Yelp Final Project Rubric

Steven Lee

1. Title:Understanding top 10 variables which determine restaurant business having more than 3 stars rating.

This is a report on determining the top 10 variables which contribute to a restaurant business having more than 3 stars. There are 3 datasets being used: business data, tips data and reviews data. A simple sentimental analysis is done to calculate a normalised sentimental star score based on tip and review comments. Hence, we have the original Yelp's stars rating, and also the normalised sentimental star rating by tips and reviews. Random forest is used to determine the important variables which let restaurant business have more than 3 stars rating. The top 3 factors (latitude, longitude and review count) are the same throughout the 4 analysis, while the next few variables varies from each other's. This analysis serves as a preliminary analysis on how sentimental analysis can be used with machine learning technique on such dataset. The full R code will be upload into my GitHub repository(leesinfook2015) at a later time, while partial r code will be shown in this report.

2. Introduction

At the beginning, I was confused on what to do with the 5 Yelp's datasets. Then I put myself in the customer's shoes on how will I use Yelp, what I will use it for, and which are the thing I will look up for. Hence, this has helped me reduce my scope to restaurant business, as I am a food lover, and I always research online to see which restaurants to dine in based on ratings and reviews.

Hence, my research problems will be:
Which are the 10 variables that determine a restaurant business having more than 3 stars rating? Will there be any changes in the top 10 variables if we will to use some sentimental analysis on tip and reviews dataset to create a new normalised sentimental star rating as compared to Yelp's star rating?.

3. Method

Based on my research problem, 3 datasets will be used: 1. Yelp Business data (dimension: 61184 x 109), 2. Yelp Tips data (dimension: 495107 x 5), 3. Yelp Reviews data (dimension: 1569264 x 9).

3.1 Extracing Restaurant business data

As I am only focusing on Restaurant business data, I will extract out the business who have "Restaurant" tagging.

```
rest_list=grepl ( "^(.*[Rr]estaurant.*)",  business_flat$cat)
rest_data=business_flat[rest_list,] # with dimension [1] 21892   109
```

3.2 Merging the 3 datasets using library(plyr)

First, I will merge Restaurant and Tips using "inner join" where they will merge by the "business id". Similarly, merging Restaurant and Reviews using "inner join" where they will merge by the "business id".

```
new_tip_data=join(rest_id, tip_flat, type = "inner") # dimension [1] 304388 x 6
new_review_data=join(rest_id, review_flat, type = "inner") # dimension [1] 990627 x 10
```

3.3 Removing all unnecessary data for sentimental analysis

For sentimental analysis, we will only need the tips/ reviews contents with the "business id", hence we removed the columns that are not useful for sentimental analysis.

```
new_tip_data2=new_tip_data[,c(1,4)]
new_review_data2=new_review_data[,c(1,7)]
#> c(names(new_review_data2), names(new_tip_data2))
#[1] "business_id" "text"         "business_id" "text"
```

3.4 Sentimental Analysis using Tips & Reviews Text

To create a sentimental analysis score, we have to first load in 2 sets of data (a list of positive words and a list of negative words), which we get in from **Hu and Liu, KDD-2004**.

```
pos.words = scan('positive-words.txt', what='character', comment.char=';')
neg.words = scan('negative-words.txt', what='character', comment.char=';')
```

Then, I will compare our tips/review text to the dictionaries of positive & negative term, using a *score sentimental* function:
"totalscore" = sum(positive words found) - sum(negative words found).

We also need to count the number of words in the each tips/reviews text using this **count.word** function:

```
count.words <- function(x) { sapply(gregexpr("\\W+", x), length) + 1 } # function to count words
```

Eventually we get a final score called **"ScoreWordsNormalized"**:
"ScoreWordsNormalized" = **totalscore** / **count.words** , for each text.

As the **"ScoreWordsNormalized"** will varies from a range of decimals, we will do **normalisation** on it to ensure the scale is from 0 to 1:
"Scorescaled2" = { scores - minimum(scores)}/ {maximum(scores) - minimum(scores)}

