The Attributes of successful and un-successful restaurants from the Yelp Database

Alejandro G. Bedregal

November 22nd 2015

Introduction

For somebody who is starting a new business, or for organizations (i.e., banks) that lend money to start such endeavors, it is very important to have an idea of what factors will make the business succeed or fail. In this spirit, I use the Yelp database and its large sample of restaurants in order to answer the following questions: what attributes make a restaurant (un-) successful?. For each restaurant I obtain (1) the average customer evaluation via the star-ranking, (2) the total number of reviews and (3) the number of check-in customers as estimators of how good a given restaurant is doing. I combine these success-tracers with the information of 37 different business attributes contained in the Yelp database, aiming to find distinctive features that characterize successful and un-successful businesses.

Methods and Data

In this work we use the Yelp database available for their *chagenges* and for academic purposes. The data set consist on 5 JSON files which include information about the business themselves, the users who evaluate the businesses, the users' reviews, tips and average weekly check-in information for many of the business listed. We make use of 2 of these files. First, the business specific data. It includes specific IDs, location details, business category, star-ranking and attributes among others for 61,184 businesses. The business attributes consist on 37 different categories, many of which with sub-categories, that the business responsible must fill while making the business profile in Yelp. The second JSON file we use is the one with average weekly check-in information for many of these businesses.

Defining Restaurant sample

In this study, we focus on business flagged as "Restaurants" in the Yelp category. This is mostly because "Restaurants" constitute the largest category sub-sample in the Yelp database, providing the largest statistical sample to run our analysis. A total of 21,892 business flagged as "Restaurants" were found (36% of all businesses in Yelp) which exceeds in almost a factor of 3 the second most common category ("Shopping").

Exploratory data Analysis: Defining "Success" tracers and Final sample

As part of our preliminary data analysis, we faced the question of which parameters to use to trace "success". Quantifying how successful a business is depends on multiple factors and it might be subjective depending on how we define "success". The Yelp database neither includes information concerning the business sale incomes nor operational costs. Without this information, we must relay on other, indirect and sometimes subjective indicators to trace the success of a business.

We use 3 indicators in order to trace success: mean number of stars from the reviews, total number of reviews, and mean number of check-in per week. At this point we proceed to make a final selection cut in our sample. Using the restaurant's ID numbers we crossmatched the selected rows in the first JSON file with the check-in information in the second JSON file. We only selected restaurants with check-in information. This reduced our number of restaurants to 18,640 (85% of the total business flagged as "Restaurants" in Yelp) and it constitutes our final sample ("the sample" from now on).

In Figure 1, we compared our three success-tracers vs each other for our sample of restaurants. In the left and central panels, we see that the number of stars has an overall positive correlation with the number of reviews and check-in. However, this tendency is not monotonic. In the fourth quantile (e.g., stars ≥ 4.5) the tendency reverses such as the restaurants with the largest average number of stars are not necessarily those with more reviews or those with more clients checked-in.

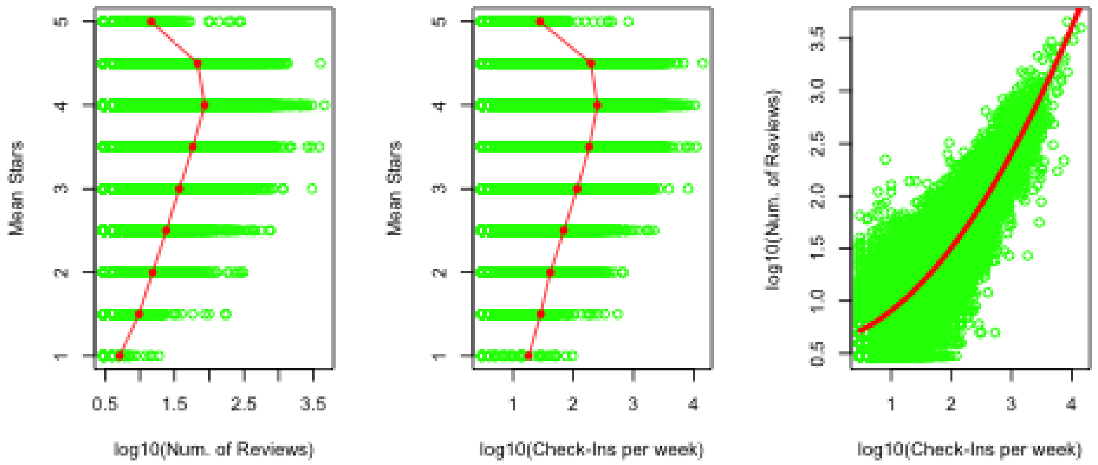

Figure 1: Our 3 success-tracers (stars, number of reviews and number of check-in) versus eachother. In red, best model fits (segments connecting medians and 3rd-order polynomial, respectively).

