# Lab 8

Christian Guaraca

11:59PM April 29, 2021

I want to make some use of my CART package. Everyone please try to run the following:

```
if (!pacman::p_isinstalled(YARF)){
  pacman::p_install_gh("kapelner/YARF/YARFJARs", ref = "dev")
  pacman::p_install_gh("kapelner/YARF/YARF", ref = "dev", force = TRUE)
}
options(java.parameters = "-Xmx4000m")
pacman::p_load(YARF)
```

For many of you it will not work. That's okay.

Throughout this part of this assignment you can use either the `tidyverse` package suite or `data.table` to answer but not base R. You can mix `data.table` with `magrittr` piping if you wish but don't go back and forth between `tbl_df`'s and `data.table` objects.

```
pacman::p_load(tidyverse, magrittr, data.table)
```

We will be using the `storms` dataset from the `dplyr` package. Filter this dataset on all storms that have no missing measurements for the two diameter variables, "ts_diameter" and "hu_diameter".

```
data(storms)

storms2 = storms %>% filter(!is.na(ts_diameter) & !is.na(hu_diameter) & ts_diameter > 0 & hu_diameter >

storms2
```

```
## # A tibble: 1,022 x 13
##    name   year month   day  hour   lat  long status    category  wind pressure
##    <chr> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <chr>     <ord>    <int>    <int>
##  1 Alex   2004     8     3     6  33   -77.4 hurricane 1           70      983
##  2 Alex   2004     8     3    12  34.2 -76.4 hurricane 2           85      974
##  3 Alex   2004     8     3    18  35.3 -75.2 hurricane 2           85      972
##  4 Alex   2004     8     4     0  36   -73.7 hurricane 1           80      974
##  5 Alex   2004     8     4     6  36.8 -72.1 hurricane 1           80      973
##  6 Alex   2004     8     4    12  37.3 -70.2 hurricane 2           85      973
##  7 Alex   2004     8     4    18  37.8 -68.3 hurricane 2           95      965
##  8 Alex   2004     8     5     0  38.5 -66   hurricane 3          105      957
##  9 Alex   2004     8     5     6  39.5 -63.1 hurricane 3          105      957
## 10 Alex   2004     8     5    12  40.8 -59.6 hurricane 3          100      962
## # ... with 1,012 more rows, and 2 more variables: ts_diameter <dbl>,
## #   hu_diameter <dbl>
```

From this subset, create a data frame that only has storm, observation period number for each storm (i.e., 1, 2, . . . , T) and the "ts_diameter" and "hu_diameter" metrics.

```r
storms2 = storms2 %>%
  select(name, ts_diameter, hu_diameter) %>%
  group_by(name) %>%
  mutate(period = row_number())

storms2
```

```
## # A tibble: 1,022 x 4
## # Groups:   name [63]
##    name  ts_diameter hu_diameter period
##    <chr>       <dbl>       <dbl>  <int>
##  1 Alex         150.        46.0      1
##  2 Alex         150.        46.0      2
##  3 Alex         190.        57.5      3
##  4 Alex         178.        63.3      4
##  5 Alex         224.        74.8      5
##  6 Alex         224.        74.8      6
##  7 Alex         259.        74.8      7
##  8 Alex         259.        80.6      8
##  9 Alex         345.        80.6      9
## 10 Alex         437.        80.6     10
## # ... with 1,012 more rows
```

