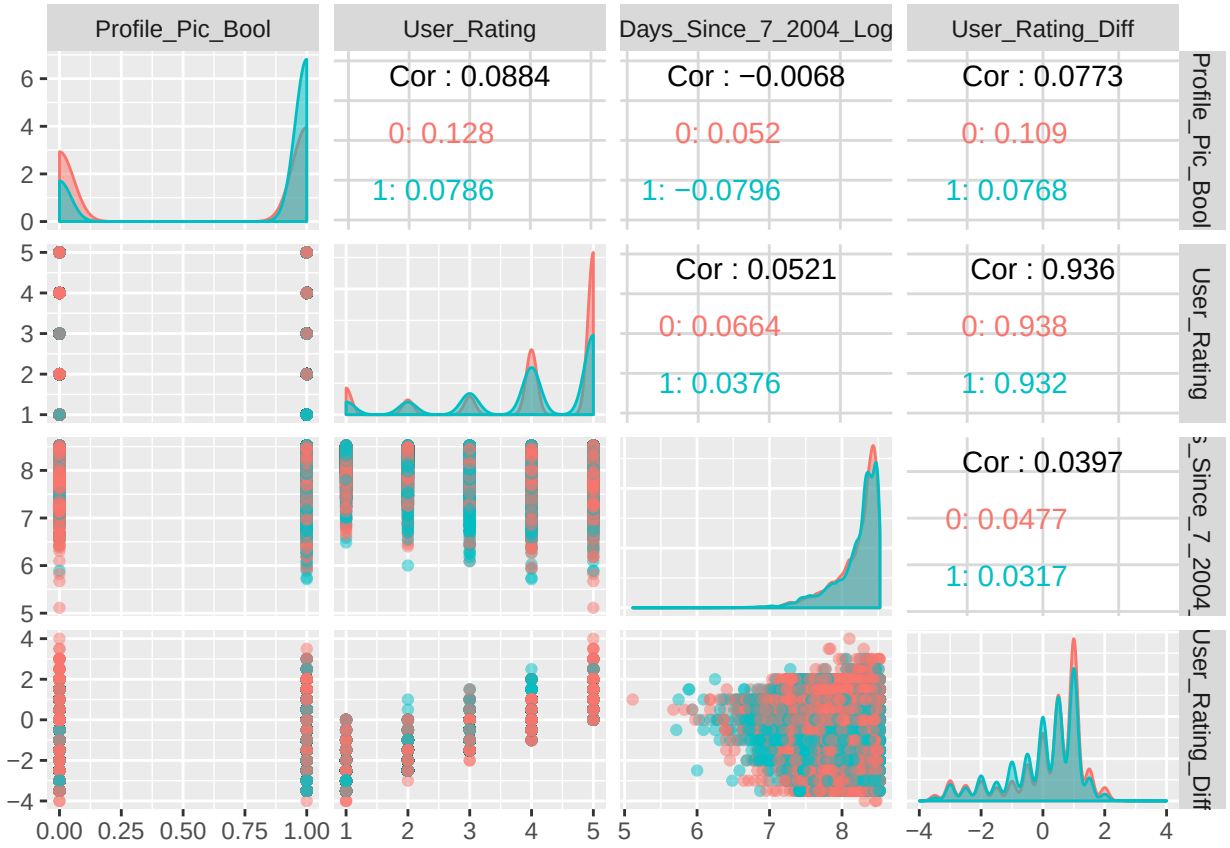
R Notebook

```
setwd("C:\\Users\\Yao\\data")
library(GGally)
analysis_balanced = read.csv("balanced_reviews.csv")
analysis_balanced$index <- NULL
summary(analysis_balanced)
```

