Restaurant business attributes affecting Yelp 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 business 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 are interested in businesses with the type "restaurant" we need to analyze only the observations with the string "Restaurants" in the variable `category`.

In addition dataset contains several business characteristics some of them will be used 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 - `cat2` variable: `ThemedAmerican`, `ThemedOther` - all other regional themes, and `NotThemed`.
5. Another attempt of categorization based on `categories` variable would be type of restaurant business - `cat1` variable: `Buffets`, `Fast Food`, `Cafes`, `Other`.
6. `attributes.Alcohol` we'll simplify to two level factor: `YES` - if there is some, otherwise - `NO` and store it in `alco` variable.

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.

We will throw away variables with more then 30% NA values count. And also variables with near zero variance which would be indicated by `nearZeroVar` function from `caret` package with default options.

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

After all data preparation we have data with 14887 observations (complete cases). We have six generated variables mentioned above. And we have some original variables which were left is data set after all filtering. These variables will be tested on affecting star rating.

To get some intuition of the data let's look at observations count for two selected variables: `region` and `cat1`:

Table 1: Table 1: Contingency table among region and restaurant category (cat1)

	Cafes	OtherRestaurants	Fast Food	Buffets	Total
Europe	27	379	72	3	481
Canada	99	1124	178	13	1414
USA	334	9075	3264	319	12992
Total	460	10578	3514	335	14887

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-analisys

Methods

There was 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. And there is no guarantee that the "distances" between star halves are 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 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 has the most value. And linear model is really easier to interpret. So we'll use linear models and refer to `star` variable like it is a real number.

Results

With variables which left after data preparation and feature engineering the process of variable choosing started. The process in fact was manual. Tries and errors. So the is not much sense describing the full path here.

After all the final linear model was built, let's look at model overall characteristics:

```
##
## 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
##
## 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
```

And then the coefficients table:

Table 2: Table 2: coefficients from linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.9837727	0.0479633	83.058807	0.0000000
regionCanada	-0.1733379	0.0324736	-5.337810	0.0000001
regionUSA	-0.2862821	0.0303892	-9.420531	0.0000000
cat1OtherRestaurants	-0.2663893	0.0380006	-7.010133	0.0000000
cat1Fast Food	-0.5371040	0.0390149	-13.766632	0.0000000
cat1Buffets	-0.8956014	0.0615645	-14.547358	0.0000000
cat2ThemedAmerican	-0.0864842	0.0158243	-5.465280	0.0000000
cat2ThemedOther	0.0570960	0.0122788	4.649973	0.0000033
attributes.Outdoor.SeatingTRUE	0.0682187	0.0104366	6.536481	0.0000000
parkingyes	0.2598963	0.0166987	15.563835	0.0000000
parkingstreet	0.3847757	0.0261457	14.716610	0.0000000
alcoYES	-0.2159753	0.0618254	-3.493312	0.0004785
I(attributes.Price.Rangeˆ4)	0.0012301	0.0001600	7.687824	0.0000000
I(noise_levelˆ2)	-0.0542882	0.0030123	-18.022405	0.0000000
cat1OtherRestaurants:alcoYES	0.2211462	0.0604004	3.661336	0.0002518
cat1Fast Food:alcoYES	0.4662057	0.0629732	7.403235	0.0000000
cat1Buffets:alcoYES	0.5105501	0.0896840	5.692768	0.0000000
parkingyes:alcoYES	-0.1304106	0.0250641	-5.203082	0.0000002
parkingstreet:alcoYES	-0.1964216	0.0350030	-5.611571	0.0000000

All variables from Table 2 have extremely low p-values - the strong evidence of their association with star rating.

The impact of variables which are absent in Table 2 was considered both small and not very sustainable.

Let's look at the coefficients 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 for different regions and different categories of restaurants. There is a tendency to rate restaurants higher 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 or buffet *exept* if you are offering alcohol there! But 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 which sounds reasonable.

There is a tendency to rate restaurants with American theme little lower then others. 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 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 then other simple functions and easier 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 * 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 to the "loud" level. And 0.16 more going one step further to the "average". We have found no correlation with other available variables, but really there could be some with different types of restaurants. And another important question is how noise levels was measured, unfortunately we don't know it.

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. And this is understandable: surely every restaurant is different and there are much more important things like quality of food, quality of service, good location spots and others which really should describe the rest.

Certainly there are numerous variable 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 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 for business owners.

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 or outdoor seatings - all are signs of some hidden "type of restaurant" variable. Really good typology of businesses is both hard vary valuable.

But despite all this we already have some objective, impersonal and measurable factors which could also affect rating. Knowing them could help in running business or help to decide which type of it where should be opened.