Take Home Test from Kreditech

Rostam Golesorkhtabar

April 1, 2016

Firstly, I loaded the data into R. Since the numbers are stored with comma instead of dot, I used read.csv2 command.

- **Step 1: Data cleaning**

  After loading the data into R, I examined the data in order to clean it.

  1. Removing duplicates using unique command.
  2. Removing NA's

     For this data set, I removed the NA's value. Of course there is more advance methods to deal with missing data such as multiple imputation, However it requires more depth knowledge about data. for instance variable V39 shows some outliers. It is not clear whether this variable can take that large values or they are typo in input of data. This leads to error when I tried to impute missing value using function mice from the package "mice". Therefore, I did not include these commands in my script.

- **Step 2: Feature selection**

  Before using any machine learning algorithm, similar features should be found and if the correlation between two features are more than 75%, then one of them must be removed. Using corrplot function (image corr_before.pdf), I found out that one variable from each of these pairs variables (V41-V20), (V18-V39),(V35-V37) and (V7-V2) must be removed. See the plot corr_after.pdf, the correlation plot after removing V41, V18, V35 and V7.

  I cleaned the validation data set also.

- **Step 3: Train the algorithm**

  I checked to algorithms lda (Linear Discriminant Analysis) and rf (random forest). The plot of ROC curves of both algorithms are provided with this file. The AUC (Area Under Curve) can be seen in the plots. Also, I checked Knn (K-nearest neighbor algorithm) but the accuracy was not good enough (about 65%), so I ignored it.

  Since the result of lda and rf are similarly good, any of them could be picked up.