INTRODUCTION TO DATA ANALYSIS

Yelp dataset challenge

Predict star rating for business using review text

Dimple Sharma

12/12/2014

Contents

Objective

The goal of this paper is to analyze semi-structured and unstructured data to predict user ratings from user review text using natural language processes techniques. In addition, we will analyze when business is busy the most weekly and daily and identify visiting patterns. We will use the user check-in information forecasting visitors and look for visiting patterns.

On the internet, social networking sites such as Yelp provide star ratings and reviews which provides insights about customer satisfaction. However, many times just a star rating is not sufficient. We will use review text, votes, and review to predict star rating for a business.

Dataset

We use Yelp dataset from the Yelp dataset challenge which contains business, reviews, and check-in data of a few cities in 3 states Arizona, Nevada, and Wisconsin.

1. **42153 Business Records**

'type': 'business',
'business_id': (encrypted business id),
'name': (business name),
'neighborhoods': [(hood names)],
'full_address': (localized address),
'city': (city),
'state': (state),
'latitude': latitude,
'longitude': longitude,
'stars': (star rating, rounded to half-stars),
'review_count': review count,
'categories': [(localized category names)]
'open': True / False (corresponds to closed, not business hours),
'hours': {
(day_of_week): {
'open': (HH:MM),
'close': (HH:MM)
},
...
},
'attributes': {
(attribute_name): (attribute_value),
...

2. **1125458 Reviews Records**

'type': 'review',

'business_id': (encrypted business id),
'user_id': (encrypted user id),
'stars': (star rating, rounded to half-stars),
'text': (review text),
'date': (date, formatted like '2012-03-14'),
'votes': {(vote type): (count)}

3. **31617 Check-in Records**

'business_id': (encrypted business id),
'checkin_info': {
'0-0': (number of checkins from 00:00 to 01:00 on all Sundays),
'1-0': (number of checkins from 01:00 to 02:00 on all Sundays),
...
'14-4': (number of checkins from 14:00 to 15:00 on all Thursdays),
...
'23-6': (number of checkins from 23:00 to 00:00 on all Saturdays)

Obtaining and Manipulation Data

The yelp dataset was obtained from yelp website for yelp data challenge, it requires consumer to fill a form before downloading the data. Here is a link to the yelp dataset challenge which will work till December 31, 2014 : http://www.yelp.com/dataset_challenge

The data in the dataset is fairly clean, however because the dataset is very large and it has many attributes there are many columns which do not have data simply because not all attribute are relevant to the entity such as business, review, or check-in.

Most NA or null values present in the dataset are because the attribute is not relevant to the entity or simply the business was closed at the hour. For these reason, all numeric columns with NA or NULL are set to zero.

The datasets are shared as a JSON files which we converted into CSV file. Hence we use CSV files for the business, reviews, and user. The datasets is large because it has thousands and some have 1.3 million rows.

Analysis

The analysis aims to analyze user sentiments and experience to predict future rating for a business. The analysis questions that will be covered in this paper are:

When is a business most busy weekly and daily? What is the trend?

How well can we predict star rating using a review's text?

Which are the best variables to predict star ratings based?

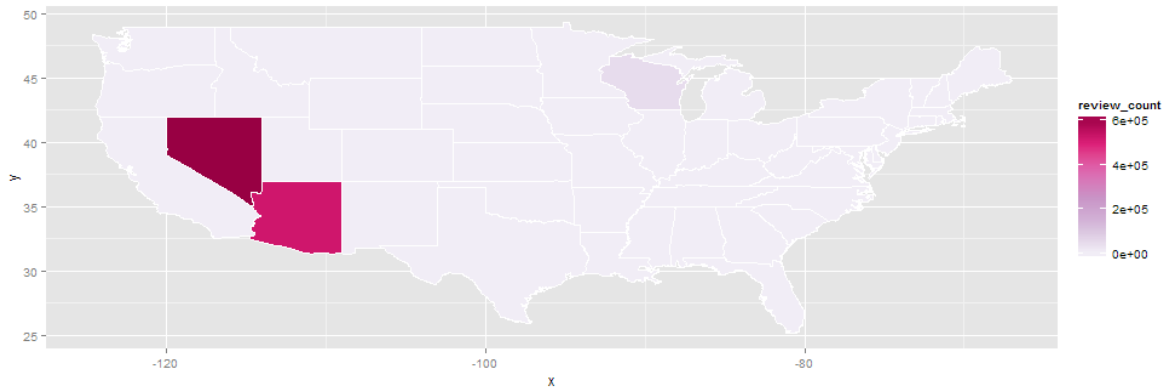
How well can we forecast user visits using check-in data? Is there a pattern?

