[Features and model description 2](#_Toc9515732)

[Introduction & Analysis 3](#_Toc9515733)

[1.1 Features Initial Analysis and Grouping: 4](#_Toc9515734)

[First Model - City Features 5](#_Toc9515735)

[2.1 Preliminary analysis: 5](#_Toc9515736)

[2.2 Regression Model: 6](#_Toc9515737)

[2.3 Model validation 9](#_Toc9515738)

[Final Model – Crime Added 10](#_Toc9515739)

[3.1 Preliminary analysis: 10](#_Toc9515740)

[3.2 Regression Model: 12](#_Toc9515741)

[3.3 Model validation 14](#_Toc9515742)

[3.4 Model Improvement 15](#_Toc9515743)

[Interpretation 16](#_Toc9515744)

[4.1 Predictor interpretation 16](#_Toc9515745)

[4.2 Correlation interpretation 19](#_Toc9515746)

[Conclusion 21](#_Toc9515747)

MSIT 423

Project Two

*Ray Liu*

# Features and model description

The data set is from a crowd-sourcing website that enables uses to submit and discuss ideas to improve the product. After compare the 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 predictors analysis

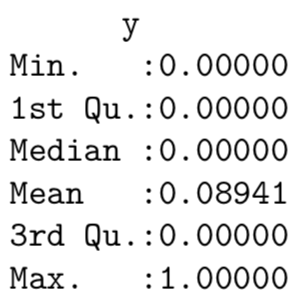
|  |  |  |
| --- | --- | --- |
| **Predictors** | **Definition** | **Category** |
| Pastaccept | Number of ideas accepted in the past | Contributor |
| commentsC | Number of comments written by the contributor | Contributor |
| X1-X11 | Summary of what is in the text | Content |
| age | How long the 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:

