Analysis on Online Customers Review

January 2019

(i) Executive Summary

What does a company care about most? Where do company's most fundamental interests come from? The answer, no doubt, is the customer. Customer behavioral analysis cases are increasing with fast-paced industry growth. In order to figure out what kind of factors have big impacts on online customer review, the analysis was performed on the data set from yelp.co.uk that records customer review behavior towards the reviews in the application, which indicates the purchases signals from customers . Let helpfulness and readership as the response variables to perform the analysis. It is important to know what the strongest predictors of helpfulness and readership towards reviews are.(H. Chen,2012) In this project, to analyze the relationship between the customer behaviors and features of reviews, and how to use this relationship to enhance the interests of the business, we have conducted statistical modeling to analyze and extract relevant information that has an impact on customer behaviors.

(ii) Basic Description of Statistical Methodology

In this project, because finding the relationship of the factors is a statistical problem, we used linear regression model to eliminate relationships between them. After getting the model, I used summary() and ANOVA() in R to have a initial check. Then we will use random forest method and varImpPlot() to have a double check, so that I can be sure what kind of factors are important to the regression model.

(iii) Analysis of Data Findings

1. Establish the hypothesis

Once we get the data, we are supposed to be clear about what the data look like and what the characteristics of the data are, which are critical for analyzing if they have important impacts on the customer behaviors.

- H1 longevity has positive impact on helpfulness and readership

  When it comes to the longevity, we can think that the longer the reviews are shown to people, more people will find them helpful if they are really helpful reviews. While considering readership, the longer reviews shown to customers, more readership they will have.

  Longevity = date set by my own("2016-12-31") – date

- H2 sentiment has positive impact on helpfulness and readership
  When we analyze the customer behaviors, we always consider customers' emotion as one of the most important factors for determining if the products got customers' attention.(X.-T.Wang,2006) Here, we consider sentiment in the review text in the same way. We think positive sentiment in the text will make people pay more attention to the products and want to go deeper with the product information. (M. Thelwall,2013)

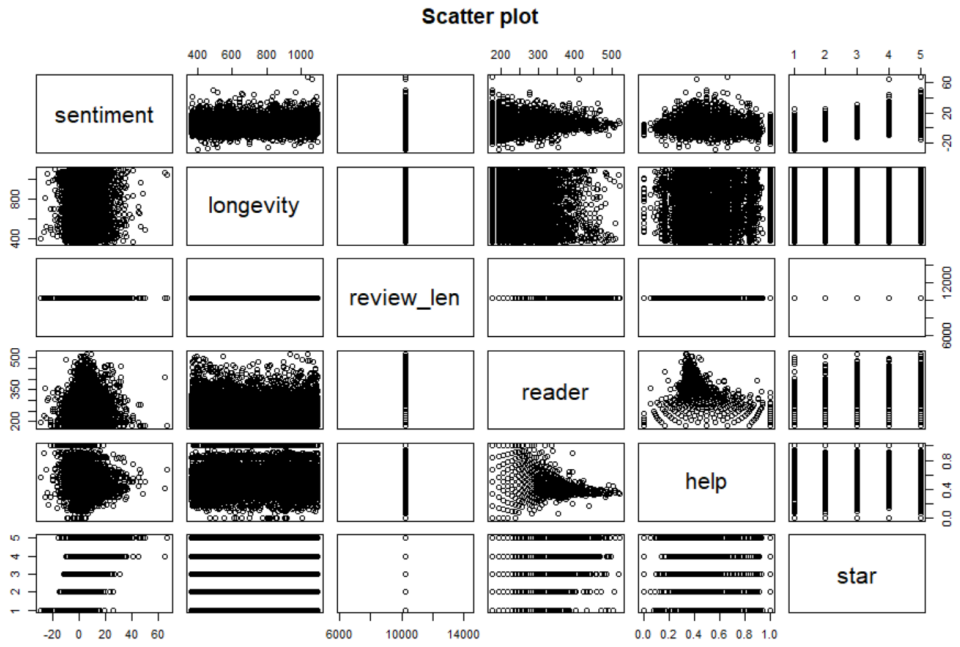
  Sentiment = (sum(positive) - sum(negative))-2

- H3 length of text has positive impact on the helpfulness and readership
  Let we consider about this situation. When a customer starts reading the reviews, short reviews often make them feel they are fake, while long reviews make people feel they are put in efforts from people, they are talking about their true feelings about the products.(D.-H.Park,2007) So maybe the length of text has positive influence on the impression customer have on the products.

- H4 stars has positive impact on the helpfulness and readership
  The reputation star system always attracts people's attention to the products. If a product has high star reputation, customers will have more trust on it and it will be larger probability for them to go deeper with the products.

