MSIT 423

Project Two

Ray Liu

Features and model description

The data set is from a crowd-sourcing website that enables users to submit and discuss ideas to improve the product. After comparing a different kind of classifier, I get the result (AUC) below:

Model	**Contributor + Content**	**All**
Lasso	0.618	0.765
Ridge	0.622	0.749
GAM	0.599	0.792
Tree	0.719	0.911
Random Forest	0.748	0.943
GBM	0.744	0.946

The predictor's analysis

Predictors	**Definition**	**Category**
Pastaccept	Number of ideas accepted in the past	Contributor
commentsC	Number of comments written by the contributor	Contributor
X1-X11	Summary of what is in the text	Content
age	How long the idea has been submitted	Content
month	The month when it was submitted	Content
diversity	How different the idea is from previous ideas	Content
comments	Number of comments written about the idea	Crowd
votes	The number of people who visited the side and voted for implementing the idea	Crowd

Data preview:

2.1 Preliminary analysis:

2.1.1 The profile of the data

When I view the dimension of the data, I found that they are heavily biased. The mean is only 0.8941, indicating that most of the response is 0. It may produce some problem in prediction, accuracy, for instance, can be extremely high even if I predict all they as 0 without any model. So, AUC may be a better criterion to evaluate the results.

```
       y
Min.   :0.00000
1st Qu.:0.00000
Median :0.00000
Mean   :0.08941
3rd Qu.:0.00000
Max.   :1.00000
```

2.1.2 Correlation analysis

According to the correlation matrix, luckily, there is not much correlation between these variables. There is only two combinations of variables have correlation more than 0.6, the age and pastideas (0.84), the pastaccept and pastideas.

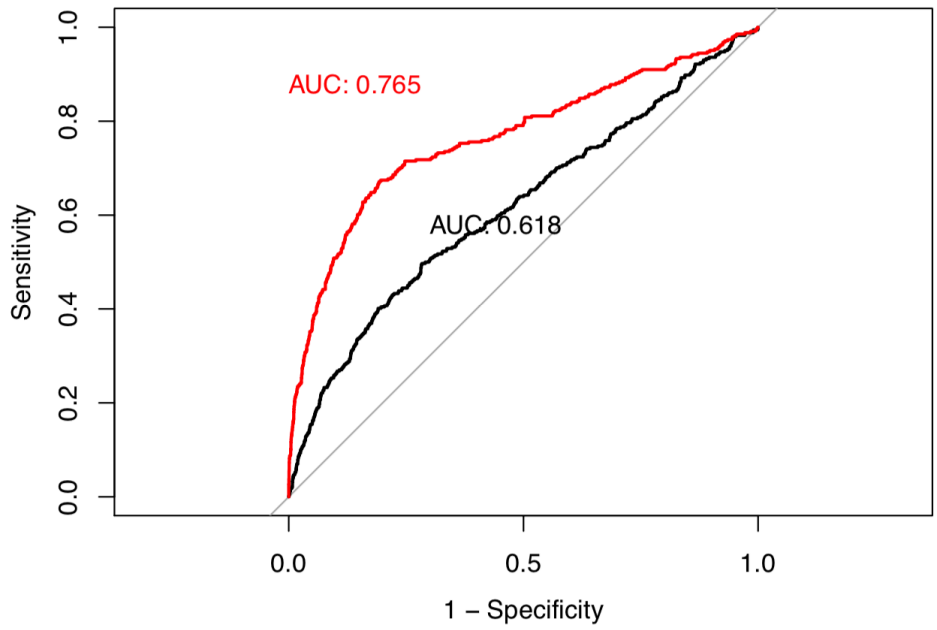
```
##            month diversity pastideas pastaccept commentsC   age votes
## month       1.00      0.02     -0.01       0.01      0.03 -0.01  0.02
## diversity   0.02      1.00     -0.03      -0.02     -0.03 -0.04 -0.01
## pastideas  -0.01     -0.03      1.00       0.64      0.40  0.84  0.01
## pastaccept  0.01     -0.02      0.64       1.00      0.33  0.48  0.04
## commentsC   0.03     -0.03      0.40       0.33      1.00  0.44  0.03
## age        -0.01     -0.04      0.84       0.48      0.44  1.00  0.01
## votes       0.02     -0.01      0.01       0.04      0.03  0.01  1.00
## comments   -0.01     -0.05      0.03       0.05      0.08  0.03  0.42
## X1          0.03      0.42     -0.09      -0.06     -0.04 -0.10 -0.01
## X2         -0.02     -0.16      0.05       0.05      0.03  0.06 -0.04
## X3         -0.01     -0.01     -0.03      -0.03     -0.01 -0.02  0.04
## X4         -0.01     -0.39     -0.01      -0.04      0.00  0.00  0.03
## X5          0.02      0.00      0.00       0.04      0.01  0.00  0.01
## X6          0.00      0.07      0.02       0.03      0.00  0.01  0.00
## X7          0.00     -0.10     -0.02       0.01      0.05  0.00  0.04
## X8          0.02     -0.23     -0.02      -0.02     -0.02 -0.02 -0.05
## X9          0.00     -0.46      0.00       0.02      0.02  0.00 -0.01
## X10        -0.01      0.27      0.01      -0.02     -0.01  0.00  0.01
## X11        -0.01      0.20     -0.03      -0.04     -0.02 -0.03 -0.03
## y          -0.01     -0.02      0.03       0.06      0.09  0.03  0.33
```