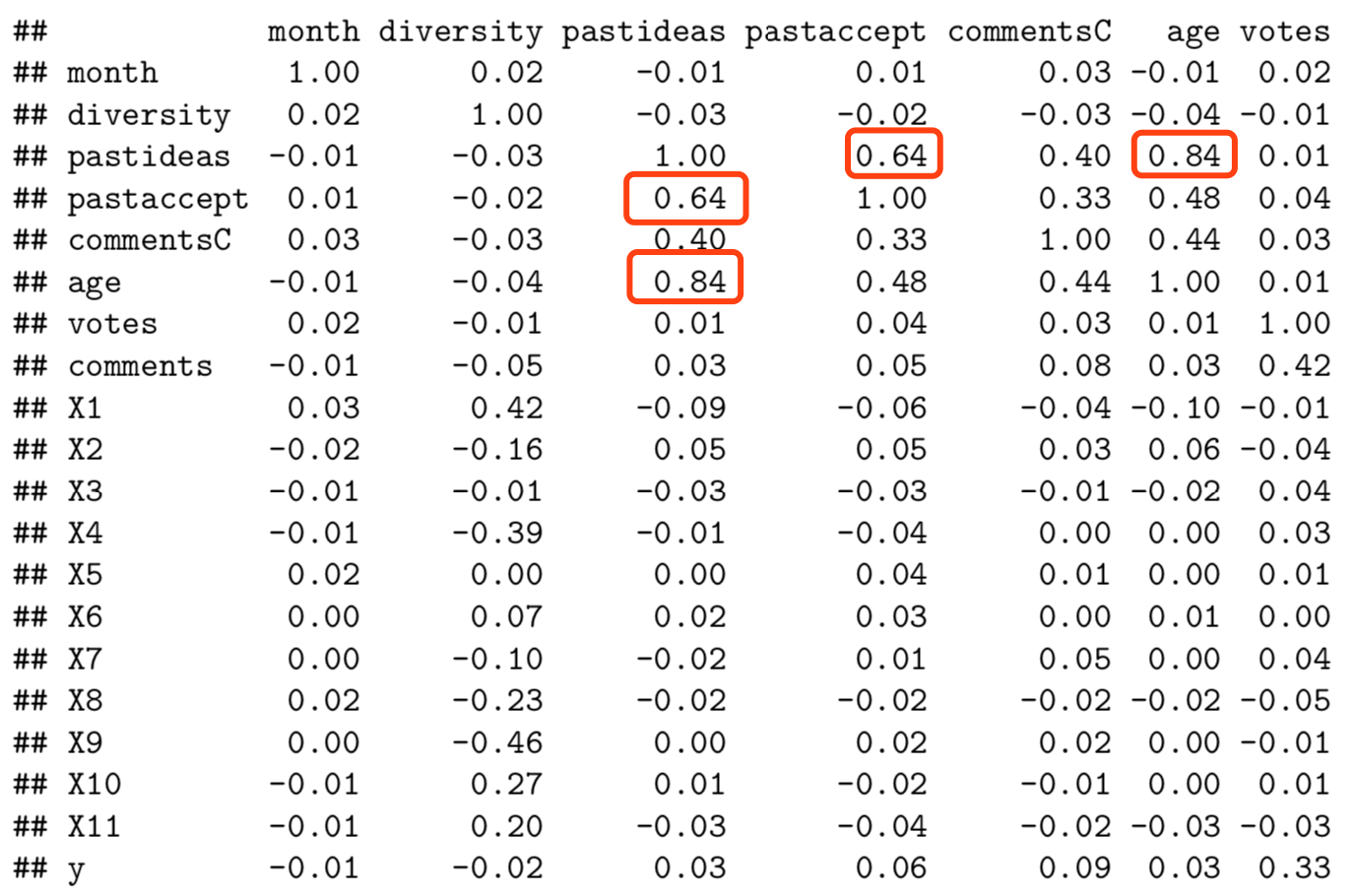
2.1.1 The profile of the data

When I view the dimension of the data, I found that the y is 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 the y as 0 without any model. So, AUC may be a better criteria to evaluate the results.