The analysis will use various types of line, scatter plot, bar graph, histogram, time series analysis, heat map and choropleth graphs for graphically representing the data and analysis.

Business

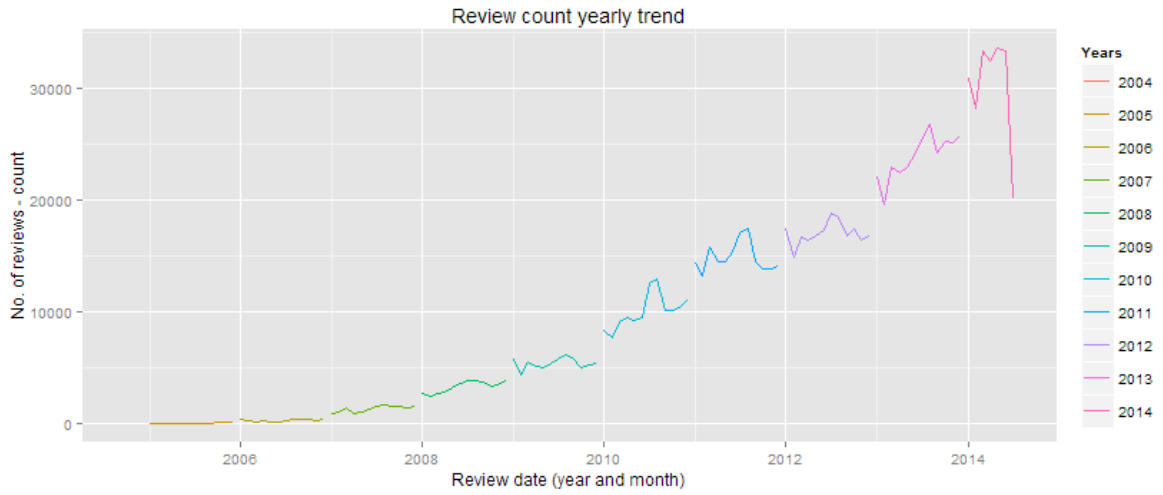
Yelp is a popular social networking website for business, which contains information and ratings for a variety of business. However, yelp contains information about restaurants and shopping places. The business dataset of yelp contains information about business of many categories, there about 707 categories for which we have location information, reviews, and check-ins. In this paper we will analyze the review and check-in for a business and predict star rating for a business using sentiment analysis, review count, and votes. We will analyze Phoenix, Las Vegas, Madison, Waterloo and Edinburgh cities in 3 states.

Below is a choropleth graph which shows the density of review counts of business in 3 states. Nevada has the most review counts, because we are analyzing data for Las Vegas city which gets a lot of tourists.



