Challenge B VF

BARRY et DUFOULEUR

08/12/2017

GITHUB : https://github.com/Fatoumat/Challenge-B---Barry-Dufouleur

Step 9 : Convert all character variables to factors

TASK 1 B

STEP 1 : Random forest technique is a machin learning algorithm, which is efficient to find the relationship between dependent and independent variables. Random Forest classify independent variables according to their relationship with the dependent variable.

STEP 3 : First, to make our codes work because there were difference between the 2 files, we had to "clean" the test file, to make them similar. So we have remove some observations that contain NA and we have remove variables. Because of a level problem, we ensure that all the variables have the same level in these 2 samples. Finally there were a problem, still remaining, of level with the variable extercond, indeed its level is different between the 2 samples. In test there's additionnaly the level PO. Because this variable don't seems to be important we get rid of this problem by deleting this variable in the test and the train sample, for all this exercise.

By comparing the 2 predictions we see that the values given by the prediction with random forest are positive instead of those find with ols prediction. Additionnaly the values predicted in OLS are very large (in the negative sens) as opposed to those of the random forest.

TASK 2 B

STEP 3 (graph on R) : The high flexibility local linear model is the model which predictions are the more variable. The predictions which have the least biais are the low flexibility local linear model predictions. Even if ll.fit.highflexi contains all training observations, we prefer to use ll.fit.lowd=flexi with which observations (of all the data) have more chance to be close to the predictions of the low flexibility model.

STEP 5 (graph on R) ; We have the same comment.The high flexibility local linear model is the model which predictions are the more variable. The predictions which have the least biais are the low flexibility local linear model predictions.

Step 10 (graph on R) : The error rate of the local linear regression on the training set increases with the bandwidth. Then, higher is the bandwidth, less good are the predictions of this model (there is more errors). But we do not want the smallest MSE of the trainig set, because it would mean that the model learns to much the observations of the training data. In other words, the model learns the "exceptions of the data" and it has no the ability to generalize.

We can see that the error rate of the local linear regression on the testing set is a convex function. So, the MSE decreases then increases with the bandwidth. Then, there is a bandwidwth from which the MSE increases, from which the model makes less good predictions. It means that the model is too based on the training set.

Thus we have to find a bandwidth such that the MSE of the training set is low enough and the MSE of the testing set is high enough.

TASK 3 :

We need less than 10 min to run this task. Unfortunately because we have to run a lot of "mini" codes to extract the observations of the data SIREN, system.time juste give us information for code run individually, so it's not significatif.

Step 2 :

In the table we have the number of organizations that has nominated a CNIL per department. With the corresponding numbers in the last column.

Step 3 :

Because the data were to large for my computer, it litteraly bug my ordinateur, i have imported 1 000 000 of observations. They are in datachunk and datachunk0. I will then continue to work only on 500 000 observations from the SIREN data. We have clean it, deleted the duplicated and keep the most recent observations of the multiple SIREN. For the CNIL data, we have kept all the observations . Then, we have merged the 2 tables, by using the number of siren as indicator

Step 4 :

To do this step at the end of the code in the step 3, we have removed the old duplications observations that we have for each SIREN. The histogrammes represent the size of the entreprises that have nominated a CNIL