1. Split data

* 30% - validation data, 70% - build data
* 30% of build data - testing set, 70% of build data - training set

1. I explored the data with the summary function and found out that there are 24 variables with the class "factor" which are really numeric. So I changed data type of theese 24 variables from factor into numeric.
2. So, there are only 9 factor variables, and no one of them influence the outcome (as we can see in cross tables). Most of them contain a lot of NA values.
3. User name, raw\_timestamp\_part\_1 and raw\_timestamp\_part\_2 are not important for a model.
4. Then I excludes near zero variables using nearZeroVar function.
5. I also excluded variables which had a lot of NA values (recognising them with summary function).
6. After this preprocessing I only had 55 possible predictors, all of them were numeric or integer. So I made plots and excluded the variables which evidently didn't influence the outcome. Now I had 20 predictors to build a model with.
7. I made several models with 20 predictors. In all models I preprocessed predictors with pca, setting the threshold to 80%.

I had different accuracy in different models:

random forest - 0.92,

lda - 0.85,

gbm - 0.73,

nb - 0.52.

1. The first idea was to make a model ensembling with all the models.

The accuracy on the test set differed depending on the model of ensembling:

rf - 0.92

lda - 0.92

gbm - 0.92

gam - 0.45

nb - 0.88

But the accuracy was not bigger than using only rf model.

1. Then I tried to make ensembling with the first two models (rf and lda), excluding gbm and nb. But the accuracy of ensembling model never was better than the only rf.
2. So, I decided that the better model for this data was rf with the 20 predictors.
3. I predicted with this model values from my validation set and the accuracy was about 0.929. I expect this accuracy in the case of a new prediction.