Lab 08

Kennly Weerasinghe

4/29/2021

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(stormss)
```

```
## Warning in data(stormss): data set 'stormss' not found
```

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

```
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
```

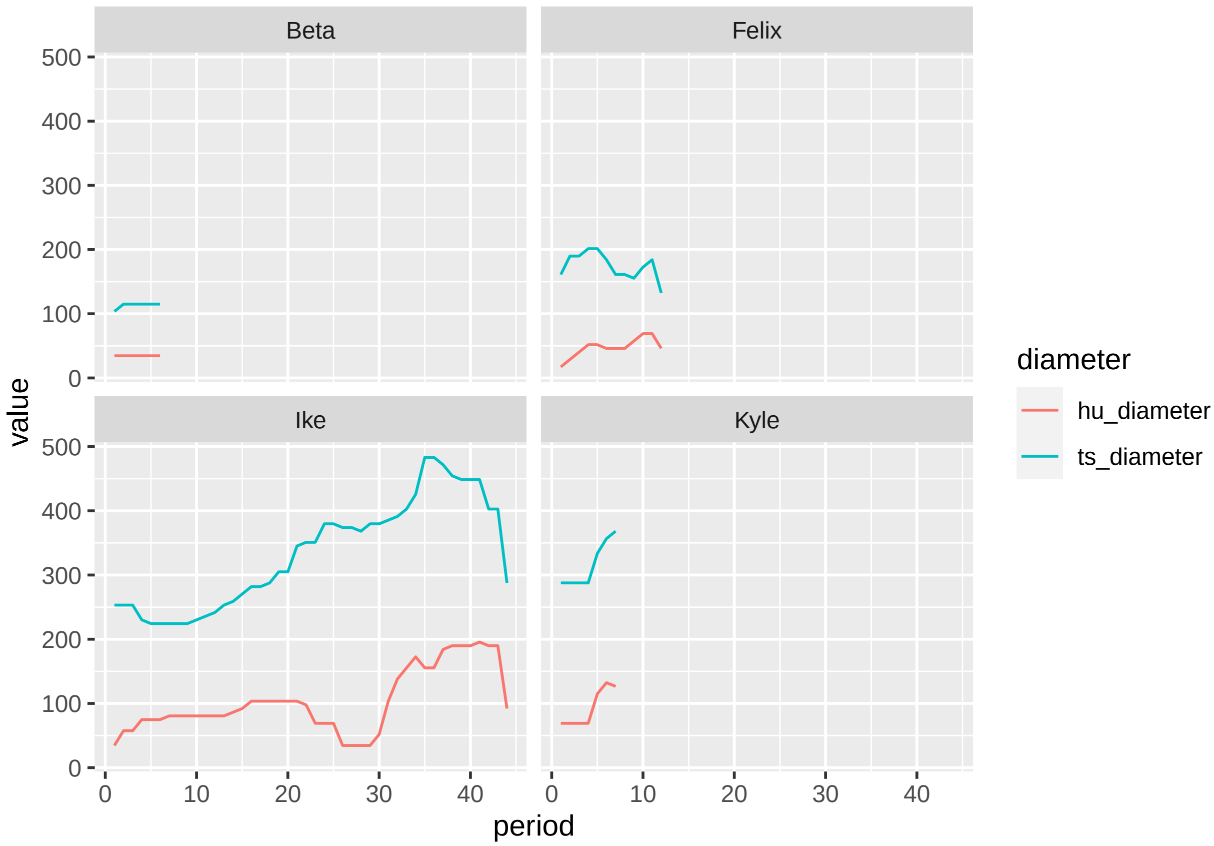
Create a data frame in long format with columns "diameter" for the measurement and "diameter_type" which will be categorical taking on the values "hu" or "ts".

