

Restaurant busyness attributes effects on business star rating

Alexandr Cherkashin

11.11.2015

Introduction

This report provides an attempt to find some objective restaurant characteristics which affects business star rating. There is no goal of building comprehensive predictive model. Instead there is an attempt to identify meaningful attributes that busyness of type “restaurants” could deal with to improve their star rating on Yelp. Understanding what is important to customers is crucial for running business and we hope this report could help.

Methods and Data

Data

In analysis we will use Yelp business dataset. This dataset contains `star` variable that is an averaged business star rating from 1 to 5 star rounded to half of a star. It will be our response variable. As we interested in businesses with type “restaurant” we’ll take into analysis only observations with string “Restaurants” in `categories` variable.

In addition dataset contain several busyness characteristics some of them we will use as predictors. There are different kinds of variables. Some of them we will transform:

1. From latitude and longitude we’ll generate `region` variable which can take three different values: “USA”, “Canada”, “Europe”.
2. `attributes.Noise.Level` will be encoded into integer, from “quiet” to “very_loud” into 0:3, and stored in `noise_level` var.
3. Attributes related to *parking* we will transform to `parking` variable with 3 different values: “no” if there is no parking available, “yes” if there are some parking option except it is not parking street, “street” if there is parking street there.
4. From `categories` var there would be an attempt to make var with some sort of geographic theme or cuisine: `ThemedAmerican`, `ThemedOther` - all other regional themes, and `NotThemed`.
5. Another attempt of categorization based on `categories` variable would be type of restaurant business: `Buffets`, `Fast Food`, `Bars`, `Cafes`, `Other`.
6. `attributes.Alcohol` we’ll simplify to two level factor: `YES` - if there is some, otherwise - `NO` and stored in `alco` variable.

Other variables will be taken from dataset without transformation.

There are some variables which contains “Good.For” or “Ambiance” substrings in their names. Such vars appears a little vague on what they mean and how they were measured. In this report we will not take them into analysis, concentrating on more clear some.

There will be no attempts to impute missing values. Only `complete.cases` will be taken.

For the details about feature engineering and other data preparation steps please refer to R code in “report.Rmd” file at: (<https://github.com/amchercashin/CapstoneProject-Yelp/tree/business-analisy>)

Methods

There were some hesitation on what model to choose. Actually, our response: `stars`, which was measured in halves of a star from 0 to 5, could be interpreted as an ordered factor variable. It has an ordered nature for sure, it is an averaged users star rating rounded to half of a star. But there is no clue that the “distances” between star halves are the same despite their location. So models like [ordered logit](#) or [ordered probit](#) could make sense.

But after all considerations choice was made in favor of linear model. The main point of this is that the goal of this work is not to build a good predictive model. Instead, we'd like to find what affects star rating most and try to infer the effect and *interpret* it. So the interpretability is what most valuable, and linear model is really easier to interpret. So will use linear models and refer to `star` variable like it was a real number.

Results

With variables which left after data preparation and feature engineering we build the linear model, let's look at `summary` table of it:

```
##
## Call:
## lm(formula = stars ~ region + cat1 + cat2 + attributes.Outdoor.Seating +
##     parking + alco + cat1:alco + parking:alco + I(attributes.Price.Range^4) +
##     I(noise_level^2), data = restaurants, subset = comp_cases)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.56484 -0.39343  0.01057  0.41175  1.99917
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.983773   0.047963  83.059 < 2e-16 ***
## regionCanada    -0.173338   0.032474  -5.338 9.55e-08 ***
## regionUSA       -0.286282   0.030389  -9.421 < 2e-16 ***
## cat1OtherRestaurants -0.266389   0.038001  -7.010 2.48e-12 ***
## cat1Fast Food   -0.537104   0.039015 -13.767 < 2e-16 ***
## cat1Buffets     -0.895601   0.061565 -14.547 < 2e-16 ***
## cat2ThemedAmerican -0.086484   0.015824  -5.465 4.70e-08 ***
## cat2ThemedOther   0.057096   0.012279   4.650 3.35e-06 ***
## attributes.Outdoor.SeatingTRUE 0.068219   0.010437   6.536 6.50e-11 ***
## parkingyes       0.259896   0.016699  15.564 < 2e-16 ***
## parkingstreet    0.384776   0.026146  14.717 < 2e-16 ***
## alcoYES         -0.215975   0.061825  -3.493 0.000478 ***
## I(attributes.Price.Range^4)  0.001230   0.000160   7.688 1.59e-14 ***
## I(noise_level^2)  -0.054288   0.003012 -18.022 < 2e-16 ***
## cat1OtherRestaurants:alcoYES  0.221146   0.060400   3.661 0.000252 ***
## cat1Fast Food:alcoYES    0.466206   0.062973   7.403 1.40e-13 ***
## cat1Buffets:alcoYES      0.510550   0.089684   5.693 1.27e-08 ***
## parkingyes:alcoYES    -0.130411   0.025064  -5.203 1.99e-07 ***
## parkingstreet:alcoYES  -0.196422   0.035003  -5.612 2.04e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.612 on 14868 degrees of freedom
```

```
## Multiple R-squared:  0.1219, Adjusted R-squared:  0.1208
## F-statistic: 114.6 on 18 and 14868 DF,  p-value: < 2.2e-16
```

