Data Science Capstone Yelp Dataset Analysis

Alok

11/22/2015

Introduction

Yelp is an American multinational corporation headquartered in San Francisco, California. It develops, hosts and markets Yelp.com and the Yelp mobile app, which publish crowd-sourced reviews about local businesses.The company also trains small businesses in how to respond to reviews, hosts social events for reviewers, and provides data about businesses, including health inspection scores. The review website Yelp not only connects customers with businesses but also allows customers to rate their experiences. There are millions of data consisting of user/business information, reviews, votes, and so on. For Johns Hopkins University Coursera capstone project, the specific Yelp dataset had been provided for the analytical purpose. From this dataset, I will attempt to answer the following questions:

1. Is the yelper community growing?
2. In which states has large business entity listed in the yelper application?
3. Are there people who only give 1 star or 5-star reviews?
4. Is there a tendency for these persons to keep giving 1 star or 5 stars, or in other words, do they only write reviews to either complain or compliment about a business? Can this behavior be apparent by looking at the dataset?
5. Is it possible at all to reasonably predict what kind of rating a business will get based on the users rating behaviour?

Methods and Data Load

Step 1: Getting and Manipulating Data

The dataset downloaded from the Data Science Capston project Coursera page into a local folder. The raw files are in JSON format and were converted into R data frames, as shown by the codes below.

```
setwd("/root/Downloads/yelp")

library(jsonlite)
library(ggplot2)
library(caret)

## Loading required package: lattice
business_dat <- stream_in(file("yelp_academic_dataset_business.json"))
```

```
f_business_dat <- flatten(business_dat, recursive = TRUE)
save(f_business_dat, file='business.RData')

checkin_dat <- stream_in(file("yelp_academic_dataset_checkin.json"))
f_checkin_dat <- flatten(checkin_dat, recursive = TRUE)
save(f_checkin_dat, file='checkin.RData')

review_dat <- stream_in(file("yelp_academic_dataset_review.json"))
f_review_dat <- flatten(review_dat, recursive = TRUE)
save(f_review_dat, file='review.RData')

tip_dat <- stream_in(file("yelp_academic_dataset_tip.json" ))
f_tip_dat <- flatten(tip_dat, recursive = TRUE)
save(f_tip_dat, file='tip.RData')

user9_dat <- stream_in(file("yelp_academic_dataset_user.json"),
pagesize = 500, verbose = TRUE, )
f_user9_dat <- flatten(user9_dat, recursive = TRUE)
save(f_user9_dat, file='user.RData')

BR <- merge(f_business_dat, f_review_dat, by="business_id")
dim(BR)
save(BR, file='BR.RData')
BRU <- merge(BR, f_user9_dat, by="user_id")
save(BRU, file='BRU.RData')
dim(BRU)

load("/root/Downloads/yelp/BRU.RData")
load("/root/Downloads/yelp/user.RData")

y_master1 <-(BRU[,c(1, 2, 6, 7,10, 11, 12, 13, 14, 109, 120)])
y_master1 [1,]
```

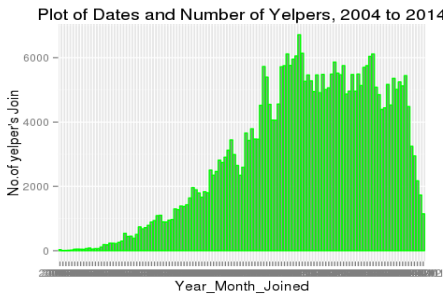
The raw files were cleaned up, dissected and crossed. The resulting dataset consists of only user ID, star rating of a business, business ID, users review count, users average stars, city, state, longitude latitude .

Step 2: Exploratory Data Analysis

In order to answer the first two questions

1 Is the yelper community growing?

```
ggplot(f_user9_dat, aes(x = yelping_since)) + geom_histogram(alpha =
.50, binwidth=.1, colour = "green" ) +labs(title ="Plot of Dates and
Number of Yelpers, 2004 to 2014", x = "Year_Month_Joined", y = "No.of
yelper's Join")
```



Note: Initially user growth increate till 2009-10 and then stable, user's growth decaling in 2014.