```
storms_long = 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.

```
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)
discounts
```

```
## # A tibble: 60 x 4
##         id num_days pct_off days_until_discount
##      <dbl>    <int>   <dbl>               <int>
##  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
##  7 6945910       75    0.75                  NA
##  8 7079442       NA   NA                     10
##  9 7197225       60   NA                     NA
## 10 7302585       NA   NA                     NA
## # ... with 50 more rows
```

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" = "i
bills_with_payments_with_discounts
```

```
## # A tibble: 279,118 x 12
##          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 5 more variables: paid_amount <dbl>,
## #   transaction_date <date>, num_days <int>, pct_off <dbl>,
## #   days_until_discount <int>
```

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.

```r
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_num = row_number()) %>%
  ungroup() %>%
  select(-customer_id) %>%
  relocate(paid_in_full, .after = last_col())
bills_data
```

```
## # A tibble: 226,434 x 5
##    tot_amount discount_id num_days_to_pay bill_num paid_in_full
##         <dbl> <fct>                 <int>    <int>        <dbl>
```

```
## 1      99480. 7397895                    45         1             0
## 2      99529. 7397895                    30         1             0
## 3      99477. 7397895                    11         1             0
## 4      99479. 7397895                     0         2             0
## 5      99477. 7397895                    30         3             0
## 6      99477. 7397895                    30         1             0
## 7      99477. 7397895                     0         1             0
## 8      99477. 7397895                    30         2             0
## 9      99485. 7397895                    30         4             0
## 10     99477. 7397895                    30         2             0
## # ... with 226,424 more rows
```

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)
mod1 = rpart(paid_in_full ~., data = bills_data_train, method = "class")
mod1
```

```
## n= 169826
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 169826 84514 1 (0.49765054 0.50234946)
##    2) discount_id=5e+06,6098612,6609438,7079442,7197225,7302585,7397895,7564949,7708050,8091042,8178
##    3) discount_id=5693147,6945910,7484907,7890372,7944439,7995732,8258097,8367296,8806662,9043051,90
##      6) tot_amount< 99476.98 117848 47885 1 (0.40632849 0.59367151)
##       12) bill_num>=1242.5 31157 13833 0 (0.55602272 0.44397728)
##         24) tot_amount>=97487.15 16151  5935 0 (0.63253049 0.36746951) *
##         25) tot_amount< 97487.15 15006  7108 1 (0.47367720 0.52632280)
##           50) bill_num< 3062.5 9657  4314 0 (0.55327742 0.44672258) *
##           51) bill_num>=3062.5 5349  1765 1 (0.32996822 0.67003178) *
##       13) bill_num< 1242.5 86691 30561 1 (0.35252794 0.64747206) *
##      7) tot_amount>=99476.98 18719  4000 1 (0.21368663 0.78631337) *
```

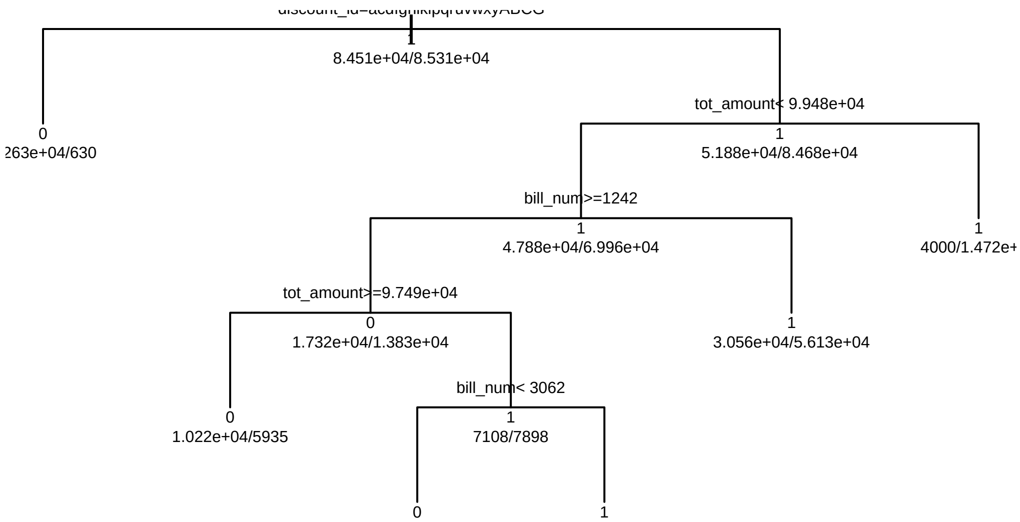
For those of you who installed `YARF`, what are the number of nodes and depth of the tree?

```
nrow(mod1$frame) #number of nodes
```

```
## [1] 11
```

