An Exploratory Study: Yelp Review vs. Checkin

L.P.

11/21/2015

1. Title

An Exploratory Study: A Connection Between the Yelp Number of Reviews and Checkin Feature. We study relation between number of reviews and number of checkins for businesses in the Yelp dataset. We find that there is a positive correlation between these two features, and we model the dependence using a linear regression. We observe an increasing variation in the number of reviews as the number of checkins increases. We explain this variation by a model coupling between the number of checkins and other features in the dataset, including WiFi, WheelchairAccessible, business weekend opening hours, or star rating. We also observe a correlation between the average star rating and average length of the review text, as well as business weekend opening hours. We predict number of reviews for a testing dataset. Our linear model prediction algorithm explains 75% of the data.

2. Introduction

Is there a way how a business can get more Yelp reviews? What has an effect on the number of reviews and/or star rating: business opening weekday/weekend hours, number of checkins, day of the week: do users write more reviews on the weekends? City? Does a particular type of a business differ from the rest of the businesses? For example, is healthcare rating in any way different than the rest of the business ratings?

The goal for studying this question is to attract more users to use Yelp review tool and to help reviewed businesses better understand their own ratings. We focus on the following two aspects: how does a business (i) increase the number of its reviews and (ii) control its star rating? In order to tackle these questions, we look at the business, checkin and review Yelp datasets. In particular, we look at the following features: (i) Is the business open on Saturday and/or Sunday, (ii) What city is the business located at, (iii) What category does the business fall into, (iv) Does the business have attributes such as: WiFi, ByAppointmentOnly, HappyHour, WheelchairAcessible, (v) What is the business's checkin on Monday - Sunday and/or the total number of checkins, (vi) What is the average star rating of the business, (vii) At what date/day has the business been reviewed, (viii) What is the text of the review, e.g., characterized by its text length.

3. Methods and Data

We upload Yelp datasets into *business*, *checkin* and *review* data frames using *jsonlite* package. We add:
1. **Open on the Weekends?** *open.Saturday*/*open.Sunday* are factor variables; TRUE if open on Sat/Sun.
2. **Location - City?** *city10* is a factor variable; its value is one of the 10 cities. We use a *kmeans* algorithm with 10 clusters corresponding to (longitude, latitude) of: Las Vegas, Phoenix, Madison, Urbana, Charlotte, Pittsburgh, Waterloo, Montreal, Edinburgh and Karlsruhe. The algorithm predicts city for each business.
3. **Business Category?** *business_category* is a factor variable; its value is one of the 23 main categories listed here: *https://www.yelp.com/developers/documentation/v2/all_category_list*
4. **Relevant Attributes?** *WiFi*, *ByAppointmentOnly*, *HappyHour* and *WheelchairAccessible* are factor variables with values: (no, free, paid) for WiFi, (TRUE, FALSE) for the rest; "Not Provided" stands for NA.
5. **Checkin for Monday - Sunday** We count number of checkins for Monday through Sunday and the total number of checkins for each business. NA counts as 0. We join the business and checkin data frames.
6. **Day** Represents the day of the week a review has been written.
7. **Text.length** Represents the (average) length of the review text (per business).

We check the number of reviews (review_count) and average stars (stars) of businesses by comparing the data from the review and tips Yelp datasets with the data from the business Yelp dataset. Although close, we find the results do not match. We choose the review dataset as the source of the business star rating and review counts. We join the review and business (plus checkin) datasets. Features of the final dataset are: business_id, stars, review_count, checkin_all, business_category, WiFi, ByAppointmentOnly, HappyHour, WheelchairAccessible, city, latitude, longitude, open, open.Saturday, open.Sunday, text.length.