Reviews for business

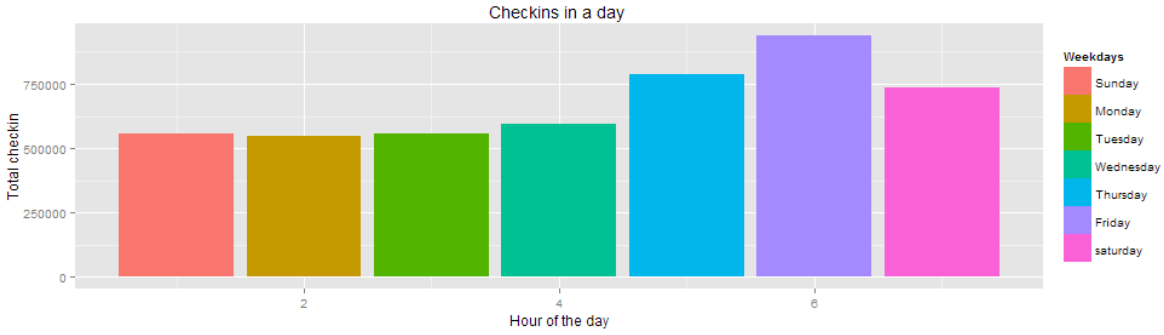
The past decade has seen significant growth in customer reviews on social networking sites such as Yelp. We know this but now we have numbers to prove the significant growth in customer reviews. Below is a graph which shows number of reviews for each month for about 10 years. For the first 5 years starting from 2004 to 2009 we can see a growing trend in number of review. The next 5 years starting from 2010 to 2014 we see a significant growth in the number of reviews.



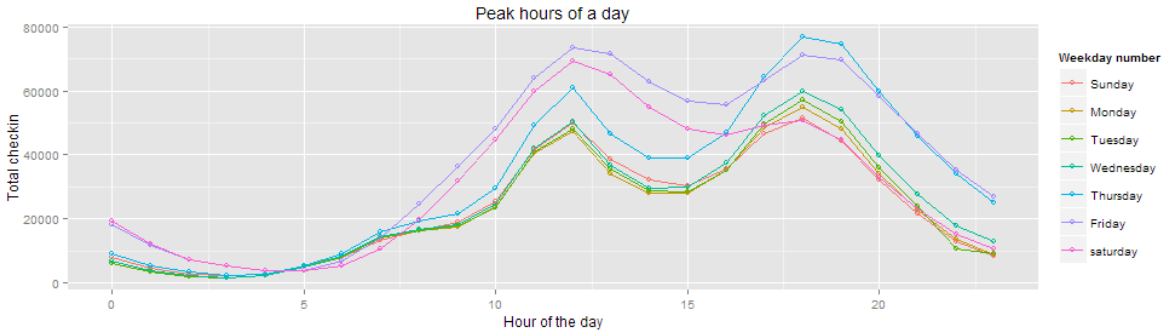
Check-ins

The Check-in dataset gives check-in count for each hour of the day i.e. 24 hours and for each day of the week from Sunday to Saturday. Although check-in information are irregular in nature because users choose to check-in using a portable device such as mobile phones, tablets, and laptops, however it provides visiting information.

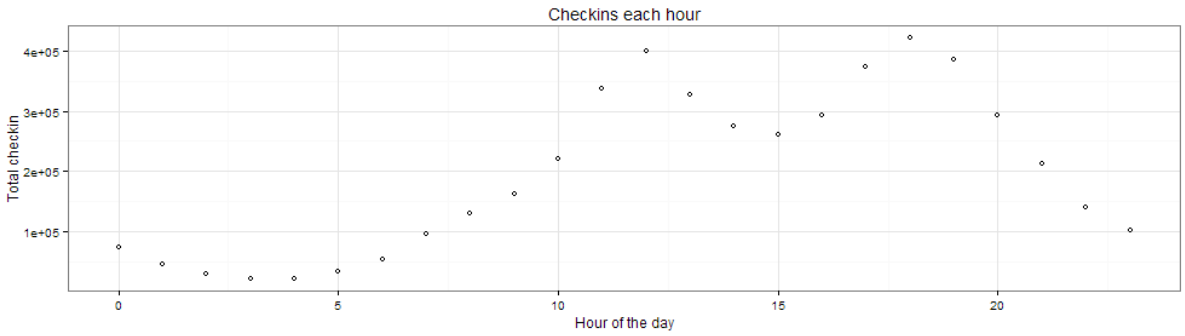
We plotted a graph to analyze which day of the week users visit businesses the most. Below is a graph which shows the check-in frequency of a day. Businesses are visited the most, in the order, on Friday, Thursday and Saturday. On days close to weekends and on weekend people visit restaurants, cafes, bars, movies, doctors, etc. and shop in the malls. While on the other days i.e. Sunday, Monday, Tuesday, and Wednesday people still go for shopping and restaurants but a little less.