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

```
plot(mod1, uniform=TRUE)
text(mod1, use.n=TRUE, all=TRUE, cex=.5)
```



Predict on the test set and compute a confusion matrix.

```
yhat = predict(mod1, bills_data_test, type = c("class"), na.action = na.pass)
oos_conf_table = table(bills_data_test$paid_in_full,yhat )
oos_conf_table
```

```
##    yhat
##         0     1
##   0 16055 12095
##   1  3572 24886
```

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

```
n = sum(oos_conf_table)
fp = oos_conf_table[1, 2]
fn = oos_conf_table[2, 1]
tp = oos_conf_table[2, 2]
tn = oos_conf_table[1, 1]
num_pred_pos = sum(oos_conf_table[, 2])
num_pred_neg = sum(oos_conf_table[, 1])
num_pos = sum(oos_conf_table[2, ])
num_neg = sum(oos_conf_table[1, ])
precision = tp / num_pred_pos
cat("precision", round(precision * 100, 2), "%\n")
```

```
## precision 67.29 %
```

```
recall = tp / num_pos
cat("recall", round(recall * 100, 2), "%\n")
```

```
## recall 87.45 %
```

```
false_discovery_rate = 1 - precision
cat("false_discovery_rate", round(false_discovery_rate * 100, 2), "%\n")
```

```
## false_discovery_rate 32.71 %
```

```
false_omission_rate = fn / num_pred_neg
cat("false_omission_rate", round(false_omission_rate * 100, 2), "%\n")
```

```
## false_omission_rate 18.2 %
```

```
misclassification_error = (fn +fp)/n
cat("misclassification_error", round(misclassification_error * 100, 2), "%\n")
```

```
## misclassification_error 27.68 %
```

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

I don't think this is a good model because the false discovery raate is 32.32 percent. This translates to roughly one third of your customers you forecast will pay the bill but they don't. Having one third of your bills unpaid would be bad for any business but especially for businesses with low margins.

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

```
C_fp = 100
C_fn = 1
oos_total_cost = (C_fn * fn) + (C_fp * fp)
oos_total_cost
```

```
## [1] 1213072
```

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

```
logistic_mod = glm(paid_in_full ~., bills_data_train, family = binomial(link = "logit"))
#phat_train = predict(logistic_mod, bills_data_train, type = "response")
#phat_test = predict(logistic_mod, bills_data_test, type = "response")
```

