HW08 REPORT

In this homework, you developed a machine learning solution in R for six real-life classification problems from finance industry. My machine learning algorithm predicted whether a credit card customer will take specific actions (six different scenarios named as TARGET\_1, TARGET\_2, TARGET\_3, TARGET\_4, TARGET\_5, and TARGET\_6) using the information given about each customer.

As a first step, I imported AUC, onehot, xgboost and caret libraries. I read the data from csv file and named them as X\_train, Y\_train and X\_test. Since both X\_train and X\_test dataset consist categorical values, to convert them into numerical ones, I applied one hot encoding both dataset and named them as X\_train\_d and X\_test\_d respectively. I created test\_predictions matrix, which consists zeros initially and column names as Y\_train, to put predicted posterior probabilities inside it according to corresponding ID and TARGET columns. I put corresponding IDs inside ID column in test\_predictions matrix from X\_test.

I created a for loop since I need to do predictions for 6 targets. I used xgboost as a model. I tried MLP Classifier, Decision Tree Classifier, Random Forest Classifier as a model, but, I got the best result from xgboost when I set its parameters nicely.

Before training my model, since there are invalid customers (in Y\_train, it corresponds to nan values), I took only non-nan values indices and worked on these customers.

As a parameter, I set trControl parameter as xgb\_control, set method “cv” (cross validation), and n=10 (do cross validation 10 times) inside xgb\_control. Inside xgboost, I set method “xgbTree” since I am going to do tree classification, set nrounds=30. As nround parameter, I tried 10, 30, 50, 100 and 300, and 30 gave the best result and 300 overfitted.

As a final step, I trained my model and made predictions on X\_test dataset. After training my model, I calculated AUROC values as 0.9117322, 0.9654076, 0.9234138, 0.9406549, 0.8893689, 0.8857272 respectively. I wrote my posterior probability predictions to a single csv file.