Data Science Capstone Project

Marc C.

2015-11-22

Marc C.

November 22, 2015

Introduction/Background

Yelp is one of the premier interfaces for restuarant reviews and guides. Often, for me I have used yelp primarily to locate a restaurant either in a new cuisine or at a new location. One of the biggest problems I have faced with Yelp is understanding the level of credence to a rated resturant with limited reviews. For this project, I will be exploring the projected ratings for restaurants given an early rating age. Specifically, I will be trying to find out that at a early timeframe of a restaurant, how accurate is its current rating be reflective of its overall rating.

Data Sources

The dataset was provided by Yelp, a website/interface where users can rate businesses with a descriptive review and a 1-5 star review. The dataset was provided on an academic license agreement. The dataset is provided by Yelp as part of a dataset challenge and consists of 1.6M reviews and 500K tips by 366K users for 61K businesses. I will be tackling restaurant businesses from this dataset.

Objective

The objective for this study is to figure out how reliable is a rating of a restaurant based on limited reviews. Specifically, if you only have a cetain amount of reviews for a restaurant, can it be a accurate indicator of its overall quality/rating?

Methodology

The way I approached this problem is to subset the data to just restaurants with at least a certain amount of reviews. For this case study, I subsetted these restaurants who have more than 125 reviews. I will then calculate a rolling average (rounded to nearest .5) throughout its time period based on user reviews from inception. From there I set certain checkpoints in the maturation of a restaurants review. I arbitrarily set a early review indicator at 15 reviews. Then compare the average rating of those restaurants then and where the overall rating stands after 100 reviews. Most of the results will be done with a students t-test assesing the mean and 95% confidence intervals. I will do a 60% 40% split on the training data and testing data. The mean and confidence intervals will be assessed in the training data and then tested in the testing data.

Loading the Data

The first step is to load the dat and the appropriate packages for this exercise. For reproducilbility, I set the seed = 12.

```
library(devtools)
```

```
## Warning: package 'devtools' was built under R version 3.1.3
```

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.1.3
```

```
library(zoo)
```

```
## Warning: package 'zoo' was built under R version 3.1.3
```

```
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Sourcing https://gist.githubusercontent.com/jbryer/4676064/raw/646d720a782efd1cdd0ead48ec
## SHA-1 hash of file is 0079a0f5e96ee15d2515312485d7865d8cdf798d
```

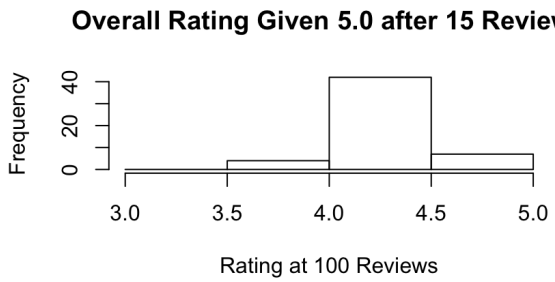
```
## [1] 2013
```

Splitting the data set for cross validation

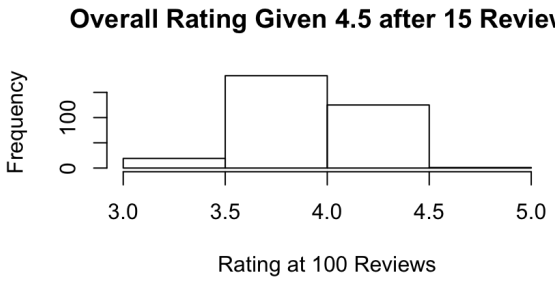
To conduct cross-validation we will subset the training data set: sub_training (60%) and sub_test (40%). Training set has about 1200 data points.

Histogram plot of a few cohorts split between their initial rating at 15 reviews. These are frequency plots of their overall average ratings after 100 reviews.

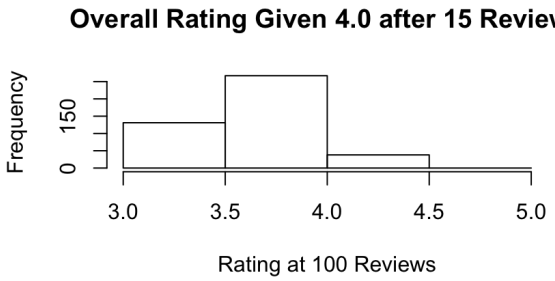
```
par(oma=c(6,6,6,6))
hist(rest_review_sample_5.0$Rating100,breaks=seq(3,5,by=.5), main = "Overall Rating Given 5.
```



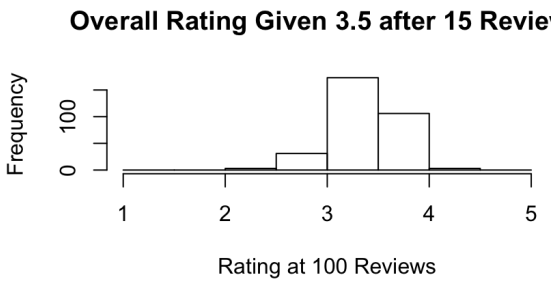

```
hist(rest_review_sample_4.5$Rating100,breaks=seq(3,5,by=.5), main = "Overall Rating Given 4.
```




```
hist(rest_review_sample_4.0$Rating100,breaks=seq(3,5,by=.5), main = "Overall Rating Given 4.
```



```r
hist(rest_review_sample_3.5$Rating100,breaks=seq(1,5,by=.5), main = "Overall Rating Given 3.
```



Calculating Means and Confidence Intervals

Below we use the students T-test against its mean rating after 15 reviews.