Pic2.1.2 Part of the correlation matrix

Classifier Model

3.1 Lasso

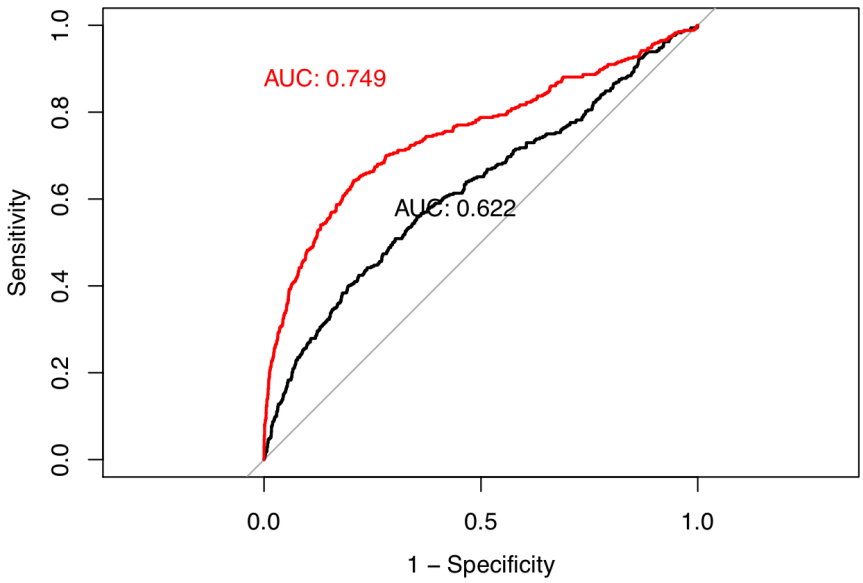
Given the optimal lambda, only X2 has been deleted from the model. We have the MSE of test dataset 0.0864 and 0.0767 for adding the predictors besides contributor and content predictors. When calculating the AUC, we find the crowd predictors increases the AUC by 0.147.



3.2 Ridge

Once again, we use Ridge to find the optimal model in the same way. When only consider contributor and content predictors, we get MSE of 0.0862 while 0.0771 for all the predictors. After calculating AUC, the crowd predictors increases the AUC by 0.127.