2.1.2 Correlation analysis

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

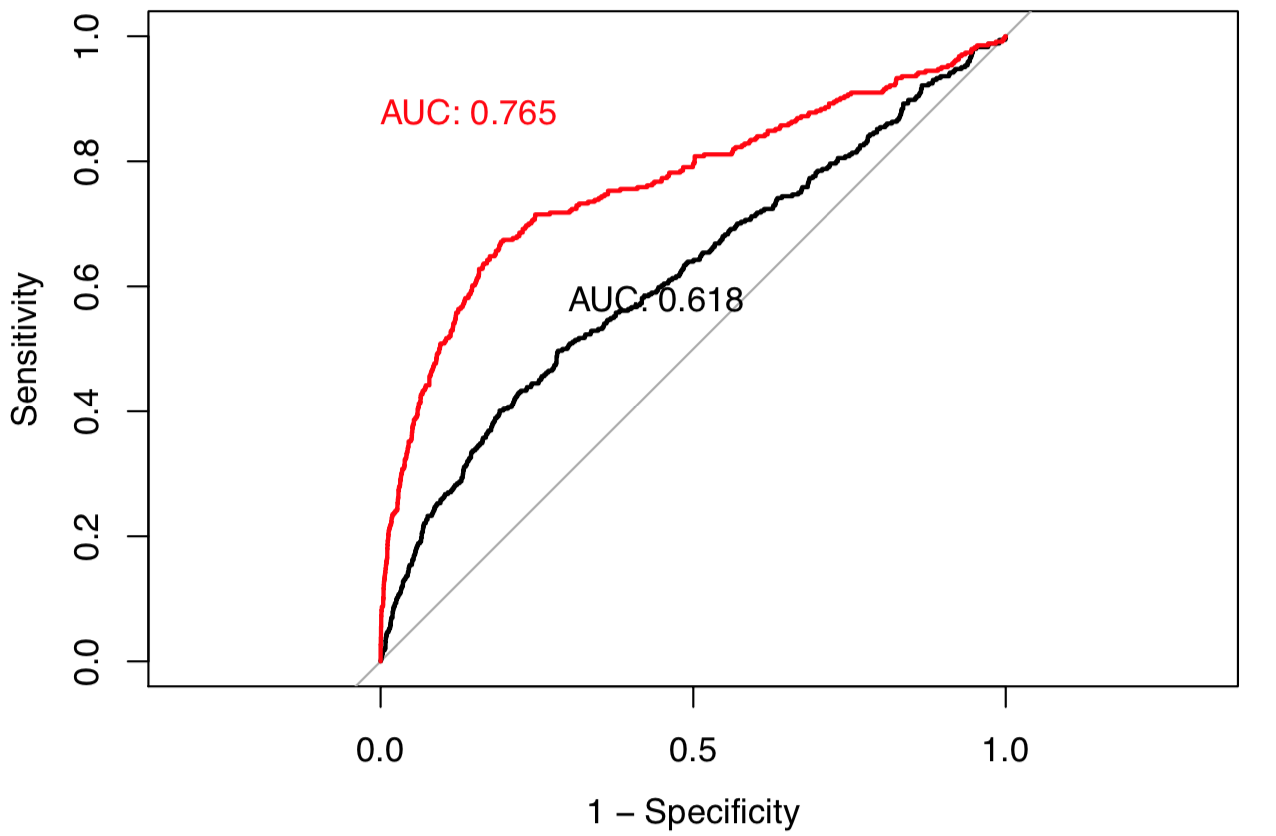


Pic2.1.2 Part of the correlation matrix

# 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 calculate 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.

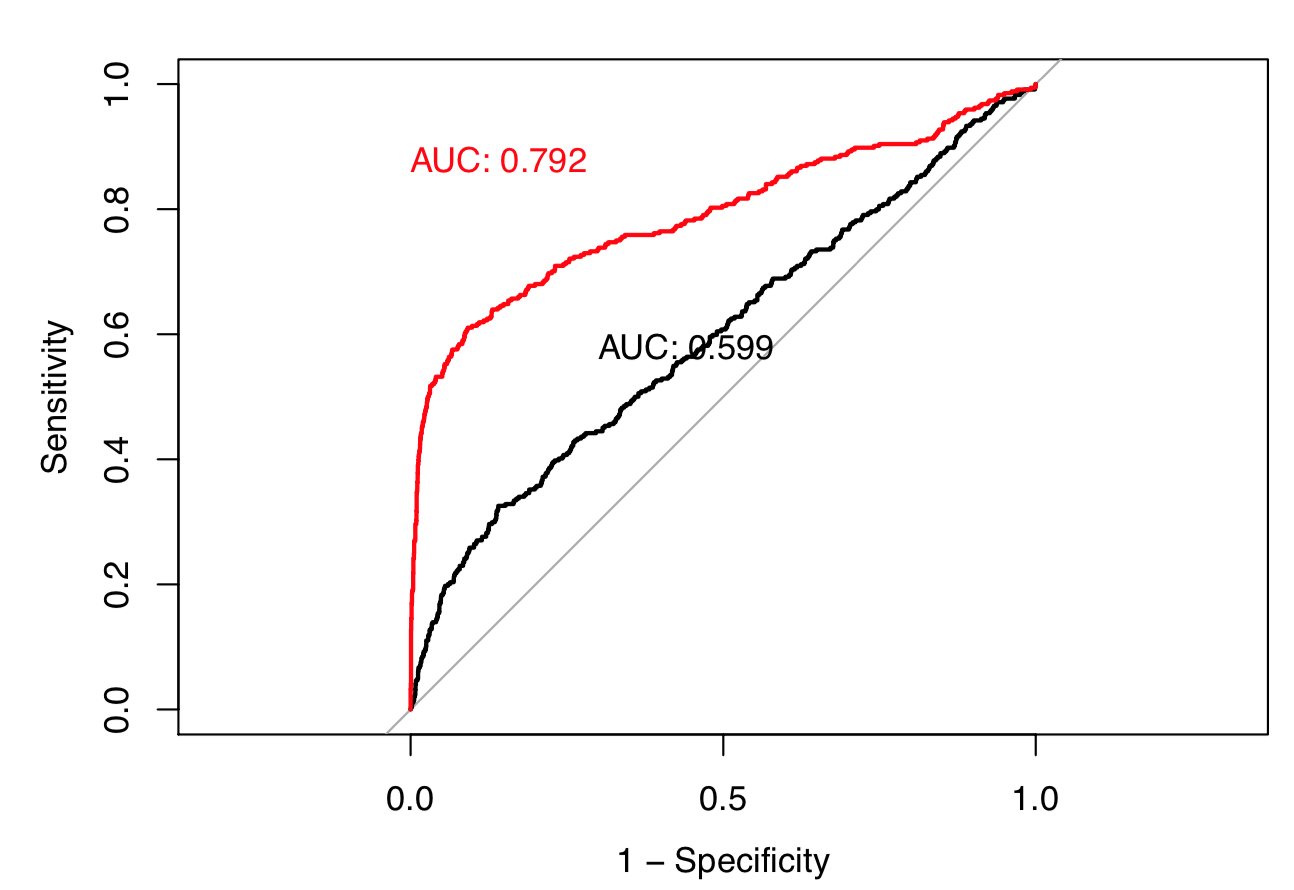
Compare the Ridge and Lasso, we find that Ridge does better when only consider contributor and content predictors. When take all the predictors into consideration, Lasso does better than Ridge.

