1.Title: Data Science Capstone - Yelp Final Project

Yelp is a company that connects people with business via users reviews or recommendations that those users made of the different business and that other users can read and consequently decide to visit or purchase. I personally decide to buy/visit a business based on the number reviews and average stars granted by previous users. I only read a reviews in case the first 2 criteria are met (enough number of reviews and more than 3,5 stars). This will normally lead to another review (mine) with similar number of stars if the review was adequate. Of course there are other criteria (like how close the business is to where I am or the type of specialization) but this is dependent on each user's situation. My project is to determine how strong is the relationship between previous reviews and next reviews (is my hypothesis/experience of cause effect demonstrated by the data provide by Yelp?)

2. Introduction

Primary question and the rationale for studying it: Is it possible to predict the average of stars a business will obtain by taking into account previous values (tips, check ins, number of reviews, minimum and maximum stars granted) ?

Would this change if I consider only influential users or reviews ? This is interesting for businesses because a high correlation of evaluations would imply consistency in the aggregation of user activity across time. Therefore, if a particular business experiences big increases or decreases, Yelp can help by highlighting problems to address or opportunities to explore.

3.Methods and data

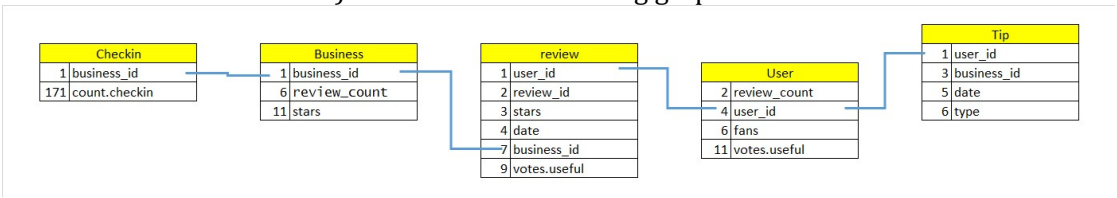
Yelp provided a set of data that would help this analysis The data provided by Yelp was 5 files in json format. Namely: reviews; users; business; checkins and tips. I read the data and coverted it by unflattening it.

I discovered the following files and number of records: 1.569.264 reviews; 366.715 users; 61.184 business; 45.116 checkins and 495.107 tips. After exploring the original data, I decided to take for my analysis only the relevant fields of each table .

In the case of checkin I just added up the numerous values into a single total by business by adding up all the columns from 3 to 170 and putting it in a new column(171)

Then I linked the data files together by using the common fields. In order to do it I had to link the different tables by the field highlighted in blue (e.g. user_id to link review and user and business_id

to link business and review). Please see the following graphic:



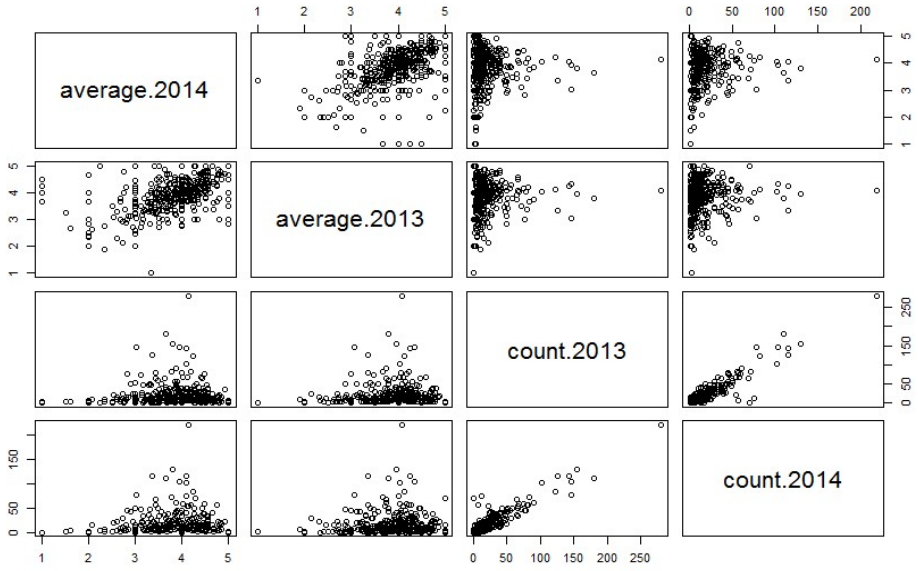
Considering that my analysis is trying to determine the impact of time on the number and average of stars, I had to transform the data into the following format:

	year 2015					year n					
Business_id	count	min	max	average	tips	count	min	max	average	tips	checkins
a											
b											
c											
d											

When we merge the data we do only a left join, This will remove those records where we have reviews for one year but not for the next or viceversa, the rationale for this is that the business that do not have consecutive years reviews will not qualify for this study and introduce noise to review given their new values after conversion. Therefore the total population to analyze is 10.446 businesses with their reviews, tips and checkins across several years.

By analyzing the data I see patterns that indicate that year 2015 is not a full year, hence I decided not to use this year in my analysis to avoid comparing years of 12 months with unfinished years.

Finally I plotted the relationship between the most interesting fields (variables) of my resulting table, obtaining the following plot that indicates some (not too clear but somewhat latent) relationships between the maximum and minimum values of the year and the average. Indicating that the average is predicted mostly by these 2 values. I decided therefor to use average.2014 as the variable I want to predict that I can extrapolate then to future years.