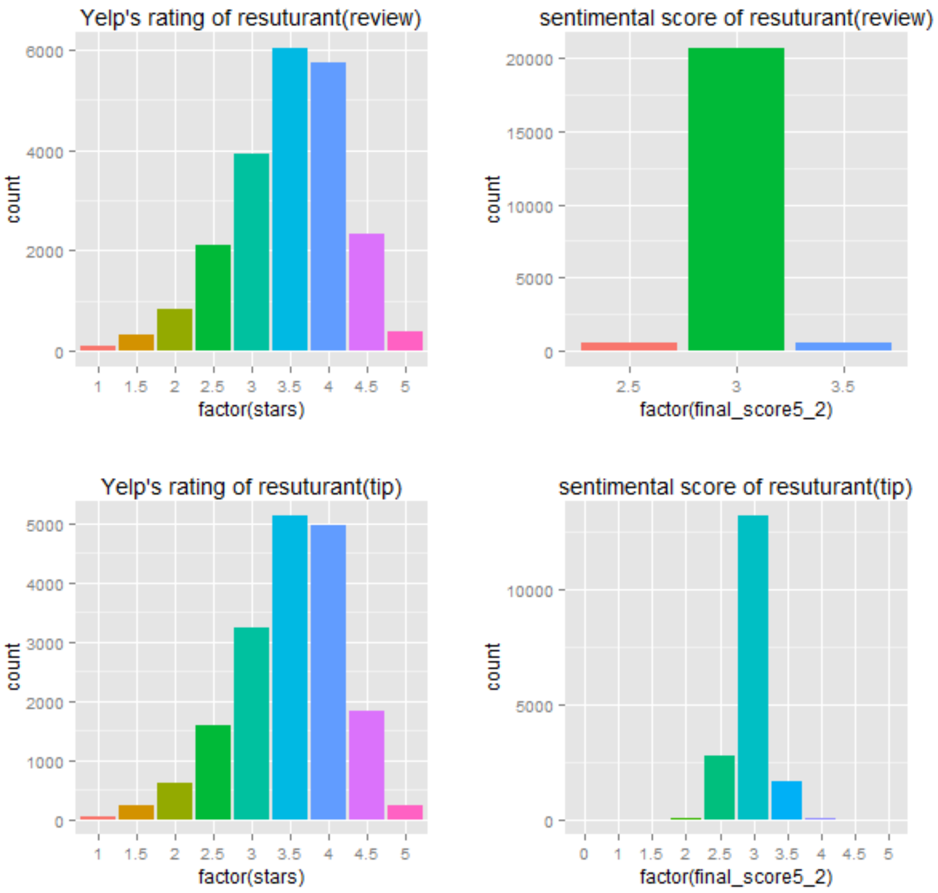
Lastly, with new normalised score, we can transform it to the Yelp's stars scoring system which is from 0 to 5 by increment of 0.5:
"finalscoreof52" = {celling(**"Scorescaled2"** * 10)} / 2

As each restaurant business id could have more than 1 reviews/ tips, we need to average up the sentimental score by business id:
"scorebybusinessID" = average of **"finalscoreof52"**, for each restaurant business id.
"scorebybusinessIDbyyelpscale" = celling(**"scorebybusinessID"** * 2) / 2, for each business id.

Lastly, we need to "inner join" the final sentimental star rating back to the restaurant business data.

```
library(plyr)
end_tip_data=join(rest_data, tipscorebygp1, type = "inner")
end_review_data=join(rest_data, reviewscorebygp1, type = "inner")
```

3.5 Visualising the Yelp's stars rating vs tips and reviews sentimental star rating



3.6 Creating new variable

Based on the barplot, it seem like the cut off at 3 star rating will be a good cut-off point. We create a new variable = **Target**, which = 1 when Yelp rating more than **3**, (similarly for sentimental score by tips and reviews). This variable = **Target**, will be use as the classifier to determine the top 10 variables for resutrant business.

3.7 Machine learning using Random Forest with 5 Cross-validation

1. We partition each of the dataset into 70% training and 30% testing dataset.
2. Removing variables that have more than 95% NAs.
3. Removing variables that are non-useful (such as business names)
4. Transform **characters** which have values as **TRUE**, else **FALSE**
5. Imputing those logical variables with **NA** as *FALSE*
6. Run Random Forest algorithiam with 5 cross-validation:

```
fitControl <- trainControl(method="cv", number=5, verboseIter=F)
fitrF4 <- train(target ~ ., data=train_4, method="rf", trControl=fitControl)
```

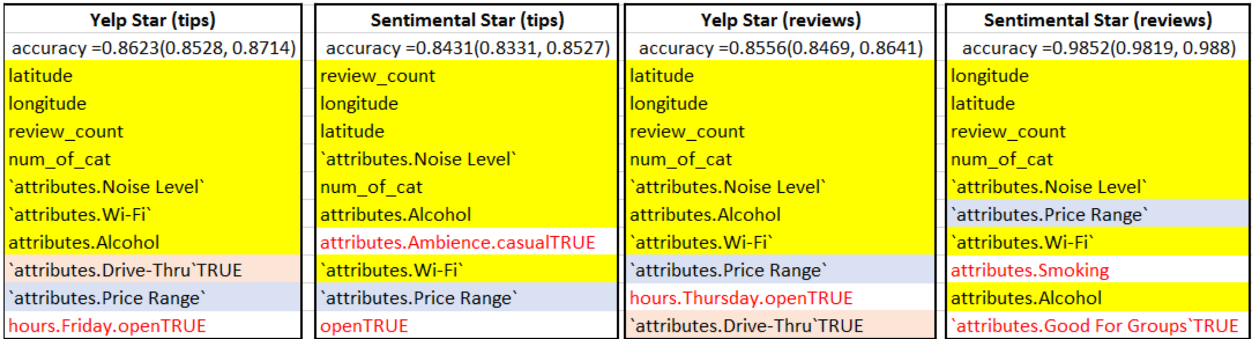
4. Results

In total, 4 Random Forest models were implemented (2 on tips (Yelp's stars vs sentimental star), 2 on reviews (Yelp's stars vs sentimental star)).

