Machine Learning
Design a Shiny App to compare the prediction results of Random Forest and Decision Tree

1. The task of the dataset

This Shiny app shows the differences between the prediction outcome of two classifiers: Decision Tree (DT) and Random Forest (RF). The dataset consists of direct marketing campaigns of a Portuguese bank, and the goal is to predict whether clients will subscribe ('yes') or not ('no') to a term deposit based on phone calls to them.

2. Design and how it works

The app presents how DT and RF work from data input, model tuning, model accuracy, and final model presentation. (Figure 1) Users can experience the whole process by adjusting parameters and result graphs.

Following functions are implemented within the app: two buttons that generate DT model and RF model; sliders and tick box for adjusting hyperparameters; complexity parameter vs metrics plot and RF plot illustrate model tuning procedure; test results sections; the plot of the DT tree; the RF model important variables ranking plot.

The app UI is divided into two sections: side panel which allows users to control model features and main panel shows the process and results of the model. Each terminology is explained below to help users understand the aim of the object.
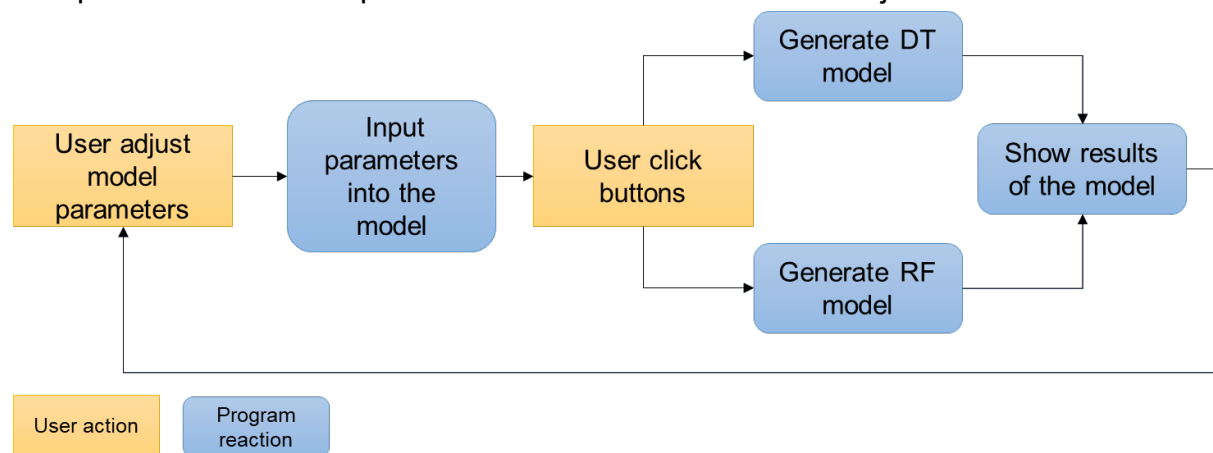

Figure 1. App workflow

3. Analysis of results from the two classifiers

According to the results from DT illustrated in Figure 2, the key features that determine whether a client would subscribe to a term deposit after a direct marketing phone call are call duration, employment variation rate, number of employees, last contact month, occupation, and education level. A client is more likely to subscribe to a term deposit if the duration of the call is over 447s and if she is being contacted during a period with higher employment variation rate. The call is more likely to fail if the duration is less

