How to Improve Your Restaurant?

Lize Du, Linquan Ma, Wenjia Xie

STAT 628

April 1, 2019

Outline

- Star Prediction
- Analyze Methods
  1. Nouns in Reviews
  2. Attributes
- Suggestions for Traditional American Restaurants
  1. Food
  2. Environment
  3. Service
  4. Facility

Star Prediction

Data: Over 5 million reviews

Method: Logistic regression

RMSE: 91.986%

Nouns in Reviews

1. Remain the words whose frequency larger than 4000, after this process, we have 1767 words left.

2. Use information gain to rank the words, choose top 1000 words.

3. Choose nouns using *nltk.pos_tag* method and in the end we have 603 nouns left.

4. Use all the review data of Traditional American Restaurants and the 603 nouns to build a linear model.

5. Use bootstrap to calculate the standard deviation of each word's coefficient in our model by rebuilding the model using random sample (size = 10000 observations) 1000 times.

6. Rank the word by coefficient/sd, focus on the top 100 positive and negative words and choose some informative words manually.

Attributes

Attributes are from the "business" dataset that reflect various characteristics of the restaurants.

There are 39 attributes in the dataset recording different aspects, such as "garage", "street", "validated", "lot" and "valet" in the "Business parking" attribute. We changed all the sub-attributes into independent columns and get 66 different attributes in total.

However, "accept insurance", "byob" and "restaurants counter service" only has one level. Hence we drop them from our analysis. A list of most important attributes for improving the stars are selected based on CART and ANOVA.

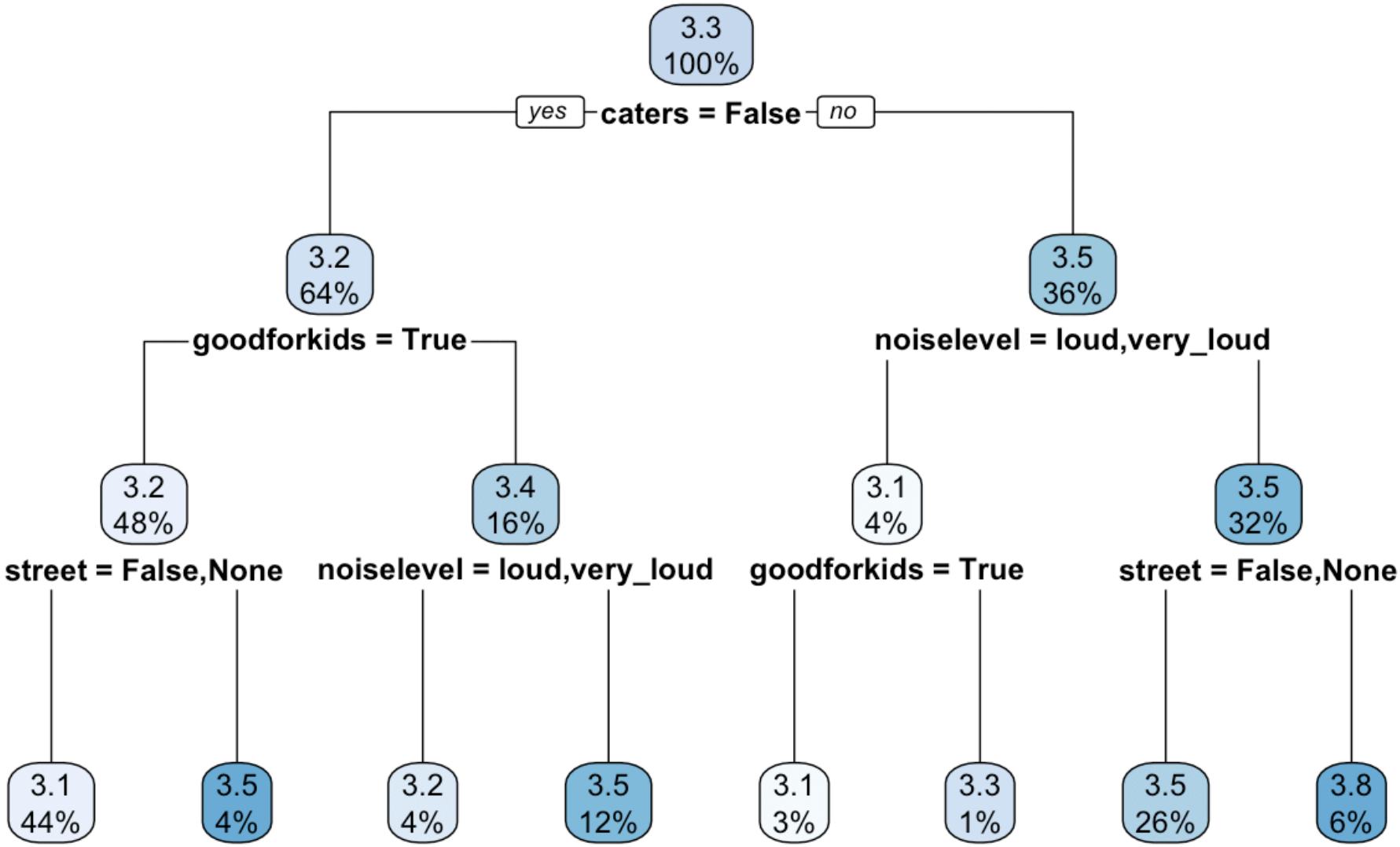
Classification and Regression Trees (CART)