## 

## 3.3 GAM

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

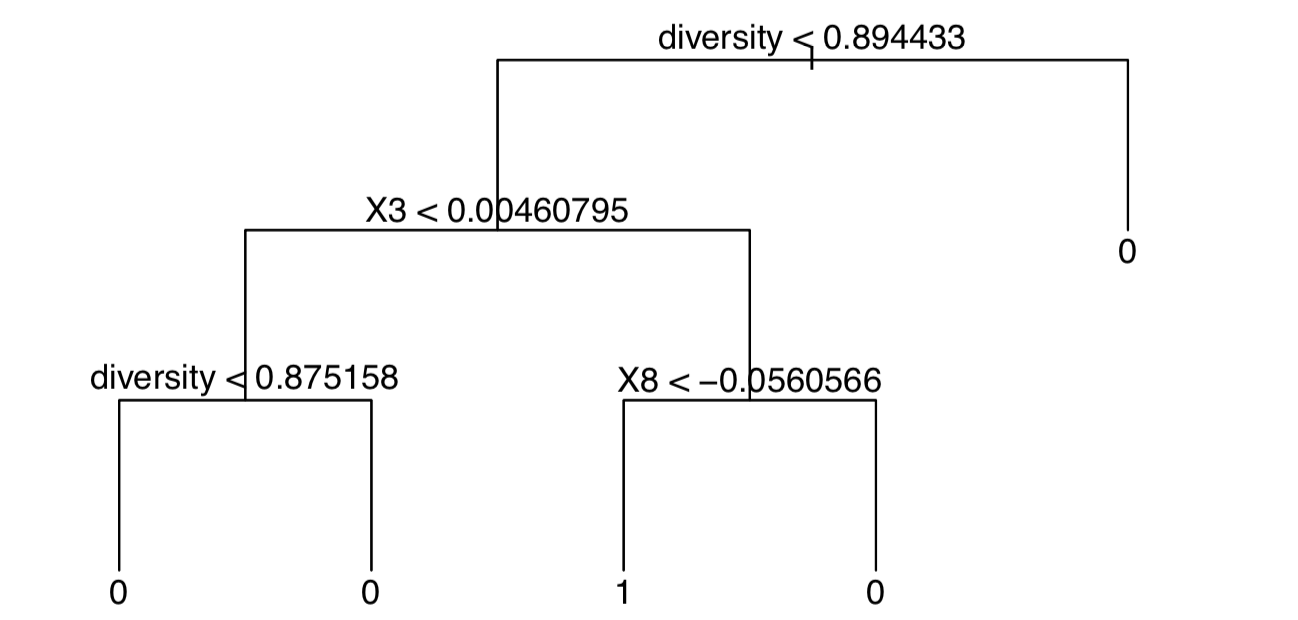
And we find that the AUC of GAM (taking all predictors into consideration) is the best now.



