Title: Why do certain businesses have an unusually high number of reviews compared with other businesses?

Introduction

Why do certain businesses have an unusually high number of reviews compared with other businesses, of the same general type, in the same geographic region (for example restaurants in the Phoenix metropolitan area)? Are there certain characteristics of the business, such as price range, relative location, or whether alcohol is served that drive the high number of reviews?

Methods

This Data Science Capstone project is based on the dataset from the Yelp Dataset Challenge Round 6 competition. This large data set is provided in 5 separate JSON formatted files:

- Business Data
- Checkin Data
- Review Data
- Tip data
- User Data

As indicated in the introduction, this analysis will attempt to determine why certain businesses of the same general type, in the same geographic area, have unusually high numbers of reviews. Put another way, are there certain characteristics present in the data set that explain the number of reviews that a business has. The analysis presented in this report specifically uses the Business Data portion of the dataset.

Additionally, in order to minimize possible effects of confounding variables, this analysis will focus on Restaurants in the Phoenix Area. One such confounding variable might be regional review volumes. For example, it can be seen from the business data set that, in general, average numbers of reviews in the Las Vegas area have order of magnitude higher reviews than businesses in the Phoenix area. This also has the nice side effect of a more manageable data set.

Data preparation for this analysis involved the following key areas:

- Subsetting The Business data by restaurants in the Phoenix area.
- Removing columns having greater than 95% NA's
- imputing the remaining columns to remove NA's
- Flattening of Categorical Variables

The imputing process only needed to be performed on the categorical features of the data set as the continuous features did not contain NAs. In general, the imputing process employed was largely a logical process depending on the type of information present in the column. For example,

NA values in the “Happy Hour” column were set to the value of FALSE. The underlying assumption here is that values of NA in the categorical columns of the data set largely imply that the restaurant does not provide that particular feature.

It is important to note that this report and its associated artifacts are completely re-producible. Random number generation seeds are set before all pertinent operations to ensure repeatability. Complete R markdown source code can be found at: https://github.com/pcomeau/DataScienceCapstone

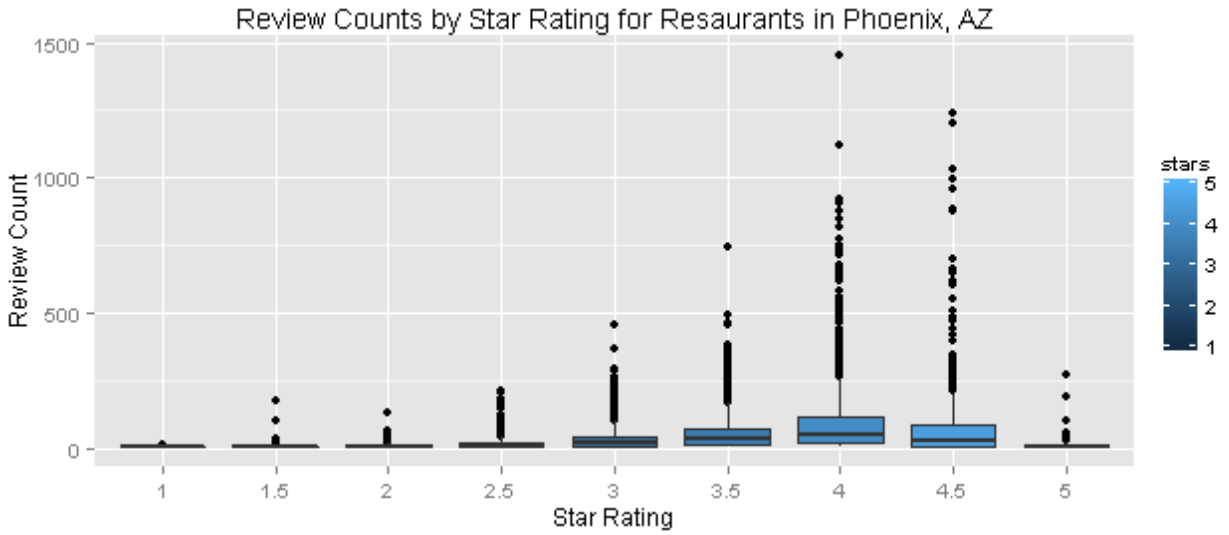
Exploratory Data Analysis

In addition to the continuous variable of interest “review_count”, the cleaned and prepared data set contains a total of 52 variables, three continuous variables and 49 categorical variables. The final data set contains 6587 rows of data.