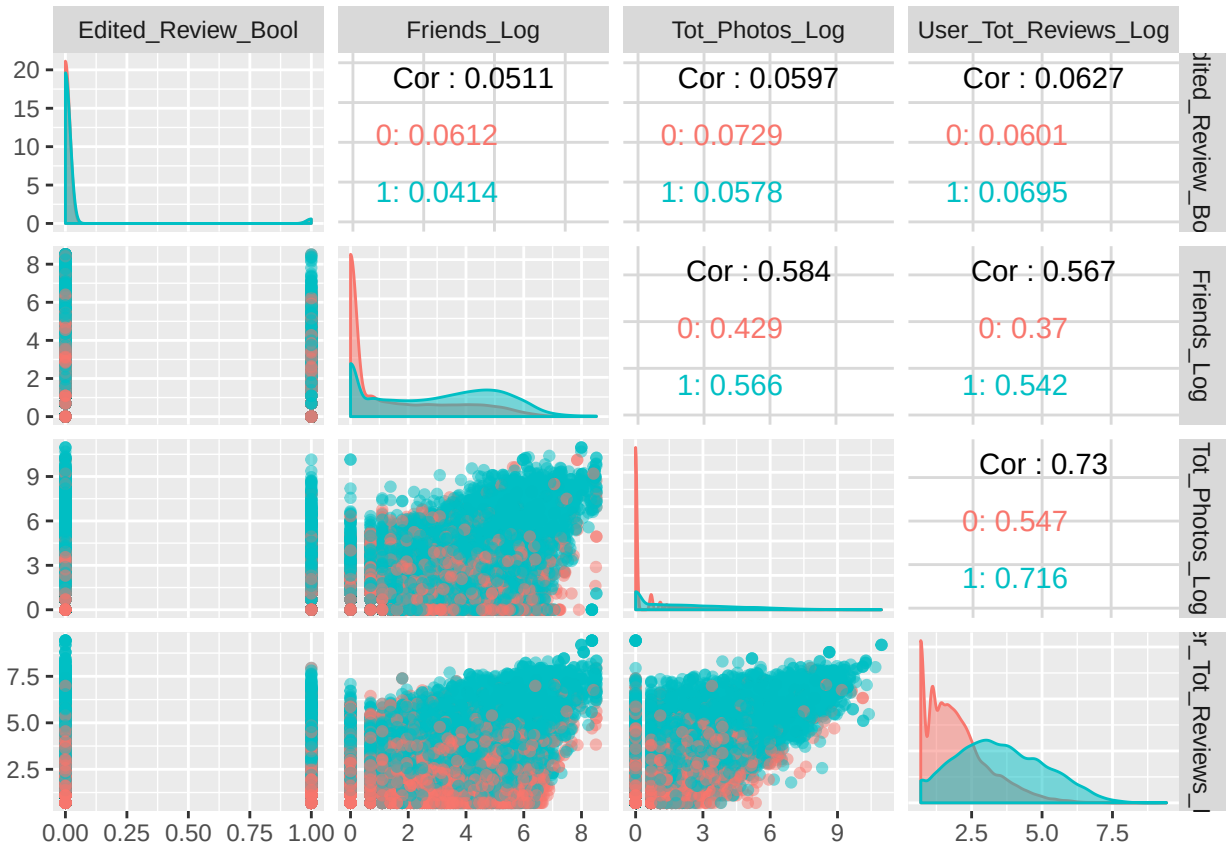
```
##  Profile_Pic_Bool  User_Rating    Rev_Avg_Sentiment  Recommended
##  Min.   :0.0000   Min.   :1.000   Min.   :0.000     Min.   :0.0
##  1st Qu.:0.0000   1st Qu.:4.000   1st Qu.:1.727     1st Qu.:0.0
##  Median :1.0000   Median :5.000   Median :2.167     Median :0.5
##  Mean   :0.6873   Mean   :4.046   Mean   :2.189     Mean   :0.5
##  3rd Qu.:1.0000   3rd Qu.:5.000   3rd Qu.:2.667     3rd Qu.:1.0
##  Max.   :1.0000   Max.   :5.000   Max.   :4.000     Max.   :1.0
##  Rev_V_Negative    Rev_V_Positive   Edited_Review_Bool Yelp_Rest_Order
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000     Min.   :  1.0
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000     1st Qu.:219.0
##  Median :0.00000   Median :0.0000   Median :0.0000     Median :419.0
##  Mean   :0.07118   Mean   :0.2681   Mean   :0.0272     Mean   :419.8
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000     3rd Qu.:574.0
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000     Max.   :990.0
##  Restaurant_ID   Ratio_Recommended Rest_Address_Words Rest_Name_Words
##  Min.   :  0.0   Min.   :0.1923    Min.   : 3.000     Min.   :1.000
##  1st Qu.:207.0   1st Qu.:0.7676    1st Qu.: 7.000     1st Qu.:2.000
##  Median :325.0   Median :0.8708    Median : 8.000     Median :3.000
##  Mean   :339.3   Mean   :0.8297    Mean   : 7.911     Mean   :2.697
##  3rd Qu.:512.0   3rd Qu.:0.9022    3rd Qu.: 9.000     3rd Qu.:4.000
##  Max.   :665.0   Max.   :0.9856    Max.   :14.000     Max.   :9.000
##  User_Rating_Diff    Friends_Log    Days_Since_7_2004_Log
##  Min.   :-4.00000   Min.   :0.000   Min.   :5.112
##  1st Qu.:-0.50000   1st Qu.:0.000   1st Qu.:8.086
##  Median : 0.50000   Median :1.386   Median :8.315
##  Mean   : 0.07073   Mean   :2.108   Mean   :8.204
##  3rd Qu.: 1.00000   3rd Qu.:4.043   3rd Qu.:8.426
##  Max.   : 4.00000   Max.   :8.517   Max.   :8.523
##  Review_Sentence_Log Tot_Photos_Log   User_Tot_Reviews_Log
##  Min.   :0.000       Min.   : 0.000   Min.   :0.6931
##  1st Qu.:1.386       1st Qu.: 0.000   1st Qu.:1.3863
##  Median :1.792       Median : 0.000   Median :2.3979
##  Mean   :1.831       Mean   : 1.316   Mean   :2.7342
##  3rd Qu.:2.197       3rd Qu.: 2.197   3rd Qu.:3.7612
##  Max.   :4.625       Max.   :10.963   Max.   :9.4221
##  Rev_Tot_Sentiment_Log Tot_Rest_Reviews_Log Tot_Rest_In_City_Log
##  Min.   :0.000         Min.   :0.6931       Min.   : 2.485
##  1st Qu.:2.079         1st Qu.:5.6836       1st Qu.: 7.511
##  Median :2.485         Median :6.6958       Median : 8.107
##  Mean   :2.457         Mean   :6.6094       Mean   : 7.963
##  3rd Qu.:2.890         3rd Qu.:7.7332       3rd Qu.: 8.459
##  Max.   :5.313         Max.   :8.4820       Max.   :10.104
##  Review_Words_Log Review_Words_No_Stopwords_Log Rev_Dist_Miles_Log
##  Min.   :0.6931   Min.   :0.000                Min.   :0.000
```

```
##  1st Qu.:3.3673   1st Qu.:2.944                1st Qu.:1.499
##  Median :4.0431   Median :3.555                Median :2.340
##  Mean   :4.0045   Mean   :3.547                Mean   :3.052
##  3rd Qu.:4.7185   3rd Qu.:4.205                3rd Qu.:4.195
##  Max.   :6.8773   Max.   :6.390                Max.   :9.361
```