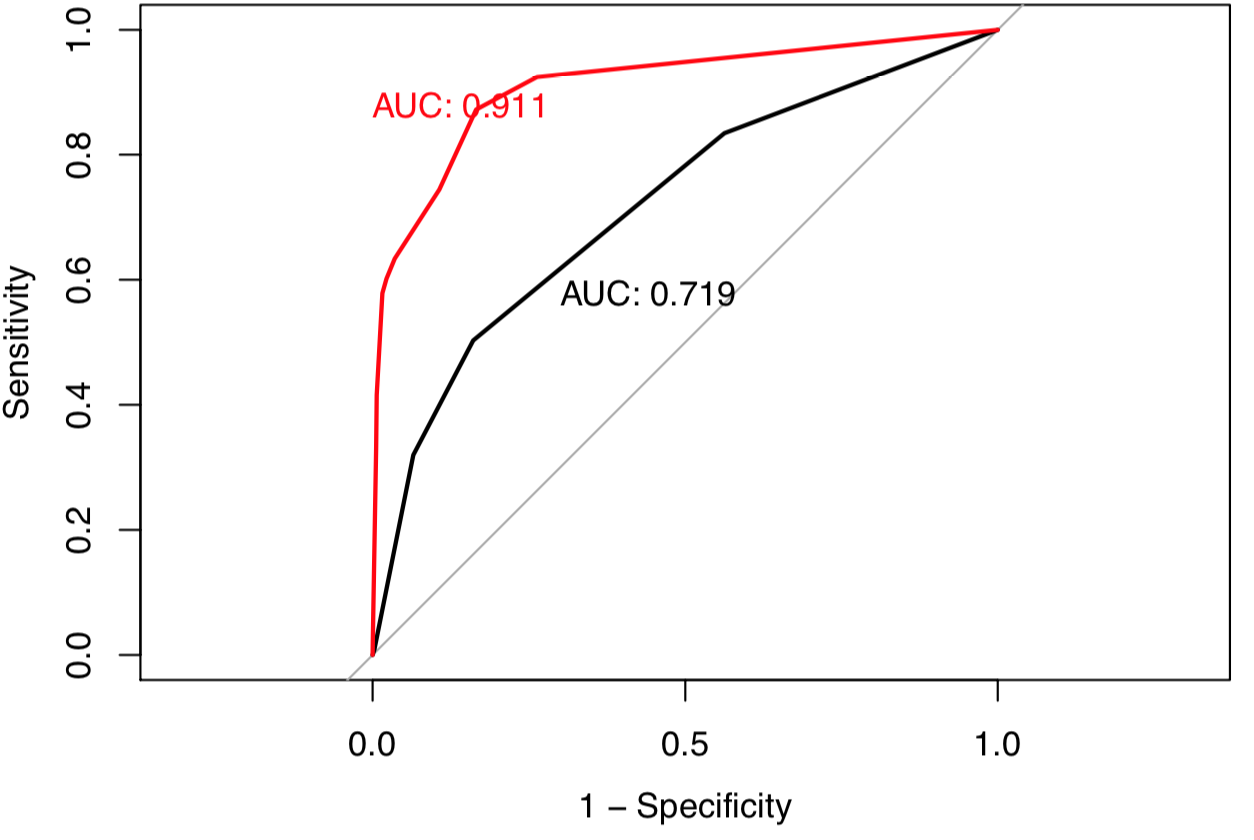
## 3.4 Tree

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 brunch can be pruned.



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



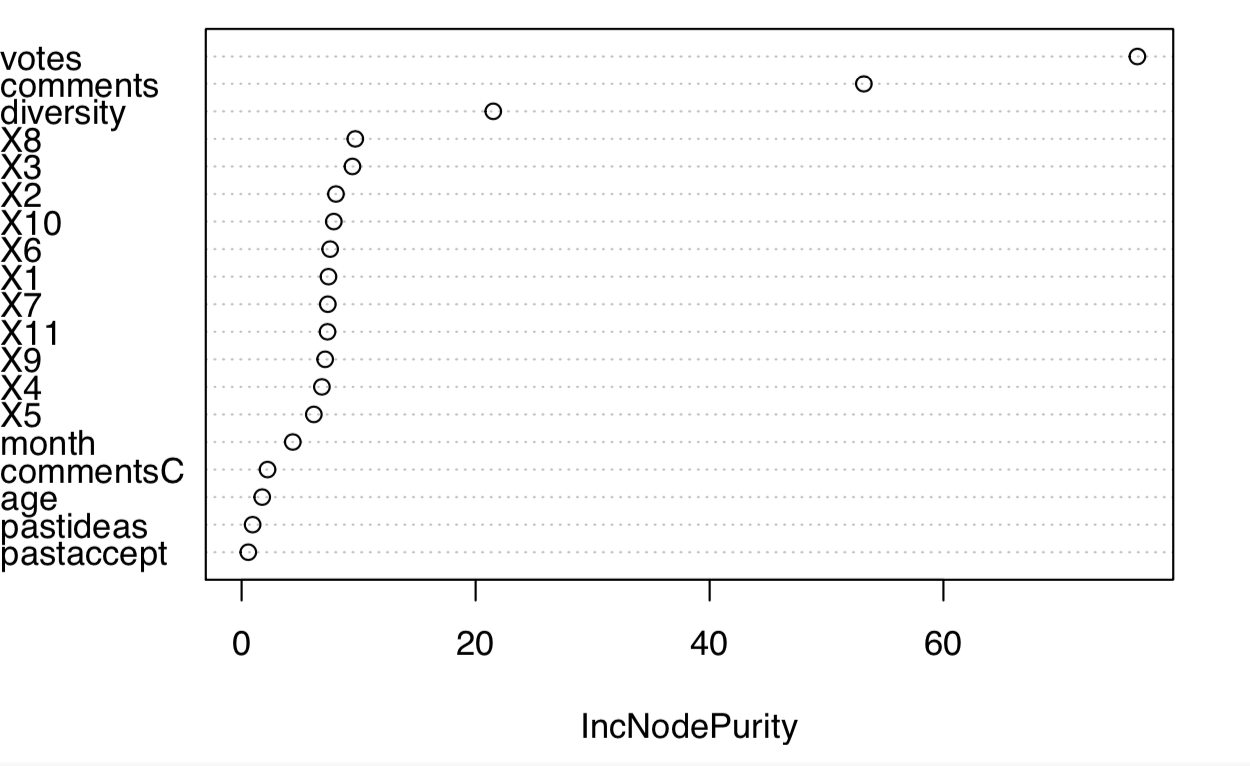
## 3.5 Random Forest

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 the 1000 has the good result in the shortest time.