The features in the final data set are:

```
##  [1] "open"                "review_count"
##  [3] "longitude"           "stars"
##  [5] "latitude"            "Happy.Hour"
##  [7] "Good.For.Groups"     "Outdoor.Seating"
##  [9] "Price.Range"         "Good.for.Kids"
## [11] "Alcohol"             "Noise.Level"
## [13] "Has.TV"              "Attire"
## [15] "Ambience.romantic"   "Ambience.intimate"
## [17] "Ambience.classy"     "Ambience.hipster"
## [19] "Ambience.divey"      "Ambience.touristy"
## [21] "Ambience.trendy"     "Ambience.upscale"
## [23] "Ambience.casual"     "Good.For.Dancing"
## [25] "Delivery"            "Coat.Check"
## [27] "Smoking"             "Take.out"
## [29] "Takes.Reservations"  "Waiter.Service"
## [31] "Wi.Fi"               "Caters"
## [33] "Good.For.dessert"    "Good.For.latenight"
## [35] "Good.For.lunch"      "Good.For.dinner"
## [37] "Good.For.breakfast"  "Good.For.brunch"
## [39] "Parking.garage"      "Parking.street"
## [41] "Parking.validated"   "Parking.lot"
## [43] "Parking.valet"       "Music.dj"
## [45] "Music.live"          "Music.video"
## [47] "Music.jukebox"       "Drive.Thru"
## [49] "Wheelchair.Accessible" "BYOB"
## [51] "BYOB.Corkage"        "Good.For.Kids"
## [53] "Dogs.Allowed"
```

Given the few continuous variables, “Star Rating” and “latitude/longitude”, it appears that restaurants having higher review counts generally fall in the range of 4.0 to 4.5 stars as can be seen from the following side by side box plot.



There appears to be no closely correlated continuous variables, so we will employ all of the continuous variables in the modeling process.

```
##              review_count    longitude       stars     latitude
## review_count  1.000000000  0.005815428 0.248863465  0.009664606
## longitude     0.005815428  1.000000000 0.001870987 -0.502285697
## stars         0.248863465  0.001870987 1.000000000  0.035921868
## latitude      0.009664606 -0.502285697 0.035921868  1.000000000
```

In keeping with data modeling best practices, a stratified random sample of the cleansed data into training, testing, and validation sets is created.

The resulting random samples contain, 1384, 5203, and 4613 rows of data respectively.

Feature Select and Modeling

In order to determine which features might be driving the overall number of reviews for restaurants in the Phoenix area, the “Boruta”“ feature selection algorithm is run against the cleansed data set. In its essence, Boruta works in an iterative manner, and in each iteration the aim is to remove features which according to a statistical test, are less relevant than what is defined by the authors as a random probe.

Next a Random Forest model using the caret package train function using the features deemed relevant by the Boruta feature selection algorithm is fit.

As a secondary approach to further understand which features might be driving the overall number of reviews, a Generalized Linear Model with Stepwise Feature Selection, (method = 'glmStepAIC') is also fit.

For both the Random Forest and the Generalized Linear Model, repeated K-fold cross-validation is used with 10 folds repeated 10 times.

Results

Given the two modeling techniques, Random forest with Boruta feature selection and Generalized Linear Model with Stepwise Feature Selection, the models generally agree that the following significant predictors drive the number of reviews for a given restaurant in the phoenix area:

- open, longitude, stars, latitude, Price.Range, Good.for.Kids, Alcohol, Noise.Level, Noise.Level, Noise.Level, Ambience.upscale, Waiter.Service, Caters, Music.dj, Music.jukebox, Wheelchair.Accessible, BYOB, Good.For.Kids, Dogs.Allowed,

Further, the Generalized Linear Model predicts that, all else being equal, for a 95% confidence interval,