All the 4 models have rather good accuracy from 84% to 98%, which implies that the Random Forest models are quite good in classifying which are good restaurant business (more than 3 Yelp's stars or 3 sentimental score).

There are a total of 7 common variables (in yellow filled) from all the 4 models in their top 10 variables which are:
“longitude”, “longitude”, “review count”, “num of cat”“,”attributes.Noise Level“,”attributes.Wi-Fi“,”attributes.Alcohol“.

The most unique variables (in red font) which only appeared once are:
“hours.Friday.openTRUE”, “attributes.Ambience.casualTRUE”, “openTRUE”, “hours.Thursday.openTRUE”, “attributes.Smoking”, “attributes.Good For GroupsTRUE”.



Yelp Star (tips)	Sentimental Star (tips)	Yelp Star (reviews)	Sentimental Star (reviews)
accuracy =0.8623(0.8528, 0.8714)	accuracy =0.8431(0.8331, 0.8527)	accuracy =0.8556(0.8469, 0.8641)	accuracy =0.9852(0.9819, 0.988)
latitude	review_count	latitude	longitude
longitude	longitude	longitude	latitude
review_count	latitude	review_count	review_count
num_of_cat	`attributes.Noise Level`	num_of_cat	num_of_cat
`attributes.Noise Level`	num_of_cat	`attributes.Noise Level`	`attributes.Noise Level`
`attributes.Wi-Fi`	attributes.Alcohol	attributes.Alcohol	`attributes.Price Range`
attributes.Alcohol	attributes.Ambience.casualTRUE	`attributes.Wi-Fi`	`attributes.Wi-Fi`
`attributes.Drive-Thru`TRUE	`attributes.Wi-Fi`	`attributes.Price Range`	attributes.Smoking
`attributes.Price Range`	`attributes.Price Range`	hours.Thursday.openTRUE	attributes.Alcohol
hours.Friday.openTRUE	openTRUE	`attributes.Drive-Thru`TRUE	`attributes.Good For Groups`TRUE

5. Discussion

Based on the results from above, when we compared across the 4 models, having 7 common variables whereby the top 5 (“longitude”, “longitude”, “review count”, “num of cat”“,”attributes.Noise Level“,) are almost in the same ranking among the 4.

From this 5, we can say that, **location** (“longitude”, “longitude”) is a very important variables that determine if a restaurant business will have more than 3 stars. Follow by the the next 3 variables (“review count”, “num of cat”“,”attributes.Noise Level“), implies that having review counts, how diversify the restaurant business are, and lastly the noise level are important to have more than 3 stars.

The next 2 variables (“attributes.Wi-Fi”, “attributes.Alcohol”), does implies that in this current social media era, having Wi-Fi and alcohol will help in making the restaurant business better.

5.1 Comparing between Yelp' star vs Sentimental Star of tips

Almost these 2 models, there isn't much difference on the common 7 variables other than the ranking. The most unique difference will be Yelp's stars model, it having drive-thru and operating on Friday, while on sentimental stars, it focus more on having casual ambience and the business is open.

5.2 Comparing between Yelp' star vs Sentimental Star of reviews

Almost these 2 models, there isn't much difference on the common 7 variables other than the ranking. The most unique difference will be Yelp's stars model, it having drive-thru and operating on Thursday, while on sentimental stars, it focus more on having smoking function and if the restaurant business is good for group gathering.

I felt that the sentimental stars model on the review dataset, give an overall of how facilities in the restaurant business will help to achieve a sentimental score of more than 3.

5.3 Limitations

As the sentimental analysis was run based on the different between the sum of positive word found and sum of negative words found divided by the total words in the text. This might not be a good gauge of the score as the number of words in the text, usually at least equal or more than the sum of positive and negative words found. In addition, the list of positive & negative words might not be effective for these Yelp's dataset as these dataset were drawn from certain demographics of people who might use different positive and negative expression of words.

Imputing variables with NAs as FALSE might not be the best way to transform variables.

Having more than 32 factor levels in those attributes of operating hours from Monday to Sunday, will cause error in running the random forest. Thus a better transformation such as grouping the factor levels into "morning", "noon", "evening", might be a better variable transformation than the current logic true or false.

Number of categories and number of reviews might have hidden correlation with the stars rating, as more review might lead to more stars achieved.

5.4 Conclusion

Overall, this report is able to address my research problem on determine the top 10 variables that contributes to restaurant business having more than 3 stars and is there any difference in the top 10 variables if we will to use sentimental analysis star rating rather than Yelp's star rating.

This report serves as the first step in using simple sentimental analysis to create a better rating system as compared to the normal Yelp's star scaling and much more exploration and improvement can be implemented.