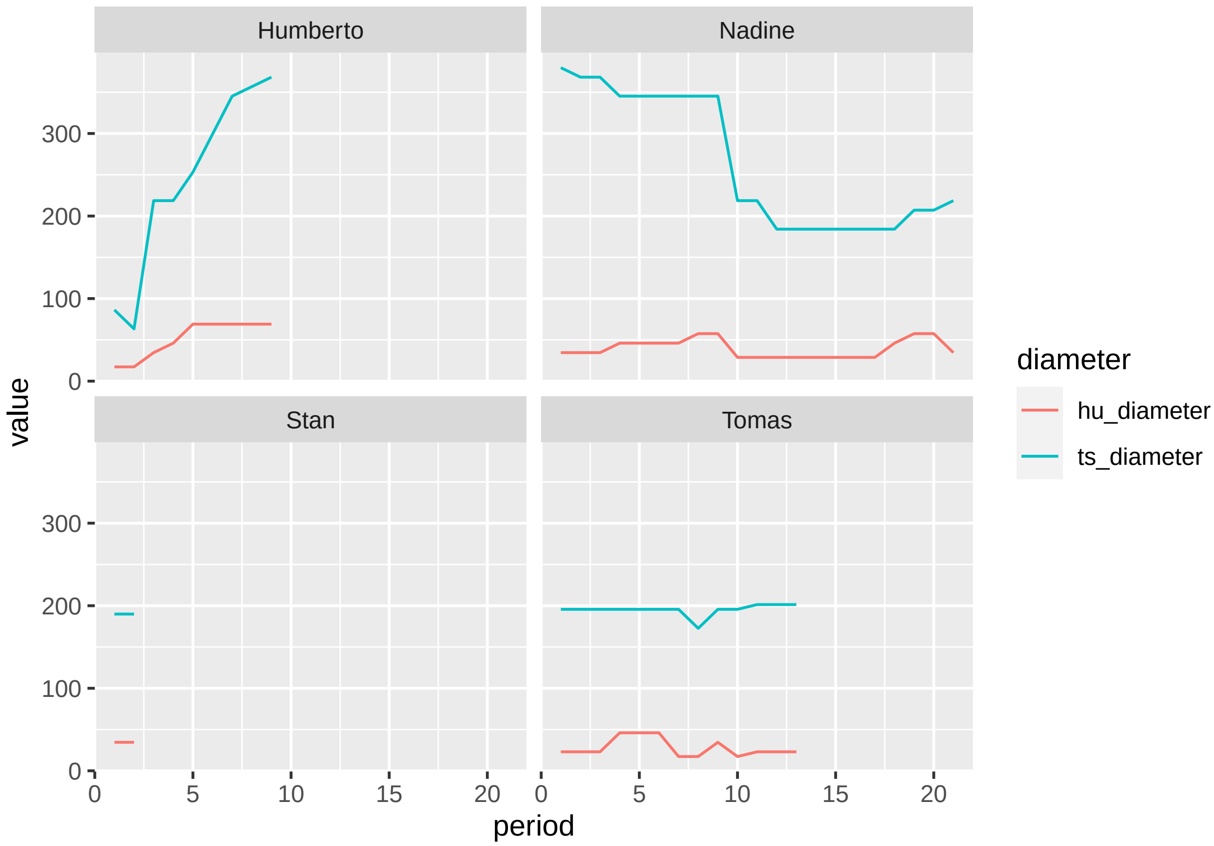
Create a data frame in long format with columns "diameter" for the measurement and "diameter_type" which will be categorical taking on the values "hu" or "ts".

```r
storms_long = pivot_longer = pivot_longer(storms2, cols = matches("diameter"), names_to = "diameter")
storms_long
```

```
## # A tibble: 2,044 x 4
## # Groups:   name [63]
##    name  period diameter    value
##    <chr>  <int> <chr>       <dbl>
##  1 Alex       1 ts_diameter 150.
##  2 Alex       1 hu_diameter  46.0
##  3 Alex       2 ts_diameter 150.
##  4 Alex       2 hu_diameter  46.0
##  5 Alex       3 ts_diameter 190.
##  6 Alex       3 hu_diameter  57.5
##  7 Alex       4 ts_diameter 178.
##  8 Alex       4 hu_diameter  63.3
##  9 Alex       5 ts_diameter 224.
## 10 Alex       5 hu_diameter  74.8
## # ... with 2,034 more rows
```

Using this long-formatted data frame, use a line plot to illustrate both "ts_diameter" and "hu_diameter" metrics by observation period for four random storms using a 2x2 faceting. The two diameters should appear in two different colors and there should be an appropriate legend.

```r
storms_sample = sample(unique(storms2$name), 4)
ggplot(storms_long %>% filter(name %in% storms_sample)) +
  geom_line(aes(x = period, y = value, col = diameter)) +
  facet_wrap(name~., nrow=2)
```