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.

```
compute_metrics_prob_classifier = function(p_hats, ytrue, 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(ytrue)
  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(ytrue, levels = c(0, 1)),
      factor(y_hats, levels = c(0, 1))
```

```r
    )

    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
}

phat_train = predict(logistic_mod, bills_data_train, type = "response")
phat_test = predict(logistic_mod, bills_data_test, type = "response")

probclassifier_metrics_in_sample = compute_metrics_prob_classifier(phat_train, bills_data_train$paid_in
probclassifier_metrics_in_sample
```

```
##       p_th    TN    FN    FN    TP miscl_err precision       recall       FDR
##   1: 0.001 10773 72759     2 85283 0.4284444 0.5396224 9.999765e-01 0.4603776
##   2: 0.002 10773 72759     2 85283 0.4284444 0.5396224 9.999765e-01 0.4603776
##   3: 0.003 10773 72759     2 85283 0.4284444 0.5396224 9.999765e-01 0.4603776
##   4: 0.004 10773 72759     2 85283 0.4284444 0.5396224 9.999765e-01 0.4603776
##   5: 0.005 10773 72759     2 85283 0.4284444 0.5396224 9.999765e-01 0.4603776
##  ---
## 995: 0.995 83526     6 85284     1 0.5022199 0.1428571 1.172539e-05 0.8571429
## 996: 0.996 83526     6 85284     1 0.5022199 0.1428571 1.172539e-05 0.8571429
## 997: 0.997 83526     6 85284     1 0.5022199 0.1428571 1.172539e-05 0.8571429
## 998: 0.998 83526     6 85284     1 0.5022199 0.1428571 1.172539e-05 0.8571429
## 999: 0.999 83526     6 85284     1 0.5022199 0.1428571 1.172539e-05 0.8571429
##               FPR          FOR    miss_rate
##   1: 8.710315e-01 0.0001856148 2.345078e-05
##   2: 8.710315e-01 0.0001856148 2.345078e-05
##   3: 8.710315e-01 0.0001856148 2.345078e-05
##   4: 8.710315e-01 0.0001856148 2.345078e-05
```

```
##   5: 8.710315e-01 0.0001856148 2.345078e-05
##  ---
## 995: 7.182876e-05 0.5052070375 9.999883e-01
## 996: 7.182876e-05 0.5052070375 9.999883e-01
## 997: 7.182876e-05 0.5052070375 9.999883e-01
## 998: 7.182876e-05 0.5052070375 9.999883e-01
## 999: 7.182876e-05 0.5052070375 9.999883e-01
```

```
probclassifier_metrics_oos = compute_metrics_prob_classifier(phat_test, bills_data_test$paid_in_full) %
probclassifier_metrics_oos
```

```
##       p_th    TN    FP    FN    TP miscl_err precision       recall       FDR
##   1: 0.001  3568 24282     2 28452 0.4289853 0.5395381 9.999297e-01 0.4604619
##   2: 0.002  3568 24282     2 28452 0.4289853 0.5395381 9.999297e-01 0.4604619
##   3: 0.003  3568 24282     2 28452 0.4289853 0.5395381 9.999297e-01 0.4604619
##   4: 0.004  3568 24282     2 28452 0.4289853 0.5395381 9.999297e-01 0.4604619
##   5: 0.005  3568 24282     2 28452 0.4289853 0.5395381 9.999297e-01 0.4604619
##  ---
## 995: 0.995 27849     1 28453     1 0.5026498 0.5000000 3.514444e-05 0.5000000
## 996: 0.996 27849     1 28453     1 0.5026498 0.5000000 3.514444e-05 0.5000000
## 997: 0.997 27849     1 28453     1 0.5026498 0.5000000 3.514444e-05 0.5000000
## 998: 0.998 27849     1 28454     0 0.5026675 0.0000000 0.000000e+00 1.0000000
## 999: 0.999 27849     1 28454     0 0.5026675 0.0000000 0.000000e+00 1.0000000
##               FPR          FOR    miss_rate
##   1: 8.718851e-01 0.0005602241 7.028889e-05
##   2: 8.718851e-01 0.0005602241 7.028889e-05
##   3: 8.718851e-01 0.0005602241 7.028889e-05
##   4: 8.718851e-01 0.0005602241 7.028889e-05
##   5: 8.718851e-01 0.0005602241 7.028889e-05
##  ---
## 995: 3.590664e-05 0.5053639302 9.999649e-01
## 996: 3.590664e-05 0.5053639302 9.999649e-01
## 997: 3.590664e-05 0.5053639302 9.999649e-01
## 998: 3.590664e-05 0.5053727155 1.000000e+00
## 999: 3.590664e-05 0.5053727155 1.000000e+00
```