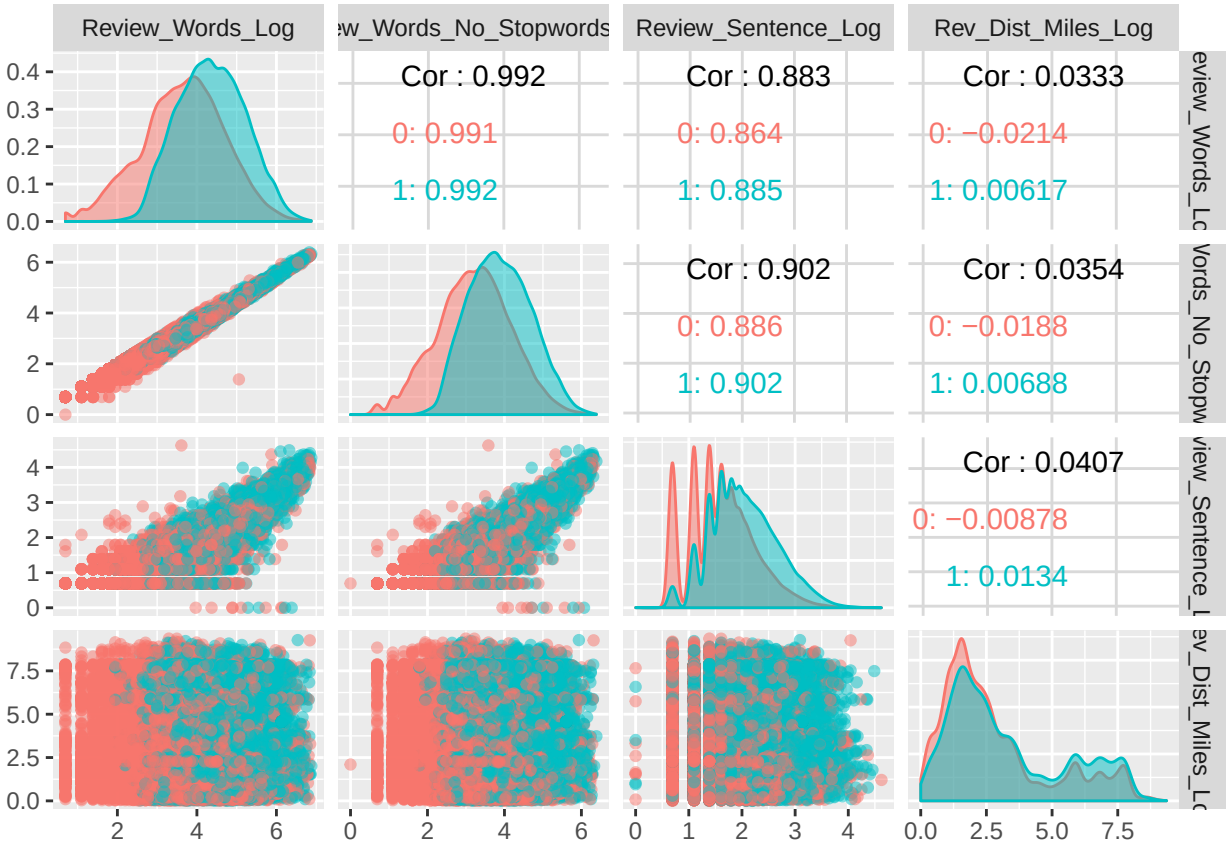
```
library(ggplot2)
my_dens <- function(data, mapping, ...) {
  ggplot(data = data, mapping=mapping) +
    geom_density(aes(fill = Recommended), alpha=0.5)
}

analysis_balanced$Recommended <- factor(analysis_balanced$Recommended, levels = c(0,1))
ggpairs(analysis_balanced, columns = c('Profile_Pic_Bool', 'User_Rating', 'Days_Since_7_2004_Log', 'Use
```



