Report Part 1

Andre Silva up201906559

2023-06-12

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.0.0 --
## v broom        1.0.4     v rsample      1.1.1
## v dials        1.2.0     v tune         1.1.1
## v infer        1.0.4     v workflows    1.1.3
## v modeldata    1.1.0     v workflowsets 1.0.1
## v parsnip      1.1.0     v yardstick    1.1.0
## v recipes      1.0.5
## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Use suppressPackageStartupMessages() to eliminate package startup messages
```

```
library(xgboost)
```

```
##
## Attaching package: 'xgboost'
##
## The following object is masked from 'package:dplyr':
##
##     slice
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following objects are masked from 'package:yardstick':
##
##     precision, recall, sensitivity, specificity
##
## The following object is masked from 'package:purrr':
##
##     lift
```

Read Data

```
train <- read_csv('../datasets/train.csv')
```

```
## Rows: 7944 Columns: 33
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): payment_type, employment_status, housing_status, source, device_os
## dbl (28): id, income, name_email_similarity, prev_address_months_count, curr...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
test <- read_csv('../datasets/test.csv')
```

```
## Rows: 2049 Columns: 32
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): payment_type, employment_status, housing_status, source, device_os
## dbl (27): id, income, name_email_similarity, prev_address_months_count, curr...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Dealing with NAN Training data

```
train$prev_address_months_count[train$prev_address_months_count == -1] <- NA
train$current_address_months_count[train$current_address_months_count == -1] <- NA
train$bank_months_count[train$bank_months_count == -1] <- NA
train$customer_age[train$customer_age == -1] <- NA
train$session_length_in_minutes[train$session_length_in_minutes == -1] <- NA
train$device_distinct_emails_8w[train$device_distinct_emails_8w == -1] <- NA
```

```r
missing_props_train = sapply(train, function(x) mean(is.na(x)))

over_threshold_train = missing_props_train[missing_props_train >= 0.6] # Remove coll with NA >= X

train_noNA <- train[, !colnames(train) %in% names(over_threshold_train)]

train_noNA <- drop_na(train_noNA)
```

Replicating NAN process to Test data

```r
test$prev_address_months_count[test$prev_address_months_count == -1] <- NA
test$current_address_months_count[test$current_address_months_count == -1] <- NA
test$bank_months_count[test$bank_months_count == -1] <- NA
test$customer_age[test$customer_age == -1] <- NA
test$session_length_in_minutes[test$session_length_in_minutes == -1] <- NA
test$device_distinct_emails_8w[test$device_distinct_emails_8w == -1] <- NA

test_noNA <- test[, !colnames(test) %in% names(over_threshold_train)] # Remove the same columns as in t
```

Separate X,y

```r
X_train = data.matrix(train_noNA[, !colnames(train_noNA) %in% c("is_fraud")])
y_train = train_noNA$is_fraud

X_test = data.matrix(test_noNA)
```