The right panel of Figure 1 shows a positive correlation between number of checks-in and total number of reviews. This is not a surprise: as more people visit a given restaurant, more people is in the position of reviewing the place.

Overall, Figure 1 shows us at least two different ways to quantify the success of a restaurant: the average number of stars and the mean number of checks-in per week. Critical for our study of the most successful restaurant's features is the fact that these two parameters are anti-correlated in the high end of the star ranking (e.g., stars ≥ 4.5). In the following sections we will select "successful" restaurants using both, stars and check-in, independently.

Results

In Figure 2, we show the histograms of sample selection by stars (left) and check-in (right) for our restaurant sample.

In the left panel of Fig.2 we see that the star-ranking distribution is not centered at the middle of the range but at 3.5. In other words, following the tendency of the whole Yelp business data set, the reviewers of restaurants tend to be "positive" at the time of assigning number of stars to a given business.

From each sample selection histogram we defined the "best" and "worst" restaurants independently. From the stars histogram, we select the "best" restaurants as those with Avg. Star ≥ 4.5 (blue, 2,145 restaurants in total). This condition is slightly more restrictive than using the 3rd-quantile of the distribution as originally planed. Similarly, we the "worst" restaurants were those with Avg. Star ≤ 2.5 (red, 2,667 restaurants in total). This selection corresponds to the 1st-quantile of the distribution. As a result of selecting best and worst restaurants in this way, the 2 sub-samples are separated in ± 1 star from the median value of 3.5.

For the check-in histogram, we selected the "best" restaurants as those with more than 158 reviews (blue, 4,637 restaurants in total, 3rd-quantile of the distribution). Worst restaurants have less than 16 reviews (red, 4,628 restaurants in total, 1st-quantile of the distribution).

Also in the right panel of Fig.2, we included two dashed histograms to compare best/worst samples selected using our two success-tracers. The cyan (orange) dashed histogram shows the best (worst) restaurants selected using the star-ranking. It is obvious that these dashed histograms differ significantly from the solid blue and red, stressing the fact that we are selecting very different sub-samples of best/worst restaurants with each of our two success-tracers.

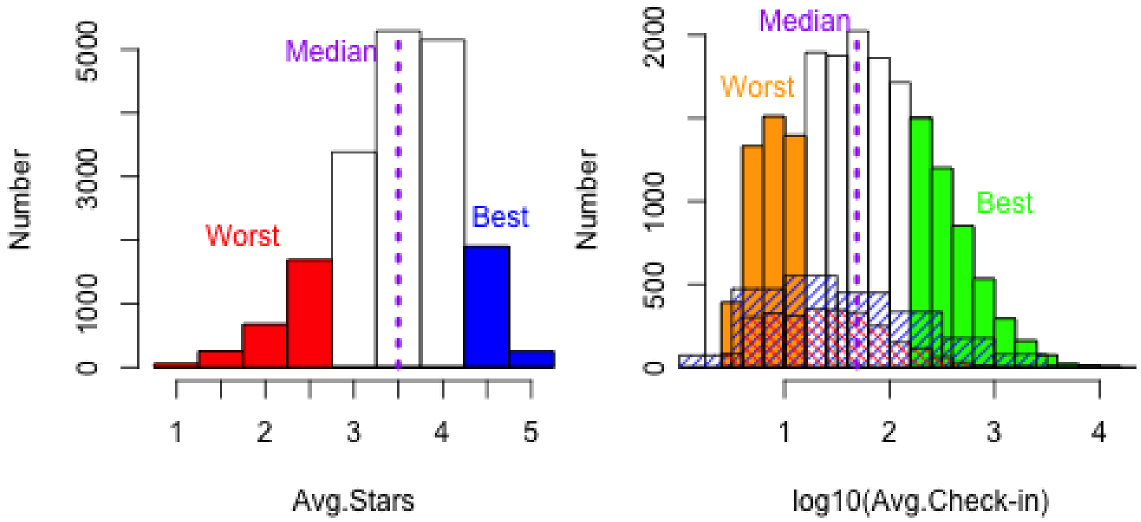

Figure 2: Histograms of sample selection by star-ranking (left) and number of check-in (right). In red and orange the corresponding worst samples. In blue and green, the best samples. median of the distributions in purple. In the right panel, shaded histograms correspond to galaxies selected as best and worst in the left panel.