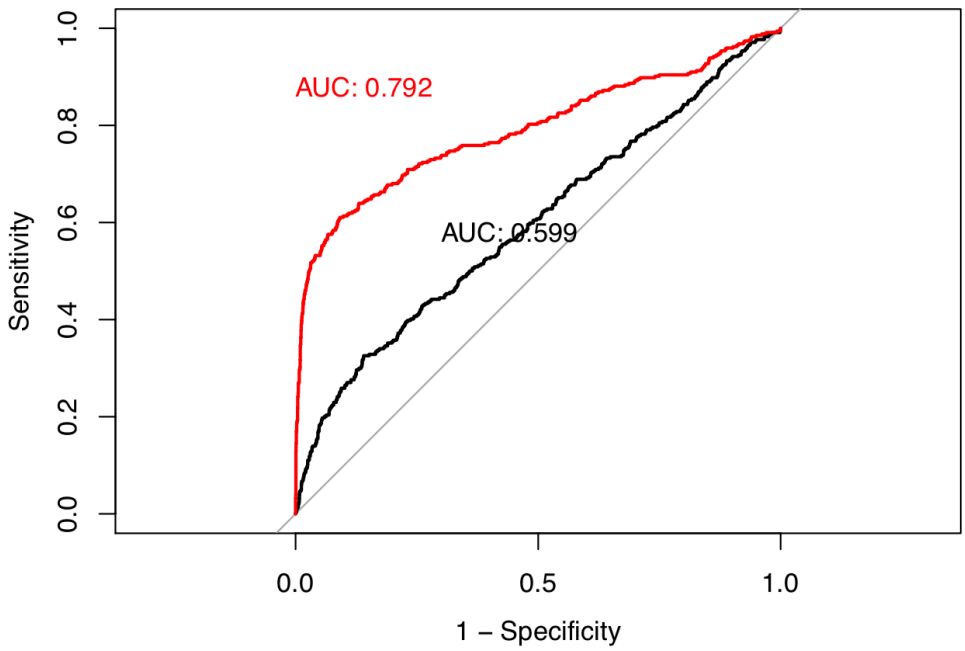
Compare the Ridge and Lasso, we find that Ridge does better when only consider contributor and content predictors. When taking all the predictors into consideration, Lasso does better than Ridge.



3.3 GAM

I tried the general additive model, assuming there is no interaction between the predictors. Through the partial depend on digraph, the votes and comments seem to be significant in a certain range. The digraph has been attached in the appendix.

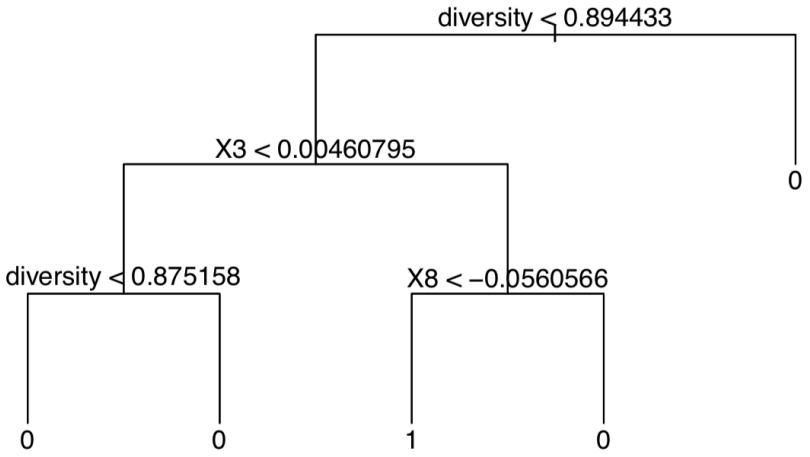
And we find that the AUC of GAM (taking all predictors into consideration) is the best now.



