Which Attributes Affect Business Stars

Ada Lee

November 20, 2015

Introduction

I will use Yelp business data to investigate which attributes affect stars of a business, so I would build a machine learning model to predict stars of business using attributes. If one want to know how to improve business, he or she will care about this.

Methods and Data

i. Preprocess the data

There are 61184 json in the business data. The format of business data is look like below. We will only use "stars" and "attributes" fields.


```
{
    "business_id" : "UsFtqoBl7naz8AVUBZMjQQ",
    "full_address" : "202 McClure St\nDravosburg, PA 15034",
    "hours" : {

    },
    "open" : true,
    "categories" : [
        "Nightlife"
    ],
    "city" : "Dravosburg",
    "review_count" : 4,
    "name" : "Clancy's Pub",
    "neighborhoods" : [ ],
    "longitude" : -79.88693,
    "state" : "PA",
    "stars" : 3.5,
    "latitude" : 40.350519,
    "attributes" : {
        "Happy Hour" : true,
        "Accepts Credit Cards" : true,
        "Good For Groups" : true,
        "Outdoor Seating" : false,
        "Price Range" : 1
    },
    "type" : "business"
}
```


There are many different values in attributes for different business json data, we only use "attributes"" field and "stars"" field. I do following things in sequence.

1. Only extract `attributes` and `stars`

2. Extract sub-fields of attributes as fields, also add `stars` as one of the field
3. Some sub-fields of `attributes` map to a sub-json, remove these sub-fields
4. Some sub-fields only exists in less than 20% of remaining data, remove them
5. read remaining data to data.frame, some rows contains missing values, remove them
6. Store the result to variable `business`

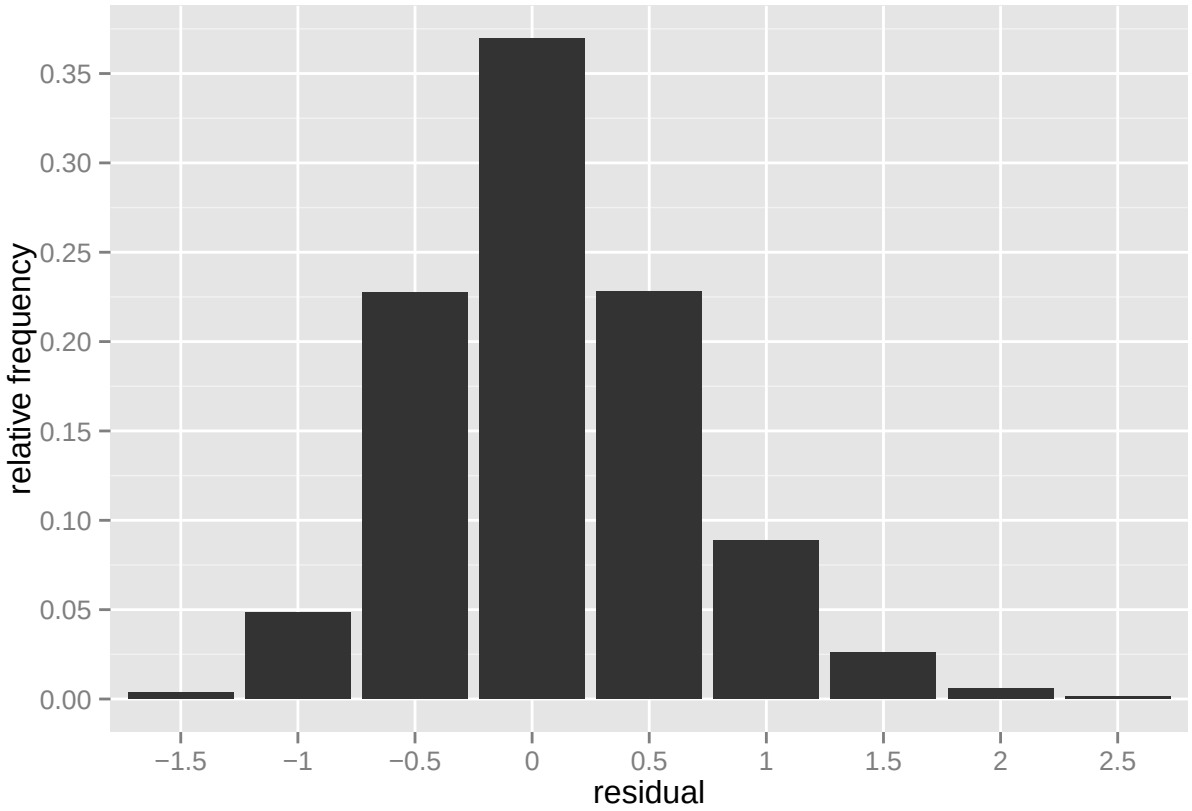
Now `business` has 6760 rows and 17 cols. And it has below columns:

