HW07 REPORT

In this homework, I developed a machine learning solution in Python for three real-life classification problems from finance industry. My machine learning algorithm predicted whether a customer will delay his/her credit card bill payment more than 1 day (named as target1), more than 31 days (named as target2), or more than 61 days (named as target3) using the information given about each customer.

As a first step, I read the data as a X\_train for train data, Y\_train for labels and X\_test for test data. I dropped ID columns from X\_train and X\_test. Since given dataset is imbalanced dataset, there was a more label corresponding to 0 compared to 1. Since my model predicts positively labeled dataset, it would be hard to classified because of lack of positively labelled data. So, to increased positively labelled data, I duplicated positively labelled data in X\_train.

After that, ratio between positively labelled data and negatively labelled data became 1 approximately. Now, I had a balanced dataset to train my model. Since X\_train and Y train consists categorical values and not available (NA) values, first, I replaced NA values with dataset’s mean, second, I did one hot encoding to convert categorical values to numerical ones.

After preprocessing my data, using train\_test\_split function from sklearn package, I divided my training set into two pieces (training set and validation set) to make cross validation and to find AUROC.

As a model, I tried MLP Classifier, Decision Tree Classifier, LDA, Random Forest Classifier and SVC form sklearn package. I got the best result from Random Tree Classifier, so I used it as a model. I tried n\_estimators parameters as 10, 20, 50, 100 and got the best result from 50, so, I used 50 as a n\_estimators parameter.

After training my model, I calculated AUROC and find 0.9641518275538894, 0.9987929993964997, 0.9827188940092166 respectively.

After that, I made predictions for X\_test and wrote my findings into the csv files. I couldn’t find posterior probabilities, but I predicted class labels and wrote them on a csv file.