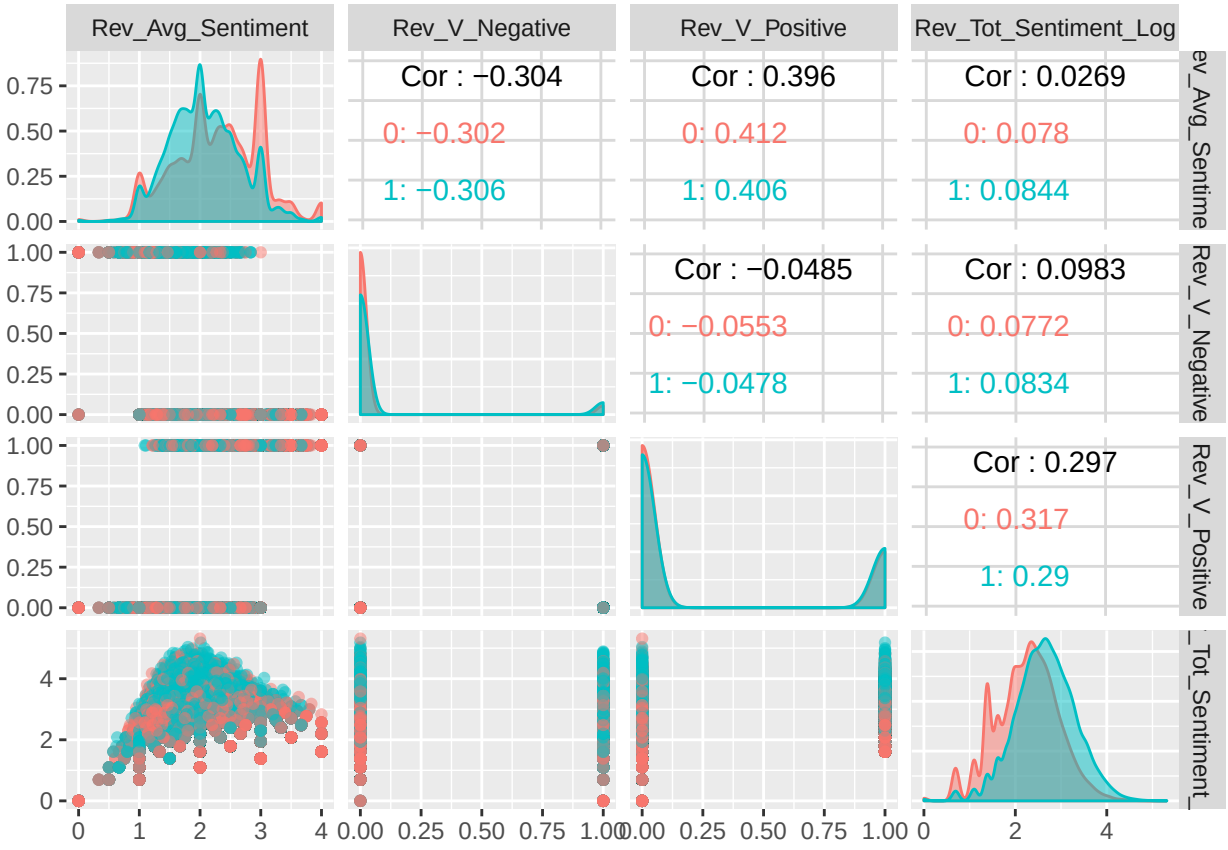

```
ggpairs(analysis_balanced, columns = c('Edited_Review_Bool', 'Friends_Log', 'Tot_Photos_Log', 'User_Tot
```