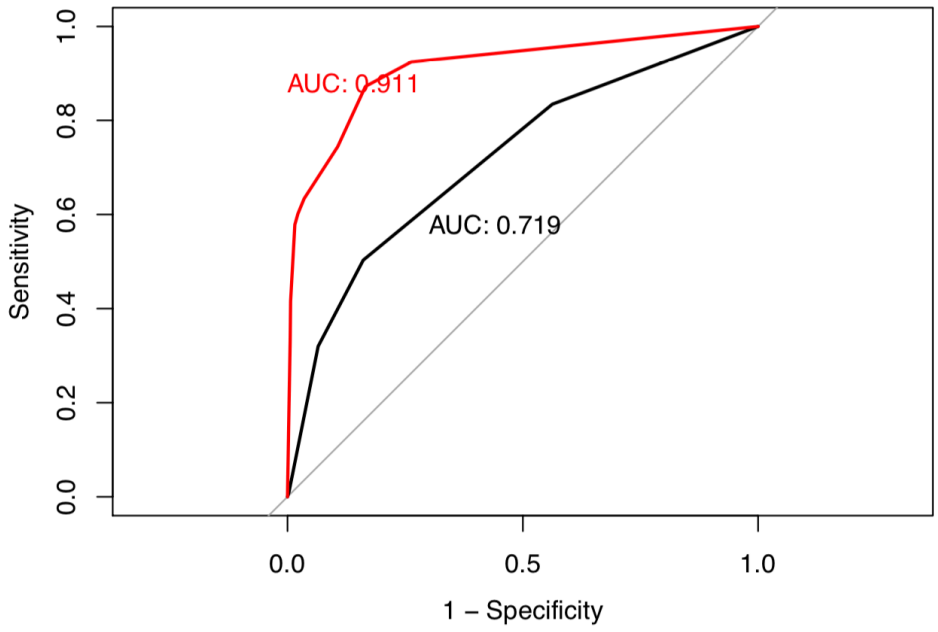
3.4 Tree

As a classifier tree, I make y as a factor in the model.

Through the summary of the tree, we find that some of the branches can be pruned.



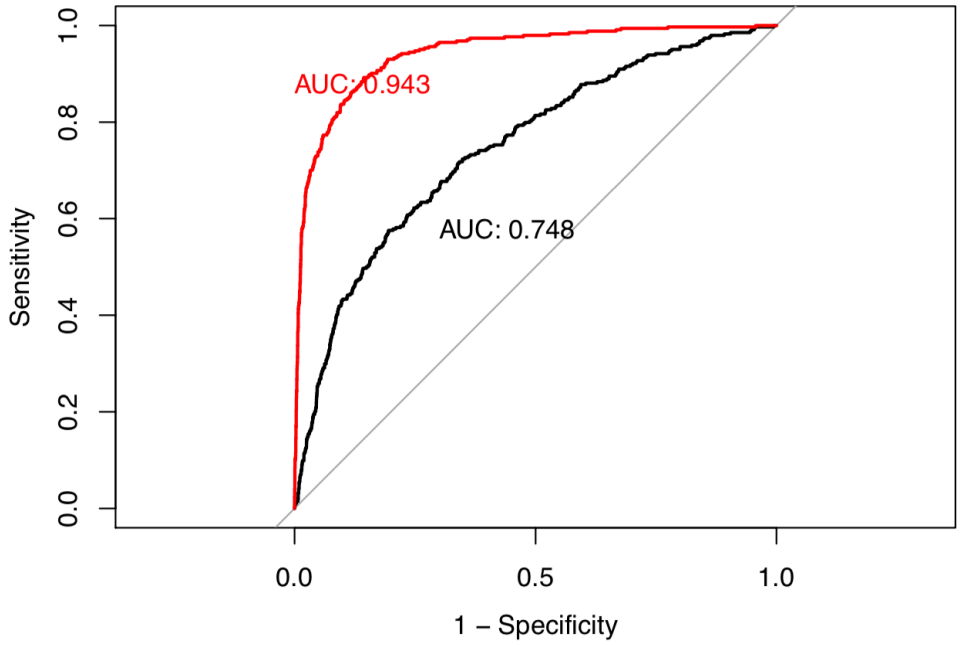
From the AUC plot, we find that the single tree has done a good job.