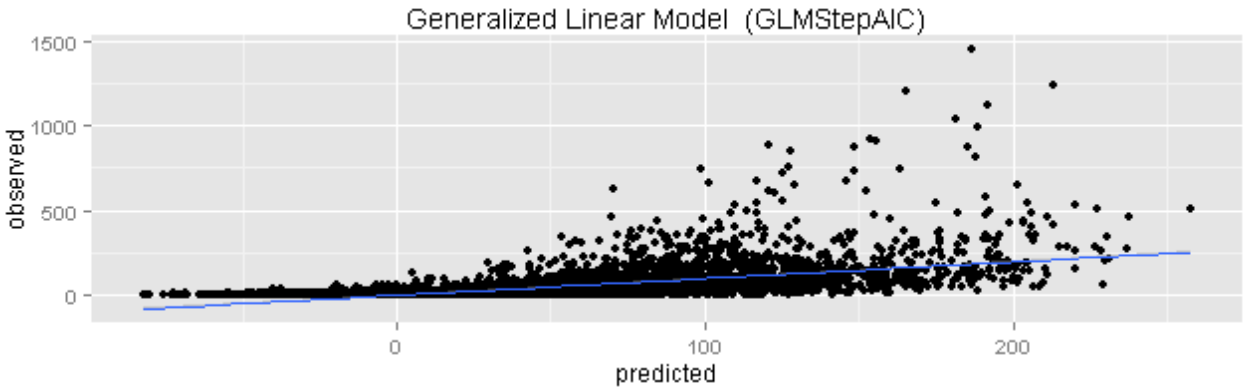
- when the "open" indicator is “TRUE”, restaurants receive 11.73 more to 31.47 more reviews
- for each additional degree of “longitude”, restaurants receive 64.97 less to 0.25 less reviews
- for each additional “stars” (rating), restaurants receive 16.46 more to 27.09 more reviews
- for each additional degree of “latitude”, restaurants receive 79.65 less to 6.54 less reviews
- for each additional increase in “Price.Range”, restaurants receive 9.72 more to 23.96 more reviews
- when the “Good.for.Kids” indicator is “TRUE”, restaurants receive 45.76 to less 7.82 less reviews
- when the “Alcohol” indicator is “full_bar”, restaurants receive 12.89 more to 33.68 more reviews
- when the “Noise.Level” indicator is “quiet”, restaurants receive 38.2 less to 18.3 less reviews
- when the “Noise.Level” indicator is “very_loud”, restaurants receive 43.72 less to 0.61 more reviews
- when the “Noise.Level” indicator is “unknown”, restaurants receive 37.6 less to 6.06 less reviews
- when the “Ambience.upscale” indicator is “TRUE”, restaurants receive 2.21 to less 26.93 more reviews
- when the “Waiter.Service” indicator is “TRUE”, restaurants receive 32.03 less to 1.25 less reviews
- when the “Caters” indicator is “TRUE”, restaurants receive 16.49 more to 37.81 more reviews
- when the “Music.dj” indicator is “TRUE”, restaurants receive 2.07 more to 46.91 more reviews
- when the “Music.jukebox” indicator is “TRUE”, restaurants receive 54.47 less to 3.46 less reviews
- when the “Wheelchair.Accessible” indicator is “TRUE”, restaurants receive 14.47 to more 31.8 more reviews
- when the “BYOB” indicator is “TRUE”, restaurants receive 22.04 more to 61.82 more reviews
- when the “Good.For.Kids” indicator is “TRUE”, restaurants receive 33.75 more to 60.73 more reviews
- when the “Dogs.Allowed” indicator is “TRUE”, restaurants receive 3.38 less to 22.24 more reviews

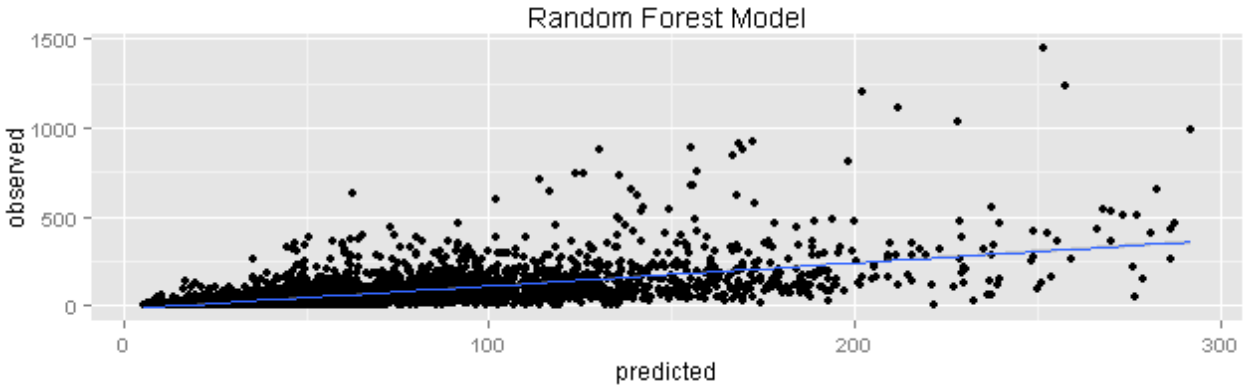
Discussion

From a modeling perspective, the predictive power of the resulting random forest model as measured by the area under the ROC curve, of 0.59, is limited (remember, a ROC area of 50% is essentially equivalent to a random guess). Also, The associated correlation of the model is 0.64

and the Root Mean Square Error (RMSE) is 73.38. The characteristics of the Generalized Linear model are similar.

Further, as can be seen from the following scatter plots of predicted review counts vs. actual review counts, both of models fall short of correctly predicting the review counts especially in cases of very high review counts. So in conclusion, the predictors present in the data set begin to explain review counts but not entirely. This can also be seen in the residuals of the 2 models.





Possible future investigations to further understand what drives very high review counts of particular restaurants might be

- to employ data around elite reviewers
- cross reference review counts to other similar web sites such as Google Reviews