than 330s, with the number of employees over 5088, and during the months of May-Aug and Nov.
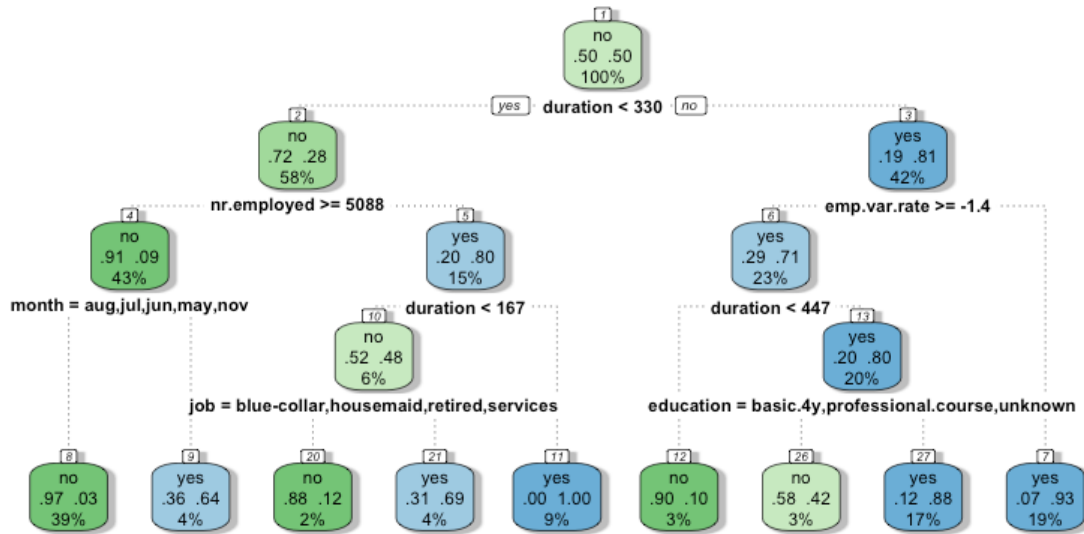

Figure 2. shows a visualisation of the decision tree. ('no' = client will not subscribe to a term deposit, 'yes' = client will subscribe to a term deposit)

The top 6 most important variables identified by RF classifiers as shown in Figure 3 are call duration, euribor 3 month rate, number of employees, occupation, employment variation rate, and age. Compared to the results from DT, last contact month and education appear to be less important than euribor rate and age.
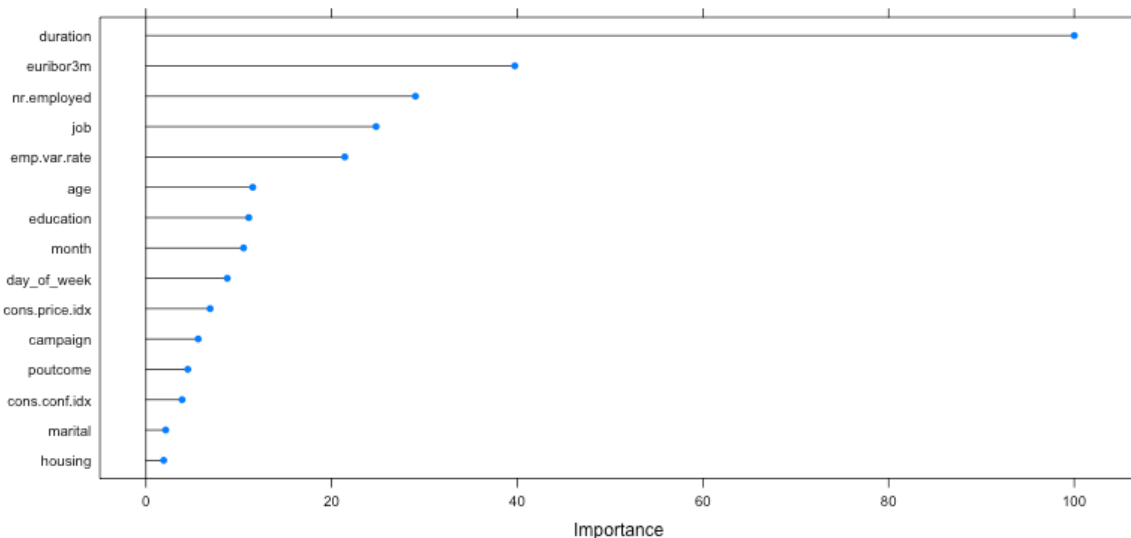

Figure 3. shows the importance of variables classified by the random forest model.

Looking at the performance measurements of both classifiers at the default setting, the overall accuracy rate of the RF classifier (88.7%) is higher than that of the DT classifier (78.7%). RF is more robust in predicting the true outcomes of the test set, with a true positive rate of 82.2% and a sensitivity of 0.99, compared to 79.4% and

0.77 for DT. The RF classifier provides higher precision to target potential clients who would subscribe to the term deposit better, and therefore can help to save the cost of calling unwilling clients.