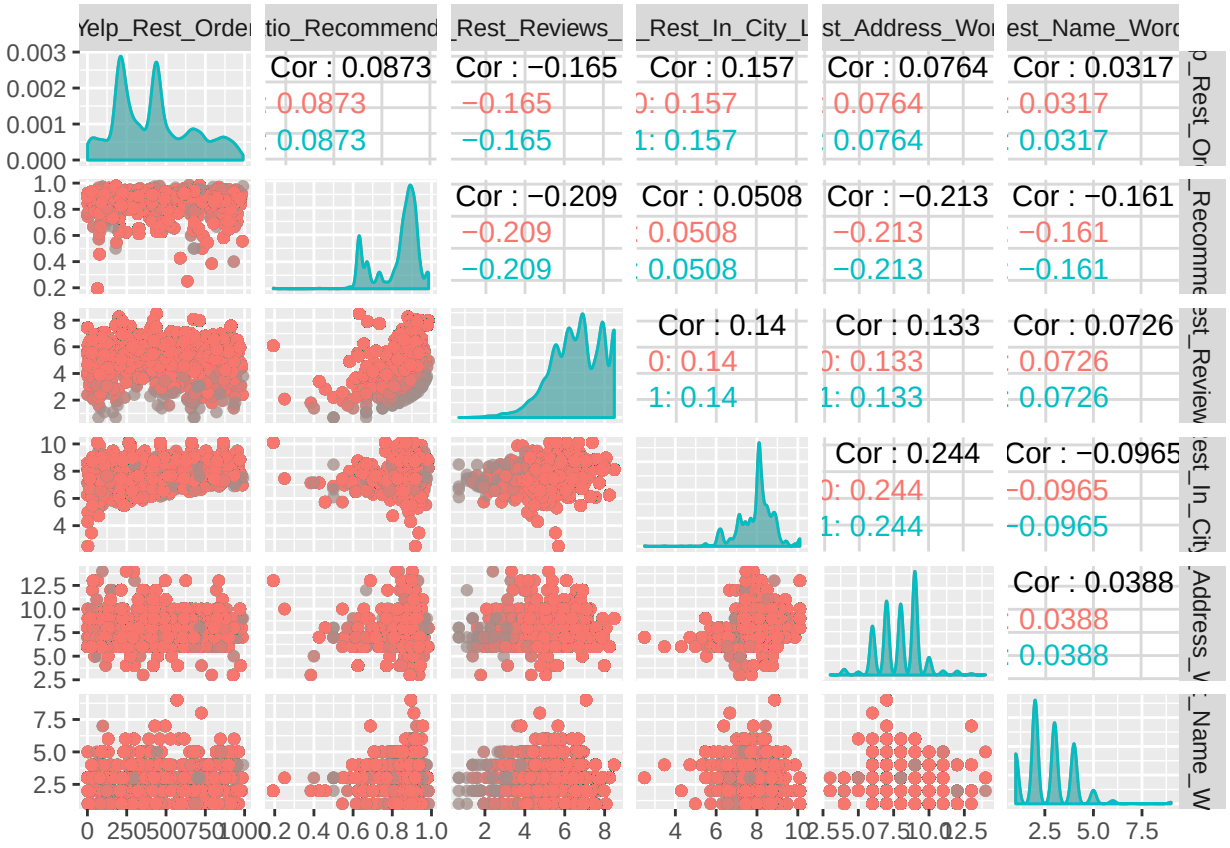

```
ggpairs(analysis_balanced, columns = c('Review_Words_Log', 'Review_Words_No_Stopwords_Log', 'Review_Sen
```




```
ggpairs(analysis_balanced, columns = c('Rev_Avg_Sentiment', 'Rev_V_Negative', 'Rev_V_Positive', 'Rev_To
```




```
ggpairs(analysis_balanced, columns = c('Yelp_Rest_Order','Ratio_Recommended', 'Tot_Rest_Reviews_Log', '
```




```
model <- glm(Recommended ~ Profile_Pic_Bool + User_Rating + Days_Since_7_2004_Log + User_Rating_Diff + E
summary(model)
```

```
##
## Call:
## glm(formula = Recommended ~ Profile_Pic_Bool + User_Rating +
##     Days_Since_7_2004_Log + User_Rating_Diff + Edited_Review_Bool +
##     Friends_Log + Tot_Photos_Log + User_Tot_Reviews_Log + Review_Words_Log +
##     Review_Words_No_Stopwords_Log + Review_Sentence_Log + Rev_Dist_Miles_Log +
##     Rev_Avg_Sentiment + Rev_V_Negative + Rev_V_Positive + Rev_Tot_Sentiment_Log +
##     Yelp_Rest_Order + Ratio_Recommended + Tot_Rest_Reviews_Log +
##     Tot_Rest_In_City_Log + Rest_Address_Words + Rest_Name_Words,
##     family = binomial(link = "logit"), data = analysis_balanced)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6400  -0.7882  -0.0286   0.7772   2.7505
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -9.272e+00  3.979e-01 -23.303  < 2e-16 ***
## Profile_Pic_Bool              1.559e-01  2.544e-02   6.129 8.85e-10 ***
## User_Rating                   1.018e-01  2.710e-02   3.757 0.000172 ***
## Days_Since_7_2004_Log         8.007e-01  3.859e-02  20.750  < 2e-16 ***
## User_Rating_Diff             -1.676e-01  2.719e-02  -6.162 7.19e-10 ***
## Edited_Review_Bool           -7.416e-01  6.909e-02 -10.734  < 2e-16 ***
```

```
## Friends_Log                    1.265e-01  6.392e-03  19.788  < 2e-16 ***
## Tot_Photos_Log                 1.016e-01  1.020e-02   9.959  < 2e-16 ***
## User_Tot_Reviews_Log           7.276e-01  1.177e-02  61.830  < 2e-16 ***
## Review_Words_Log               3.556e-01  9.130e-02   3.895 9.82e-05 ***
## Review_Words_No_Stopwords_Log  6.438e-01  1.068e-01   6.028 1.66e-09 ***
## Review_Sentence_Log           -3.857e+00  2.126e-01 -18.140  < 2e-16 ***
## Rev_Dist_Miles_Log             7.517e-02  5.057e-03  14.863  < 2e-16 ***
## Rev_Avg_Sentiment             -1.332e+00  7.623e-02 -17.468  < 2e-16 ***
## Rev_V_Negative                 5.152e-02  4.866e-02   1.059 0.289768
## Rev_V_Positive                 2.841e-03  2.845e-02   0.100 0.920443
## Rev_Tot_Sentiment_Log          3.188e+00  1.819e-01  17.525  < 2e-16 ***
## Yelp_Rest_Order               -3.902e-06  4.739e-05  -0.082 0.934387
## Ratio_Recommended             -1.014e+00  1.099e-01  -9.223  < 2e-16 ***
## Tot_Rest_Reviews_Log          -8.143e-02  9.250e-03  -8.803  < 2e-16 ***
## Tot_Rest_In_City_Log          -5.818e-02  1.355e-02  -4.295 1.75e-05 ***
## Rest_Address_Words             6.596e-03  7.907e-03   0.834 0.404142
## Rest_Name_Words               -1.315e-02  8.947e-03  -1.470 0.141556
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 74558  on 53781  degrees of freedom
## Residual deviance: 51897  on 53759  degrees of freedom
## AIC: 51943
##
## Number of Fisher Scoring iterations: 5
```

```
#Remove Not significant features.
model2 = update(model, ~.-Parch-Fare-Embarked)
summary(model2)
```

```
##
## Call:
## glm(formula = Recommended ~ Profile_Pic_Bool + User_Rating +
##     Days_Since_7_2004_Log + User_Rating_Diff + Edited_Review_Bool +
##     Friends_Log + Tot_Photos_Log + User_Tot_Reviews_Log + Review_Words_Log +
##     Review_Words_No_Stopwords_Log + Review_Sentence_Log + Rev_Dist_Miles_Log +
##     Rev_Avg_Sentiment + Rev_V_Negative + Rev_V_Positive + Rev_Tot_Sentiment_Log +
##     Yelp_Rest_Order + Ratio_Recommended + Tot_Rest_Reviews_Log +
##     Tot_Rest_In_City_Log + Rest_Address_Words + Rest_Name_Words,
##     family = binomial(link = "logit"), data = analysis_balanced)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6400  -0.7882  -0.0286   0.7772   2.7505
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -9.272e+00  3.979e-01 -23.303  < 2e-16 ***
## Profile_Pic_Bool              1.559e-01  2.544e-02   6.129 8.85e-10 ***
## User_Rating                   1.018e-01  2.710e-02   3.757 0.000172 ***
## Days_Since_7_2004_Log         8.007e-01  3.859e-02  20.750  < 2e-16 ***
## User_Rating_Diff             -1.676e-01  2.719e-02  -6.162 7.19e-10 ***
## Edited_Review_Bool           -7.416e-01  6.909e-02 -10.734  < 2e-16 ***
```

```
## Friends_Log                    1.265e-01  6.392e-03  19.788  < 2e-16 ***
## Tot_Photos_Log                 1.016e-01  1.020e-02   9.959  < 2e-16 ***
## User_Tot_Reviews_Log           7.276e-01  1.177e-02  61.830  < 2e-16 ***
## Review_Words_Log               3.556e-01  9.130e-02   3.895 9.82e-05 ***
## Review_Words_No_Stopwords_Log  6.438e-01  1.068e-01   6.028 1.66e-09 ***
## Review_Sentence_Log           -3.857e+00  2.126e-01 -18.140  < 2e-16 ***
## Rev_Dist_Miles_Log             7.517e-02  5.057e-03  14.863  < 2e-16 ***
## Rev_Avg_Sentiment             -1.332e+00  7.623e-02 -17.468  < 2e-16 ***
## Rev_V_Negative                 5.152e-02  4.866e-02   1.059 0.289768
## Rev_V_Positive                 2.841e-03  2.845e-02   0.100 0.920443
## Rev_Tot_Sentiment_Log          3.188e+00  1.819e-01  17.525  < 2e-16 ***
## Yelp_Rest_Order               -3.902e-06  4.739e-05  -0.082 0.934387
## Ratio_Recommended             -1.014e+00  1.099e-01  -9.223  < 2e-16 ***
## Tot_Rest_Reviews_Log          -8.143e-02  9.250e-03  -8.803  < 2e-16 ***
## Tot_Rest_In_City_Log          -5.818e-02  1.355e-02  -4.295 1.75e-05 ***
## Rest_Address_Words             6.596e-03  7.907e-03   0.834 0.404142
## Rest_Name_Words               -1.315e-02  8.947e-03  -1.470 0.141556
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 74558  on 53781  degrees of freedom
## Residual deviance: 51897  on 53759  degrees of freedom
## AIC: 51943
##
## Number of Fisher Scoring iterations: 5
```

