Building the Ultimate Restaurant for Yelpers

S. P.

November 9, 2015

Introduction: What Attributes Lead to High Ratings For Restaurants on Yelp?

This paper is an exploration of ratings of restaurants on Yelp. The primary question is simple: what attributes of restaurants will lead to highly rated restaurants? In other words, if one were to build a new restaurant with the sole goal of getting the highest possible rating on Yelp, how should the restaurant be constructed? This question may be of great interest to restauranteurs and investors.

Methods Used to Conduct This Analysis

Our approach to answering the question put forth possessed 5 key elements:

1. Data Acquisition. I started by downloading the data that Yelp provided. Specifically, I utilized the Yelp business data set, as well as the reviews set. I conducted a group_by operation on the reviews set so that I could obtain the average rating for each business, as well as the sum of useful, cool, and funny votes, and the sum of total reviews for each specific business. I then conducted an inner join (using the dplyr package) so as to create a single dataframe in which each row reflected the attributes of each unique business in Yelp's data set, coupled with its aggregated review data. As the focus of our study is on restaurants, I then filtered the set so as to exclude businesses that did not identify themselves as restaurants.

2. Attribute Filtering. The next step was to filter out the attributes that did not possess enough values in the Yelp set. This was something I was unsure about -- how to define an attribute that had too much missing data? To resolve this issue, we ran two filtering options: Option #1 was to look for attributes that had at least 85% completion rate; meaning only 15% of restaurants in our set had NA values for that specific attribute. For Option #2, we lowered the threshold ot 50%. The result was that Option #1 had 13,294 restaurants with 37 potential attributes, while Option #2 had 6,293 restaurants with 50 attributes.

3. Attribute Conversion. To facilitate a linear regression study, in which the aggregated average reviews would be the dependent variable and the attributes would be the independent variables, I first needed to ensure all the attributes were in a sensible numeric format. The vast majority of attributes were of a logical class, and so converting them to numeric was simple; if the attribute was logical and had a value of FALSE, it could simply be converted to a numeric class with a value of 0. Likewise, logical attributes with a value of TRUE were converted to numeric items with a value of 1.

The situation became slightly more nuanced with some attributes that were of a factor class. For instance, the Alcohol attribute -- whether or not a business served alcohol -- possessed three values: no, beer and wine, or full bar. In this scenario, values were converted to 0, 1, and 2 respectively. The same framework was applied to the Attire attribute, in which values were initially assigned a value of either none, casual or formal; this too was converted to 0, 1, or 2. Because these factor values constituted a scale from absence to increasing definition, it made sense that a numeric conversion would be justifiable.

4. Test Set Creation. After ensuring a sufficient amount of the correct variables were in place and were in numeric format so as to facilitate a linear regression model, the next step was to divide the data set into training, validation, and testing sets. This was done using the functions made available in the caret package.

5. Linear Regression Modelling. Finally, the stage is set for regression modelling to occur. The dependent variable is "avgstars"", which is the average rating given across all reviews for that particular restaurant. This rating is more precise than the "stars" value included in business data set that Yelp has provided, which increments only in 0.5 star increments. "avgstars" is thus far more conducive to linear regression modelling, and is what was used; "stars" was dropped from the data set.

I then looked for other variables that could be dropped. This was done primarily through the creation of a correlation matrix; attributes that possessed a correlation of 0.9 or more with each other were evaluated and cut on a subjective basis so as to ensure the independent variables driving the regression model were not overly correlated with one another. After this was completed, models were run on the training set with the remaining variables. The output of the regression was then evaluated, with independent variables whose p-values were greater than .05 eliminated. The model was then run again on the training sets, saved, and then applied to the validation and test sets.

Results

The finding was that Option #2 -- the model that had more independent variables but a lower amount of total observations (because it required that each attribute to have only a 50% completion rate) produced a higher adjusted r-score and lower residuals on both the training and validation sets.

Below are the regression summaries for Option #1 and Option #2. The coefficient values for each attribute tell us the most powerful attributes, both in terms of what restaurants should strive to possess (positive values) as well as those that should be avoided (negative values).

Option #1 Regression Table

```
##
## Call:
## lm(formula = avgstars ~ ., data = d1lmset1)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.42983 -0.35625  0.05463  0.39953  1.86328 
## 
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       3.8671375  0.0598974  64.563  < 2e-16
## longitude                         0.0026739  0.0004239   6.307 3.08e-10
## `attributes.Accepts Credit Cards` -0.2257907  0.0491643  -4.593 4.48e-06
## attributes.Alcohol                -0.0617984  0.0109513  -5.643 1.76e-08
## attributes.Attire                  0.2743314  0.0442565   6.199 6.12e-10
## `attributes.Takes Reservations`    0.0928933  0.0206048   4.508 6.67e-06
## `attributes.Good For.dessert`      0.1349673  0.0566557   2.382 0.017243
## `attributes.Good For.latenight`   -0.1666878  0.0341112  -4.887 1.06e-06
## `attributes.Good For.breakfast`   -0.0701891  0.0268914  -2.610 0.009077
## `attributes.Good For.brunch`       0.1144966  0.0335740   3.410 0.000654
## attributes.Parking.garage         -0.0934986  0.0326163  -2.867 0.004165
## attributes.Parking.street          0.2417459  0.0243146   9.942  < 2e-16
## attributes.Parking.lot             0.2075721  0.0196899  10.542  < 2e-16
## allvu                              0.0007191  0.0000561  12.818  < 2e-16
##                                   
## (Intercept)                       ***
## longitude                         ***
## `attributes.Accepts Credit Cards` ***
## attributes.Alcohol                ***
## attributes.Attire                 ***
## `attributes.Takes Reservations`   ***
## `attributes.Good For.dessert`     *  
## `attributes.Good For.latenight`   ***
## `attributes.Good For.breakfast`   ** 
## `attributes.Good For.brunch`      ***
## attributes.Parking.garage         ** 
## attributes.Parking.street         ***
## attributes.Parking.lot            ***
## allvu                             ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.5847 on 5305 degrees of freedom
## Multiple R-squared:  0.09792,    Adjusted R-squared:  0.09571 
## F-statistic:  44.3 on 13 and 5305 DF,  p-value: < 2.2e-16
```