`stars`, `Alcohol`, `Noise.Level`, `Has.TV`, `Attire`, `Good.for.Kids`, `Wheelchair.Accessible`, `Caters`, `Delivery`, `Accepts.Credit.Cards`, `Take.out`, `Price.Range`, `Outdoor.Seating`, `Takes.Reservations`, `Waiter.Service`, `Wi.Fi`, `Good.For.Groups`

ii. Build Random Forest Models

1. Divide the 6760 data to 20% training dataset, and 80% testing dataset.
2. Use training dataset to train random forest, use `stars` (ordinal data) as independent variable, use `Noise.Level`, `Has.TV`,`Good.for.Kids`, `Wheelchair.Accessible`, `Caters`, `Delivery`, `Take.out`, `Price.Range`,`Outdoor.Seating`, `Takes.Reservations`, `Waiter.Service`, `Wi.Fi`, `Good.For.Groups` as dependent variables.

iii. Residual plot



Accuracy is 36.9%. And 82.5% of prediction has error that is between -0.5 and 0.5. So our the performance of our model is good.

Results

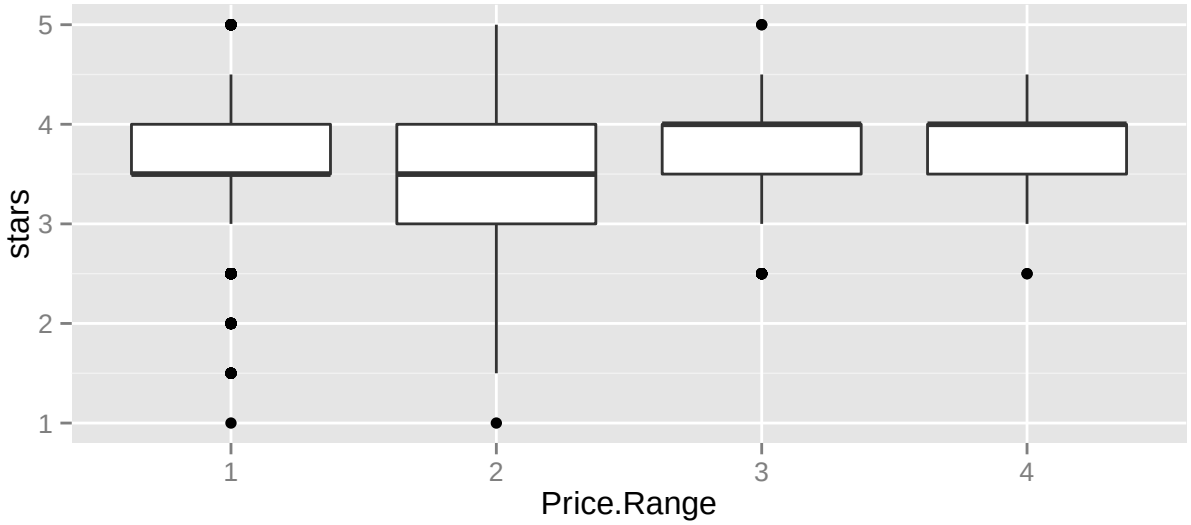
i. Variable Importance

```
library(caret)
varImp(model)
```

```
## rf variable importance
##
##                             Overall
## Price.Range                  100.00
## Has.TVTRUE                    72.57
## Outdoor.SeatingTRUE           69.48
## CatersTRUE                    57.31
## Wi.Fino                       55.97
## Takes.ReservationsTRUE        52.32
## Good.for.KidsTRUE             50.26
## Noise.Levelloud               48.66
## Waiter.ServiceTRUE            47.60
## DeliveryTRUE                  45.93
## Noise.Levelquiet              41.85
## Wheelchair.AccessibleTRUE     41.74
## Alcoholnone                   39.11
## Alcoholfull_bar               39.10
## Take.outTRUE                  38.40
## Good.For.GroupsTRUE           32.25
## Noise.Levelvery_loud          12.23
## Wi.Fipaid                      0.00
```