Methods
- We use the *base::cut* function to cut a vector variable into n intervals determined by the variable's quantiles: *text.interval <- cut(sumdftext.length, breaks = quantile(sumdf text.length, probs=seq(0,1,0.125)))*
- We use the *stats::cor* function for computing the correlation of x and y variables: *cor(x, y)*
- We use the *stats::lm* function for fitting a linear model of reviews vs checkins: *lmfit <- lm(review ~ checkin)*
- We use the *base::summary(lmfit)*, *base::plot(lmfit)* and *caret::varImp(lmfit)* functions to summarize the results of a fitting function (including statistics, p-values), to plot the results of the fitted lm (including plots of residuals and normal Q-Q plots) and to calculate variable importance for the model.
- We use the *stats::t.test* function to perform two sample t-test and to obtain the test *p-value*: *t.test(weekendyes, weekendno, paired=FALSE, var.equal=FALSE, alternative="greater")*
- We use the *stats::predict* functions to predict values for the test set.
- We use the *caret::createDataPartition* for a series of test/train partitions and test/train sets: *inTrain <- createDataPartition(review, p=0.6, list = FALSE)*; *train.set <- data[inTrain,]*, *test.set <- data[-inTrain,]* We split the data into training (60%) and testing set (40%) by review count. In the next section, we perform an exploratory analysis on the training data. We apply our final prediction model on the testing set.
- We use the *stats::anova* function to compute analysis of variance tables for model objects; we compare several fitting models and choose our final model with significant predictors. Our final model is:

```
lmfit <- lm( review_count ~ checkin_all*stars + checkin_all*text.length + checkin_all*WiFi
             + checkin_all*open.Saturday + checkin_all*open.Sunday + checkin_all*HappyHour
             + checkin_all*business_category + checkin_all*WheelchairAccessible +
             checkin_all*open + checkin_all*city + checkin_all*longitude, data = training)
```

For further inquires about the used methods and code, please inspect the Rmd file on the following website: *https://github.com/lpalova/Data-Science—Final-Project.*

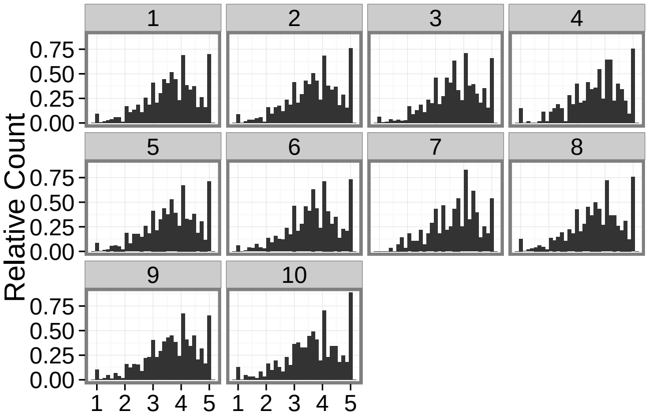
4. Results

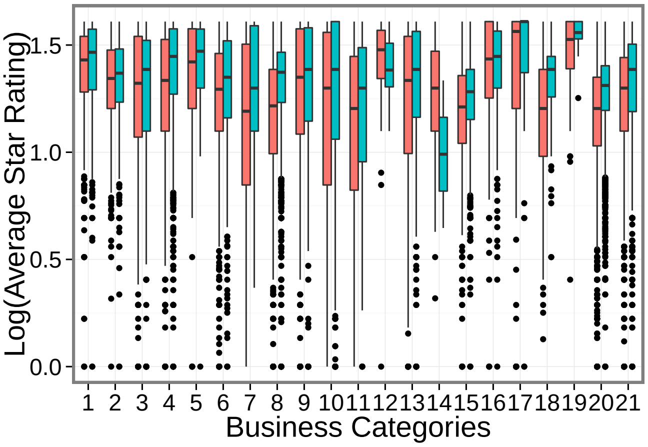
A) Exploratory Analysis

i.) Average Star Ratings We look at the average review text length for each business as a function of star rating. The table below shows counts of reviews for a given star rating 1-5 (row) and review text length interval (column). We observe that most reviews show a trend of shorter review texts with better star ratings.

