UCLA Extension Data Science Intensive

Instructor: William Yu

Project 5

A. Predicting Blood Donation

- Import blood_traindata.xlsx and blood_testdata.xlsx dataset. In this train dataset, we want to train the model to predict whether the person will donate blood this month with four possible predictors: "Months since last donation", "Number of donations", "Total volume donated (c.c.)", "Months since first donation".
- (1) *Selecting variables: features engineering*
- First, use a logistic model with the whole trainset to select the best number and format of variables/predictors by comparing AIC. Hint: if your model cannot run well, check the correlation of the variables.
- (2) *Check other machine learning models*
- Based on (1), run all the major machine learning models you learned in the class (e.g. D05a_mlmodels and p04) with the 10-fold cross validation. Choose the best model based on **accuracy** and use this model to predict whether the people in the blood_testdata.xlsx will donate or not (donate: 1; not: 0). Export and submit your prediction in a csv file. ***I will calculate the accuracy rate of your prediction.*** Note: There should be 176 rows.
- (3) *Optimal cut-off point on probability* for the whole trainset
- In D04c_churn and Project 5, we have learned and practiced ROC and the search of optimal cut-off point for the Logistic model. Now try to do the same thing for the xgbTree model here. What is the value of AUC? What is the optimal threshold?
- Hint: you could try applying the following code:

```
pred = predict(fit.xgb, type = "prob", blood)
pred.1 = as.numeric(pred[,2])
xgb.roc = roc(response = blood$donate, predictor = pred.1)
plot(xgb.roc, legacy.axes = TRUE, print.auc.y = 1.0, print.auc = TRUE)
coords(xgb.roc, "best", "threshold")
```

B. Predicting Credit Card Default

- Go import and check the data: creditcard.csv into R and read the data description in creditcard_description.

(1) Supervised Learning

- Do necessary data management and some brief EDAs (exploratory data analysis); For EDAs, you can check D04c_churn.R.
- Here you might want to convert some variables (imported as numeric) into factors (simply using as.factor…)
- Run a logistic regression to predict the probability of the variable "default.payment.next.month". Assume that you need to decide a cut-off point based on

the predicted probability. Do you think 0.5 is the best point? If not, present a better cut-off point and explain why.

(2) Unsupervised Learning - PCA

- Remove "default.payment.next.month" from the data; called the data frame a different name, say credit2.
- Run a simple principal component analysis and select the first 5 PCs and check how these 5 PCs load all the variables.

(3) Unsupervised Learning - Clustering

- Run K-means cluster on the data with k = 4. And plot the chart with 2 variables: LIMIT_BAL, and BILL_AMT1. Briefly explain what these 4 clusters might mean in the chart.