**Data Science HW4 (Due on 12/11 24:00)**

% Remember to optimize the parameters in the training set in advance

1. Based on bank dataset, remove Job, Month, Day and treat “y” as the label: (a) construct a decision tree (CART) to identify the important predictors. (b) derive the response by including a new sample: discrete features are just the mode of all samples and numeric features are just the median of all samples. (c) using partial predictors identified by the CART to construct random forest and gradient boosting and compare their performances (in terms of overall accuracy and respective accuracy for y-yes and y-no).
2. Based on the churn dataset, remove the category, “InVol”, and apply CART, random forest, and adaboost to select the common predictors that have higher priorities. Derive the confusion matrix and the overall accuracy (70% for training and 30% for testing). Note that this is a binary problem with only two levels (Current/Vol). Further, calculate recall, precision, and *F*-measure (positive means Vol and negative means Current). Finally, construct the ROC (receiver operating characteristics) and compare three models (CART, random forest, adaboost) by showing the AUC (area under curve).
3. Based on the stock dataset (response is Taiwan, Korea, Japan, India), split the dataset into training (2011~2021) and testing (2022~2023). Then, apply random forest to determine key performance indicators. Finally, construct XGB (gbm is fine) to conduct forecasting and compare their performances in terms of RMSE, MAE, and MAPE. Draw a plot to demonstrate the actual value and the predictive results.
4. Based on the ETF dataset (00662 00668 00646 00645 00652), apply T test and F test to justify their annual rate of return and variance are equal or not (please convert daily return and variances into annual return and variances). Apply mathematical programming to derive the optimal portfolios by considering two scenarios: maximizing return but requiring annual risk less than 0.25 or minimizing risk but requiring annual return greater than 0.12.
5. Please visit the Kaggle website, discuss with your team members, download an interesting dataset, and then apply the methodologies you learn in this semester to solve a real problem and get managerial insights (following a PDCA loop).