled 21 variables (which were 1/3 of all the predictors that we had) at each split. We also tried the bagging method where we sampled all 63 predictors in the data set at each split to make sure this model produced similar results to what we obtained from the random forest model.

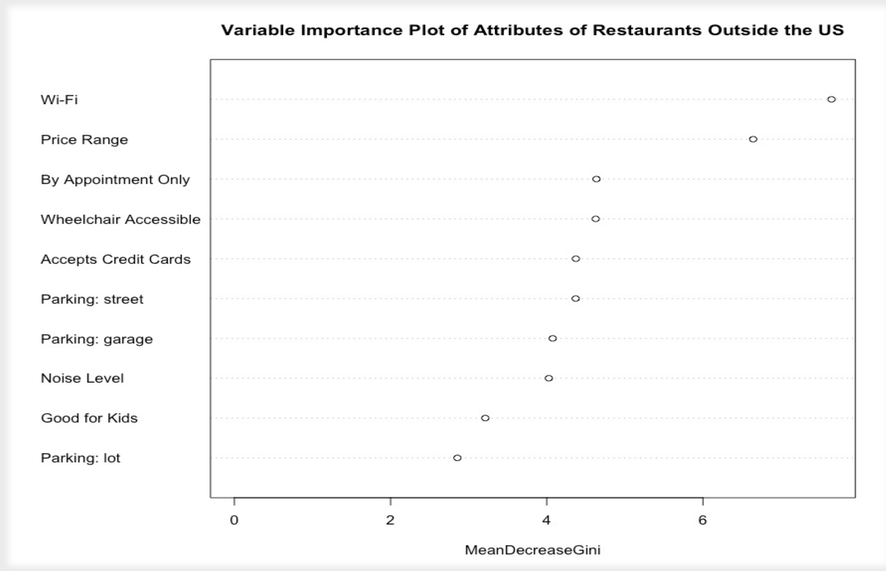
In Python, we modified the data set slightly different. We used ***city***, instead of ***state,*** as the variable of location. In addition, we recoded all categorical variables to numerical arrays. The factor levels in each categorical variable were encoded from 0 to n-1.

Results



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 we used the bagging method.

From fitting random forest in Python, it appears 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.



Conclusions & Future Work

Although our results in R and Python were slightly different, ***price range*** and ***location (states/city)*** were consistently among the most important factors for both U.S. and international restaurants to achieve overall reviews greater than or equal to 3 stars. ***Price range*** came up in every model we fitted while ***location*** appeared in three out of four models. On the other hand, the levels of importance of other attributes varied from model to model. This suggests that local businesses should pay special attention to site selection and price strategy.

We would 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, we would like to look into 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.

References

Data source: https://www.yelp.com/dataset_challenge/dataset

Acknowledgements

This project was completed in partial fulfillment of the requirements of SDS293: Machine Learning. This course is offered by the Statistical and Data Sciences Program at Smith College, and was taught by R. Jordan Crouser during Spring 2016.