Data Science: Capstone Project

Pritam Dey

Nov 20, 2015

Introduction

The goal of the capstone project is to apply the data science skills that we have learned in this course, and therefore mimic the experience of being a data scientist.

The specific objective of this capstone project is to analyze the Yelp data set to answer the question that I have formulated. The question that I have posed is:

What correlation can we draw between a user's attributes (e.g. no. of fans, no. of friends, no. of compliments received) and his/her review of a business? Does an active user (user with higher no.of attributes) tend to review a particular type of business more?

The rationale for choosing this question is to understand whether a user's personality (determined by his/her attributes such as no. of fans, friends, votes & compliments received) has any dependency on the type of reviews he/she provides. Lets assume that a user with lots of friends, fans and compliments is more socially-outgoing personality than a user with minimal friends. So does such people tend to visit specific types of business more (e.g. restaurants, bars, etc.), and provide more reviews to such businesses? Answering these kind of questions can help marketeers to come up with targeted marketing strategies to attract specific crowd to visit the business.

Methods and Data

Overall Approach My question deals with **User Attributes** and **Business Review**. I am interested in the impact of user attributes on his/her review of business. Hence I will run a regression with 'review_count' (from user dataset) as dependent (outcome) variable and other user attributes (votes, friends, fans, compliments) as independent (preditcot) variables. There is "review_count" variable in the 'user' dataset against which the prediction will be made.

The key steps of my approach are:

1. I extract ONLY required columns from 'user' data set. The columns I am interested in are: user_id, review_count, average_stars, votes, friends, compliments, fans.
2. Since 'friends' is a list of user friends (with unique user_id), I add them into total No. of friends.
3. Similarly I add the total number of compliments and votes received.
4. Since the size of the datasets are huge, I take a subset of 10000 complete data (rows).
5. I them partition data into **Training** and **Testing** datasets.
6. Against the above predictor variables, I run multi-variable regression. I run two regressions:
    1. To check the impact of user attributes on review count (data from 'user' dataset)
    2. Next run another regression on - User Attributes + Business Attributes vs. Business Review count (data from 'business', 'review' and 'user' datasets).

    This will tell me how much of dependency does user attributes and business attributes (individually and collectively) have on the business review.
7. Next I will run the following prediction models:

a. Prediction model with Decision Tree (using both Training and Testing datasets)
b. Prediction model with Random Forest (using both Training and Testing datasets)

8. Prove/disprove my hypothetis based on above results.
9. Summarize the results.

Below is the key data processing steps I took to perform my analysis:

1. Downloaded, extracted, and loaded the data (json files):

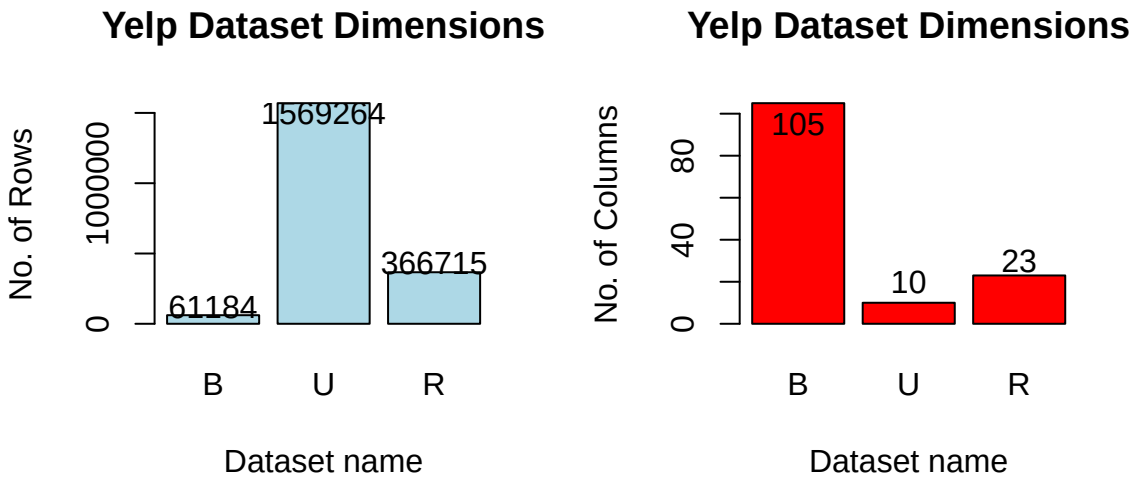
```
a. Downloaded the Yelp dataset as per the instuctions provided by the course.
b. There are five files in JSON format - business, review, user, checkin, and user.
c. I extracted the data using jsonlite package, and stored the data in data frames.
I used 'flatten' functionality to flatten hierarchical structure in the data sets.
d. For my question scope, only 'user', 'business', and 'review' files are used.
I ignored 'checkin' and 'tip' data set.
```