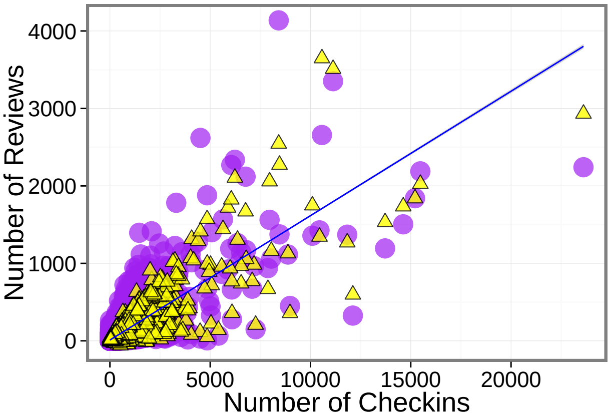
	(10,356]	(356,448]	(448,522]	(522,590]	(590,665]	(665,763]	(763,921]	(921,5e+03]
1	134	82	78	71	57	74	119	301
1.5	120	95	100	98	107	109	124	252
2	290	289	294	274	302	306	381	584
2.5	511	556	507	501	595	580	704	819
3	872	1061	1033	1071	1154	1244	1247	1172
3.5	1092	1267	1493	1646	1677	1703	1671	1380
4	1366	1636	1809	1900	1898	1930	1773	1526
4.5	1164	1192	1247	1200	1071	1008	1006	876
5	2054	1418	1041	830	740	641	573	688

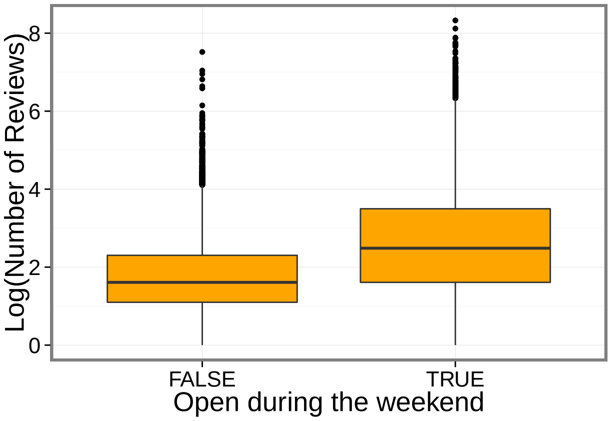
There is no apparent relation between the star rating and location of a business in a particular city; the shape of the distribution of average stars (left figure) remains unchanged for different locations. We note that that there are 16469 and 25225 businesses located in Las Vegas (1) and Phoenix (2), resp., but only between 351 to 5149 businesses in other cities. Nevertheless, the star distribution shape is the same regardless of the volume of businesses. We have performed an analysis of the dependence of star ratings (and review counts) on the day when a business is reviewed. We note that there is no apparent dependence, and the shape of the distribution(s) remains the same regardless of the day of the week. However, we observe dependence of the star ratings (and review counts) on business being open/closed during the weekend; businesses open during the weekend show higher star ratings (green boxplots in the right figure). We also observe dependence on business categories; medians of the average star ratings change as we go from one category to another.





ii.) Review count We plot the number of reviews as a function of total number of checkins for each business (left, purple circles). We observe a correlation of 0.66 between the two variables. We fit a linear regression model (blue line): $review.count = 9.84 + 0.16 \times checkin$. R squared is 0.62. We also show fitted values for our final prediction algorithm as yellow triangles. Next, we plot a boxplot of the (log of) number of reviews vs business being open/closed on Saturday or Sunday (right). We observe median of 12 and 5, and mean of 39 and 10.9, for the number of reviews for businesses open during the weekend and businesses closed during the weekend, respectively. We perform a two-sample t-test on the mean of these two sets and conclude that businesses open during the weekend have greater number of reviews.