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

```
C_fp = 100
C_fn = 1

probclassifier_metrics_in_sample = probclassifier_metrics_in_sample %>%
  mutate(total_cost = (C_fn * FN) + (C_fp * FP))

probclassifier_metrics_oos = probclassifier_metrics_oos %>%
  mutate(total_cost = (C_fn * FN) + (C_fp * FP))

probclassifier_metrics_in_sample
```

```
##       p_th    TN    FP    FN    TP miscl_err precision       recall       FDR
##   1: 0.001 10773 72759     2 85283 0.4284444 0.5396224 9.999765e-01 0.4603776
##   2: 0.002 10773 72759     2 85283 0.4284444 0.5396224 9.999765e-01 0.4603776
##   3: 0.003 10773 72759     2 85283 0.4284444 0.5396224 9.999765e-01 0.4603776
##   4: 0.004 10773 72759     2 85283 0.4284444 0.5396224 9.999765e-01 0.4603776
##   5: 0.005 10773 72759     2 85283 0.4284444 0.5396224 9.999765e-01 0.4603776
##  ---
```

```
## 995: 0.995 83526     6 85284     1 0.5022199 0.1428571 1.172539e-05 0.8571429
## 996: 0.996 83526     6 85284     1 0.5022199 0.1428571 1.172539e-05 0.8571429
## 997: 0.997 83526     6 85284     1 0.5022199 0.1428571 1.172539e-05 0.8571429
## 998: 0.998 83526     6 85284     1 0.5022199 0.1428571 1.172539e-05 0.8571429
## 999: 0.999 83526     6 85284     1 0.5022199 0.1428571 1.172539e-05 0.8571429
##               FPR          FOR    miss_rate total_cost
##   1: 8.710315e-01 0.0001856148 2.345078e-05    7275902
##   2: 8.710315e-01 0.0001856148 2.345078e-05    7275902
##   3: 8.710315e-01 0.0001856148 2.345078e-05    7275902
##   4: 8.710315e-01 0.0001856148 2.345078e-05    7275902
##   5: 8.710315e-01 0.0001856148 2.345078e-05    7275902
##  ---
## 995: 7.182876e-05 0.5052070375 9.999883e-01      85884
## 996: 7.182876e-05 0.5052070375 9.999883e-01      85884
## 997: 7.182876e-05 0.5052070375 9.999883e-01      85884
## 998: 7.182876e-05 0.5052070375 9.999883e-01      85884
## 999: 7.182876e-05 0.5052070375 9.999883e-01      85884
```

```
probclassifier_metrics_oos
```

```
##       p_th    TN    FP    FN    TP miscl_err precision       recall       FDR
##   1: 0.001  3568 24282     2 28452 0.4289853 0.5395381 9.999297e-01 0.4604619
##   2: 0.002  3568 24282     2 28452 0.4289853 0.5395381 9.999297e-01 0.4604619
##   3: 0.003  3568 24282     2 28452 0.4289853 0.5395381 9.999297e-01 0.4604619
##   4: 0.004  3568 24282     2 28452 0.4289853 0.5395381 9.999297e-01 0.4604619
##   5: 0.005  3568 24282     2 28452 0.4289853 0.5395381 9.999297e-01 0.4604619
##  ---
## 995: 0.995 27849     1 28453     1 0.5026498 0.5000000 3.514444e-05 0.5000000
## 996: 0.996 27849     1 28453     1 0.5026498 0.5000000 3.514444e-05 0.5000000
## 997: 0.997 27849     1 28453     1 0.5026498 0.5000000 3.514444e-05 0.5000000
## 998: 0.998 27849     1 28454     0 0.5026675 0.0000000 0.000000e+00 1.0000000
## 999: 0.999 27849     1 28454     0 0.5026675 0.0000000 0.000000e+00 1.0000000
##               FPR          FOR    miss_rate total_cost
##   1: 8.718851e-01 0.0005602241 7.028889e-05    2428202
##   2: 8.718851e-01 0.0005602241 7.028889e-05    2428202
##   3: 8.718851e-01 0.0005602241 7.028889e-05    2428202
##   4: 8.718851e-01 0.0005602241 7.028889e-05    2428202
##   5: 8.718851e-01 0.0005602241 7.028889e-05    2428202
##  ---
## 995: 3.590664e-05 0.5053639302 9.999649e-01      28553
## 996: 3.590664e-05 0.5053639302 9.999649e-01      28553
## 997: 3.590664e-05 0.5053639302 9.999649e-01      28553
## 998: 3.590664e-05 0.5053727155 1.000000e+00      28554
## 999: 3.590664e-05 0.5053727155 1.000000e+00      28554
```