We wanted to analyze which are the peak hours of the business, at which hour businesses gets the most visitors. Below is a graph which shows total of check-in for each hour of the day and for all days of the week. This is an interesting graph which shows spikes and peaks at interesting hours. The maximum check-in is at approximately between 6:00-7:00 PM on Thursday, which is very interesting and we can say that this is the top peak hour of the whole week. Another peak hour is on Friday between 12 to 1 PM and 6 to 7PM when users visit the most.

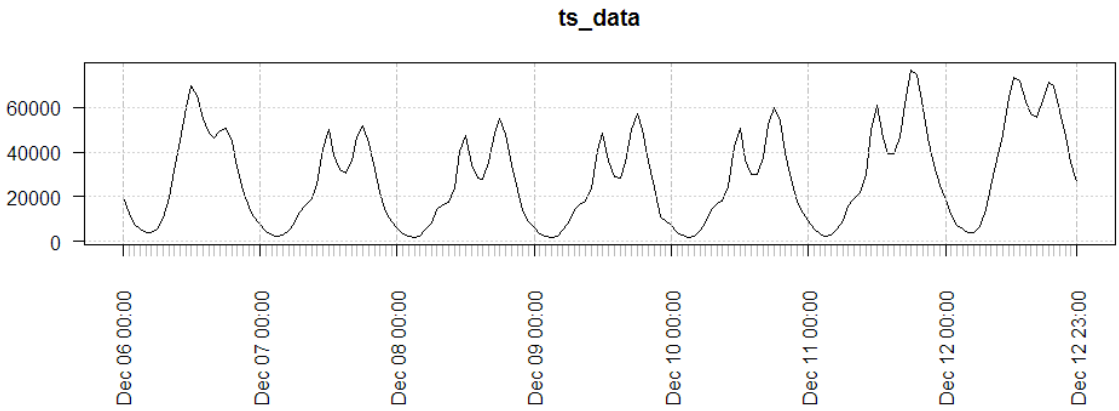


Below shows a distribution of the graph for total check-in for each hour for all days. We can see that 12:00 PM to 13:00 PM, which is most likely the lunch hour, is a peak hours for business. We can also see that 6:00 to 7:00 PM, which is most likely the dinner time, is a peak hour for business.

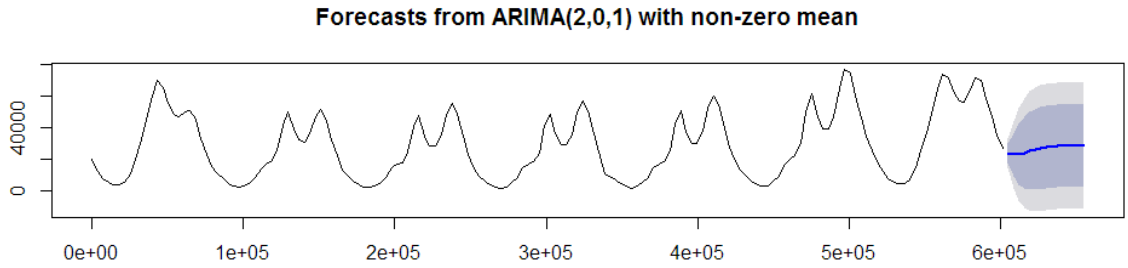


Time Series analysis

The check-in dataset provides check-in for each hour of the week so we will use this data to forecast user visits. Below is a time series which shows total check-in each hour for each day of a week. From the hourly time series we see an hourly pattern for each day. The hourly time series plot shows that the businesses are visited the most on Thursday evenings. On Friday, business are visited the most at early afternoon and then in the evening.