l1 regularization (and l2)

```
fitted.results <- predict(model2,newdata=analysis_balanced,type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != analysis_balanced$Recommended)
print(paste('Accuracy',1-misClasificError))
```

```
## [1] "Accuracy 0.771484883418244"
```

```
anova(model2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Recommended
##
## Terms added sequentially (first to last)
##
##
##                               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                          53781      74558
## Profile_Pic_Bool               1   3238.2     53780      71319 < 2.2e-16
## User_Rating                    1    370.7     53779      70949 < 2.2e-16
## Days_Since_7_2004_Log          1      2.7     53778      70946  0.097907
## User_Rating_Diff               1     11.1     53777      70935  0.000861
## Edited_Review_Bool             1      2.5     53776      70933  0.117424
## Friends_Log                    1   5035.8     53775      65897 < 2.2e-16
```

```
## Tot_Photos_Log                 1   4147.6     53774      61749 < 2.2e-16
## User_Tot_Reviews_Log           1   4234.5     53773      57515 < 2.2e-16
## Review_Words_Log               1   4843.7     53772      52671 < 2.2e-16
## Review_Words_No_Stopwords_Log  1      5.1     53771      52666  0.023830
## Review_Sentence_Log            1     22.4     53770      52643 2.230e-06
## Rev_Dist_Miles_Log             1    219.9     53769      52424 < 2.2e-16
## Rev_Avg_Sentiment              1      9.2     53768      52414  0.002375
## Rev_V_Negative                 1     18.0     53767      52396 2.224e-05
## Rev_V_Positive                 1      0.0     53766      52396  0.849520
## Rev_Tot_Sentiment_Log          1    325.8     53765      52070 < 2.2e-16
## Yelp_Rest_Order                1      0.3     53764      52070  0.555218
## Ratio_Recommended              1     62.0     53763      52008 3.349e-15
## Tot_Rest_Reviews_Log           1     91.8     53762      51916 < 2.2e-16
## Tot_Rest_In_City_Log           1     16.7     53761      51900 4.462e-05
## Rest_Address_Words             1      0.6     53760      51899  0.428583
## Rest_Name_Words                1      2.2     53759      51897  0.141470
##
## NULL
## Profile_Pic_Bool              ***
## User_Rating                   ***
## Days_Since_7_2004_Log         .
## User_Rating_Diff              ***
## Edited_Review_Bool
## Friends_Log                   ***
## Tot_Photos_Log                ***
## User_Tot_Reviews_Log          ***
## Review_Words_Log              ***
## Review_Words_No_Stopwords_Log *
## Review_Sentence_Log           ***
## Rev_Dist_Miles_Log            ***
## Rev_Avg_Sentiment             **
## Rev_V_Negative                ***
## Rev_V_Positive
## Rev_Tot_Sentiment_Log         ***
## Yelp_Rest_Order
## Ratio_Recommended             ***
## Tot_Rest_Reviews_Log          ***
## Tot_Rest_In_City_Log          ***
## Rest_Address_Words
## Rest_Name_Words
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(pscl)
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