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

```
best_prob_threshold_index_in_sample = which.min(probclassifier_metrics_in_sample$total_cost)
best_prob_threshold_metrics_in_sample = probclassifier_metrics_in_sample[best_prob_threshold_index_in_sa

cat("The total cost of the winning probability threshold in sample is", min(best_prob_threshold_metrics_
```

```
## The total cost of the winning probability threshold in sample is 85826
```

```
best_prob_threshold_index_oos = which.min(probclassifier_metrics_oos$total_cost)
best_prob_threshold_metrics_oos = probclassifier_metrics_oos[best_prob_threshold_index_oos, ]
```

```r
cat("\n\nThe total cost of the winning probability threshold OOS is", min(best_prob_threshold_metrics_oo
```

```
##
##
## The total cost of the winning probability threshold OOS is 28540
```

```r
probclassifier_metrics_oos[best_prob_threshold_index_oos, ]
```

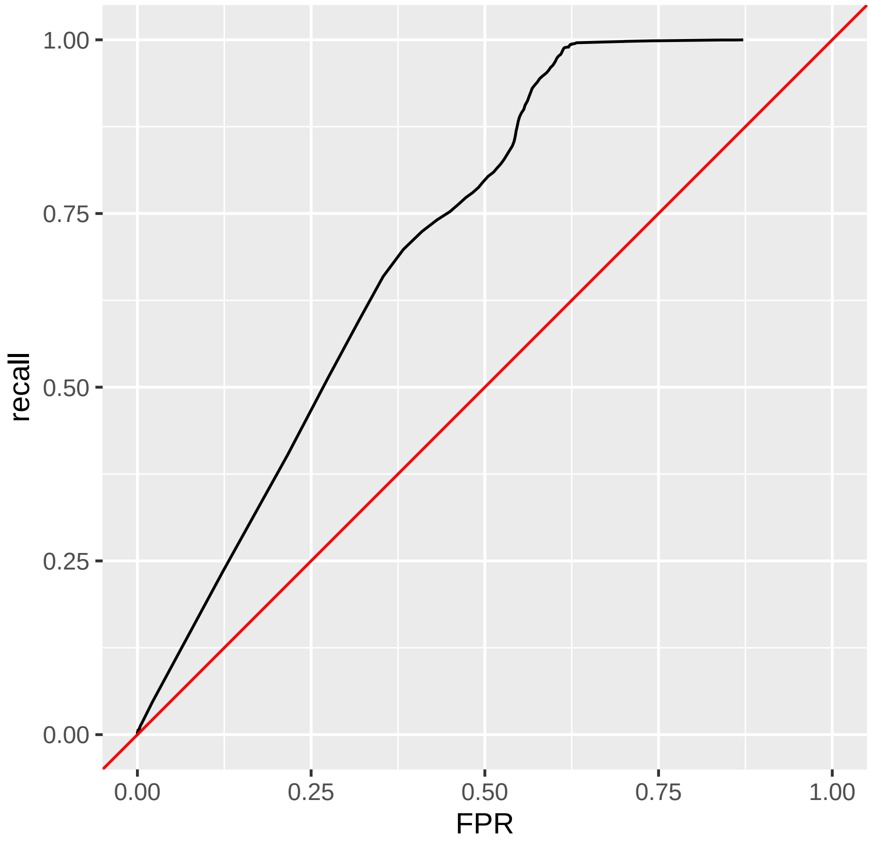
```
##     p_th    TN FP    FN TP miscl_err precision       recall        FDR
## 1: 0.952 27849  1 28440 14 0.5024202 0.9333333 0.0004920222 0.06666667
##              FPR       FOR miss_rate total_cost
## 1: 3.590664e-05 0.5052497  0.999508      28540
```

```r
probclassifier_metrics_in_sample[best_prob_threshold_index_in_sample, ]
```

```
##     p_th    TN FP    FN TP miscl_err precision       recall        FDR
## 1: 0.952 83526  6 85226 59 0.5018784 0.9076923 0.0006917981 0.09230769
##              FPR      FOR miss_rate total_cost
## 1: 7.182876e-05 0.505037 0.9993082      85826
```