We use auto arima to forecast user visit for the next fifteen hours. Below graph shows a graph of the forecast model which forecasts visits for next 15 hours. The forecast model predicts a total 23500 visitors in the next hour with 95% confidence.



Linear regression analysis

Yelp contains data for business and user reviews which share reviews and experience. This is a crowdsourcing approach of generating reviews about a business. However, a single star rating is not sufficient to know the success of a business. Users share personalized experience about a business which gives more insights about a place.

In this analysis, we use linear regression to quantify a relationship between star rating of a business and unstructured and semi structured predictors. We use sentiment analysis to estimate a sentiment score for each user review and we use sentiment score to predict rating for each business. In addition, we use other columns that users also use for voting including cool, funny, and useful vote. We will create 3 linear regression models to allow comparison between models and find the best fit model.

Following are the features used for linear regression:

Total count of reviews for each business

Total of sentiment score of each business

Average sentiment score of each business

Minimum sentiment score of each business

Maximum sentiment score of each business

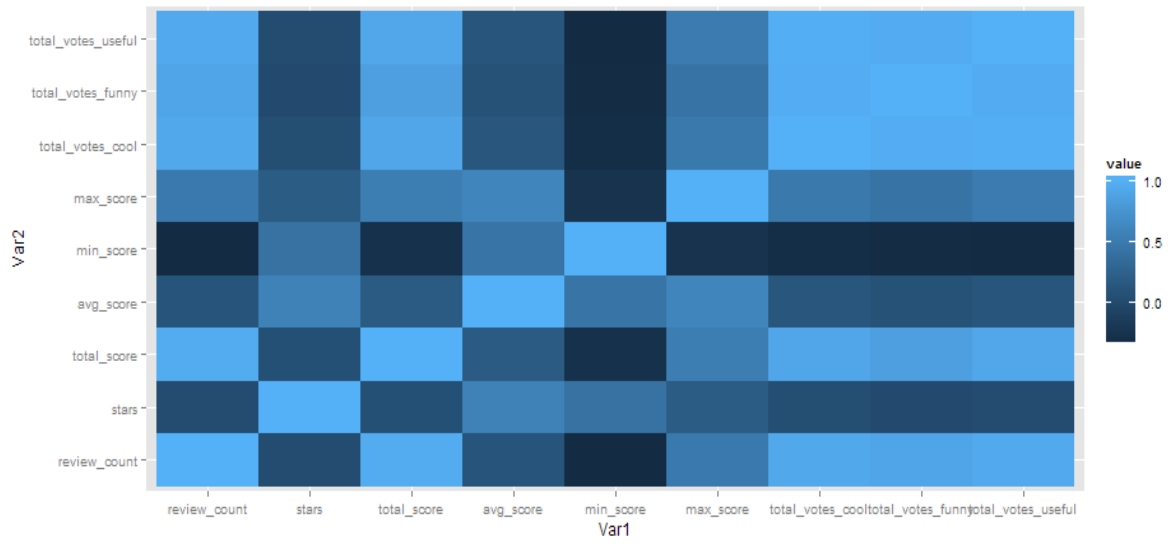
Number of cool votes of each business

Number of funny votes of each business

Number of useful votes of each business

Heat map graph with correlation

We create a correlation graph to check for correlation of the features. Following is a graph of correlation between features:



The heat map graph shows good correlation between all the features. We can see that there is high correlation of total score, review count, cool votes, funny votes, and useful votes.

We modeled 3 linear regressions with multiple predictors and regression subset to compare different linear regression fit.