Simulations: Errors and sub-sample completeness levels

For the best, worst and *average* sub-samples (the latest being all the restaurants not selected as best or worst) from each of the success-tracers (6 sub-samples in total) we run multiple mock-catalog simulation in order to define (1) the error in the estimated incidence of a given restaurant attribute, and (2) the minimum completeness required for an attribute to be statistically significant at the estimated error.

For each sub-sample and attribute we run 1,000 simulations creating mock catalogs by randomly picking up restaurants from the sub-sample. This allowed us to retrieve a distribution of the total incidence of an attribute within a sub-sample and therefore we could estimate the uncertainty (1-sigma distribution width). We found that, for the different attributes and sub-samples, a statistically significant (3-sigma) difference in a given attribute must be of, at least, 10%.

As above mentioned, we also estimate the minimum sub-sample completeness required for an attribute to be considered reliable. This require some additional explanation. When the restaurant responsible is filling the attributes to make the business profile, he/she is not obliged to fill all attributes to finish the process. This implies that for a given attribute in a sub-sample there will be a certain number of missing data (usually flagged as NA). If the fraction of missing data in an attribute is too large (for example, 80%) the results we obtain from it might not be truly representative of the whole sub-sample. Therefore, we must determine via modelling what is the maximum fraction of NA data we can allow in order to consider the attribute information truly representative of the whole sub-sample. We performed more mock catalog models, this time increasing the number of NA data in steps of 5%. Summarizing, we found that we need sub-sample completeness of, at least, 85% for a given attribute in order to consider the results representative.

The above criterion removed several attributes from our analysis as we cannot consider them as representative. In Table 1 we present our main results for 27 attributes, all reliable for our 6 restaurant sub-samples. The results are presented as a percent incidence of a given attribute for each sub-sample. The percentages were obtained by normalizing the number of incidences by the total number of valid (not NA) data per attribute and sub-sample (not by normalizing by the total sub-sample size as if would bias our relative comparisons between sub-samples).

Discussion

Num.	Attribute	B.Star	Avg.Star	W.Star	B.Check-in	Avg.Check-in	W.Check-in
		[%]	[%]	[%]	[%]	[%]	[%]
1	Accepts Credit Cards	89.7	95.6	93	96.8	86.9	89.2
2	Good For Groups	80.5	87.7	74.5	88.6	83.4	81.7
3	Outdoor Seating	38.9	42	26.3	37.4	30.7	29.3
4	Price Range 1	49.3	41.1	51.1	50.6	47.8	56.4
5	Price Range 2	35.9	49.8	37.1	43.2	44.7	43.6
6	Price Range 3	7.4	4.8	2.6	4.9	5.9	0
7	Price Range 4	2.8	0.8	0.7	1.2	1.6	0
8	Good for Kids	77.5	80	76.3	82.1	79.8	83.8
9	Has TV	35.2	48.8	33.7	54.9	51.3	61.4
10	Attire casual	87.8	93	87.3	94.8	96.9	100
11	Attire dressy	5.8	2.6	1.2	4.4	3.1	0
12	Attire formal	0.3	0.1	0.5	0.7	0	0
13	Take-out	82.7	88.1	84.9	95.3	93.6	84
14	Takes Reservations	32.9	34.1	16.8	30.2	33	19.9
15	Waiter Service	49	59.2	37.2	65	68.4	84.7
16	Caters	36.8	31.8	14.4	50.7	52.9	29
17	Good For dessert	2.7	1.3	0.7	1	2	0
18	Good For latenight	2	6.1	4.3	6.3	8.9	29.2
19	Good For lunch	31.9	40.3	21	41.1	33.2	6.3
20	Good For dinner	19.8	30.4	13.1	36.7	39.1	18.2
21	Good For breakfast	6.8	8	7.7	9.1	8.7	19
22	Good For brunch	6.4	4.8	2.4	3.2	6.7	17.6
23	Parking garage	2.4	7.3	5.4	5	2	0
24	Parking street	19.4	13.2	4.9	23.7	41.4	22.4
25	Parking validated	0.5	0.5	0.2	0.8	0.7	0
26	Parking lot	37.8	47.3	25.2	48.8	38.6	34.8
27	Parking valet	2.1	3.3	1.4	3	2	0

Table 1: Percent incidence of a given attribute for our 6 restaurant sub-samples. Made for the 27 attributes with completeness > 85% in all 6 sub-samples.