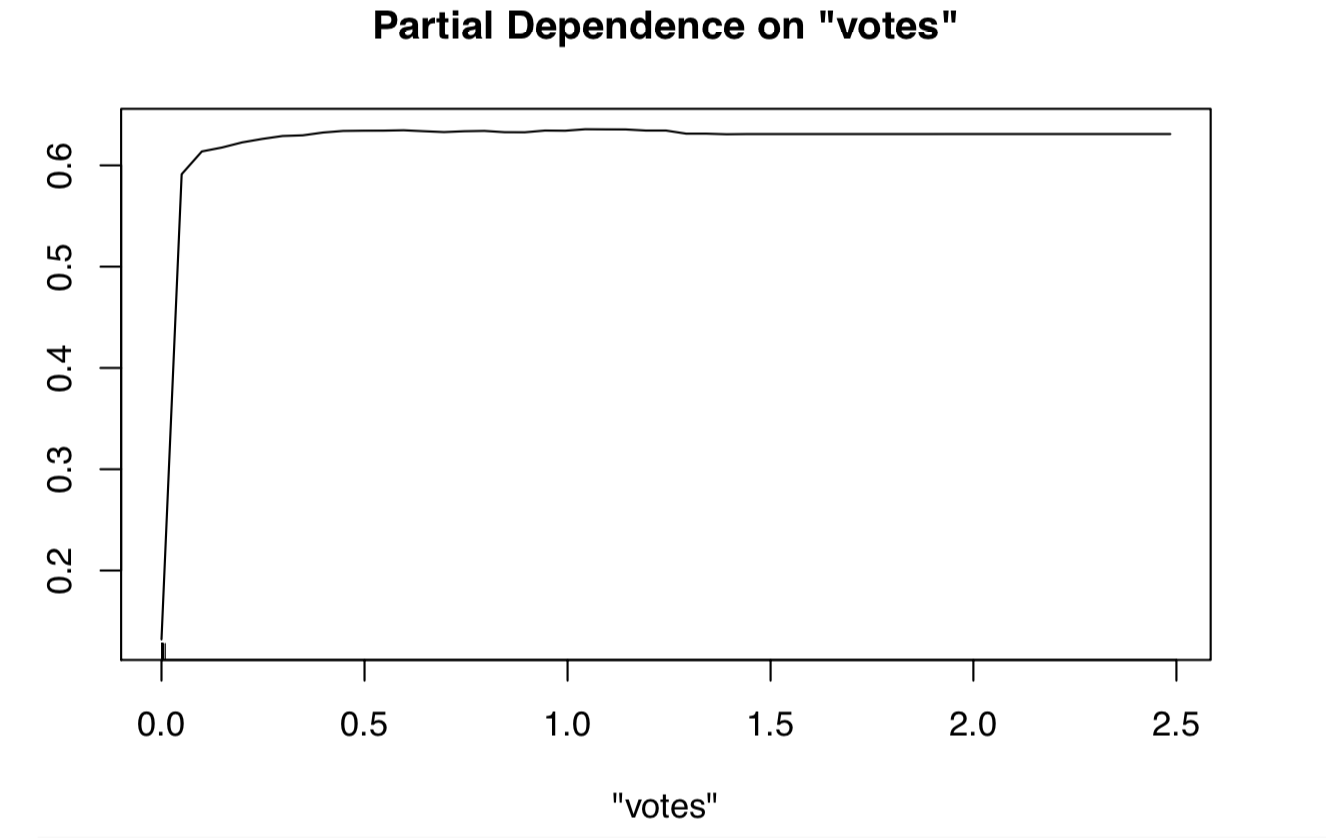
Random Forest has done a pretty good job with an AUC of 0.943.