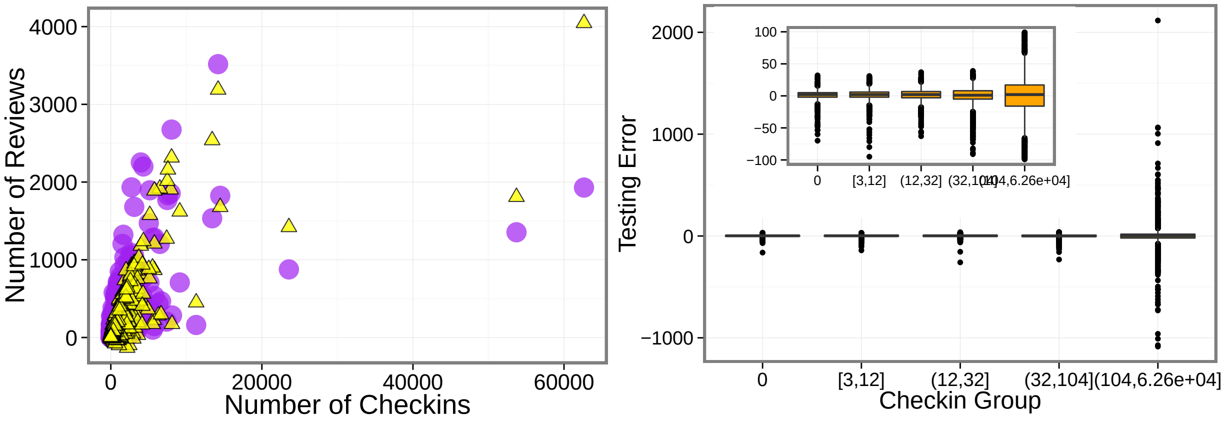
B) Prediction Algorithm

Review count We observe the largest correlation values of: 78%, -22%, 20%, 20%, 17%, -10%, -9%, -6%, 8% between the number of reviews and the number of checkins, WiFi, WheelchairAccessible, open.Sunday, open.Saturday, city, longitude, latitude and business_category in our training set, respectively. We use a liner regression model to fit the number of reviews. We find that the simple linear model: $review.count = 9.84 + 0.16 \times checkin$ explains about 62% of the data. We observe in figure in section 4Aii that the number of reviews shows larger variation about the linear regression line as the number of checkins increases. In order to explain this variance in the data, we need to consider a coupling between the number of checkins and yet another feature. If this other feature is bounded, such as star rating (values between 1-5) or a factor with discrete levels, then a coupling of the form $checkin * stars$ or $checkin * WiFi$ gives a larger interval of possible responses as the *checkin* variable increases. We explore this idea by adding new couplings to the starting simple linear model and using the *anova* function to keep only significant interactions. We note that there are correlations of 78%, -16%, 15%, 12%, 16%, -11%, -10%, -8%, 7% between the number of checkins and the number of reviews, WiFi, WheelchairAccessible, open.Saturday, open.Sunday, city, longitude, latitude and HappyHour, respectively. Our final model includes interactions with WiFi, WheelchairAccessible, open.Saturday, open.Sunday, city, longitude, HappyHour, business_category, open, stars and text.length.

We train our final *lmfit* model on the training set. The model's R squared is 0.83. We test the model on the testing set; we obtain predictions for review counts and plot these predicted values (yellow triangles) on top of the actual values (purple circles) in figure below (left). We also calculate the error for each prediction as a discrepancy between predicted and actual number of reviews. The error's 1st quartile, median, mean and 3rd quartile are: -4, 2, 1 and 7, respectively. We show boxplot of the testing error for different checkin groups in the figure below (right), where the checkin counts are divided by 0%, 25%, 50%, 75% and 100% quartiles forming 4 groups; in addition, we consider the case with zero number of checkins as a separate checkin group. We see that the error between the 1st and 3rd quartiles falls between -20 and 18; the size of this interval slightly increases as number of checkins increases; the lower/upper box boundary moves towards smaller (more negative)/larger (more positive) error values. This trend corresponds to the trend in figure 4Aii (left) where the number of reviews has a larger variation about the linear regression line as the number of checkins increases. The model's test set R squared is 0.75 and explains about 75% of the data.

```
predict_test <- round( predict( lmfit, newdata=testing ) )
error_test <- predict_test - testing$review_count
summary(error_test)
```

```
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-1086.0000    -4.0000     2.0000     0.7024     7.0000  2116.0000
```



5. Discussion

We find that number of reviews depends, to a great degree, on the number of checkins a business experiences. We observe a correlation of 78% between these two features. Number of reviews increases with the checkins.