In the right panel of Fig.2 we observed how different the best and worst samples are depending on the way they have been selected (e.g., via star-ranking or through number of check-in). Our results from Table 1 show that, for statistically significant differences (>10%) between sub-samples (best, average or worst), our different success tracers show both consistent and inconsistent results between them. In particular, the star-ranking selection shows more often unclear trends in a given attribute for best-average-worst sub-samples. In contrast, the check-in classification shows more clear differences for the same attributes, particularly between the combined best+average group versus the worst sub-sample.

Attributes of un-successful business

The features of un-successful businesses are the easiest to identify in our data set. We define *strong* features as those that appear using our two success-tracers. *Weak* features are those that appear with only one of our success-tracers. So, the *strong* features of un-successful restaurants are:

- Take Reservations: Worst restaurants tend to take *less* reservations than the group of best+average restaurants (by >16% and >10% difference depending of success-tracer used).
- Cater: Worst restaurants have comparatively *less* cater services than best+average business (by >16% and >20% difference depending of success-tracer used).
- Good for Lunch: Worst restaurants are *less* often flagged as good for lunch than best+average business (by >10% and >27% difference depending of success-tracer used)

Now, some *weak* features of un-successful business:

- Good for Late Night: Worst businesses are flagged >20% *more* often as good for late night that best+average in the check-in selection. No indication of a trend in the star-ranking selection.
- Good for Dinner: Worst businesses are flagged >20% *less* often as good for dinner that best+average in the check-in selection. In the star-ranking selection we observe a similar trend, but it is not statistically significant.
- Good for Brunch: Worst restaurants are flagged >10% *more* often as being good for brunch that best+average in the check-in selection. No trend observed in the star-ranking selection.

Attributes of successful restaurants

It was difficult to identify specific features that take successful businesses apart from average and worst restaurants. As we described above, best restaurants have many features that are indistinguishable from those of average businesses. No *strong* attributes were found for the best restaurant samples. Only one, *weak* attribute was identified:

- Parking in Lot: For the check-in selection method we found that the best businesses are flagged >10% more often as having their own parking lot. No such trend is observed if star-ranking is used instead. However, for this selection method we observe that *worst* business have *less* often their own parking lot (>12% less compared to best+average).

Therefore, the data suggests that customers care about having an easier parking spot when they chose a restaurant. And also, the data hints that those businesses that have their own parking lot do slightly better than those who don't. Interestingly, the data for *Parking in street* seem to point in the same direction: in the star-ranking selection, worst businesses have less often street parking than best+average restaurants. This difference, however, is slightly smaller than our 10% threshold in order to be considered statistically significant.

Conclusions

We briefly summarize our main conclusions: - It is challenging to trace *restaurant business success* using the Yelp star-ranking and number of check-in. These tracers select significantly different sub-samples of *best* and *worst* restaurants. Also, comparing the (percent) incidence of a given attribute between sub-samples does not allowed us to conclude that one of the selection methods is superior than the other in identifying a distinctive population of restaurants.

- Relative to the best and average restaurants, some *strong* features of the worst businesses are: not taking reservations, not having a cater service and not being oriented for lunch. Some *weaker* attributes of the worst businesses are: being too oriented to late night, brunch, and not so much for dinner as their best and average counterparts.

- Relative to average and worst businesses, the only feature we found that distinguishes best restaurants is the availability of parking. In particular, having your own parking lot seems to be desirable by customers.

Overall, these results must be interpreted with caution. We **do not** claim that the attributes we found for best and worst businesses represent them in *absolute* terms, but only *relative* to the other sub-samples. In fact, for most of the attributes presented in Table 1, the percent incidence is <50%. This means that a large fraction of those business **does not** possess the given attribute but they are still (un-) successful restaurants.

Finally, we highlight that this study does not include the attribute of *location* in the analysis which might be fundamental for the success of a business. In the future we are planning to introduce this variable by using the geographical coordinates of these businesses and study their clustering around metropolitan/downtown areas.