```r
###Calculate training statistics for each cohort####

t.test(rest_review_sample_5.0$Rating100,mu=5,alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  rest_review_sample_5.0$Rating100
## t = -15.0493, df = 52, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 5
## 95 percent confidence interval:
##  4.465406 4.591197
## sample estimates:
## mean of x
##  4.528302
```

```r
t.test(rest_review_sample_4.5$Rating100,mu=4.5,alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  rest_review_sample_4.5$Rating100
## t = -20.4797, df = 327, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 4.5
## 95 percent confidence interval:
##  4.130749 4.195471
## sample estimates:
## mean of x
##   4.16311
```

```
t.test(rest_review_sample_4.0$Rating100,mu=4.0,alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  rest_review_sample_4.0$Rating100
## t = -7.7465, df = 435, p-value = 6.713e-14
## alternative hypothesis: true mean is not equal to 4
## 95 percent confidence interval:
##  3.854787 3.913561
## sample estimates:
## mean of x
##  3.884174
```

```
t.test(rest_review_sample_3.5$Rating100,mu=3.5,alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  rest_review_sample_3.5$Rating100
## t = 6.2575, df = 315, p-value = 1.276e-09
## alternative hypothesis: true mean is not equal to 3.5
## 95 percent confidence interval:
##  3.581358 3.655984
## sample estimates:
## mean of x
##  3.618671
```

```
t.test(rest_review_sample_3.0$Rating100,mu=3.0,alternative="two.sided")
```

```
##
##  One Sample t-test
```

```
##
## data:  rest_review_sample_3.0$Rating100
## t = 8.261, df = 118, p-value = 2.416e-13
## alternative hypothesis: true mean is not equal to 3
## 95 percent confidence interval:
##  3.239587 3.390666
## sample estimates:
## mean of x
##  3.315126
```

```
t.test(rest_review_sample_2.5$Rating100,mu=2.5,alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  rest_review_sample_2.5$Rating100
## t = 6.472, df = 35, p-value = 1.854e-07
## alternative hypothesis: true mean is not equal to 2.5
## 95 percent confidence interval:
##  2.795501 3.065610
## sample estimates:
## mean of x
##  2.930556
```

Here we can see with 95% confidence that most restaurants with an early rating of 5.0 will end up at 4.5. Those that start at a 4.5 or a 4.0 rating will end up at a 4.0 rating. Now let's see what happens when we apply these predictions to the rest of the dataset (testing set).

Results/Prediction

Here we will take our testing dataset and also calculate their means and confidence intervals given their early ratings.

```
###Calculate testing statistics for each cohort####
```

```
t.test(rest_review_testing_5.0$Rating100,mu=5,alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  rest_review_testing_5.0$Rating100
## t = -9.9188, df = 29, p-value = 7.938e-11
```

```
## alternative hypothesis: true mean is not equal to 5
## 95 percent confidence interval:
##  4.336592 4.563408
## sample estimates:
## mean of x
##      4.45
```

```
t.test(rest_review_testing_4.5$Rating100,mu=4.5,alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  rest_review_testing_4.5$Rating100
## t = -15.5808, df = 166, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 4.5
## 95 percent confidence interval:
##  4.128925 4.212392
## sample estimates:
## mean of x
##  4.170659
```

```
t.test(rest_review_testing_4.0$Rating100,mu=4.0,alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  rest_review_testing_4.0$Rating100
## t = -6.3459, df = 267, p-value = 9.432e-10
## alternative hypothesis: true mean is not equal to 4
## 95 percent confidence interval:
##  3.845995 3.918930
## sample estimates:
## mean of x
##  3.882463
```

```
t.test(rest_review_testing_3.5$Rating100,mu=3.5,alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  rest_review_testing_3.5$Rating100
## t = 4.8227, df = 149, p-value = 3.464e-06
## alternative hypothesis: true mean is not equal to 3.5
## 95 percent confidence interval:
```

```
##  3.580670 3.692663
## sample estimates:
## mean of x
##  3.636667
```

```
t.test(rest_review_testing_3.0$Rating100,mu=3.0,alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  rest_review_testing_3.0$Rating100
## t = 6.6515, df = 61, p-value = 9.171e-09
## alternative hypothesis: true mean is not equal to 3
## 95 percent confidence interval:
##  3.197403 3.367113
## sample estimates:
## mean of x
##  3.282258
```

```
t.test(rest_review_testing_2.5$Rating100,mu=2.5,alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  rest_review_testing_2.5$Rating100
## t = 5.2207, df = 29, p-value = 1.374e-05
## alternative hypothesis: true mean is not equal to 2.5
## 95 percent confidence interval:
##  2.753436 3.079898
## sample estimates:
## mean of x
##  2.916667
```

The result of the testing data set also is consistent with what we saw in the training data set. Most restaurants with an early rating of 5.0 will end up at 4.5. Those that start at a 4.5 or a 4.0 rating will end up at a 4.0 rating.

Conclusion/Discussion

We can see that even with limited reviews, we can see that there are certain trends to be deduced. Most restaurants that start with a 5.0 rating tend to mature with a high rating of 4.5. My early guess is that the disparity will be greater since most early reviews tend to be more bias (friends and family giving

a supportive review). I personally tend to look for restaurants that are 4.5 or higher in rating. So one with similar tastes can be weary with restaurants that have a more modest early rating (4.0 or 4.5) and more safely assume that a 5.0 early rated restaurant will be a winner.

As a separate project I would like to gain more understanding of how early rated 5.0 restaurants will end up in their matured rating. There are instances where a early 5.0 rating can end up well short of it. With more data and setting even earlier thresholds for early indications, I think I can find some interesting indicators for predicting where these restaurants will end up.