2. Basic exploration of the data:

```
a. A quick overview of the dimensions of the dataset is shown in the below table.
```

Table 1: Dimensions of Yelp Dataset

	Rows	Columns
Business	61184	105
User	1569264	10
Review	366715	23
Checkin	45166	170
Tip	495107	6



Legend: B: Business U: User R: Review

3. Processing the Data:

```
I performed the following data processing:
```

1. **Imputation:** Replaced all NAs by zero

2. **Summation:** For user dataset, created new columns with sum of compliments, votes, friends. This helps me to consider only this summed columns for my analysis, thereby helping me to ignore 20+ variables.

3. **Extractation:** Due to large data size, I extracted only specific columns needed for my modeling.

4. **Subsetting:** Due to large size of data set, I extracted 10000 rows randomly using 'subset' function. This is the final data set for my analysis.

5. **Merging:** I also merged 'business', 'review', and 'user' datasets since I would need these combined dataset at some point to run my regression.

4. Partitioning the data:

1. To perform the analysis, I needed to split the data into a training sample to build my model, and a separate testing data set to validate and test my model.

2. Since the size of training data set is very large (more than 350,000 rows for user and 1.5 million for review), I made the choice to partition the data into 50:50 ratio on the 10000 records that I extracted above.

3. After partitioning, the new data set size is: 5000 rows for training set, and 5000 rows for testing set. The number of variables have been brought down to 7.

5. Running the FIRST regression model with only user attributes: I ran the regression model with the following variables and got the following output:

fit0 <- lm(review_count ~ fans + average_stars + tot_compliments + tot_votes + tot_friends, data=myTraining)

6. Running the SECOND regression model with only user + business attributes: The above regression with only 'user attributes' could only explain about 64% of variation in the review count. So I run another regression that includes both 'user attributes' and 'business attributes'. The idea is review count should also depend on the business attributes such as quality of service, type of service, etc.

Using the merged data as explained above (Sec 3.5), I run the following regression this time:

fit1 <- lm(review_count.y ~ stars.y + tot_attributes + fans + average_stars + tot_votes + tot_compliments + review_count.x, data=myTraining)

The output of the two regressions models is shown below:

Regression 1 (only user attributes):

fit0 <- lm(review_count ~ fans + average_stars + tot_compliments + tot_votes + tot_friends, data=myTraining)

Regression Statistics	
Multiple R	79.78%
R Square	63.66%
Adjusted R Squar	63.62%
Observations	10000

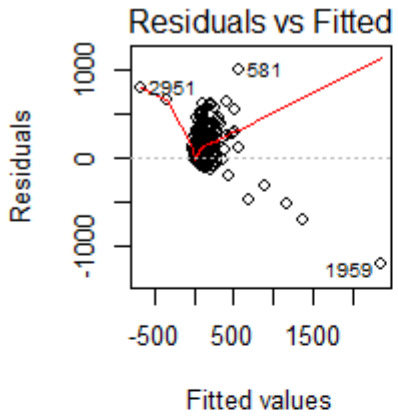
	Coefficients	*Standard Error*	*P-value*
Intercept	17.4952468	2.924862194	0.00
fans	6.310229473	0.197975329	0.00
average_stars	0.311364363	0.75797395	0.68
tot_compliments	-0.129789809	0.005650371	0.00
tot_votes	0.055139558	0.002559716	0.00
tot_friends	0.05272216	0.037852437	0.16

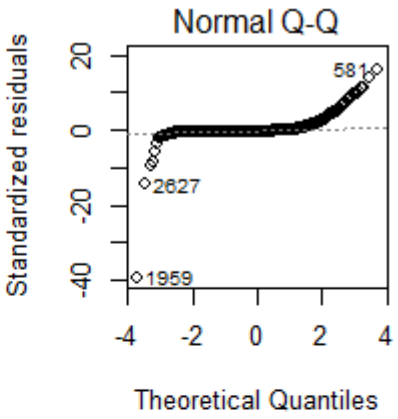
Regression 2 (user + business attributes):