In this next first part of this lab, we will be joining three datasets in an effort to make a design matrix that predicts if a bill will be paid on time. Clean up and load up the three files. Then I'll rename a few features and then we can examine the data frames:

```
rm(list = ls())
pacman::p_load(tidyverse, magrittr, data.table, R.utils)
bills = fread("https://github.com/kapelner/QC_MATH_342W_Spring_2021/raw/master/labs/bills_dataset/bills
payments = fread("https://github.com/kapelner/QC_MATH_342W_Spring_2021/raw/master/labs/bills_dataset/pa
discounts = fread("https://github.com/kapelner/QC_MATH_342W_Spring_2021/raw/master/labs/bills_dataset/d
setnames(bills, "amount", "tot_amount")
setnames(payments, "amount", "paid_amount")
head(bills)
```

```
##          id   due_date invoice_date tot_amount customer_id discount_id
## 1: 15163811 2017-02-12   2017-01-13   99490.77    14290629     5693147
## 2: 17244832 2016-03-22   2016-02-21   99475.73    14663516     5693147
## 3: 16072776 2016-08-31   2016-07-17   99477.03    14569622     7302585
## 4: 15446684 2017-05-29   2017-05-29   99478.60    14488427     5693147
## 5: 16257142 2017-06-09   2017-05-10   99678.17    14497172     5693147
## 6: 17244880 2017-01-24   2017-01-24   99475.04    14663516     5693147
```

```
head(payments)
```

```
##          id paid_amount transaction_date  bill_id
## 1: 15272980    99165.60       2017-01-16 16571185
## 2: 15246935    99148.12       2017-01-03 16660000
## 3: 16596393    99158.06       2017-06-19 16985407
## 4: 16596651    99175.03       2017-06-19 17062491
## 5: 16687702    99148.20       2017-02-15 17184583
```

```
## 6: 16593510    99153.94       2017-06-11 16686215
```

```r
head(discounts)
```

```
##         id num_days pct_off days_until_discount
## 1: 5000000       20      NA                  NA
## 2: 5693147       NA       2                  NA
## 3: 6098612       20      NA                  NA
## 4: 6386294      120      NA                  NA
## 5: 6609438       NA       1                   7
## 6: 6791759       31       1                  NA
```

```r
bills = as_tibble(bills)
payments = as_tibble(payments)
discounts = as_tibble(discounts)
```

The unit we care about is the bill. The y metric we care about will be "paid in full" which is 1 if the company paid their total amount (we will generate this y metric later).

Since this is the response, we would like to construct the very best design matrix in order to predict y.

I will create the basic steps for you guys. First, join the three datasets in an intelligent way. You will need to examine the datasets beforehand.

```r
bills_with_payments = left_join(bills, payments, by = c("id"= "bill_id"))
bills_with_payments
```

```
## # A tibble: 279,118 x 9
##          id due_date   invoice_date tot_amount customer_id discount_id     id.y
##       <dbl> <date>     <date>            <dbl>       <int>       <dbl>    <dbl>
##  1 15163811 2017-02-12 2017-01-13       99491.    14290629     5693147 14670862
##  2 17244832 2016-03-22 2016-02-21       99476.    14663516     5693147 16691206
##  3 16072776 2016-08-31 2016-07-17       99477.    14569622     7302585       NA
##  4 15446684 2017-05-29 2017-05-29       99479.    14488427     5693147 16591210
##  5 16257142 2017-06-09 2017-05-10       99678.    14497172     5693147 16538398
##  6 17244880 2017-01-24 2017-01-24       99475.    14663516     5693147 16691231
##  7 16214048 2017-03-08 2017-02-06       99475.    14679281     5693147 16845763
##  8 15579946 2016-06-13 2016-04-14       99476.    14450223     5693147 16593380
##  9 15264234 2014-06-06 2014-05-07       99480.    14532786     7708050 16957842
## 10 17031731 2017-01-12 2016-12-13       99476.    14658929     5693147       NA
## # ... with 279,108 more rows, and 2 more variables: paid_amount <dbl>,
## #   transaction_date <date>
```

```r
bills_with_payments_with_discounts = left_join(bills_with_payments, discounts, by = c("discount_id"="id
```

