Porto Seguro

Krishna

5 July 2018

```
str(train2)
```

```
## 'data.frame':    595212 obs. of  59 variables:
##  $ id            : int  7 9 13 16 17 19 20 22 26 28 ...
##  $ target        : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ ps_ind_01     : int  2 1 5 0 0 5 2 5 5 1 ...
##  $ ps_ind_02_cat : int  2 1 4 1 2 1 1 1 1 1 ...
##  $ ps_ind_03     : int  5 7 9 2 0 4 3 4 3 2 ...
##  $ ps_ind_04_cat : int  1 0 1 0 1 0 1 0 1 0 ...
##  $ ps_ind_05_cat : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ps_ind_06_bin : int  0 0 0 1 1 0 0 1 0 0 ...
##  $ ps_ind_07_bin : int  1 0 0 0 0 0 1 0 0 1 ...
##  $ ps_ind_08_bin : int  0 1 1 0 0 0 0 0 1 0 ...
##  $ ps_ind_09_bin : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ ps_ind_10_bin : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ps_ind_11_bin : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ps_ind_12_bin : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ps_ind_13_bin : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ps_ind_14     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ps_ind_15     : int  11 3 12 8 9 6 8 13 6 4 ...
##  $ ps_ind_16_bin : int  0 0 1 1 1 1 1 1 1 0 ...
##  $ ps_ind_17_bin : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ ps_ind_18_bin : int  0 1 0 0 0 0 0 0 0 1 ...
##  $ ps_reg_01     : num  0.7 0.8 0 0.9 0.7 0.9 0.6 0.7 0.9 0.9 ...
##  $ ps_reg_02     : num  0.2 0.4 0 0.2 0.6 1.8 0.1 0.4 0.7 1.4 ...
##  $ ps_reg_03     : num  0.718 0.766 -1 0.581 0.841 ...
##  $ ps_car_01_cat : int  10 11 7 7 11 10 6 11 10 11 ...
##  $ ps_car_02_cat : int  1 1 1 1 1 0 1 1 1 0 ...
##  $ ps_car_03_cat : int  -1 -1 -1 0 -1 -1 -1 0 -1 0 ...
##  $ ps_car_04_cat : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ ps_car_05_cat : int  1 -1 -1 1 -1 0 1 0 1 0 ...
##  $ ps_car_06_cat : int  4 11 14 11 14 14 11 11 14 14 ...
##  $ ps_car_07_cat : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ ps_car_08_cat : int  0 1 1 1 1 1 1 1 1 1 ...
##  $ ps_car_09_cat : int  0 2 2 3 2 0 0 2 0 2 ...
##  $ ps_car_10_cat : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ ps_car_11_cat : int  12 19 60 104 82 104 99 30 68 104 ...
##  $ ps_car_11     : int  2 3 1 1 3 2 2 3 3 2 ...
##  $ ps_car_12     : num  0.4 0.316 0.316 0.374 0.316 ...
##  $ ps_car_13     : num  0.884 0.619 0.642 0.543 0.566 ...
##  $ ps_car_14     : num  0.371 0.389 0.347 0.295 0.365 ...
##  $ ps_car_15     : num  3.61 2.45 3.32 2 2 ...
##  $ ps_calc_01    : num  0.6 0.3 0.5 0.6 0.4 0.7 0.2 0.1 0.9 0.7 ...
##  $ ps_calc_02    : num  0.5 0.1 0.7 0.9 0.6 0.8 0.6 0.5 0.8 0.8 ...
##  $ ps_calc_03    : num  0.2 0.3 0.1 0.1 0 0.4 0.5 0.1 0.6 0.8 ...
##  $ ps_calc_04    : int  3 2 2 2 2 3 2 1 3 2 ...
##  $ ps_calc_05    : int  1 1 2 4 2 1 2 2 1 2 ...
##  $ ps_calc_06    : int  10 9 9 7 6 8 8 7 7 8 ...
##  $ ps_calc_07    : int  1 5 1 1 3 2 1 1 3 2 ...
##  $ ps_calc_08    : int  10 8 8 8 10 11 8 6 9 9 ...
##  $ ps_calc_09    : int  1 1 2 4 2 3 3 1 4 1 ...
##  $ ps_calc_10    : int  5 7 7 2 12 8 10 13 11 11 ...
##  $ ps_calc_11    : int  9 3 4 2 3 4 3 7 4 3 ...
##  $ ps_calc_12    : int  1 1 2 2 1 2 0 1 2 5 ...
##  $ ps_calc_13    : int  5 1 7 4 1 0 0 3 1 0 ...
##  $ ps_calc_14    : int  8 9 7 9 3 9 10 6 5 6 ...
##  $ ps_calc_15_bin: int  0 0 0 0 0 0 0 1 0 0 ...
##  $ ps_calc_16_bin: int  1 1 1 0 0 1 1 0 1 1 ...
##  $ ps_calc_17_bin: int  1 1 1 0 0 0 0 1 0 0 ...
```

```
##  $ ps_calc_18_bin: int  0 0 0 0 1 1 0 0 0 0 ...
##  $ ps_calc_19_bin: int  0 1 1 0 1 1 1 1 0 1 ...
##  $ ps_calc_20_bin: int  1 0 0 0 0 1 0 0 1 0 ...
```