2. Review factors and Customer Behaviors

Once we have our own hypothesis, the first thing we should do is to recreate the variables we need in the hypothesis through the formula I mentioned above. At first, I put all the variables including helpfulness and readership into the scatter plot, trying to find out their correlations between each other. We can find out that longevity, sentiment and stars have impacts on both helpfulness and readership. (*fig is below*)



3. Build the linear regression model

Researching only factors in a scatter plot is not enough to determine the importance of factors for the response variables. We need to do more attempts. In this part, I build two linear regression models. One if for helpfulness and the other is for readership. I use ANOVA() to check the importance of the factors. Through the tables, we can say that the length of text has nothing to do with the customer behaviors, while the other factors have big impacts on them. (*fig is below*)

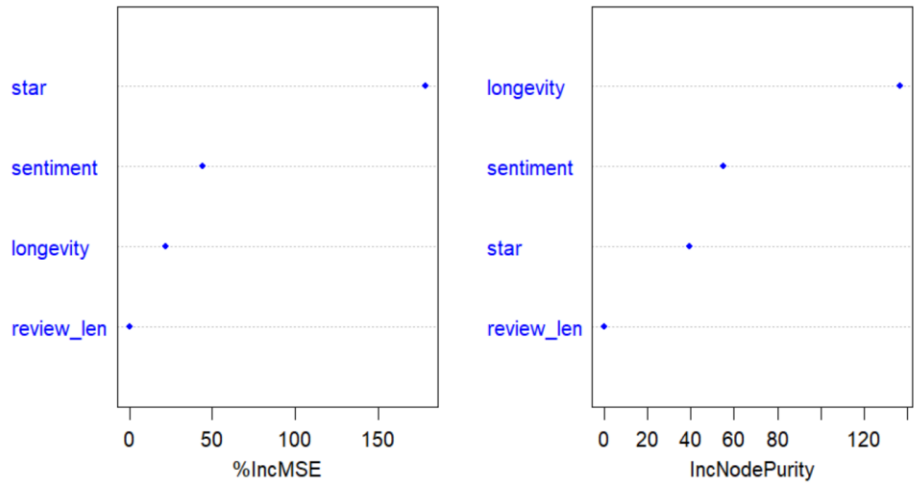
```
> anova(fit,test="Chisq")
Analysis of Variance Table

Response: help
             Df  Sum Sq Mean Sq  F value Pr(>F)
star          1  25.515 25.5146 910.1479 <2e-16 ***
sentiment     1   0.012  0.0120   0.4279  0.513
longevity     1   1.966  1.9661  70.1355 <2e-16 ***
Residuals 10234 286.895  0.0280
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> anova(fit,test="Chisq")
Analysis of Variance Table

Response: reader
             Df   Sum Sq Mean Sq F value    Pr(>F)
star          1   331760  331760 113.153 < 2.2e-16 ***
sentiment     1   330718  330718 112.798 < 2.2e-16 ***
longevity     1   122214  122214  41.683 1.122e-10 ***
Residuals 10234 30005542    2932
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In addition, I also conducted random forest model to have a double check if these factors are important for the helpfulness and readership. Here, I used varImpPlot() to achieve that. (*fig is below*) So we can conclude that longevity, stars and sentiment may have impacts on the customer behaviors, but the length of text may not have impact on them directly.



(iv) How to improve the result of readership and helpfulness of reviews

In order to achieve the best trade-off between customers and products, establishing a reputation system may be a sustainable development strategy, which allows customers read the reviews more directly and easily. For example, we can establish a system including stars, longevity and sentiment score for every review. Since the factors mentioned above have big impacts on the customer behaviors. Therefore, people may have a good platform to know the products well and select the reviews to read more efficiently.

In addition, the reviews should be voted by customers who have already bought the same products, so that we can make a ranking of the reviews according to the number of voters. That would improve the efficiency for people reading the reviews.

(v) Conclusions

After we reviewed the parts above we determined that it would be a good idea for a company to go deeper with the data to improve their business. The first major thing is that the data could provide the basic information, such as what kind of factor has the biggest importance for helpfulness and readership, after getting the information of the data, where can be improved between products review system and customer behaviors to satisfy the market. So coming up some brand-new strategies may lead to higher profits and more success for the long-term.

(vi) Appendix

R code

```
str(review_sample)

### calculate positive and negative
library(stringr)
#setup the working directory
setwd("C:/STONY/Practice/R (No.12)")
nk.text<- sapply(review_sample$text,function(row) iconv(row, "latin1", "ASCII", sub=""))

#loading Hu Liu's opinion lexicon
hu.liu.pos<- scan("positive-words.txt",what="character", comment.char="");
hu.liu.neg<- scan("negative-words.txt",what="character", comment.char="");

