STT 811

In-Class Assignment 15

This problem will use the Weekly dataset, with Direction as the target.

1. Take a look at the data.

   View(weekly)

2. Split the data into training and test datasets (with a 75/25 split).

   ```
   split_pct <- 0.75
   n <- length(weekly$Direction)*split_pct
   row_samp <- sample(1:length(weekly$Direction), n, replace = FALSE)
   train <- weekly[row_samp,]
   test <- weekly[-row_samp,]
   ```

3. Build a xgboost based on the lagged values and volume, using nrounds = 50, eta = 0.1, and max_depth = 3. Calculate the accuracy from the confusion matrix.

   ```
   weekly_xgb <- xgboost(data = data.matrix(train[,c(2:7)]), nrounds = 50, max_depth = 3,
   eta = 0.1, label = train$Direction, objective = "binary:logistic")

   pred <- predict(weekly_xgb, data.matrix(test[,c(2:7)]))
   confusionMatrix(as.factor(as.integer(2*pred)), as.factor(test$Direction))
   ```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 112  97
         1  37  27

               Accuracy : 0.5092
                 95% CI : (0.4482, 0.5699)
    No Information Rate : 0.5458
    P-Value [Acc > NIR] : 0.8989

                  Kappa : -0.0319

 Mcnemar's Test P-Value : 3.454e-07

            Sensitivity : 0.7517
            Specificity : 0.2177
         Pos Pred Value : 0.5359
         Neg Pred Value : 0.4219
             Prevalence : 0.5458
         Detection Rate : 0.4103
   Detection Prevalence : 0.7656
      Balanced Accuracy : 0.4847

       'Positive' Class : 0
```

4. Calculate the variable importance from this model

   xgb.importance(colnames(train[,c(2:7)]), model = weekly_xgb)

Feature <chr>	Gain <dbl>	Cover <dbl>	Frequency <dbl>
Lag1	0.2634499	0.21869945	0.2544170
Lag5	0.1927768	0.22062463	0.2014134
Lag2	0.1749421	0.21754331	0.1766784
Lag4	0.1335290	0.15325715	0.1236749
Lag3	0.1303521	0.08502465	0.1201413
Volume	0.1049502	0.10485082	0.1236749

5. Try different values for the hyperparameters, calculate the confusion matrix for each, and compare the results. Let the class know which set works particularly well.

   Using: nrounds = 75, max_depth = 5, eta = 0.8, the accuracy was boosted to 0.5531136.