ACTL4305/5305 Actuarial Data Analytic Application

Week 7: Random Forest

Learning Objectives

In this tutorial, we use the credit data of the credit card clients in Taiwan to predict if a client will default or not. The data set is the customers' default payments which include 30000 instances described over 24 attributes. This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. The risk involved in lending business or credit card can be the business loss by not approving the good client or financial loss by approving the client who is at bad risk.

- We use the tree-based methods namely the classification trees, Bagging, and Random Forest (RF) to predict the creditworthiness of the client.
- We introduce the basic implementation of classification trees in R and explore the tuning process, which involves understanding the hyperparameters that can be adjusted for each method and performing grid or random searches.
- We explore two methods for assessing feature importance.

For illustrative purposes, we will work with a subset of 1000 observations, as running these methods on the complete dataset could be time-consuming. However, we encourage you to explore the full dataset or a larger sample after the lab to facilitate comprehensive result comparison.

```
# load packages

library(dplyr)
library(randomForest)
library(caret)
library(pROC)
library(ROCR)
library(tidyr)
library(PRROC) #roc.curve

# load data

credit <- read.csv("credit.csv")%>% dplyr::select(-X, -ID)

payamt_colsnames <- paste0("PAY_", c(1, 2:6))

credit <- credit%>%dplyr::mutate_at(vars(EDUCATION, MARRIAGE,SEX, default,
                                         payamt_colsnames), funs(factor))

credit$default <- as.factor(ifelse(credit$default == 1, "Yes", "No"))

#Extract a sample from the training set to speed up the computation
```

```r
set.seed(310)
credit <- credit[sample(nrow(credit),size = 1000, replace = FALSE),]

# reproduciblity

set.seed(123)

# data splitting

index <- createDataPartition(credit$default, p = 0.7, list = FALSE)
train <- credit[index, ]; test <- credit[-index, ]
```

Lab Tasks

1. Train classification and regression trees (CART), Bagging, and Random Forest models using the `caret` package. Utilize the `train` function from the caret package to construct predictive tree-based models. Refer to the online book for The caret Package for available hyperparameters for each method within the `train` function.
2. For the best-selected model from each method, assess the out-of-sample predictive accuracy using metrics such as Area Under the Curve (AUC), recall, precision, and F-score. Consider the suitability of each metric for the task and contemplate additional factors that should be taken into account.
3. Using the Random Forest method, plot the out-of-bag (OOB) error versus the test set error. Identify the number of trees providing the lowest error rate.
4. Using the Random Forest method, assess the features importance using the variable importance plot. Retrain all the models using only the most important features. Explain how you decided how many important variables to choose.