XGBoost

```r
xgboost_train = xgb.DMatrix(data=X_train, label=y_train)
xgboost_test = xgb.DMatrix(data=X_test)

model <- xgboost(data = xgboost_train, max.depth=2, nrounds=50, objective = "reg:logistic")
```

```
## [1]  train-rmse:0.359974
## [2]  train-rmse:0.267537
## [3]  train-rmse:0.204415
## [4]  train-rmse:0.161497
## [5]  train-rmse:0.133009
## [6]  train-rmse:0.114698
## [7]  train-rmse:0.103464
## [8]  train-rmse:0.096767
## [9]  train-rmse:0.092839
## [10] train-rmse:0.090629
## [11] train-rmse:0.089311
```

```
## [12] train-rmse:0.088543
## [13] train-rmse:0.087803
## [14] train-rmse:0.087397
## [15] train-rmse:0.087140
## [16] train-rmse:0.086803
## [17] train-rmse:0.086230
## [18] train-rmse:0.085945
## [19] train-rmse:0.085556
## [20] train-rmse:0.085086
## [21] train-rmse:0.084864
## [22] train-rmse:0.084257
## [23] train-rmse:0.084013
## [24] train-rmse:0.083619
## [25] train-rmse:0.082807
## [26] train-rmse:0.082381
## [27] train-rmse:0.081780
## [28] train-rmse:0.081433
## [29] train-rmse:0.080904
## [30] train-rmse:0.080532
## [31] train-rmse:0.080133
## [32] train-rmse:0.079839
## [33] train-rmse:0.079461
## [34] train-rmse:0.078601
## [35] train-rmse:0.078313
## [36] train-rmse:0.077441
## [37] train-rmse:0.077015
## [38] train-rmse:0.076288
## [39] train-rmse:0.076151
## [40] train-rmse:0.075759
## [41] train-rmse:0.075422
## [42] train-rmse:0.075116
## [43] train-rmse:0.074599
## [44] train-rmse:0.073928
## [45] train-rmse:0.073644
## [46] train-rmse:0.072657
## [47] train-rmse:0.071861
## [48] train-rmse:0.071158
## [49] train-rmse:0.070329
## [50] train-rmse:0.069866
```

```r
pred_test = predict(model, xgboost_test)

is_fraud = factor(round(pred_test, digits=2))

id = test$id

df <- data.frame(id , is_fraud)
colnames(df) <- c("id", "is_fraud")

i = 1
f = "../predictions/R_attempt1.csv"
while (file.exists(f)){
  i = i + 1
  f = paste(c("../predictions/R_attempt", i, ".csv"), collapse="")
```

```
}
f
```

```
## [1] "../predictions/R_attempt5.csv"
```

```
write.csv(df, f, row.names = FALSE)
is_fraud
```

```
##    [1] 0    0    0    0    0    0.01 0    0.03 0    0    0    0.01 0.02 0
##   [15] 0.01 0    0    0    0    0    0    0.05 0.02 0    0    0    0    0
##   [29] 0    0    0    0.01 0    0    0    0    0    0    0    0    0    0
##   [43] 0    0    0    0    0.01 0    0    0    0    0.02 0.01 0.01 0    0
##   [57] 0    0    0    0.21 0.05 0    0    0.01 0.01 0    0    0    0    0
##   [71] 0    0    0.01 0    0.02 0    0.01 0    0    0    0    0    0.06 0
##   [85] 0.02 0    0    0    0    0    0    0    0    0    0    0.01 0    0.04
##   [99] 0    0    0.01 0.01 0    0.07 0    0    0    0    0    0.01 0    0
##  [113] 0.01 0.05 0    0    0.01 0    0    0.01 0.02 0    0.01 0    0.02 0
##  [127] 0    0    0.01 0    0.1  0    0    0.05 0    0    0    0    0    0.01
##  [141] 0    0    0    0    0    0.08 0    0    0    0    0    0    0    0
##  [155] 0.05 0.01 0    0.5  0    0    0    0    0    0    0.01 0    0    0
##  [169] 0    0    0    0.08 0    0.01 0    0    0    0    0    0    0    0
##  [183] 0    0.08 0.3  0    0.01 0.01 0    0    0.09 0.1  0.01 0.01 0    0
##  [197] 0.06 0    0    0    0    0    0    0    0    0    0    0    0    0
##  [211] 0    0    0    0    0.01 0    0    0    0    0    0.01 0.01 0.07 0
##  [225] 0    0.06 0.04 0    0    0    0    0    0.02 0.01 0    0    0    0
##  [239] 0    0.03 0.03 0    0    0.04 0.02 0    0    0.01 0    0.05 0.01 0
##  [253] 0.02 0    0    0    0    0.03 0.01 0.02 0.04 0    0    0    0    0
##  [267] 0    0    0.01 0    0.01 0.01 0.01 0    0.02 0    0    0    0    0.01
##  [281] 0    0    0    0    0.02 0    0    0    0    0    0    0    0    0
##  [295] 0    0    0.01 0    0.1  0.06 0.01 0    0    0    0    0    0.01 0
##  [309] 0    0    0.01 0    0    0    0.02 0    0    0    0    0.05 0.07 0.01
##  [323] 0    0    0    0.03 0    0    0    0.01 0.02 0    0    0    0    0.01
##  [337] 0    0    0    0    0.17 0    0    0    0.07 0    0    0    0    0.01
##  [351] 0    0.01 0    0    0    0    0.01 0    0    0.04 0    0.01 0    0.07
##  [365] 0.01 0    0.01 0.06 0    0.02 0.02 0    0    0    0.11 0    0    0.01
##  [379] 0.01 0    0    0.01 0    0    0    0.03 0    0    0.07 0    0    0
##  [393] 0    0.01 0    0    0    0    0    0    0.03 0    0.02 0    0    0
##  [407] 0    0    0    0    0.01 0    0.01 0    0    0.04 0.01 0    0    0
##  [421] 0    0.01 0    0.01 0    0.01 0    0    0    0    0    0    0    0
##  [435] 0.02 0    0    0.03 0    0    0    0    0    0    0    0    0    0
##  [449] 0    0    0.04 0    0    0    0    0.01 0    0    0.05 0    0    0
##  [463] 0    0.01 0.01 0    0    0    0    0    0    0    0    0.02 0    0
##  [477] 0.02 0    0    0    0.08 0    0    0.01 0.03 0    0    0    0.09 0.01
##  [491] 0.01 0.02 0.01 0    0    0    0.06 0    0    0.01 0    0.01 0    0.01
##  [505] 0    0    0    0    0    0    0    0.01 0    0    0.18 0    0    0
##  [519] 0.01 0    0    0    0    0    0    0.01 0    0.02 0    0    0    0
##  [533] 0    0    0.01 0    0.02 0.01 0    0    0    0    0    0    0    0
##  [547] 0    0    0    0    0.01 0    0    0.01 0    0    0    0.01 0.01 0
##  [561] 0    0    0    0    0    0    0    0.01 0    0.02 0    0    0    0
##  [575] 0    0.01 0    0.01 0.01 0    0    0.01 0    0    0    0.01 0    0
##  [589] 0    0    0    0    0    0    0    0    0.01 0.1  0    0    0    0
##  [603] 0.02 0    0.01 0    0    0    0    0.01 0.02 0.04 0    0    0.01 0.01
##  [617] 0    0    0    0    0.01 0.07 0    0    0    0.02 0    0    0    0
```

```
##  [631] 0    0.02 0    0    0    0.01 0    0    0    0    0    0    0    0.01
##  [645] 0    0    0.01 0    0    0    0    0    0    0    0.01 0    0.02 0.01 0.09
##  [659] 0    0    0    0    0    0    0    0    0    0    0    0    0    0    0.03
##  [673] 0    0    0.03 0    0.03 0.01 0    0.01 0    0.01 0.01 0    0    0.01
##  [687] 0.02 0    0    0    0.01 0    0    0.01 0    0    0    0    0.01 0.03
##  [701] 0    0.03 0    0    0    0    0.01 0    0    0    0    0    0    0.01
##  [715] 0    0.02 0    0    0    0    0.01 0    0    0.01 0    0    0    0.01
##  [729] 0    0    0    0    0    0    0    0    0.01 0    0.01 0    0    0
##  [743] 0.01 0.05 0    0    0    0.07 0    0.02 0    0    0    0    0    0
##  [757] 0    0    0    0    0    0    0    0.01 0    0    0.13 0    0    0
##  [771] 0    0.01 0    0    0    0    0    0    0    0    0.01 0    0.02 0    0
##  [785] 0.01 0.01 0    0    0.01 0.03 0    0    0.01 0    0    0    0.03 0
##  [799] 0    0    0    0.01 0    0    0    0    0.01 0.01 0.01 0    0    0.04
##  [813] 0    0    0    0    0    0.01 0    0    0.02 0.01 0    0.02 0.02 0
##  [827] 0.05 0    0    0    0    0    0    0.08 0.02 0.01 0    0    0    0
##  [841] 