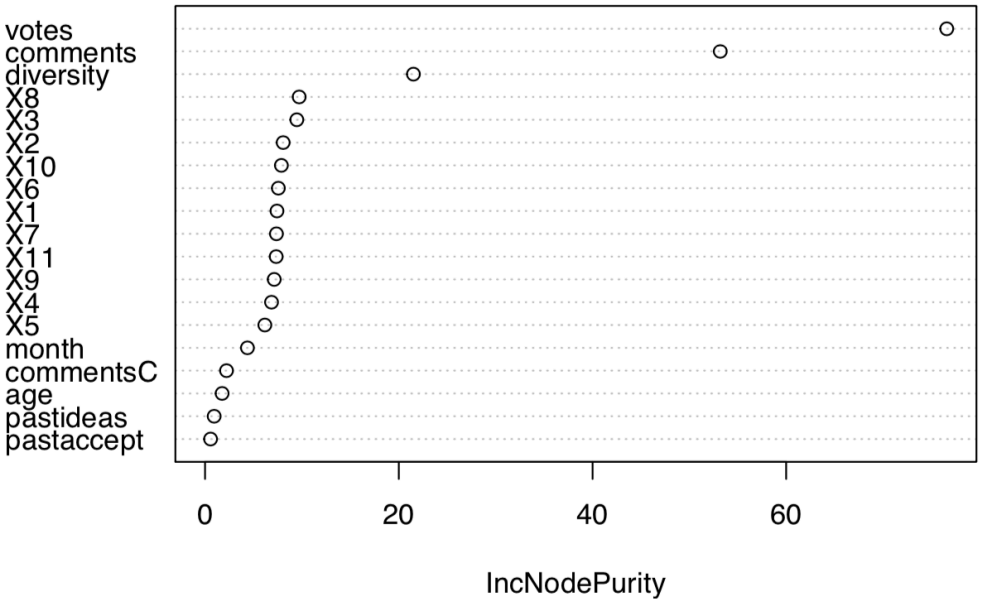
3.5 Random Forest

Then I tried the Random Forest. I have fit the Random forest with the number of tree 500, 1000, 5000, and 10000. Finally, find that 1000 has a good result in the shortest time.

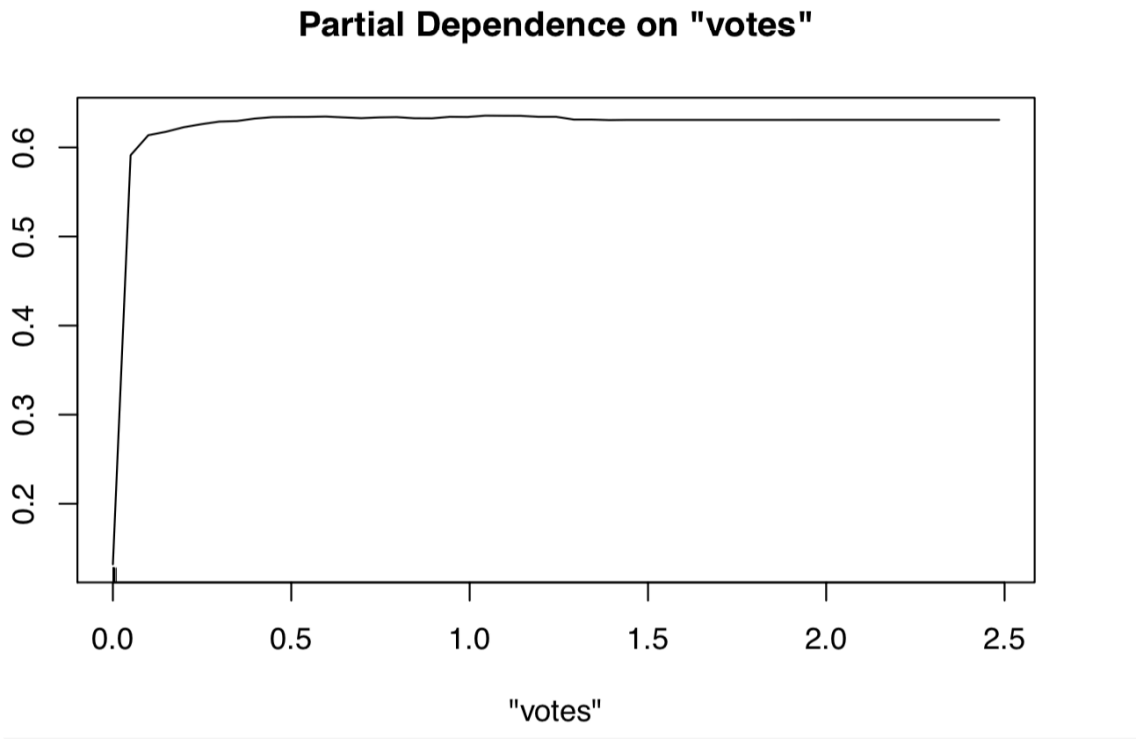
Random Forest has done a pretty good job with an AUC of 0.943.



