Prediction_Assignment_Writeup

Andres Caso

15-07-2020

Introduction

Using devices such as Jawbone Up, Nike FuelBand, and Fitbit it is now possible to collect a large amount of data about personal activity relatively inexpensively. These type of devices are part of the quantified self movement – a group of enthusiasts who take measurements about themselves regularly to improve their health, to find patterns in their behavior, or because they are tech geeks. One thing that people regularly do is quantify how much of a particular activity they do, but they rarely quantify how well they do it.

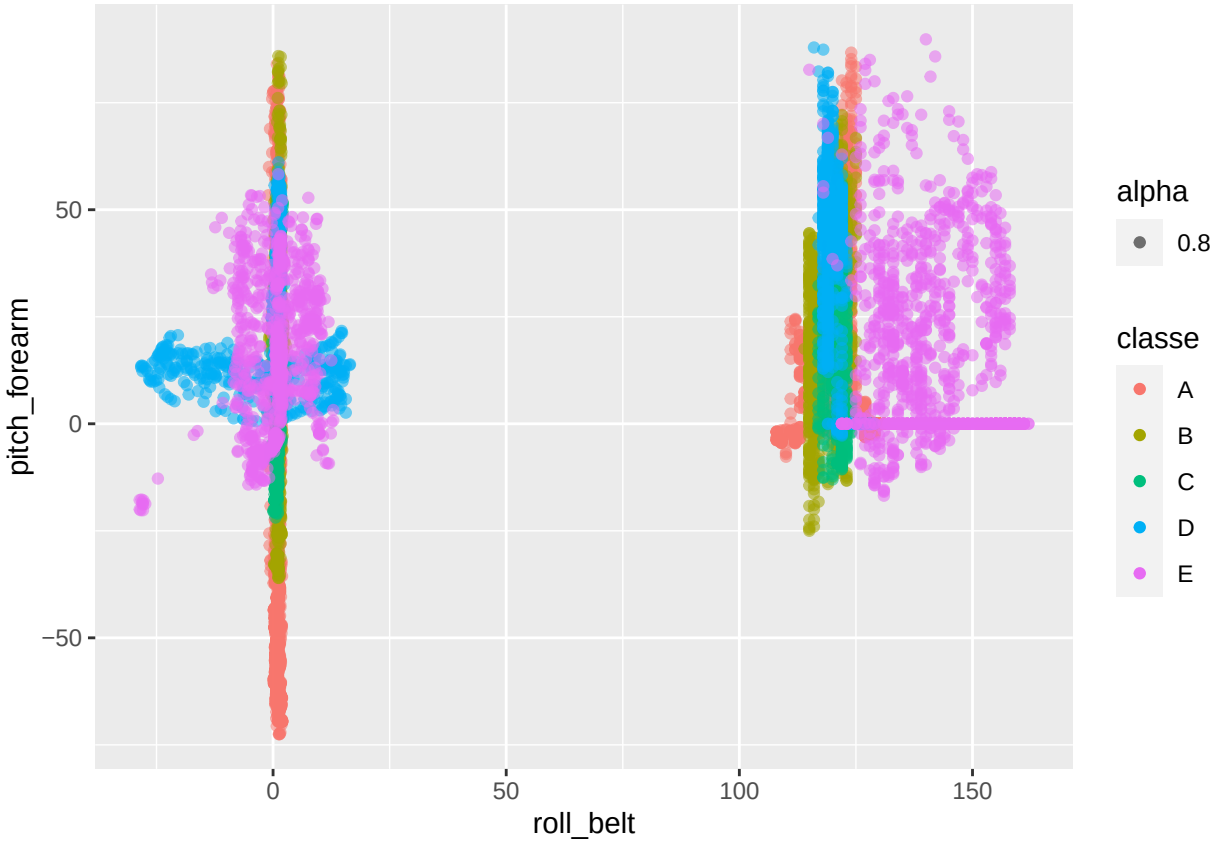
The objective is use data from accelerometers on the belt, forearm, arm, and dumbell of 6 participants. They were asked to perform barbell lifts correctly and incorrectly in 5 different ways

Load and prepare the DATA

1) Checking variables with NA, training set have 67 features with all the values equal to NA, and the validate set have 100 features with all the values equal to NA.
2) Subsetting the data set, for simplicity will be deleted the 100 features of validate data set with all the values equal to NA from both data set, so both data set have a 59 features plus the outcome.
3) Check features with zero variance or near, only new_window have almost zero variance, this feature don't contribute probably to the model, therefore is deleted.
4) Change to factor the outcome variable classe and the user_name variable
5) Change to numeric to some predictors.
6) Delete time features because they not important.
7) Create a Training and testing data set, split by 60% - 40%