```
pR2(model)
```

```
##           llh       llhNull            G2      McFadden          r2ML
```

```
## -2.594840e+04 -3.727884e+04  2.266089e+04  3.039376e-01  3.438376e-01
##          r2CU
##  4.584502e-01
```

```
fitted.results <- predict(model,newdata=analysis_balanced,type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != analysis_balanced$Recommended)
print(paste('Accuracy',1-misClasificError))
```

```
## [1] "Accuracy 0.771484883418244"
```

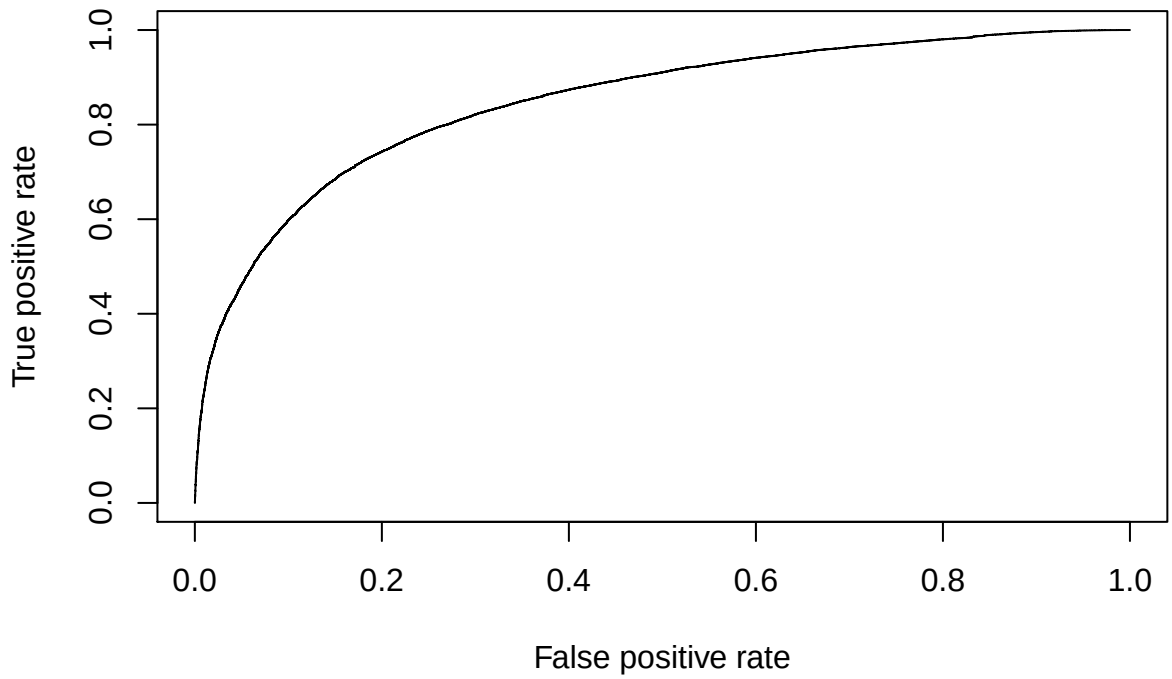
```
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
p <- predict(model, newdata=analysis_balanced, type="response")
pr <- prediction(p, analysis_balanced$Recommended)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
```




```
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
```

```
auc
```

```
## [1] 0.8476807
```

glmnet, regularization

```
require("MASS")
```

```
## Loading required package: MASS
```

```
attach(analysis_balanced)

lda.model = lda(factor(Recommended)~Profile_Pic_Bool + User_Rating + Days_Since_7_2004_Log + User_Ratin
lda.model
```