Option #2 Regression Table

```
## 
## Call:
## lm(formula = avgstars ~ ., data = d1lmset2)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.02462 -0.27526  0.04111  0.33813  1.51897
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      3.521e+00  5.417e-02  65.009  < 2e-16 ***
## review_count                     5.940e-04  5.493e-05  10.814  < 2e-16 ***
## `attributes.Good for Kids`      -7.832e-02  3.481e-02  -2.250 0.024546 *
## attributes.Alcohol              -9.703e-02  1.289e-02  -7.527 7.22e-14 ***
## `attributes.Noise Level`        -9.223e-02  2.138e-02  -4.315 1.66e-05 ***
## attributes.Attire                3.434e-01  6.167e-02   5.569 2.83e-08 ***
## attributes.Caters                1.126e-01  2.087e-02   5.393 7.59e-08 ***
## attributes.Ambience.romantic     1.926e-01  6.953e-02   2.770 0.005653 **
## attributes.Ambience.intimate     2.606e-01  8.098e-02   3.219 0.001305 **
## attributes.Ambience.classy       1.629e-01  6.221e-02   2.619 0.008871 **
## attributes.Ambience.hipster      3.996e-01  7.550e-02   5.293 1.31e-07 ***
## attributes.Ambience.divey        1.537e-01  4.912e-02   3.130 0.001768 **
## attributes.Ambience.touristy    -3.156e-01  1.223e-01  -2.581 0.009915 **
## attributes.Ambience.trendy       2.710e-01  5.291e-02   5.121 3.27e-07 ***
## attributes.Ambience.upscale      1.880e-01  8.019e-02   2.344 0.019140 *
## attributes.Ambience.casual       1.695e-01  3.077e-02   5.508 4.01e-08 ***
## `attributes.Good For.latenight` -2.029e-01  4.115e-02  -4.929 8.79e-07 ***
## `attributes.Good For.lunch`     -6.262e-02  2.321e-02  -2.697 0.007035 **
## `attributes.Good For.breakfast` -8.892e-02  3.565e-02  -2.494 0.012691 *
## attributes.Parking.garage       -1.842e-01  3.911e-02  -4.710 2.61e-06 ***
## attributes.Parking.street        2.257e-01  3.252e-02   6.941 4.94e-12 ***
## attributes.Parking.lot           1.211e-01  3.194e-02   3.790 0.000154 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4884 on 2497 degrees of freedom
## Multiple R-squared:  0.1723, Adjusted R-squared:  0.1654
## F-statistic: 24.76 on 21 and 2497 DF,  p-value: < 2.2e-16
```

Discussion

I took away the following key points from the regression analysis:

1. Accepting credit cards is clearly a negative attribute for the business, as it had a relatively high correlation in both models. I would guess that accepting credit cards is a proxy that reveals other key elements about the business. For instance, perhaps food trucks or other businesses that have a very small real estate footprint, or deal with small portions, do not accept credit cards and are the real attribute that Yelp customers value. This might also be why alcohol was seen as a negative coefficient in both models.

2. Restaurants that had a hipster ambience experienced a lift of .39 stars -- the biggest lift any attribute could provide. Ambience attributes in general far outperformed other attributes in terms of the lift they provided, suggesting that it is vital for restaurants to have a clear ambience they cater towards in order to generate a high Yelp rating. Many

restaurants appear to lack an ambience rating, as the ambience attributes only appeared in Option #2 -- the regression model that required only 50% of restaurants on Yelp to have a value for it to be considered for inclusion in the model.

3. In both models that were run, parking was a very influential attribute. Interestingly, having a parking garage had a negative coefficient and thus contributed to lower ratings, while having a parking lot or street parking were associated with positive coefficients and thus contributed to higher ratings. I believe the parking attributes, like I suspected with the the credit card attribute, may be reflective of other causal attributes. Does a parking garage denote a lack of an appealing ambience? Does accessible street parking bring the right combination of accessibility and style? In my opinion this hypothesis is corroborated by the highly negative coefficient (-0.239) given to restaurants with a touristy ambience in Option #2 regression model. A touristy place with a garage is almost the polar opposite of the qualities that create a memorable ambience and a correspondingly high rating on Yelp.

4. Attire was also one of the attributes that could significantly increase average ratings, as it had a coefficient greater than 0.22 in both models. I suspect attire may be serving as a proxy for price and elegance here. It also fits in with the broader theme of ambience being a hugely important factor.