Estimating sentiment

To perform a sentiment analysis, we need to identify an approach to estimate review provided by users. There are many text mining methods that provide different ways to get an estimate for a given text, however this is an interesting area where a lot of active research is being done hence we will use a simple approach to estimate text.

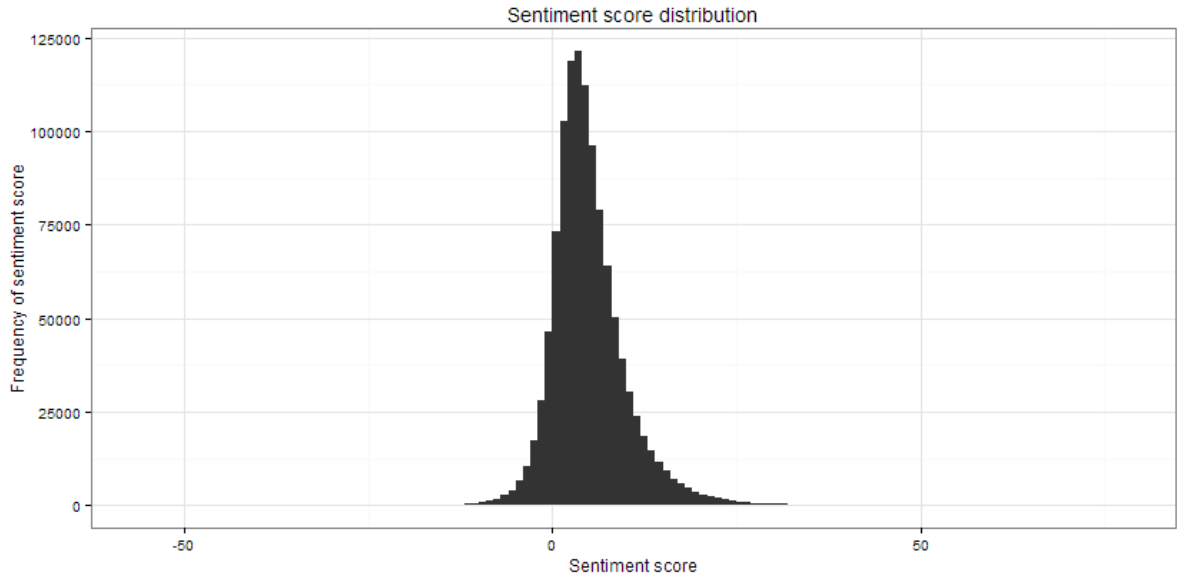
This analysis will use reviews provided by user for sentiment analysis to predict user ratings from the text. We estimate a sentiment score for each review provided by user by counting the number of occurrences of positive and negative words.

The sentiment score will be a numeric score which will be estimated for each review by subtracting the number of occurrences of negative words from the positive words. Hence a large negative score will correspond to a negative expression of sentiment, a large positive score will correspond to a positive expression of sentiment, and a neutral expression should net to zero.

We use Hu and Liu's "opinion lexicon" categorizes which contains nearly 6,800 words list of positive and negative sentiments. It can be downloaded from Bing Liu's web site:
http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

Sentiment score

We plot a histogram of the scores for the distribution of the scores. We use ggplot to plot a histogram plot of sentiment scores.



The histogram of the scores shows that most responses are positive but there are also some responses that are negative. The mean is 4.48 and median is 4, the different between mean and median is small 0.48. The distribution of the sentimental score shows that the data overall follows normal distribution however it is slightly skewed to right, because the tail is extended to the right. In conclusion, user reviews are largely positive.

1. Linear regression with sentiment score, review, and votes

We create a model which quantifies the relationship of predictors total sentiment score, review counts, number of cool votes, number of funny votes, and number of useful votes of a business as a function that predicts star rating for a business. *See appendix A for a summary of linear regression fit.*

The F-statistics is 247.7 which is a high number and p-value is 2.2e-16 which is a small number. Hence we can conclude that the model is statistically significant. The coefficients of the variables and the intercept are not close to zero and the t values of all coefficients have 3 stars. So we can say that all the variables and the intercept are significant. The R squared is 0.030 and the adjusted R squared is 0.030, both values are very small. The adjusted R squared tells us that 3% of the star rating is predicted by the predictor variables.