(i) We model the dependence by fitting a simple linear regression in section 4Aii (blue line). We find a slope of 0.16 representing an increase of 16 in the number of reviews for every 100 checkins, on average.
(ii) We employ a linear model with couplings between the number of checkins (as the main predictor) and other features, including WiFi, WheelchairAccessible, open.Saturday, open.Sunday, city, longitude, HappyHour, business_category, open, stars and text.length. We observe that this model captures the increasing variance of the number of reviews as a function of the number of checkins; we observe a relatively nice match between the dataset values (purple circles) and the fitted or predicted review values (yellow triangles) in figures in sections 4Aii and 4B. We train this model on our training dataset (60% of the available data), R squared of 83%, and apply the model to our testing dataset (remaining 40% of the data). The model's performance is measured by R squared of 75%; the model explains about 75% of the test data. We also look at the error, a discrepancy between the predicted number of reviews and actual number of reviews. We find the testing error mean of about 1 and median of 2; we overestimate the number of reviews by 1 (or by 2) on average (or for the most cases). The 1st and 3rd quartiles are -4 (we underestimate the count by 4) and 7 (we overestimate the count by 7), respectively. Our prediction error density is centered at 1 with a standard deviation of 40.
(iii) We note that we employed several other models including poisson glm, glmboost or gbm models. These models do not perform significantly better (some of them perform worse) than our final model.

We perform a p-value test on our training set to address the significance of attributes like business being open/closed during the weekend (open.Saturday or open.Sunday), WheelchairAccessible, WiFi or HappyHour. We find that businesses open during the weekend (19613 businesses) have greater number of reviews than those closed during the weekend (16859 businesses); p-value ≈ 0. Similarly, we find that businesses that are wheelchair accessible (10638 businesses) have greater number of reviews than those without a wheelchair access (998 businesses); p-value ≈ 0. We note, however, that there is a large amount of missing data in this attribute. We also look at businesses with WiFi attributes (only non-NA values) and conclude that, interestingly, free, paid or no WiFi has no effect on the number of reviews. Again, there is a considerable amount of missing information about WiFi; only 4391, 236 and 6066 businesses are recorded to have a free, paid and no WiFi out of 36472 total number of businesses in our working dataset. Finally, businesses offering a HappyHour (1583 businesses) do not receive significantly more reviews than businesses not offering this service (662 businesses); p-value of 0.12 is not significant. Again, much of the HappyHour information is missing.

In this report, we study business, checkin and review Yelp datasets. We look at features including review_count, checkin, stars, business category, WiFi, ByAppointmentOnly, HappyHour, WheelchairAccessible, city, latitude, longitude, open, open.Saturday, open.Sunday and review text length. We examine dependence of the number of reviews a business receives on the various features; we find number of checkins, stars, text.length, WiFi, WheelchairAccessible, open.Saturday/Sunday, HappyHour, business category and city, longitude and open (in coupling with the number of checkins only) to be significant predictors. Our main findings are: (i) number of reviews increases with number of checkins, (ii) number of reviews has a larger variance as the number of checkins increases, and this variance is explained by further couplings between the number of checkins with other features like WiFi, HappyHour, business weekend opening hours, star rating and text.length, or location/city, (iii) businesses open during the weekend receive more reviews and better star ratings; wheelchair accessible businesses also receive more reviews, (iv) business category and average review text length are significant predictors for number of reviews and star ratings; we observe that most reviews show shorter texts with higher ratings, (v) day of the week a review is written and location/city of the business does not have a significant impact on the number of reviews and star ratings as per se; however, these factors enter our model via coupling with other features. We note, however, that we have not studied location as a function of proximity to other businesses, and this task is left for further studies.

In summary, a business may get more Yelp reviews by providing more checkins, weekend opening hours or wheelchair access. Weekend opening hours may contribute also to better star ratings. In general, we observe differences in star ratings among different business categories (see 4Ai). For example, the health and medical category (9) receives higher ratings compared to other catogeries like financial services (7), hotels & travel (11), mass media (14), nightlife (15), public services & government (18), restaurants (20) or shopping (21). On the other hand, services like active life (1), beauty & spas (4), education (5), local flavor (12), pets (16), professional services (17) and religious organizations (19) receive higher ratings.