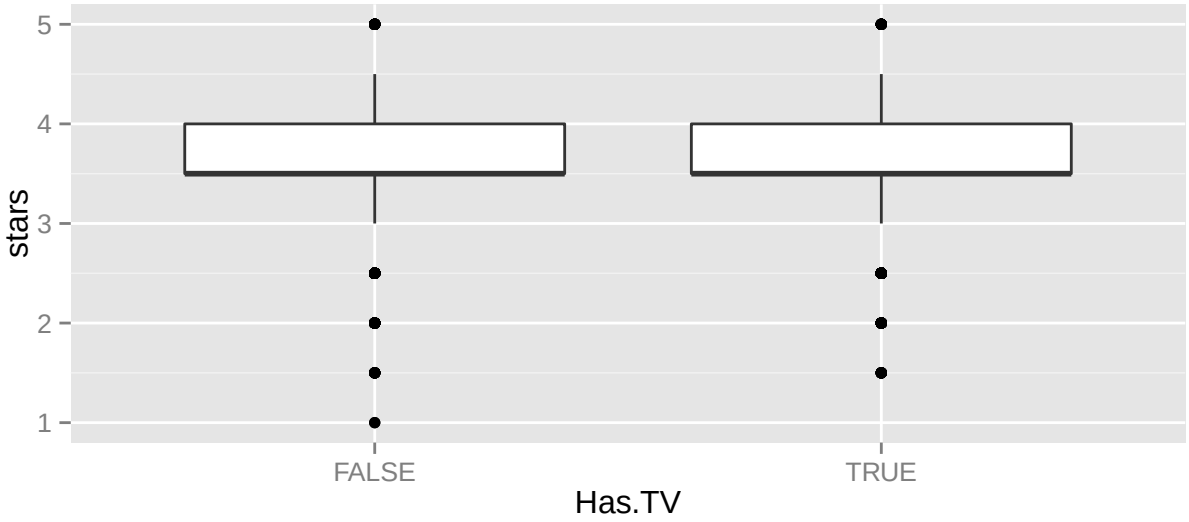
We see that `Price.Range`, `Has.TV`, and `Outdoor.Seating` and `Caters` affect stars most.