# 

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



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



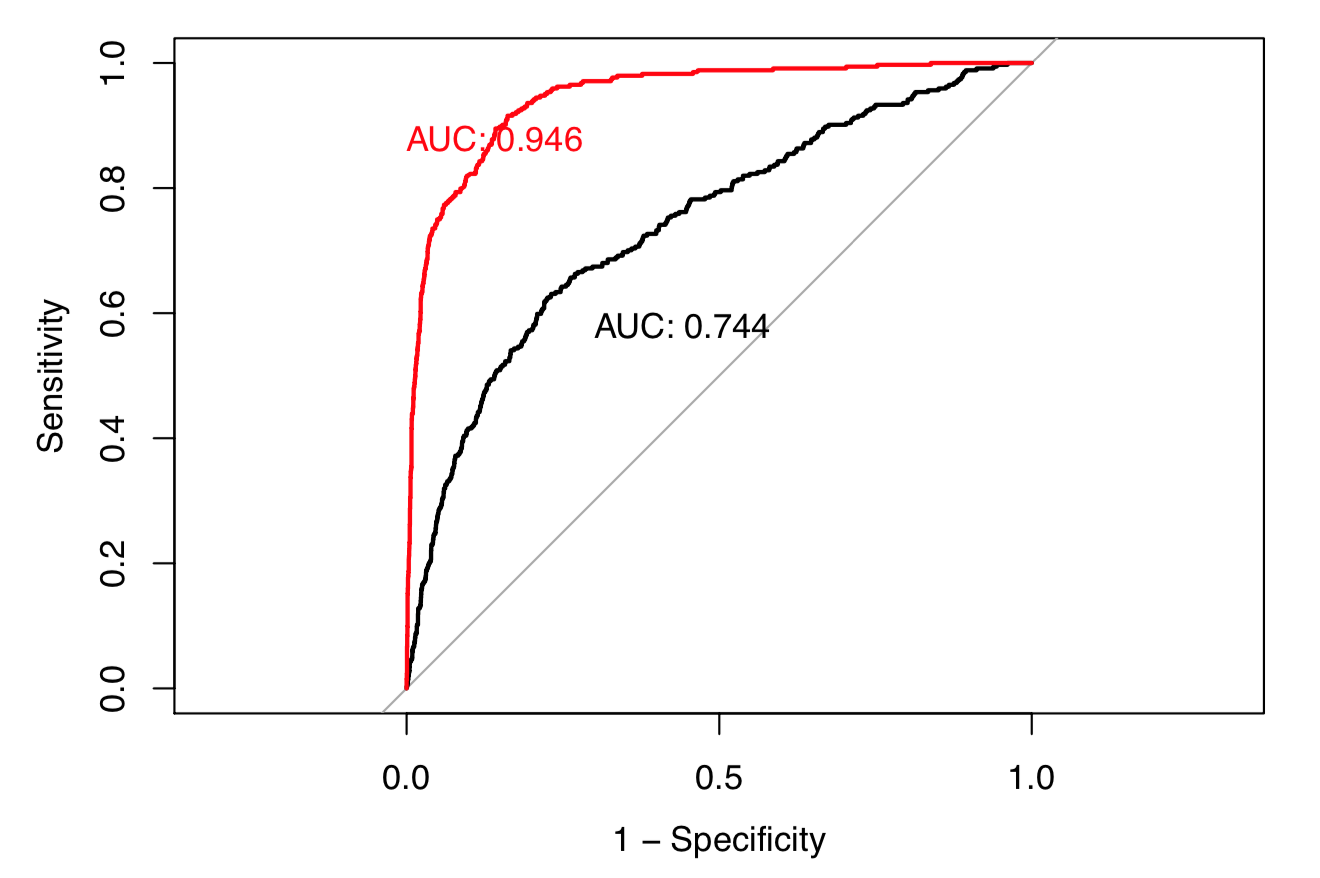
From the graphic, there seems to be a threshold around 0.1, we have to take care of the prediction with a votes number below 0.1, because it can have a huge influence.

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.

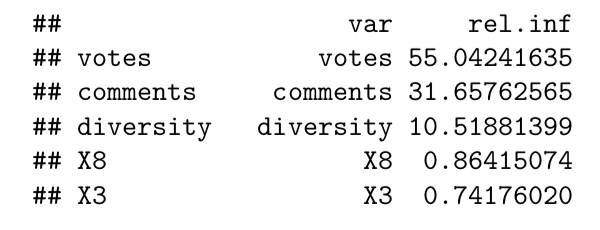
## 3.6 GBM

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.