```
## Call:
## lda(factor(Recommended) ~ Profile_Pic_Bool + User_Rating + Days_Since_7_2004_Log +
##     User_Rating_Diff + Edited_Review_Bool + Friends_Log + Tot_Photos_Log +
##     User_Tot_Reviews_Log + Review_Words_Log + Review_Words_No_Stopwords_Log +
##     Review_Sentence_Log + Rev_Dist_Miles_Log + Rev_Avg_Sentiment +
##     Rev_V_Negative + Rev_V_Positive + Rev_Tot_Sentiment_Log +
##     Yelp_Rest_Order + Ratio_Recommended + Tot_Rest_Reviews_Log +
##     Tot_Rest_In_City_Log + Rest_Address_Words + Rest_Name_Words,
##     data = analysis_balanced)
##
## Prior probabilities of groups:
##   0   1
## 0.5 0.5
##
## Group means:
##   Profile_Pic_Bool User_Rating Days_Since_7_2004_Log User_Rating_Diff
## 0        0.5745417    4.119222              8.204076      0.144341973
## 1        0.8001190    3.971998              8.204644     -0.002882005
##   Edited_Review_Bool Friends_Log Tot_Photos_Log User_Tot_Reviews_Log
## 0         0.02521290    1.304466      0.5226719             1.954569
## 1         0.02919192    2.912021      2.1096964             3.513751
##   Review_Words_Log Review_Words_No_Stopwords_Log Review_Sentence_Log
## 0         3.640520                     3.206761            1.614650
## 1         4.368569                     3.887173            2.046566
##   Rev_Dist_Miles_Log Rev_Avg_Sentiment Rev_V_Negative Rev_V_Positive
## 0           2.802159          2.298472     0.05154141      0.2566286
## 1           3.302444          2.079943     0.09081105      0.2796103
##   Rev_Tot_Sentiment_Log Yelp_Rest_Order Ratio_Recommended
## 0              2.255813        419.7993         0.8297025
## 1              2.658207        419.7993         0.8297025
##   Tot_Rest_Reviews_Log Tot_Rest_In_City_Log Rest_Address_Words
## 0             6.609357             7.962609             7.9109
## 1             6.609357             7.962609             7.9109
##   Rest_Name_Words
## 0        2.696813
## 1        2.696813
##
## Coefficients of linear discriminants:
##                                       LD1
## Profile_Pic_Bool              2.062236e-01
## User_Rating                   6.153856e-02
```

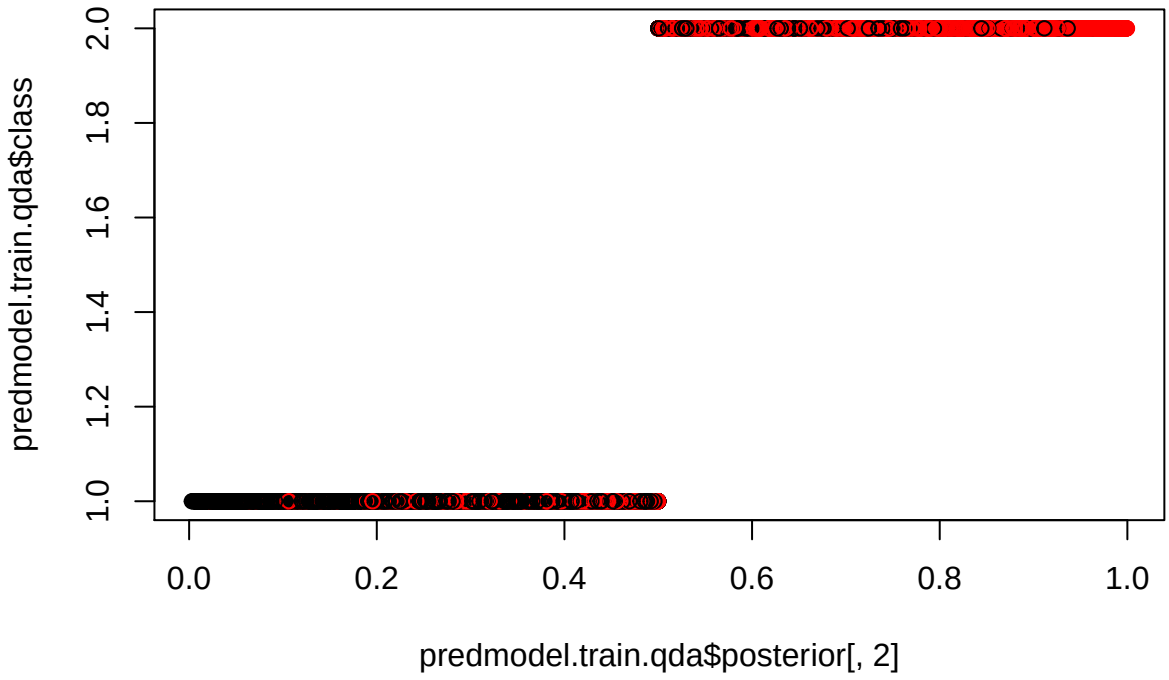
```
## Days_Since_7_2004_Log          5.511388e-01
## User_Rating_Diff              -1.083602e-01
## Edited_Review_Bool            -4.507645e-01
## Friends_Log                    8.474661e-02
## Tot_Photos_Log                 1.894959e-03
## User_Tot_Reviews_Log           5.204469e-01
## Review_Words_Log               2.558694e-01
## Review_Words_No_Stopwords_Log  3.936127e-01
## Review_Sentence_Log           -2.597891e+00
## Rev_Dist_Miles_Log             5.293344e-02
## Rev_Avg_Sentiment             -9.134882e-01
## Rev_V_Negative                 1.265028e-02
## Rev_V_Positive                 9.237084e-03
## Rev_Tot_Sentiment_Log          2.163890e+00
## Yelp_Rest_Order               -1.003287e-05
## Ratio_Recommended             -6.205591e-01
## Tot_Rest_Reviews_Log          -4.974336e-02
## Tot_Rest_In_City_Log          -3.777242e-02
## Rest_Address_Words             3.634780e-03
## Rest_Name_Words               -6.993509e-03
```

https://datascienceplus.com/how-to-perform-logistic-regression-lda-qda-in-r/ https://rpubs.com/ryankelly/LDA-QDA http://uc-r.github.io/discriminant_analysis

```
#Predicting training results.
predmodel.train.qda = predict(lda.model, data=analysis_balanced)
table(Predicted=predmodel.train.qda$class, Recommended=Recommended)
```

```
##          Recommended
## Predicted     0     1
##         0 22270  7721
##         1  4621 19170
```

```
par(mfrow=c(1,1))
plot(predmodel.train.qda$posterior[,2], predmodel.train.qda$class, col=analysis_balanced$Recommended)
```


predmodel.train.qda$class
1.0
1.2
1.4
1.6
1.8
2.0
0.0
0.2
0.4
0.6
0.8
1.0
predmodel.train.qda$posterior[, 2]