- Missing data

  The overall missing rate is around 0.5, first drop the attributes whose missing rate is greater than 0.5. Then we have 36 attributes left with overall missing rate 16.5%.

  The R package *rpart* applies surrogate splits to deal with missing data comfortably. With *rpart*, we can directly applies the dataset with missing values to the rpart function.

Classification and Regression Trees (CART)



ANOVA test

1. Performed one-way ANOVA marginally for each attribute to see if an attribute is significant.

2. Ranked the p-values of each attribute from lowest to the highest, to discover that the previously top 5 attributes selected using rpart have the top 10 smallest p-values.

3. Confirms that the top 5 attributes are important indicators for review stars.

Food

Figure: Positive Foods

Figure: Negative Foods

Food

You'd better have

- Menu: Include burgers, pancakes, pork, tacos, filet and bbqs;
- Meat: Juicy and crispy;
- Dessert: Cakes, waffles or chocolates;
- Drinks: Tea and cocktails;
- Specials on your menu every few days.

What you should take care of is you shouldn't make the food too salty, too chewy or even burnt, raw, frozen or tasteless food.

Another tip is you should know if the customer wants lettuce in their sandwisches and burgers and they want what knid of lettuce in it.

Environment

Figure: Positive Environment

Figure: Negative Environment

Environment

You'd better

- Keep everything in their places all the time;
- Come up with some ideas to make people eating here happily;
- Prepare some big tables because some people may come with their family;
- Keep the restroom clean all the time;
- Make the air in your restaurant fresh;
- Keep the noise level low.

Service

Figure: Positive Service

Figure: Negative Service

Service

You'd better

- Train your employees well so that they can properly answer customers' questions and satisfy their requests, such as give the customers right recommendations.
- Prepare food quickly;
- Smile to customers;
- Make your team professional to serve with loves.

Facility

You'd better prepare

- Caters;
- Parking lots, especially street parking;
- Facilities that are good for kids, such as children seats;
- Restaurants delivery.

Advantages and Disadvantages

- Advantages
  1. For reviews, we focus on nouns and the result seems reasonable.
  2. For business, we use both ANOVA and Cart tree method to calculate the importance of features, which is reliable.

- Disadvantages
  1. For the prediction model, we only use logistic regression with 2000 covariates which can not predict too well.
  2. We are not able to quantify the effect of wach word or feature to the rate of review.


not_neg
back_neg
today
when_neg
chang
toast
use
eat_neg
sweet
an_neg
size
lot
spot
hope
larg
long
absolut


do_neg
half
walk
order_neg
excel
fill
quit
fresh
fun
select
might
just_neg
front

were_neg
run
instead
again_neg
going_neg
did_neg
me_neg
hear
yummi
star
thank
still


talk
ask
plu
felt
pay
us_neg
ye
fine
slow
pm
la
disappoint
dri
bill
less
els
averag
mean
howev
roll
mayb
guess
leav
manag
up_neg
highli


oh
bad
job
saw
minut
host
decent
ok
okay
seem
worst
money
tell
call
let
later
spici
tasti
cold
crispi
final
piec
noth
if_neg
leave
our_neg
fantast


although
perfectli
will_neg
town
he
late
season
someon
smooth
bite
music
sit
line
complet
money
put
least
amaz
water
awesom
patio
worth
ice
___neg
service_neg
arriv
share


cool
favorit
care
last
wine
notder
red
them_neg
may
perfect
almost
their_neg
or_neg
light
total
how_neg
bbq
at_neg
away
do
stay