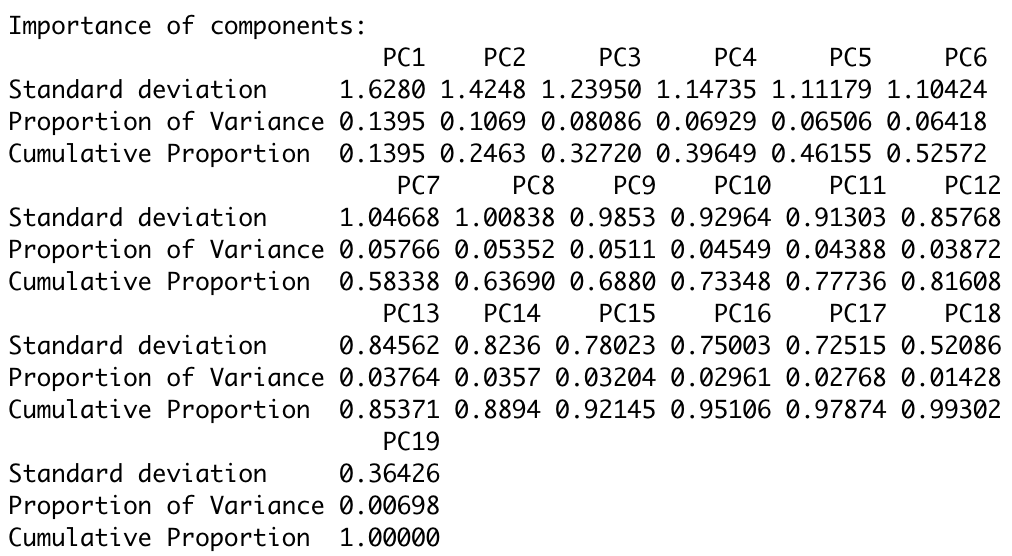
When summary the GBM model, we find the same results when we analyze the random forest model that, the votes is the most important predictors. The three most important predictors are votes, comments, and diversity. As the summary below, the relevant influence of forth predictor is much smaller than the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



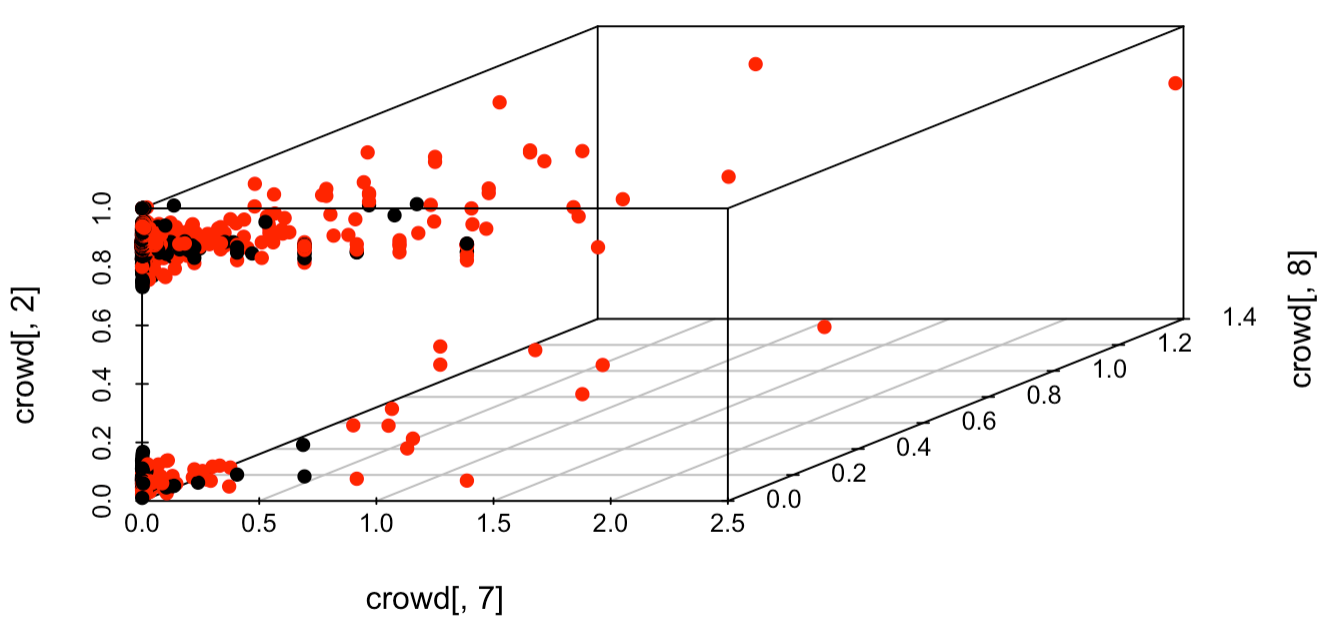
# Classifier Plot

In order to have an intuitive diagraph, I tried PCA at first. However, according to the summary of PCA, I found that each eigenvectors 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.



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.