0    0    0.02 0.02 0    0.02 0    0    0    0    0    0    0.01 0.01
##  [855] 0    0    0    0    0.01 0.01 0    0    0.02 0    0    0.05 0    0
##  [869] 0.01 0    0    0    0    0    0    0.01 0.04 0    0    0    0    0
##  [883] 0.11 0.02 0    0    0.03 0    0    0    0    0    0    0    0    0
##  [897] 0    0    0    0    0    0    0    0    0    0    0    0    0    0
##  [911] 0.03 0    0.01 0.04 0    0    0    0    0    0    0.01 0    0    0
##  [925] 0    0    0.01 0    0    0.03 0    0    0    0    0    0.01 0    0
##  [939] 0    0    0    0.01 0    0    0    0    0.04 0.01 0    0    0    0
##  [953] 0    0    0.07 0.01 0    0    0    0.01 0.03 0    0    0.03 0    0
##  [967] 0    0    0.01 0    0.01 0    0    0.01 0.01 0    0    0    0.07 0
##  [981] 0.02 0    0    0    0.02 0    0.07 0    0    0    0    0.03 0.04 0.01
##  [995] 0    0    0    0.04 0    0.01 0    0    0    0    0    0    0.01 0
## [1009] 0    0    0    0    0    0    0    0.01 0    0    0    0    0    0
## [1023] 0.01 0    0.02 0    0    0    0    0    0.01 0    0    0    0    0
## [1037] 0.01 0.01 0.08 0    0    0    0    0    0.05 0    0    0    0    0.01
## [1051] 0.01 0    0    0    0.01 0    0    0    0.09 0    0.07 0    0.01 0.02
## [1065] 0    0.01 0    0    0    0    0    0.01 0    0    0    0    0    0
## [1079] 0    0    0    0    0    0    0    0    0.01 0    0    0    0    0
## [1093] 0    0    0    0    0    0    0.02 0.02 0    0    0    0    0    0
## [1107] 0    0    0    0    0.01 0    0    0    0.05 0.02 0    0.14 0    0
## [1121] 0    0.01 0.01 0.03 0    0    0    0.02 0.01 0    0    0    0    0
## [1135] 0    0    0    0.01 0.11 0    0    0    0.01 0    0    0    0.01 0
## [1149] 0.03 0    0    0    0    0    0    0.01 0    0    0    0    0    0
## [1163] 0.02 0.03 0.02 0    0    0.01 0    0    0    0    0    0    0    0
## [1177] 0.09 0    0.01 0    0    0    0.15 0.01 0    0.01 0    0    0    0
## [1191] 0    0    0.05 0.01 0    0.03 0.09 0    0.03 0    0.02 0    0    0.01
## [1205] 0    0    0    0    0    0.01 0    0    0.01 0    0    0.01 0    0
## [1219] 0    0    0    0    0    0    0.01 0    0.01 0    0    0    0    0
## [1233] 0    0    0.02 0    0    0    0    0    0    0    0    0    0.05 0
## [1247] 0    0    0    0    0    0.01 0    0    0    0    0    0    0    0
## [1261] 0    0.16 0    0    0    0    0    0    0    0.01 0    0    0    0
## [1275] 0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [1289] 0    0    0.02 0    0    0.01 0.15 0    0    0    0    0    0    0
## [1303] 0    0    0    0    0    0    0.07 0    0    0    0    0.01 0    0
## [1317] 0    0.01 0.01 0    0    0    0    0    0    0    0    0.01 0    0.03
## [1331] 0    0    0    0    0    0    0    0    0.03 0    0    0    0    0
## [1345] 0    0    0    0.06 0    0    0    0    0.01 0    0    0    0    0
## [1359] 0    0    0.01 0    0.02 0.01 0    0.04 0.01 0    0    0.01 0.07 0
## [1373] 0    0.04 0    0.01 0    0    0    0.01 0.01 0    0    0    0.01 0
```

```
