* How you build model

Firstly, clean the data. The original training data set has a lot of blank and NA figure. There are 406 samples have completed data, other samples have same missing data. I deleted columns with NA and blank to avoid meaningless variables. At least, there are 53 variables in the training data set.

Secondly, predictor creation. Removing zero covariate using function. The result shows there are no more variables with zeroVar or nzv is positive. So there are no selected variable in this process. Principal component analysis is another preprocess. By using pca function, there are 26 variables be processed. But the pca method cannot get good outcome in tree model, so I abandon this preprocess.

Thirdly, preparing data. The original training data is divided in three parts. First step is to divide the whole data into validation and build set. Second step is divided the build data into training set and testing set. The validation set is for the fitting of the parameters.

Fourthly, building model. Classification tree is the simplest model to deal with category prediction. In cart model, I tried to use principal component analysis to decrease influence of unrelated variables, but the preprocess inside the train model didn’t work. Bagging model directly deal with the weak influence variables and the accuracy rate of the model increased a lot. In random forest model, I set the cross validation call in the function. Random forest model has the best in the sample and out of sample rate. Boosting model also performs well in predict testing set. Model Stacking is a method that I used to combining all the previous model to form a new model. After combined all the models or several good performance models together, I found that random forest alone has the best accuracy.

* How you used cross validation

I set the cross validation method in random forest model.

* What you think the expected out of sample error

The accuracy rates of bagging, random forest model and boosting model are high in both in the sample and out of sample. The combined model using gam method performed not good in training data set and testing set. The possible reason may be the tree predict violate the performance. However, when put boosting, bagging and random forest together to get the model, the accuracy rate still lower than fifty percent.

* Why you made the choice

By comparing the accuracy rate, the highest accuracy model is random forest model. So random forest model is the best model to do the prediction.

* Predict 20 test case

B A B A A E D B A A B C B A E E A B B B