Now create the binary response metric `paid_in_full` as the last column and create the beginnings of a design matrix `bills_data`. Ensure the unit / observation is bill i.e. each row should be one bill!

```r
bills_data = bills_with_payments_with_discounts %>%
  mutate(tot_amount = if_else(is.na(pct_off), tot_amount, tot_amount*(1-pct_off/100)))%>%
  group_by(id) %>%
  mutate(sum_of_payment_amount = sum(paid_amount))%>%
  mutate(paid_in_full = if_else(sum_of_payment_amount >= tot_amount, 1, 0, missing = 0))%>%
  slice(1) %>%
  ungroup()
table(bills_data$paid_in_full, useNA = "always")
```

```
##
##      0      1   <NA>
```

```
## 112664 113770      0
```

How should you add features from transformations (called "featurization")? What data type(s) should they be? Make some features below if you think of any useful ones. Name the columns appropriately so another data scientist can easily understand what information is in your variables.

```
pacman::p_load("lubridate")
bills_data = bills_data%>%
  select(-id, -id.y, -num_days, -transaction_date, -pct_off, -days_until_discount, -sum_of_payment_amou
  mutate(num_days_to_pay = as.integer(ymd(due_date) - ymd(invoice_date))) %>%
  select(-due_date, -invoice_date) %>%
  mutate(discount_id = as.factor(discount_id)) %>%
  group_by(customer_id) %>%
  mutate(bill_number = row_number()) %>%
  ungroup() %>%
  select(-customer_id)%>%
  relocate(paid_in_full, .after = last_col())
```

Now let's do this exercise. Let's retain 25% of our data for test.

```
K = 4
test_indices = sample(1 : nrow(bills_data), round(nrow(bills_data) / K))
train_indices = setdiff(1 : nrow(bills_data), test_indices)
bills_data_test = bills_data[test_indices, ]
bills_data_train = bills_data[train_indices, ]
```

Now try to build a classification tree model for `paid_in_full` with the features (use the `Xy` parameter in `YARF`). If you cannot get `YARF` to install, use the package `rpart` (the standard R tree package) instead. You will need to install it and read through some documentation to find the correct syntax.

Warning: this data is highly anonymized and there is likely zero signal! So don't expect to get predictive accuracy. The value of the exercise is in the practice. I think this exercise (with the joining exercise above) may be one of the most useful exercises in the entire semester.

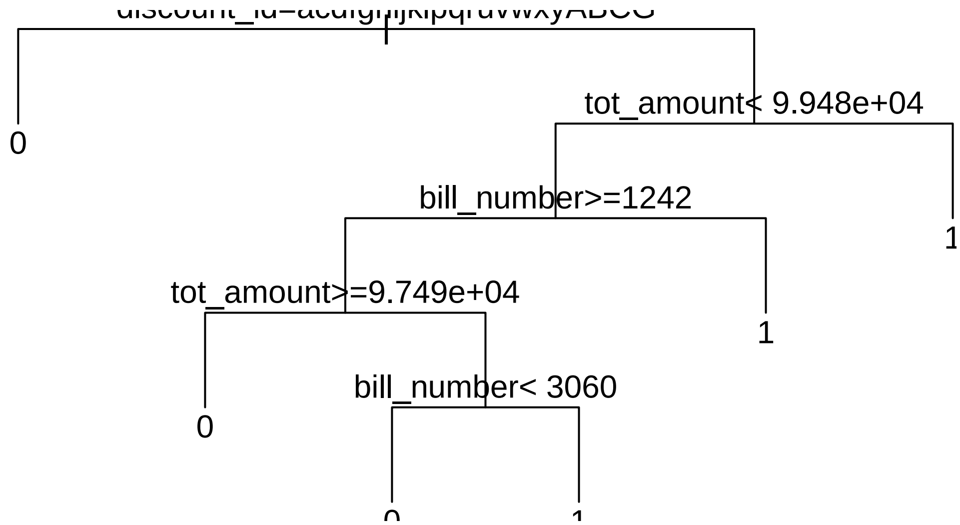
```
pacman::p_load(rpart)
?rpart
tree_mod = rpart(paid_in_full~., data=bills_data, method = 'class')
```