2) In which states has large business entity listed in the yelper application ?

```
splot <- barplot(table(y_master1$city, y_master1$state ),  main = "No of business unit in yelp per states vs city" , col=c(colors()), xlab = "Number of States", ylab="Counts of Bsuiness Listed in Yelp", ylim = c(0, 1000000),  xlim = c(1, 30), xpd=TRUE )
```

3) Regarding whether there are people who only give either 1 star or 5 star reviews, a histogram of users average rating is plotted.

```
ggplot(f_user9_dat, aes(x = average_stars)) + geom_histogram(alpha = .50, binwidth=.1, colour = "green")
```

4) To answer the next question on these users tendency to write only 1 star or 5 star reviews, the relationship between users average stars and review count is plotted.

```
ggplot(f_user9_dat, aes(average_stars, review_count)) + geom_point()
```

#Step 3: Model Building

The last question of interest for this project is whether it is possible to predict how many star a user will rate a business based on the users rating behaviour, which is extremely subjective. It would be impossible to be able to predict with 100% accuracy what a user would rate a business based on past rating history, but we might be able to produce a model with reasonable prediction (for example, accurate at least half of the time, to be conservative). We will evaluate if we can achieve this by making the model as simple as possible. In this case, only the users average stars and number of reviews will be used to build the model. We are going to perform nested likelihood ratio test on three linear regression fits: stars against users average stars, against users review counts, and against both.

```
fit1 <- lm(stars.x ~ average_stars, data = y_master1)
fit2 <- lm(stars.x ~ review_count.x, data = y_master1)
fit3 <- lm(stars.x ~ average_stars + review_count.x, data = y_master1)
anova(fit1, fit2, fit3)
```

The fit comparison shows that either users average stars or review counts are terrible predictors of users ratings, but a decent fit if both are taken into account. They will therefore be the basis for the model. To test this model, the dataset was split into training and testing data sets. Since the sample size is large, the dataset was split equally.

```
inTrain = createDataPartition(y_master1$stars.x, p = 0.5, list=FALSE)
training = y_master1[ inTrain,]
testing = y_master1[-inTrain,]
dim(training)
dim(testing)
```

Step 4: Model Validation

The model is built upon the testing dataset, which is then applied to predict on the testing dataset. The output will be in the form of a confusion matrix as well as accuracy percentage.

```
modFit <- train(stars.x ~ average_stars +
review_count.x,data=training,method="lm")
predictions <- round(predict(modFit, testing))
u = union(predictions, testing$stars.x)
t = table(factor(predictions, u), factor(testing$stars.x, u))
confusionMatrix(t)
```

Results

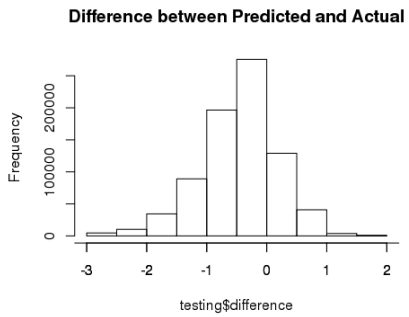
The first histogram shows that there are substantial number of users with average rating of either 1 or 5 star (in fact, a large number of people have average rating of 5 star, more so than the others). The second plot looks at the relationship between the number of reviews and average star ratings of the users in this dataset. The plot shows that users who give average rating of 1 or 5 star write only small number of reviews, and it gets increasingly more for users that are in between. As a validation for the prediction model created, a matrix comparing the number of accurately (and inaccurately) predicted rating was generated. The calculated accuracy based on the generated confusion matrix is ~38%.

Discussion

The first histogram readily answers the first question that there are users who only write 1 or 5 star reviews. The second plot deduces that these users likely write reviews for the sole purpose of either strongly complaining or complimenting about their experiences (which is why there are not that many of them). It does answer the second question that there are

people with the tendency to write only positive or negative reviews. However, note that the plot is also densely populated across all average stars, meaning there are also users who write only a few reviews that average anywhere between 1 and 5 star, so not exclusively for strongly negative or positive reviews. We must therefore state that there are users who exclusively write reviews on Yelp solely to give either 1 or 5 star, as this is apparent from the frequency of the review, but we cannot make a similar statement that infrequent reviewers are users that tend to give only 1 or 5 star. The confusion matrix shows that the oversimplified model, which only takes into account the users average stars and review count (omitting all other factors), is unfortunately far from perfect in predicting a business rating based on the users average stars and review count. However, if we are willing to relax the accuracy requirement, the model may be good enough as a rough predictor. To check this, the difference between the predicted and actual values are calculated and plotted as a histogram. It is apparent that a majority of predicted values are within 1 star away from actual values.

```
testing$predictions <- predictions
testing$difference <- testing$stars.x - testing$predictions
hist(testing$difference, breaks = 10, main = "Difference between
Predicted and Actual")
axis(side=1, at=seq(-4,4, 1), labels=seq(-4,4,1))
```



To be more precise, roughly 82% of predicted values are only off by +/- 1 star, as shown by the calculation below.

```
difference <- as.data.frame(table(testing$difference))
(difference[4,2]+difference[5,2]+difference[6,2])/sum(difference[,2])

## [1] 0.4077535
```

Considering the subjectivity of this dataset and oversimplification of the variables, the model is quite acceptable as a rough estimating tool. The answer to the third question is that it is possible to reasonably predict the rating of a business based on the users rating behavior, which in this case only involves the users average stars and review count. A much more robust model can potentially be built by taking into account all aspects of the dataset.