The mean of residual error is close to zero and Q1 and Q3 are similar with reversed signs. Residual standard error is high 0.89.

The summary of ANOVA says that probability of the event occurring by chance for 4 variables is less, however funny votes variable has a high probability 0.134 which means that the coefficient is not zero

by chance. There is a probability that the coefficient of votes funny can be zero and the coefficient is not significant. 4 variables have 3 star rating and 1 variable has no star rating. Hence we can conclude that 4 variables are significant and vote funny is not significant.

2. Multiple Linear model with multiple variables

This model will quantify a relationship between a star rating and predictor variables which will include total review counts, total sentiment score, average sentiment score, minimum sentiment score, maximum sentiment score, number of cool votes, number of funny votes, and number of useful votes. *See appendix B for summary of the model.*

The F-statistics is 2796 which is a high number and p-value is 2.2e-16 which is a small number. Hence we can conclude that the model is statistically significant. The coefficients of the variables and the intercept are not close to zero and the t values of most coefficients have 3 stars and a few coefficients have 2 and 1 stars. So we can say that all the variables and the intercept are significant. The R squared is 0.3654 and the adjusted R squared is 0.3653, both values are not very big. The adjusted R squared tells us that 36.53% of the star rating is predicted by the predictor variables.

The mean of residual error is close to zero and Q1 and Q3 are small and similar values with reversed signs. Residual standard error is high 0.72.

The summary of ANOVA says that probability of the event occurring by chance for all variables is less, In addition, most variable have 3 stars and 1 variable has 1 star, so we can say that all variable are significant.

3. Multiple regression with a subset

We create a third model that will take a subset of the data to find a fit for a regression with subset of the data. In this model, we select data for business that have 100 or more reviews.

This model will quantify a relationship between a star rating and predictor variables which will include total review counts, total sentiment score, average sentiment score, minimum sentiment score, maximum sentiment score, number of cool votes, number of funny votes, and number of useful votes. See appendix C for summary of the model.

The F-statistics is 257.2 which is a high number and p-value is 2.2e-16 which is a small number. Hence we can conclude that the model is statistically significant. The coefficients of the variables and the intercept are not close to zero and the t values of most coefficients have 3 stars and one coefficient has no stars. So we can say that 8 variables and the intercept are significant, but review count variable is not significant. The R squared is 0.4641 and the adjusted R squared is 0.4623, both values are fairly big. The adjusted R squared tells us that 46.23% of the star rating is predicted by the predictor variables.

The mean of residual error is close to zero and Q1 and Q3 are small and similar values with reversed signs. Residual standard error is high 0.37, which is less.

The summary of ANOVA says that probability of the event occurring by chance for all variables is less, In addition, all variable have 3 stars, so we can say that all variable are significant.

Conclusion

We modeled 3 different types of fit to find the best fit model that predicts star ratings for business.

The F-statistics for all 3 models are significant and the coefficients are also significant. But the last model has the best adjusted r squared error of 46.23% and the residual standard error is also less 0.37. Hence we conclude that the third model which is a multiple regression model that uses a subset of the data to be the best fit model.

Conclusion

In conclusion, we can predict with reasonable accuracy the average star rating that will be assigned to a business, as long as we know sentiment score, number of reviews, and cool, funny and useful votes. We see a significant growth in reviews in the past decade. Businesses are visited most on Friday, Thursday, and Saturday in respective order. Business gets most visits on Thursday between 6:00 to 7:00 PM in the evenings, Friday between 12:00 to 1:00 PM in the afternoon, and 6:00 to 7:00 PM on Friday.