The impact of variables which are absent in table was considered both small and not very sustainable. The process of choosing variables was in fact manual. Tries and errors, so there is not much sense describing it here.

Let's look at the model closer. The *intercept* - the baseline - is an Cafe in Europe without alcohol, with no regional theme, without outdoor seating and without any parking option.

We see that there are different intercepts of star rating for different regions and different categories of restaurants. There is a tendency to rate them more in Europe, in Canada mean rating is smaller by 0.17 of a star and in USA even less by next 0.12. What is interesting that it is such a negative effect of being fast food and buffet *except* you are offering alcohol there! As for cafes alcohol is a bad idea and for other restaurants it is OK.

Parking usually gives significant boost to score, especially if there is a parking street. But.. the effect shrinks by half if you offer alcohol at place.

There is a tendency to rate restaurants with American theme little less than other. The effect is small, but nevertheless. Outdoor seatings gives a little plus to overall rating.

It is interesting how noise level affects rating. It looks like that negative effect of increasing noise is accelerative in nature: every next level of loudness gives a more and more negative effect on rating. From different approximations we choose a one simple: quadratic.

```
## Analysis of Variance Table
##
## Model 1: stars ~ I(noise_level)
## Model 2: stars ~ I(noise_level^2)
## Model 3: stars ~ I(noise_level^3)
## Model 4: stars ~ I(noise_level^4)
## Model 5: stars ~ exp(noise_level)
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1  14885 6209.6
## 2  14885 6182.5  0    27.122
## 3  14885 6197.7  0   -15.235
## 4  14885 6214.8  0   -17.025
## 5  14885 6193.3  0    21.486
```

It is better than other and simpler to interpret than a large polynome.

So it looks like that people tolerance to the noise is falling rapidly. There is a half of a star between “quite” and “very_loud” levels: $3^2 \cdot 0.0543 = 0.49$, and most of it is between very loud and average levels. If you have a very loud environment you can gain 0.27 of a star going one step towards quite and 0.16 more going one step further. We have found no correlation with other available variables, but there could be some with different types of restaurants. And another important question is how noise levels was measured, it's a pity we don't know.

There is a small but sustainable positive effect of price range. It is non linear too, but still it is small.

Discussion

First of all, an adjusted R-squared of presented model is 0.1208. So the model describes only 12% of variance in restaurant ratings. Surely every restaurant is different and there are much more important things like quality of food, quality of service, good location spots and others and which really should describe the rest.

But as we see we already have some objective, impersonal and measurable factors which could also affect rating. It looks like that working on them could increase business appeal.

Certainly there are numerous interaction possibilities which are hard to explore. One of the main conclusions that we draw from this analysis was that it is better to concentrate on even more narrow theme. For an example like exploring only for one narrow type of restaurant in specific city. And if you succeed you can compare you results to other types or cities. Also it looks like that such separate analysis could have more real value.

There was some simplifications in feature design. The main reason was to decrease feature space for better interpretability. That was the price for choosing a broad question. `cat2` variable is an example: surely it should be more specific. Typology based on `categories` original variable is a deep question by itself.

Another point: there are some doubts about several variables. Actually it could be that price range variable with combinations for an example with alcohol and categories - all are signs of some hidden “type of restaurant” variable. Really good typology of businesses again is a hard question by itself.