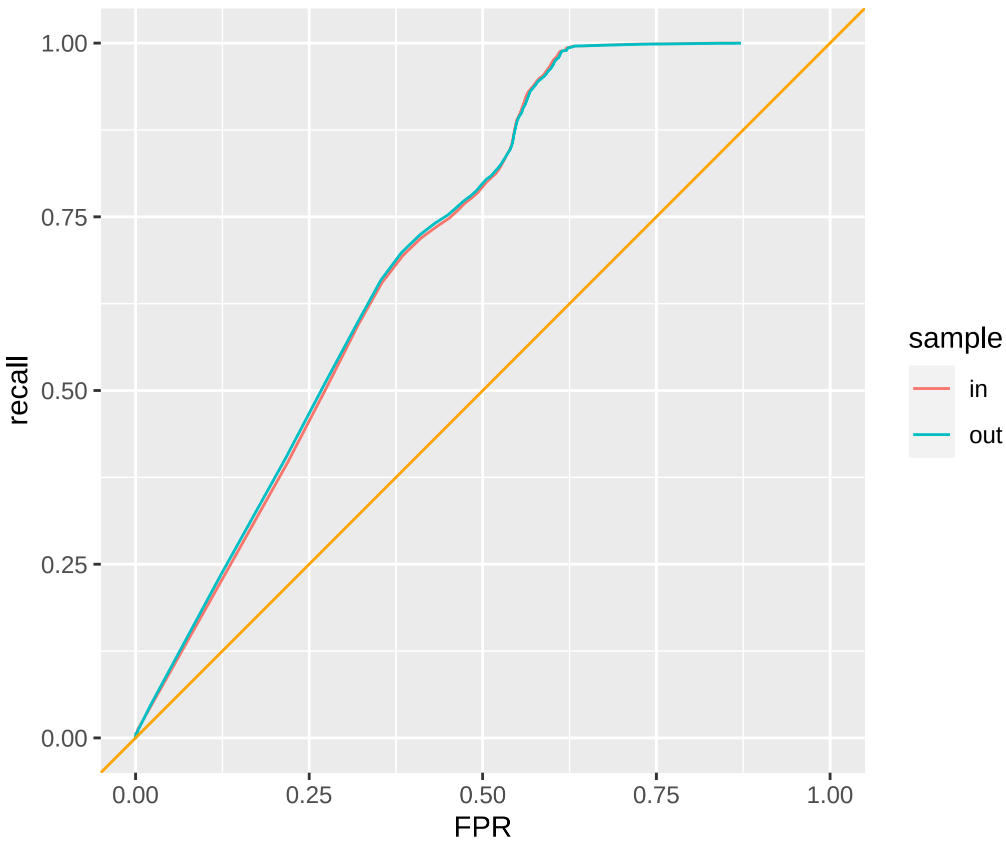
Plot an ROC curve and interpret.

```r
pacman::p_load(ggplot2)
ggplot(probclassifier_metrics_oos) +
  geom_line(aes(x = FPR, y = recall)) +
  geom_abline(intercept = 0, slope = 1, col = "Red") +
  coord_fixed() + xlim(0, 1) + ylim(0, 1)
```



```
pacman::p_load(ggplot2)
probclassifier_metrics_in_and_oos = rbind(
    cbind(probclassifier_metrics_in_sample, data.table(sample = "in")),
    cbind(probclassifier_metrics_oos, data.table(sample = "out"))
)

ggplot(probclassifier_metrics_in_and_oos) +
  geom_line(aes(x = FPR, y = recall, col = sample)) +
  geom_abline(intercept = 0, slope = 1, col = "orange") +
  coord_fixed() + xlim(0, 1) + ylim(0, 1)
```



ROC stands for the ”receiver operator curve and this gives a line that can be used to compare probability estimation models by calculating the area under the curve to assess predictive power.