For further investigations, we can use other variables such as other business attributes, opening hours, and location proximity to get a better accuracy at predicting star rating. Furthermore, we could extend this to predict individual user ratings rather than the average, since people tend to respond differently. A K-means clustering of users can also be used to get more insights about the user demographics.

Acknowledgement

I would like to thank our instructor Dr. Ram Narasimhan for a wonderful class at University of California Santa Cruz Silicon Valley Extension College. The material was well organized, and the lectures included practical code and were interactive.

References

http://www.inside-r.org/howto/mining-twitter-airline-consumer-sentiment

http://www.yelp.com/dataset_challenge

http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

Appendix

A. Linear regression with sentiment score, review count, and votes

Call:

```
lm(formula = stars ~ ., data = data_analyze_2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3783 -0.6376 -0.0678  0.7900  4.1374

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.676e+00  4.858e-03 756.609   <2e-16 ***
review_count    -3.597e-03  1.997e-04 -18.013   <2e-16 ***
total_score     8.299e-04  3.978e-05  20.863   <2e-16 ***
total_votes_funny  -3.082e-03  3.707e-04  -8.313   <2e-16 ***
total_votes_cool    7.419e-03  4.093e-04  18.124   <2e-16 ***
total_votes_useful -2.981e-03  2.517e-04 -11.845   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8884 on 38854 degrees of freedom
Multiple R-squared:  0.03089,    Adjusted R-squared:  0.03077
F-statistic: 247.7 on 5 and 38854 DF,  p-value: < 2.2e-16
```

B. Multiple linear regression with multiple variable

```
Call:
lm(formula = stars ~ ., data = data_analyze[, -1])

Residuals:
    Min      1Q  Median      3Q     Max
-5.4981 -0.4456  0.0260  0.4847  3.6727

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.1369453  0.0084954 369.250  < 2e-16 ***
review_count    0.0008399  0.0001652   5.083 3.74e-07 ***
total_score     -0.0001085  0.0000332  -3.268  0.00108 **
avg_score       0.2223839  0.0028058  79.260  < 2e-16 ***
min_score       0.0303201  0.0014063  21.561  < 2e-16 ***
max_score       -0.0212118  0.0008666 -24.477  < 2e-16 ***
total_votes_cool    0.0046764  0.0003328  14.050  < 2e-16 ***
total_votes_funny  -0.0036999  0.0003032 -12.204  < 2e-16 ***
total_votes_useful -0.0004469  0.0002108  -2.120  0.03399 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.719 on 38851 degrees of freedom
```

Multiple R-squared: 0.3654, Adjusted R-squared: 0.3653
F-statistic: 2796 on 8 and 38851 DF, p-value: < 2.2e-16

C. Multiple linear regression with subset regression

```
lm(formula = stars ~ ., data = data_analyze[data_analyze$review_count >
## 100, -c(1)])
##
## Residuals:
## Min 1Q Median 3Q Max
## -1.4578 -0.2495 -0.0035 0.2383 1.3899
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.22e+00 4.36e-02 73.87 < 2e-16 ***
## review_count 1.02e-04 1.13e-04 0.90 0.3668
## total_score -6.32e-05 2.33e-05 -2.72 0.0066 **
## avg_score 1.96e-01 8.81e-03 22.29 < 2e-16 ***
## min_score 2.68e-02 2.28e-03 11.74 < 2e-16 ***
## max_score -1.05e-02 1.14e-03 -9.21 < 2e-16 ***
## total_votes_cool 3.13e-03 2.03e-04 15.43 < 2e-16 ***
## total_votes_funny -1.35e-03 1.86e-04 -7.27 4.7e-13 ***
## total_votes_useful -7.90e-04 1.28e-04 -6.15 8.8e-10 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.378 on 2376 degrees of freedom
## Multiple R-squared: 0.464, Adjusted R-squared: 0.462
## F-statistic: 257 on 8 and 2376 DF, p-value: <2e-16
```