For those of you who installed `YARF`, what are the number of nodes and depth of the tree?

```
#YARF not installed
```

For those of you who installed `YARF`, print out an image of the tree.

```
plot(tree_mod, uniform = TRUE)
text(tree_mod)
```




Predict on the test set and compute a confusion matrix.

```
y_hat = predict(tree_mod, bills_data, type = c('class'))
con_table = table(bills_data$paid_in_full, y_hat)
con_table
```

```
##    y_hat
##         0     1
##   0 64268 48396
##   1 14482 99288
```

Report the following error metrics: misclassifcation error, precision, recall, F1, FDR, FOR.

```
n = sum(con_table)
fp = con_table[1,2]
fn = con_table[2,1]
tp = con_table[2,2]
tn = con_table[1,1]
misclass_error = (fn+fp)/n
F1 = 2*tp / (2*tp+fp+fn)
num_pred_p = sum(con_table[,2])
num_pred_n = sum(con_table[,1])
num_p = sum(con_table[2,])
num_n = sum(con_table[1,])
precision = tp/num_pred_p
recall = tp/num_p
FDR = 1 - precision
FOR = fn/num_pred_n
```

Is this a good model? (yes/no and explain).

This is not a good model because the FDR is a bit high at around 32%. This means that 32% is false truths and that is too high of an error to continue with this model.

There are probability asymmetric costs to the two types of errors. Assign the costs below and calculate oos total cost.

```
cost_fp = 50
cost_fn = 1
cost = cost_fp*fp + cost_fn*fn
cost
```

```
## [1] 2434282
```

We now wish to do asymmetric cost classification. Fit a logistic regression model to this data.

```r
log_reg_mod = glm(paid_in_full~., bills_data, family = binomial(link = 'logit' ))
```

Use the function from class to calculate all the error metrics for the values of the probability threshold being 0.001, 0.002, . . . , 0.999 in a data frame.

```r
compute_metrics_prob_classifier = function(p_hats, y_true, res = 0.001){
  #we first make the grid of all prob thresholds
  p_thresholds = seq(0 + res, 1 - res, by = res) #values of 0 or 1 are trivial

  #now we create a matrix which will house all of our results
  performance_metrics = matrix(NA, nrow = length(p_thresholds), ncol = 12)
  colnames(performance_metrics) = c(
    "p_th",
    "TN",
    "FP",
    "FN",
    "TP",
    "miscl_err",
    "precision",
    "recall",
    "FDR",
    "FPR",
    "FOR",
    "miss_rate"
  )

  #now we iterate through each p_th and calculate all metrics about the classifier and save
  n = length(y_true)
  for (i in 1 : length(p_thresholds)){
    p_th = p_thresholds[i]
    y_hats = factor(ifelse(p_hats >= p_th, 1, 0))
    confusion_table = table(
      factor(y_true, levels = c(0, 1)),
      factor(y_hats, levels = c(0, 1))
    )

    fp = confusion_table[1, 2]
    fn = confusion_table[2, 1]
    tp = confusion_table[2, 2]
    tn = confusion_table[1, 1]
    npp = sum(confusion_table[, 2])
    npn = sum(confusion_table[, 1])
    np = sum(confusion_table[2, ])
    nn = sum(confusion_table[1, ])

    performance_metrics[i, ] = c(
      p_th,
      tn,
      fp,
      fn,
      tp,
      (fp + fn) / n,
```

```
      tp / npp, #precision
      tp / np,  #recall
      fp / npp, #false discovery rate (FDR)
      fp / nn,  #false positive rate (FPR)
      fn / npn, #false omission rate (FOR)
      fn / np   #miss rate
    )
  }

  #finally return the matrix
  performance_metrics
}
p_hat_train = predict(log_reg_mod, bills_data, type = 'response')
p_hat_test = predict(log_reg_mod, bills_data, type = 'response')
y_1 = bills_data$paid_in_full
y_2 = bills_data$paid_in_full
c_metric_IS = compute_metrics_prob_classifier(p_hat_train, y_1)
c_metric_OOS =
compute_metrics_prob_classifier(p_hat_test, y_2)
```