Calculate AUC and interpret.

```
pacman::p_load(pracma)
auc_in_sample = -trapz(probclassifier_metrics_in_sample$FPR, probclassifier_metrics_in_sample$recall)
cat("AUC in-sample: ", auc_in_sample)
```

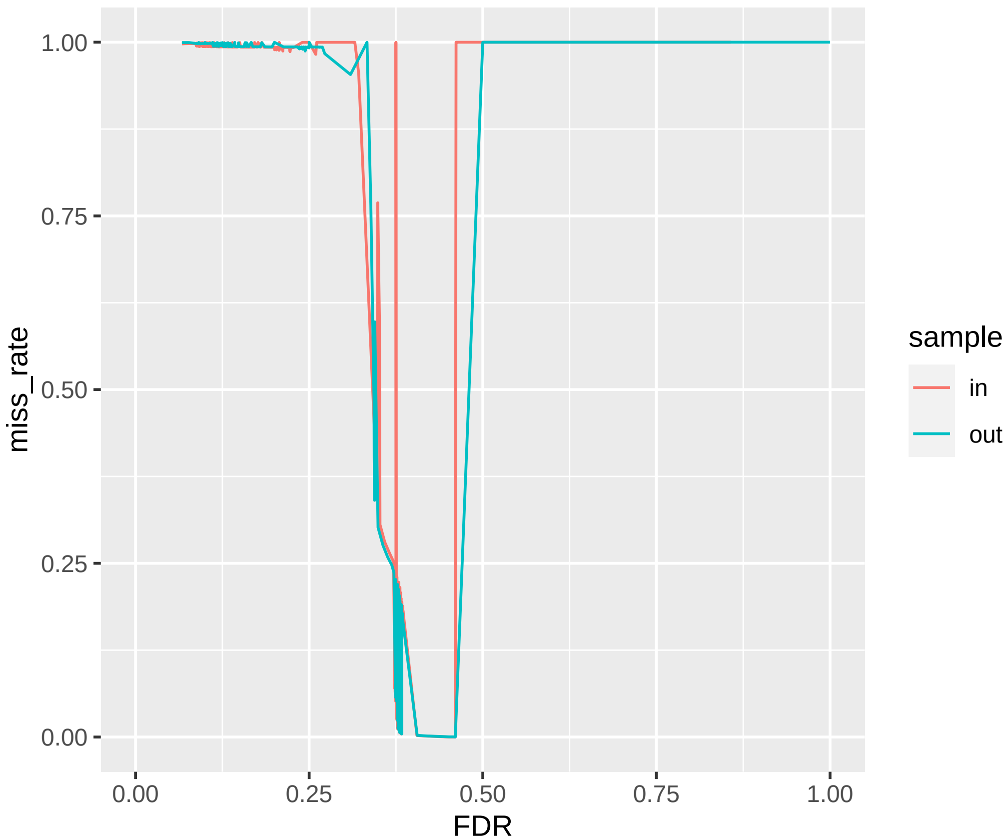
```
## AUC in-sample:  0.580102
```

```
auc_oos = -trapz(probclassifier_metrics_oos$FPR, probclassifier_metrics_oos$recall)
cat("\n\nAUC OOS: ", auc_oos)
```

```
##
##
## AUC OOS:  0.5843712
```

AUC values are 0.5815653 for in sample and 0.58085 for out of sample. Since AUC is greater than 0.5, this means that this model has predictive power.

```
ggplot(probclassifier_metrics_in_and_oos) +
  geom_line(aes(x = FDR, y = miss_rate, col = sample)) +
  coord_fixed() + xlim(0, 1) + ylim(0, 1)
```



DET is the detection error tradeoff and is traced out by varying the p_th from [0,1]. This graph would seem to indicate that FDR from around 37.5% gives FOR close to 0 percent. The rest of this graph is not very interpretable because of the zigzag lines.