We can also do former analysis with random forest. The importance of the predictors, for example, can be found from the model.



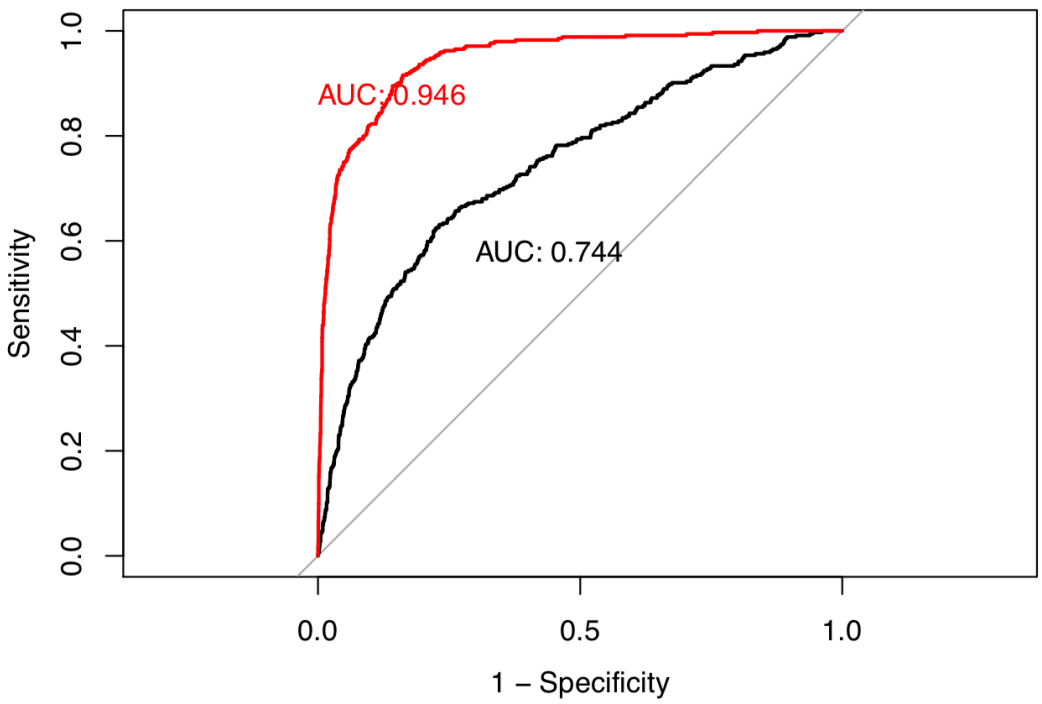
Then we can also do the partial dependence on certain variables. Take the most important predictors votes for instance.



From the graphic, there seems to be a threshold around 0.1, we have to take care of the prediction with a votes number below 0.1, because it can have a huge influence.
Meanwhile, the comments predictor has a threshold at around 0.05, the diversity is stable between 0.1 and 0.88. The graphics are attached in the appendix.

3.6 GBM

When I try to model in GBM, I have tried some different number in certain arguments. Finally find that the interaction depth of 1, tree number of 500, with a shrinkage number of 0.02, are good enough for the boosted tree model.
Finally, I got the best AUC up until now, 0.946 when taking all the predictors into consideration.