Exploratory analisys in the training data set

1) Searching patterns in the predictors roll_belt and pitch_forearm, it's apreciate the difference of the outcome "classe" separetion



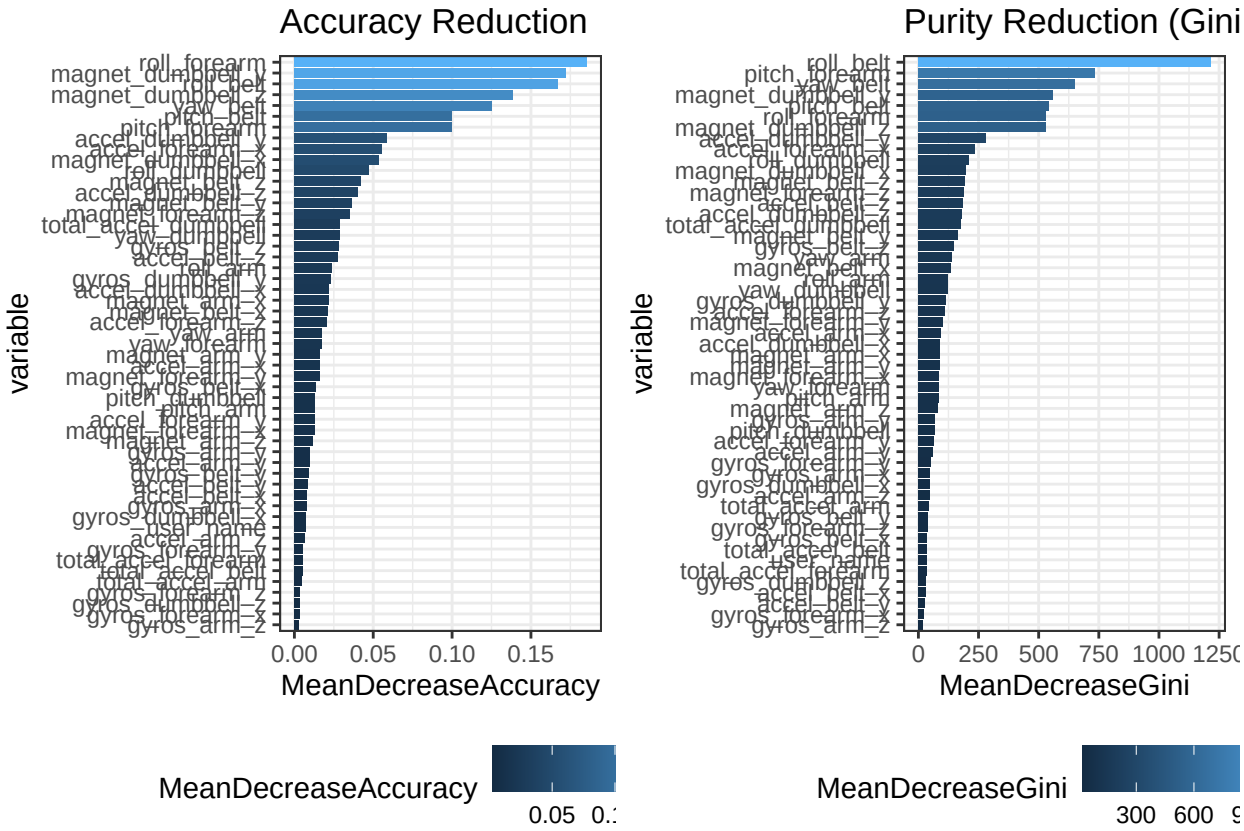
Create Models, since it is a classification problem, will be searched the best model between: Random Forest, Decision Tree and K-mean.

1) Random Forest

```
## Accuracy
## 0.989039
```

After running the model, it can see the accuracy (>99%) of prediction of the model with mtry = 29 in the testing data set.

Now look the importance of the predictors, to confirm the exploratory analisys, roll_belt, roll_forearm and pitch_forearm are very important predictors.



2) CART (Decision tree)

```
##  Accuracy
## 0.4942646
```

After running the model, it can see the accuracy (>47%) of prediction of the model in the testing data set.

3) K-mean

```
##  Accuracy
## 0.8912822
```

After running the model, it can see the accuracy (>89%) of prediction of the K means model in the testing data set.

Conclusion

The Random Forest model it's for lnger the most accuracy model, and predict almost all the case right, > 99% of accuracy.