3.1 Description of the statistical model and prediction algorithm

My initial model was that I could predict the average of stars a business would receive in 2014 (average.2014) by the following 15 variables: count.2014 + min.2014 + max.2014 + count.2013 + min.2013 + max.2013 + average.2013 + count.2012 + min.2012 + max.2012 + average.2012 + count.chicken + tip.2012 + tip.2013 + tip.2014

However, before testing the model I had to remove the correlated variables that introduce noise to the model. So I used the function correlationMatrix and I discovered that the variables count.2013, count.2012, tip.2012, tip.2013 and count.2014 were highly correlated. Therefore I removed them from this model.

I partitioned the data into testing of 75% of the businesses and training of 25%

I tweaked the data of averages to make them only rounded number (integers), This will make the varialbe as categoritcal and it will allow me to have less intermediate values and use a very clear confusion matrix.

For the prediction I used the following statistical models with the following results:

Statistical Model	Accuracy	Accuracy bs(*)	P-Value
Recursive Partitioning	0.4475	0.4475	
Naive Bayes		0.6893	<2.2e-16
Random Forest	0.6893	0.6912	<2.2e-16

Given that the model was not giving accurate values of prediction, I applied bootstrapping (*) to improve the values, however, they still remain rather low and with very small increases.

Then I came to the conclusion that I should modify the population to those users that are more influential and only consider influential reviews.

In the case of reviews I considered influential only those that had a votes.useful >=1 that means that the review was taken into account by at least another user to make decisions hence this other user influenced at least another one. Furthermore, if the qualification of useful review was given by the next user after he purchased or visit the business, there are strong chances that he agreed with the first reviewer. The fans, as per my theory will also have a strong tendency to agree and hence have the same number of stars. The users will be more influential if they have more fans and if they have more votes useful. We will therefore only consider the reviews of those users with fans>=5 and votes.useful>1 The disadvantage is that this will reduce the number of records to consider, however these remaining records are more relevant at the time of influencing other users to buy or visit and perhaps also the average of future reviews.

I divide the data in training and testing records.

This approach reduces the number of records from 10.446 businesses to 4.772, however the consistency of the data should be more predictable if my hypothesis is right. I inspect again if I have values that have high correlation and I ended up removing almost the same fields.

If I run the same but now with this influential data I obtain:

Statistical Model	Accuracy	P-Value
Linear Model	0.83	<2.2e-16
Recursive Partitioning	0.72	<2.2e-16
Random Forest	0.89	<2.2e-16

4.Results

The results indicate that the average quantity of starts given by influential users and deemed as useful by at least another user is closely related to previous values (tips, checkins, maximum and minimum number of stars and number of reviews).

The primary question of interest was answered by the model. So we can conclude that I can predict with almost 90% of accuracy the average quantity of starts a business will receive if I know in advance what is the number of stars, tips, check ins, minimum and maximum values obtained before.

I produced the following confusion matrix to summarize the results of the model:

```
          Reference
Prediction   1   2   3   4   5
         1_ 20   0   0   0   0
         2   0  83   4   0   0
         3   0  16 201  21   0
         4   0   1  53 630  30
         5   0   0   0   5 128

Overall Statistics

               Accuracy : 0.8909
                 95% CI : (0.8719, 0.9081)
    No Information Rate : 0.5503
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.82
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity           1.00000  0.83000   0.7791   0.9604   0.8101
Specificity           1.00000  0.99631   0.9604   0.8133   0.9952
```

Throught the analysis of the data and the model, I conclude that the accuracy obtained is a bit bigger than 89% with a p-value of 2.2 x 10^-16, therefore the classifier is predicting quite well. I used accuracy as a messure gvien that the predictable variable is categorical and not numeric after my conversion of averages into integers.

The model predicts the reviews in 5 classes. And each of them represents the number of stars in the reviews.

From the above data we can see that our classifier correctly identified (sensitivity or true positive) the average number of stars by 100%, 83%, 78%, 96% and 81% of the times respectively. When we shouldn't have predicted a number of average stars we didn't (specificity or true negative) by 100%, 100%, 96%, 84% and 100% respectively. Therefore for all the classes, the model did a good estimation.

5.Discussion

My interpretation of the result is that there is some correlation between the maximum, minimum values and previous averages and the average stars the users grant in a particualr year. But the correlation becomes stronger and it is predictable with a 90% of confidence if I only consider influential users and reviews. This implies, in my opinion that the influential reviews are strongly related to other reviews. The user that is influenced by this review qualified the review as useful or is a fan of the reviewer because he has a high level of agreement in taste and opinions. If the qualification of useful review was given by the next user after he purchased or visit the business, there is a strong possibility that he/she agreed with the first reviewer. The fans, as per my theory will also have a strong tendency to agree and hence have the same number of stars. Another implication I see is that the business would do well to know who are the influential reviewers (reviewers with lots of fans and with important number of useful reviews). They could target them in special promotions to encourage them to visit their business if they did not do it and influence his followers. In case they did visit, they could obtain an updated and better review.

NOTE: Reproducibility of work: see the markdown html version of this document that contains the code explained step by step. This is stored in

https://github.com/pcmolinari