## [1387] 0    0    0    0    0    0    0.01 0    0.13 0.27 0    0.01 0    0
## [1401] 0.05 0    0.01 0    0    0    0    0    0    0    0    0    0    0
## [1415] 0    0    0.1  0.01 0.01 0    0    0    0    0    0.03 0    0.01 0.01
## [1429] 0    0.01 0.05 0    0    0.03 0    0    0    0    0    0.01 0.04 0.01
## [1443] 0    0    0    0    0.03 0    0.08 0    0    0.03 0    0    0    0
## [1457] 0    0    0.01 0    0    0.01 0    0    0    0.01 0    0.01 0    0.01
## [1471] 0    0    0.1  0    0    0.05 0.14 0.01 0    0    0.02 0.02 0    0
## [1485] 0    0    0    0    0    0    0    0    0.24 0    0    0.01 0.01 0.01
## [1499] 0    0    0    0    0    0    0.12 0    0    0.01 0.01 0    0    0
## [1513] 0    0    0    0    0    0.03 0    0.01 0    0    0    0    0    0
## [1527] 0    0    0    0    0    0    0    0.01 0    0    0    0    0    0
## [1541] 0    0.03 0    0.01 0.09 0    0    0    0    0    0    0    0    0
## [1555] 0.01 0    0    0    0    0    0    0    0    0.03 0    0    0    0
## [1569] 0    0    0.02 0    0.01 0    0    0    0.01 0    0    0    0    0
## [1583] 0    0    0    0    0    0    0    0.16 0    0.01 0    0    0    0
## [1597] 0    0.01 0    0    0    0.02 0.01 0.01 0    0    0    0    0    0
## [1611] 0    0    0    0    0.15 0    0.04 0    0    0.05 0.02 0    0    0
## [1625] 0    0    0    0.01 0.01 0    0    0    0    0    0    0    0    0.01
## [1639] 0.02 0    0    0.01 0    0    0    0    0    0.01 0.03 0    0    0
## [1653] 0    0    0.01 0    0.01 0.04 0.01 0    0.01 0    0.01 0.01 0    0
## [1667] 0    0    0    0    0    0    0    0    0    0    0    0    0    0.01
## [1681] 0    0    0    0    0.04 0    0    0.02 0    0    0    0    0    0.02
## [1695] 0    0.01 0.01 0    0    0    0    0.03 0    0    0    0    0    0.13
## [1709] 0    0    0    0    0.02 0    0    0    0.02 0    0.02 0    0    0.01
## [1723] 0.01 0    0    0    0.14 0    0.01 0    0    0    0.02 0    0    0
## [1737] 0    0    0    0    0    0    0.06 0    0    0    0    0    0    0
## [1751] 0    0    0    0    0    0    0    0.01 0    0    0    0.01 0    0
## [1765] 0    0    0    0    0    0    0    0    0    0    0    0.02 0    0
## [1779] 0    0    0.06 0    0    0    0    0    0    0    0    0.02 0    0
## [1793] 0    0    0.11 0    0    0    0.01 0    0    0    0    0.02 0    0
## [1807] 0    0    0    0    0    0    0.01 0    0    0.01 0    0    0    0
## [1821] 0    0.07 0    0    0    0    0    0.01 0    0    0    0    0    0
## [1835] 0.01 0    0    0    0    0    0    0    0    0    0.01 0.15 0    0
## [1849] 0    0    0    0    0.01 0    0.01 0    0    0    0    0.01 0    0
## [1863] 0    0    0    0    0.08 0.16 0    0    0    0    0.14 0.09 0.06 0
## [1877] 0    0    0    0    0    0.01 0    0    0    0    0    0    0    0
## [1891] 0    0    0.19 0.05 0    0    0    0.01 0    0    0.01 0    0    0.01
## [1905] 0.07 0.01 0    0.01 0    0    0.02 0    0    0    0    0    0    0
## [1919] 0    0.18 0    0    0.05 0    0    0    0    0    0.02 0    0    0
## [1933] 0    0.14 0    0.02 0.15 0    0    0    0    0    0.32 0.05 0    0.01
## [1947] 0    0    0    0.31 0    0    0    0    0    0.01 0    0    0.12 0.01
## [1961] 0    0    0    0    0    0    0    0    0.06 0    0    0.04 0    0.02
## [1975] 0    0    0    0    0    0    0.01 0.01 0    0    0    0    0    0
## [1989] 0    0    0    0    0    0    0    0    0    0    0    0.2  0    0
## [2003] 0    0    0    0    0.02 0    0.05 0    0.04 0    0    0    0    0
## [2017] 0    0    0    0    0    0    0.02 0    0.01 0    0    0.02 0    0.01
## [2031] 0.07 0.01 0    0.08 0    0    0    0.01 0    0.07 0    0    0    0
## [2045] 0    0.11 0.03 0    0.2
## 28 Levels: 0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1 0.11 ... 0.5
```