When summary the GBM model, we find the same results when we analyze the random forest model that, the votes is the most important predictors. The three most important predictors are votes, comments, and diversity. As the summary below, the relevant influence of forth predictor is much smaller than the top 3.

```
##                  var     rel.inf
## votes          votes 55.04241635
## comments    comments 31.65762565
## diversity  diversity 10.51881399
## X8                X8  0.86415074
## X3                X3  0.74176020
```

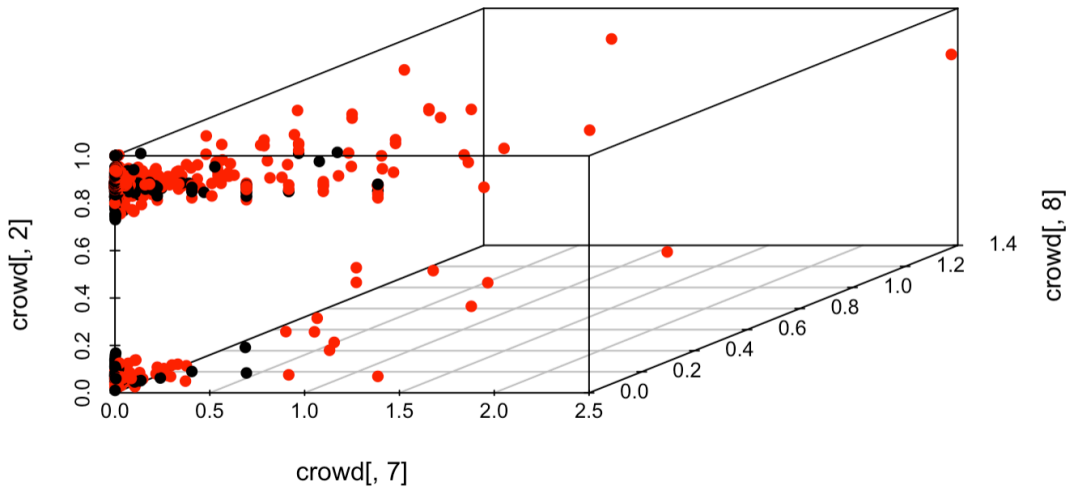
Classifier Plot

In order to have an intuitive digraph, I tried PCA at first. However, according to the summary of PCA, I found that each eigenvector seems to explain a little proportion of the variance (the top 3 totally explain 0.3272 of variance only).

Calling back the correlation matrix mentioned in Section 2.1, there is little correlation among the predictors, PCA cannot help much.

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6
Standard deviation     1.6280 1.4248 1.23950 1.14735 1.11179 1.10424
Proportion of Variance 0.1395 0.1069 0.08086 0.06929 0.06506 0.06418
Cumulative Proportion  0.1395 0.2463 0.32720 0.39649 0.46155 0.52572
                          PC7     PC8    PC9    PC10    PC11    PC12
Standard deviation     1.04668 1.00838 0.9853 0.92964 0.91303 0.85768
Proportion of Variance 0.05766 0.05352 0.0511 0.04549 0.04388 0.03872
Cumulative Proportion  0.58338 0.63690 0.6880 0.73348 0.77736 0.81608
                          PC13   PC14    PC15    PC16    PC17    PC18
Standard deviation     0.84562 0.8236 0.78023 0.75003 0.72515 0.52086
Proportion of Variance 0.03764 0.0357 0.03204 0.02961 0.02768 0.01428
Cumulative Proportion  0.85371 0.8894 0.92145 0.95106 0.97874 0.99302
                          PC19
Standard deviation     0.36426
Proportion of Variance 0.00698
Cumulative Proportion  1.00000
```

As mentioned in Section 3.6, the most important predictors, votes, comments, and diversity have high relevant influence. They can visualize the observations.



From the 3d scatterplot, votes and comments are useful predictors.

Conclusion

1. Random Forest and boosted trees give a pretty good result. We have the best classifier with AUC of 0.946.
2. The most important predictors are votes (rel.inf=55.0424), comments (rel.inf=31.6576), and diversity(10.5188). While the fourth predictor (0.8642) has a long distance from them.
3. Votes and comments are important in classification.