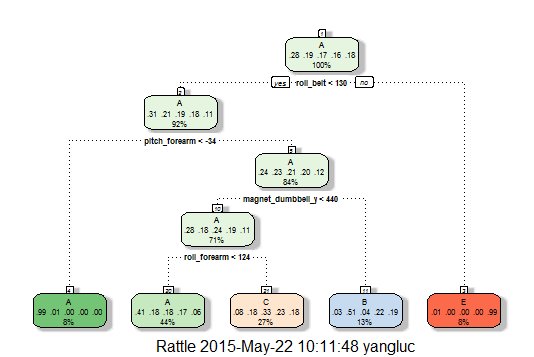
In order to predict classe variable, the first step is to understand the data. Certain variables may have little predictive power. I exclude columns with more than half of the missing values, and I exclude columns of user names, time stamps and new window indicator. There are 52 covariates left.

I build tree model, treebag model, random forest model and gbm model. Firstly, using tree model building as an example. I divide the cleaned pml-training data into 10-fold so that I could use 10-fold cross validation. It is done through “trainControl” function. Then I build the model using “train” function with “rpart” method in caret package. The final model is shown in below graph. 

The accuracy(1-out-of-sample error) is not impressive with this model as shown in the below cross validation results. Almost half of the predictions are likely to be wrong.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Accuracy** | **Kappa** | **Resample** |
| 1 | 0.516556 | 0.3682306 | Fold02 |
| 2 | 0.517839 | 0.3705311 | Fold01 |
| 3 | 0.495923 | 0.338879 | Fold03 |
| 4 | 0.491586 | 0.335856 | Fold06 |
| 5 | 0.488277 | 0.3313197 | Fold05 |
| 6 | 0.494903 | 0.3399763 | Fold04 |
| 7 | 0.512494 | 0.3642522 | Fold07 |
| 8 | 0.522426 | 0.3776729 | Fold10 |
| 9 | 0.533096 | 0.3921132 | Fold09 |
| 10 | 0.490576 | 0.3343398 | Fold08 |

After the tree model, I build treebag model. I also use 10-fold cross validation. The results are shown below. The out-of-sample error is much smaller, around 1-2%.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Accuracy** | **Kappa** | **Resample** |
| 1 | 0.980632 | 0.9754965 | Fold01 |
| 2 | 0.98369 | 0.9793638 | Fold02 |
| 3 | 0.99083 | 0.9884009 | Fold03 |
| 4 | 0.988277 | 0.9851719 | Fold04 |
| 5 | 0.98318 | 0.9787203 | Fold05 |
| 6 | 0.99134 | 0.9890466 | Fold06 |
| 7 | 0.990311 | 0.9877431 | Fold07 |
| 8 | 0.990321 | 0.9877556 | Fold08 |
| 9 | 0.985729 | 0.9819452 | Fold09 |
| 10 | 0.988277 | 0.9851691 | Fold10 |

Then I further build random forest and gbm models. I use 2-fold cross validation for these two models.

Random forest model cross validation results are shown below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Accuracy** | **Kappa** | **Resample** |
| 1 | 0.9926613 | 0.9907166 | Fold1 |
| 2 | 0.9887881 | 0.9858171 | Fold2 |

Gbm model cross validation results are as follows.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Accuracy** | **Kappa** | **Resample** |
| 1 | 0.9625892 | 0.952666 | Fold1 |
| 2 | 0.9596413 | 0.9489313 | Fold2 |

Random forest model achieves less than 1.2% error and gbm model achieves less than 5% error.

From the cross validation performance, random forest model and treebag model are the best. It is dangerous that the model can be over fitting if we do not perform enough out-of-sample test. Treebag model is selected in the end due to more extensive testing it goes through, and it is not as computational intensive as random forest model. The predicted values for the 20 test cases are below.

"B","A","B","A","A","E","D","B","A","A","B","C","B","A","E","E","A","B","B","B"