Calculate the column `total_cost` and append it to this data frame.

```
cost_fp = 50
cost_fn = 1
c_table_IS = as_tibble(c_metric_IS) %>% mutate(total_cost = cost_fp * fp + cost_fn * fn)
c_table_IS
```

```
## # A tibble: 999 x 13
##     p_th    TN    FP    FN     TP miscl_err precision recall   FDR   FPR     FOR
##    <dbl> <dbl> <dbl> <dbl>  <dbl>     <dbl>     <dbl>  <dbl> <dbl> <dbl>   <dbl>
##  1 0.001 14136 97246     2 113737     0.429     0.539  1.00  0.461 0.873 1.41e-4
##  2 0.002 14136 97246     2 113737     0.429     0.539  1.00  0.461 0.873 1.41e-4
##  3 0.003 14136 97246     2 113737     0.429     0.539  1.00  0.461 0.873 1.41e-4
##  4 0.004 14136 97246     2 113737     0.429     0.539  1.00  0.461 0.873 1.41e-4
##  5 0.005 14136 97246     2 113737     0.429     0.539  1.00  0.461 0.873 1.41e-4
##  6 0.006 14136 97246     2 113737     0.429     0.539  1.00  0.461 0.873 1.41e-4
##  7 0.007 14299 97083     3 113736     0.429     0.539  1.00  0.461 0.872 2.10e-4
##  8 0.008 14300 97082     3 113736     0.429     0.539  1.00  0.461 0.872 2.10e-4
##  9 0.009 15035 96347     3 113736     0.426     0.541  1.00  0.459 0.865 1.99e-4
## 10 0.01  28426 82956   137 113602     0.367     0.578  0.999 0.422 0.745 4.80e-3
## # ... with 989 more rows, and 2 more variables: miss_rate <dbl>,
## #   total_cost <dbl>
```

```
c_table_OOS = as_tibble(c_metric_IS) %>% mutate(total_cost = cost_fp * fp + cost_fn * fn)
c_table_OOS
```

```
## # A tibble: 999 x 13
##     p_th    TN    FP    FN     TP miscl_err precision recall   FDR   FPR     FOR
##    <dbl> <dbl> <dbl> <dbl>  <dbl>     <dbl>     <dbl>  <dbl> <dbl> <dbl>   <dbl>
##  1 0.001 14136 97246     2 113737     0.429     0.539  1.00  0.461 0.873 1.41e-4
##  2 0.002 14136 97246     2 113737     0.429     0.539  1.00  0.461 0.873 1.41e-4
##  3 0.003 14136 97246     2 113737     0.429     0.539  1.00  0.461 0.873 1.41e-4
##  4 0.004 14136 97246     2 113737     0.429     0.539  1.00  0.461 0.873 1.41e-4
##  5 0.005 14136 97246     2 113737     0.429     0.539  1.00  0.461 0.873 1.41e-4
##  6 0.006 14136 97246     2 113737     0.429     0.539  1.00  0.461 0.873 1.41e-4
```

```
##   7 0.007 14299 97083     3 113736     0.429     0.539  1.00  0.461 0.872 2.10e-4
##   8 0.008 14300 97082     3 113736     0.429     0.539  1.00  0.461 0.872 2.10e-4
##   9 0.009 15035 96347     3 113736     0.426     0.541  1.00  0.459 0.865 1.99e-4
## 10 0.01  28426 82956   137 113602     0.367     0.578  0.999 0.422 0.745 4.80e-3
## # ... with 989 more rows, and 2 more variables: miss_rate <dbl>,
## #   total_cost <dbl>
```