ii. How `Price.Range` affect `stars`



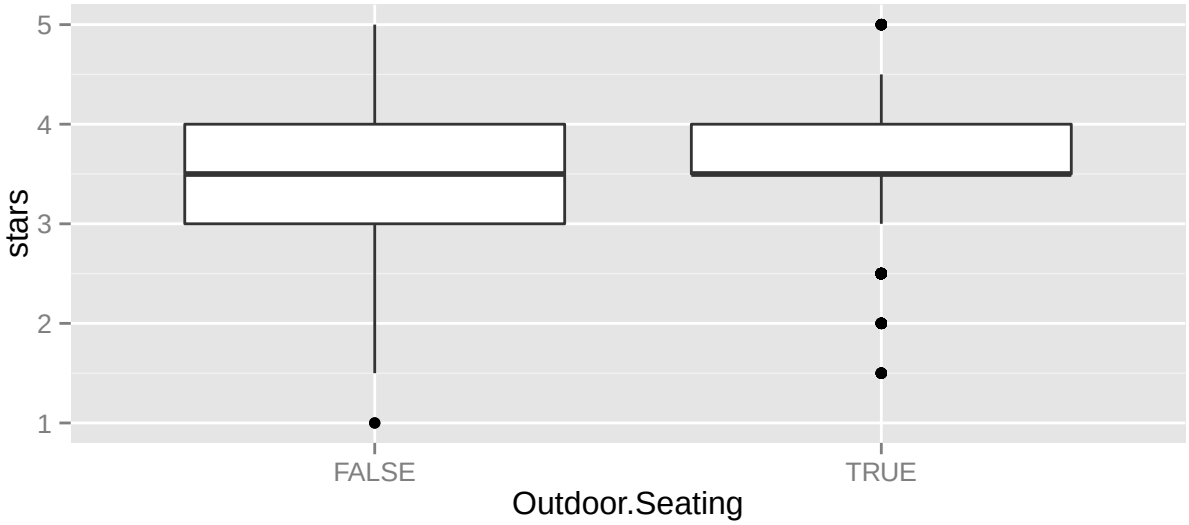
We can see that higher price business get larger stars from customers, this means that higher price business tend to have better quality of service.

iii. How `Has.TV` affect `stars`



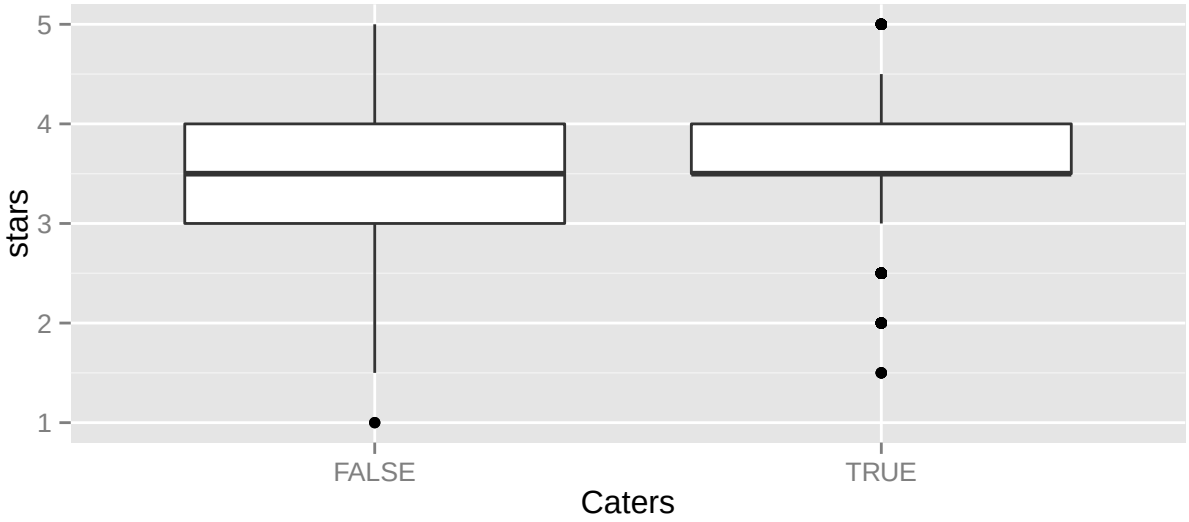
From the plot we see that whether have TV does not affect stars much.

iv. How `Outdoor.Seating` affect `stars`



From the plot we see that Among business that has outdoor seating, very few of them get stars that is less than 3. For business that have 5 stars, most of them do not offer outdoor seating, this may be that highest quality business often does not offer outdoor seating.

v. How `Cater` affect `stars`



From the plot we see that `Caters` has similar affect on `stars` with `Outdoor.Seating`.

Discussion

From our analysis we see that `Price.Range`, `Outdoor.Seating` and `Caters` affect business stars most. Higher price business get higher stars, I think higher price is due to higher quality, so if you want your business get high stars you can open high quality business, and it does not matter that the price is sometimes high. If you want to start a normal quality business, you can improve you business by offering outdoor seating and caters, if your want to start a highest quality business, then weather offering outdoor seating and caters does not matter.