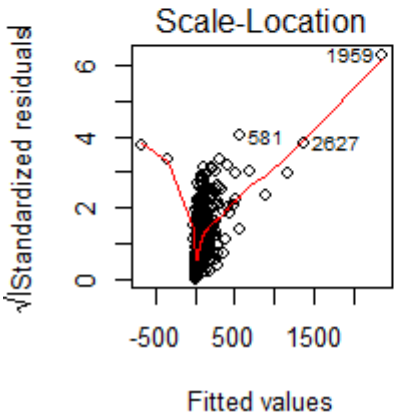
fit1 <- lm(review_count.y ~ stars.y + tot_attributes + fans + average_stars + tot_votes + tot_compliments + review_count.x, data=myTraining)

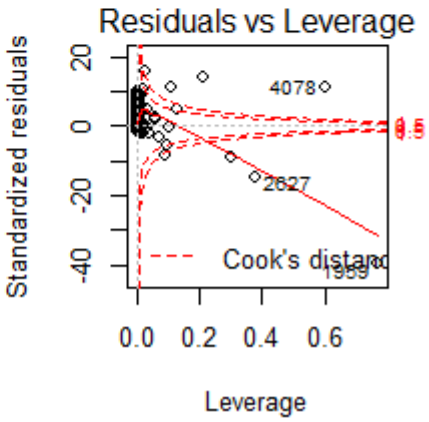
Regression Statistics	
Multiple R	77.58%
R Square	60.18%
Adjusted R Squar	60.15%
Observations	10000

	Coefficients	*Standard Error*	*P-value*
Intercept	113.7247905	12.77309083	0.00
review_count.x	0.036586163	0.006454183	0.00
stars.y	-8.614123436	2.675424672	0.00
tot_attributes	1.740871605	0.719014346	0.02
fans	1.595940332	0.091666208	0.00
average_stars	-1.754867488	2.932475105	0.55
tot_votes	0.076195587	0.00115321	0.00
tot_compliments	-0.084092584	0.002755201	0.00









The plot above indicate that the residuals are NOT normally distributed and homoskedastic.

7. Building the prediction model:

```
a. I chose to build my model using two approaches: **Decision Tree** and **Random Forest**
```

Results

1. In Regression Model 1, Adjusted R-Square is 63.66% and p-value is less than 0.05.
2. In Regression Model 2, Adjusted R-Square is 60.18% and p-value is less than 0.05.
3. 95% confidence interval shows there is no zero.
4. Regression Model 1 shows that there is some relationship between review_count and user attributes.
5. Regression Model 2 shows that there is some relationship between review_count vs. user attributes **&** business attributes.
6. In combination of these two models, R-square value indicates that about **64% of the variation in business review can be explained by various user and business attributes**.
7. Using Decision Tree, I got accuracy of 52.00% and out of sample error of 0.12% (1-0.9888)
8. Using Random Forest, I got accuracy of 52.00% and out of sample error of 0.08% (1-0.9992)
9. As compared to prediction with Decision Tree, prediction with Random Tree yielded better accuracy percentage, and low out of sample error. The difference between the two models is marginal; nevertheless Random Tree provided slightly better prediction result. Hence I chose Random Forest model for my further prediction.

Final Result

Primary Question: What correlation can we draw between a user's attributes (e.g. no. of fans, no. of friends, no. of compliments received) and his/her review of a business? Does an active user (user with higher no. of attributes) tend to review a particular type of business more?

Conclusion: Regression output shows there is some correlation between a user's attributes and his/her review of a business.

Discussion

The various attributes of the user explains for 64% for the variation (Regression 1) in his/her review of the business. That the value of adjusted R-square is low is along expected lines because intuitively we expect the business review to depend largely on the quality and type of business service, and not only on user attributes. That a user's attributes plays a significant role in the review process should be a good marketing input to the business. It basically means that user's friends circle, and his positive ratings plays a role in how he review a business. This is not surprising as our social ecosystem plays a big role in our behavior.

What is indeed surprising is that the adjusted R-square did not improve much in Regression 2 as well. I think this could be because of the underlying data. I factored only business attributes such as smoking facilities, Wi-Fi presence, Credit Card, Noise Level, Attire, etc. The regression output shows these attributes do not experience much or none at all. What ultimately should matter is the user experience of consuming the business service. This is a subjective experience and may not have been captured adequately in the datasets provided.