Which is the winning probability threshold value and the total cost at that threshold?

```
W_prob_IS = which.min(c_table_IS$total_cost)
W_prob_IS_metric = c_table_IS[W_prob_IS, ]
W_prob_OOS = which.min(c_table_OOS$total_cost)
W_prob_OOS_metric = c_table_OOS[W_prob_OOS, ]
c_table_OOS[W_prob_OOS, ]
```

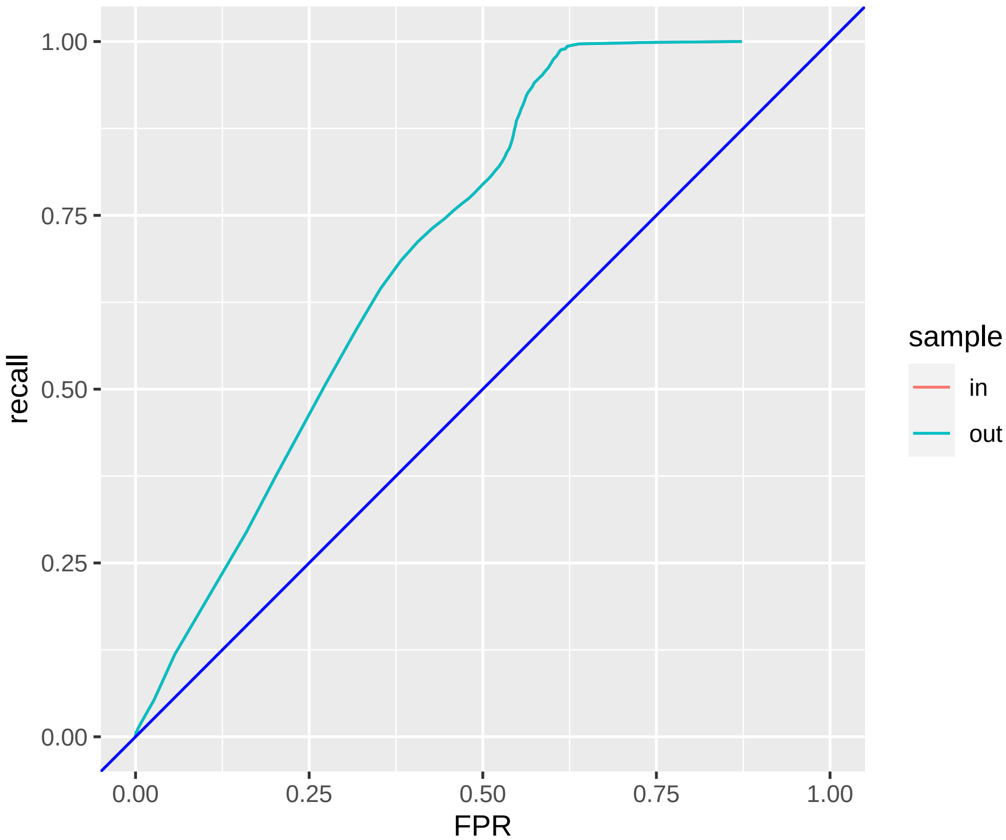
```
## # A tibble: 1 x 13
##    p_th    TN    FP    FN     TP miscl_err precision recall   FDR   FPR      FOR
##   <dbl> <dbl> <dbl> <dbl>  <dbl>     <dbl>     <dbl>  <dbl> <dbl> <dbl>    <dbl>
## 1 0.001 14136 97246     2 113737     0.429     0.539   1.00 0.461 0.873 0.000141
## # ... with 2 more variables: miss_rate <dbl>, total_cost <dbl>
```

```
c_table_IS[W_prob_IS, ]
```

```
## # A tibble: 1 x 13
##    p_th    TN    FP    FN     TP miscl_err precision recall   FDR   FPR      FOR
##   <dbl> <dbl> <dbl> <dbl>  <dbl>     <dbl>     <dbl>  <dbl> <dbl> <dbl>    <dbl>
## 1 0.001 14136 97246     2 113737     0.429     0.539   1.00 0.461 0.873 0.000141
## # ... with 2 more variables: miss_rate <dbl>, total_cost <dbl>
```

Plot an ROC curve and interpret.

```
pacman::p_load(ggplot2)
ggplot(rbind(
  cbind(c_table_IS, data.table(sample = 'in')),
  cbind(c_table_OOS, data.table(sample = 'out'))
)) +
  geom_line(aes(x = FPR, y = recall, col = sample)) +
  geom_abline(intercept = 0, slope = 1, col = "Blue") +
  coord_fixed() + xlim(0,1) + ylim(0, 1)
```




By taking the area under the curve, we are able to measure the models predictive power.