#loading some industry-specific and/or especially emphatic terms
pos.words<- c(hu.liu.pos, 'prize')
neg.words<- c(hu.liu.neg, 'late')

#function for the score.sentiment
score.sentiment = function(sentences, pos.words, neg.words, .progress='none')
{
  require(plyr)
  require(stringr)

  # we got a vector of sentences. plyr will handle a list
  # or a vector as an "l" for us
  # we want a simple array ("a") of scores back, so we use
  # "l" + "a" + "ply" = "laply":
  scores = laply(sentences, function(sentence, pos.words, neg.words) {

    # clean up sentences with R's regex-driven global substitute, gsub():
    sentence = gsub('[[:punct:]]', '', sentence)
    sentence = gsub('[[:cntrl:]]', '', sentence)
    sentence = gsub('\\d+', '', sentence)
    # and convert to lower case:
    sentence = tolower(sentence)
```

```
    # split into words. str_split is in the stringr package
    word.list = str_split(sentence, '\\s+')
    # sometimes a list() is one level of hierarchy too much
    words = unlist(word.list)

    # compare our words to the dictionaries of positive & negative terms
    pos.matches = match(words, pos.words)
    neg.matches = match(words, neg.words)

    # match() returns the position of the matched term or NA
    # we just want a TRUE/FALSE:
    pos.matches = !is.na(pos.matches)
    neg.matches = !is.na(neg.matches)

    # and conveniently enough, TRUE/FALSE will be treated as 1/0 by sum():
    score = (sum(pos.matches) - sum(neg.matches))-2

    return(score)
  }, pos.words, neg.words, .progress=.progress )

  scores.df = data.frame(score=scores, text=sentences)
  return(scores.df)
}



#score the tweets
nk.scores=score.sentiment(nk.text,pos.words,neg.words,.progress='text')
class(nk.scores)
#save the score in a csv file
write.csv(nk.scores, "review_scores.csv")

### read the sentiment information and create new data frame
sentiment <- read.csv("review_scores.csv")[,2]
longevity = difftime(as.Date("2016-12-31"), as.Date(review_sample$date), units="days")
review_len = log(length(review_sample$text))
reader = log(review_sample$votes_sum)*100
help = review_sample$votes$useful/(review_sample$votes_sum)
```

```
review <-
cbind(sentiment,longevity,review_len,reader,help,star=review_sample$stars) %>%
as.data.frame()


### scatter plot among the variables
plot(review, main="Scatter plot")

### correlation function among the width and length
cor <- cor(review) %>% as.matrix()
install.packages("corrplot")
library(corrplot)
corrplot(cor, method="circle") # the darker color is, the greater the correlation will be
corrplot(cor, method="number")


# Prediction Model   #
######################
# Preparing regression function for helpness
regression <- help~
  star+
  sentiment+
  longevity+
  review_len


# build the regression for the model and check the variable importance
fit <- lm(regression,data=review)
summary(fit)
anova(fit,test="Chisq")

# build the random forest model and double check the variable importance
set.seed(1234)
library(randomForest)
data_fac=review %>% mutate_if(is.character, as.factor)
rf.fit <- randomForest(regression,data=data_fac,mtry=3,ntree=1000,importance=TRUE)
varImpPlot(rf.fit,color="blue",pch=20,cex=1.25,main="")
```

```
# Preparing regression function for readership
regression <- reader~
   star+
   sentiment+
   longevity+
   review_len


# build the regression for the model and check the variable importance
fit <- lm(regression,data=review)
summary(fit)
anova(fit,test="Chisq")

# build the random forest model and double check the variable importance
set.seed(1234)
library(randomForest)
data_fac=review %>% mutate_if(is.character, as.factor)
rf.fit <- randomForest(regression,data=data_fac,mtry=3,ntree=1000,importance=TRUE)
varImpPlot(rf.fit,color="blue",pch=20,cex=1.25,main="")
```

(vii) Reference

[1] X.-T. Wang, Emotions within reason: resolving conflicts in risk preference, Cognition and Emotion 20 (8) (2006) 1132–1152.
[2] H. Chen, R.H. Chiang, V.C. Storey, Business intelligence and analytics: from big data to big impact, MIS Quarterly 36 (4) (2012) 1165–1188.
[3] D.-H. Park, J. Lee, I. Han, The effect of on-line consumer reviews on consumer purchasing intention: the moderating role of involvement, International Journal of Electronic Commerce 11 (4) (2007) 125–148.
[4] M. Thelwall, K. Buckley, Topic-based sentiment analysis for the social web: the role of mood and issue-related words, Journal of the American Society for Information Science and Technology 64 (8) (2013).