Calculate AUC and interpret.

```
pacman::p_load(pracma)
AUC_IS = -trapz(c_table_IS$FPR, c_table_IS$recall)
AUC_IS
```

```
## [1] 0.5835963
```

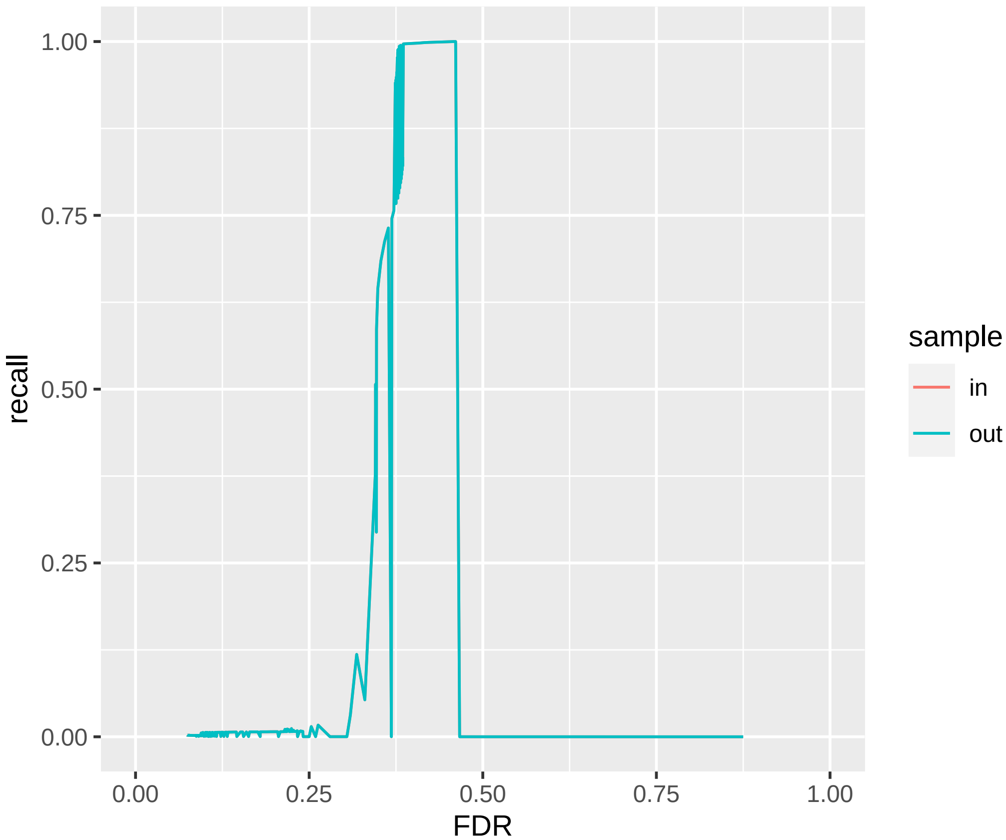
```
AUS_OOS = -trapz(c_table_OOS$FPR, c_table_OOS$recall)
AUS_OOS
```

```
## [1] 0.5835963
```

The model seems to be wrong seens they are both equal.

Plot a DET curve and interpret.

```
ggplot(rbind(
  cbind(c_table_IS, data.table(sample = 'in')),
  cbind(c_table_OOS, data.table(sample = 'out'))
)) +
  geom_line(aes(x = FDR, y = recall, col = sample)) +
  coord_fixed() + xlim(0,1) + ylim(0, 1)
```




It can be assumed that around 40% FDR is close to 100%. While the other most part is close to 0%.