Replcing “-1” with NA

```
train2[train2== -1] = NA
test[test == -1] = NA
```

```
t = sapply(train2 , function(x) sum(is.na(x)))
t[t>100000]
```

```
##     ps_reg_03 ps_car_03_cat ps_car_05_cat
##        107772        411231        266551
```

```
t1 = sapply(test , function(x) sum(is.na(x)))
t1[t1>100000]
```

```
##     ps_reg_03 ps_car_03_cat ps_car_05_cat
##        161684        616911        400359
```

Removing Columns Where Missing Value is more than 1lakh

```
train = train2 %>% select(-ps_reg_03,-ps_car_03_cat,-ps_car_05_cat)
test = test %>% select(-ps_reg_03,-ps_car_03_cat,-ps_car_05_cat)
```

```
sapply(train, function(x) sum(is.na(x)))
```

```
##             id         target      ps_ind_01  ps_ind_02_cat      ps_ind_03
##              0              0              0            216              0
##  ps_ind_04_cat  ps_ind_05_cat  ps_ind_06_bin  ps_ind_07_bin  ps_ind_08_bin
##             83           5809              0              0              0
##  ps_ind_09_bin  ps_ind_10_bin  ps_ind_11_bin  ps_ind_12_bin  ps_ind_13_bin
##              0              0              0              0              0
##      ps_ind_14      ps_ind_15  ps_ind_16_bin  ps_ind_17_bin  ps_ind_18_bin
##              0              0              0              0              0
##      ps_reg_01      ps_reg_02  ps_car_01_cat  ps_car_02_cat  ps_car_04_cat
##              0              0            107              5              0
##  ps_car_06_cat  ps_car_07_cat  ps_car_08_cat  ps_car_09_cat  ps_car_10_cat
##              0          11489              0            569              0
##  ps_car_11_cat      ps_car_11      ps_car_12      ps_car_13      ps_car_14
##              0              5              1              0          42620
##      ps_car_15     ps_calc_01     ps_calc_02     ps_calc_03     ps_calc_04
##              0              0              0              0              0
##     ps_calc_05     ps_calc_06     ps_calc_07     ps_calc_08     ps_calc_09
##              0              0              0              0              0
##     ps_calc_10     ps_calc_11     ps_calc_12     ps_calc_13     ps_calc_14
##              0              0              0              0              0
## ps_calc_15_bin ps_calc_16_bin ps_calc_17_bin ps_calc_18_bin ps_calc_19_bin
##              0              0              0              0              0
## ps_calc_20_bin
##              0
```

```
sapply(test, function(x) sum(is.na(x)))
```

```
##             id      ps_ind_01  ps_ind_02_cat      ps_ind_03  ps_ind_04_cat
##              0              0            307              0            145
##  ps_ind_05_cat  ps_ind_06_bin  ps_ind_07_bin  ps_ind_08_bin  ps_ind_09_bin
##           8710              0              0              0              0
##  ps_ind_10_bin  ps_ind_11_bin  ps_ind_12_bin  ps_ind_13_bin      ps_ind_14
##              0              0              0              0              0
##      ps_ind_15  ps_ind_16_bin  ps_ind_17_bin  ps_ind_18_bin      ps_reg_01
##              0              0              0              0              0
##      ps_reg_02  ps_car_01_cat  ps_car_02_cat  ps_car_04_cat  ps_car_06_cat
##              0            160              5              0              0
##  ps_car_07_cat  ps_car_08_cat  ps_car_09_cat  ps_car_10_cat  ps_car_11_cat
##          17331              0            877              0              0
##      ps_car_11      ps_car_12      ps_car_13      ps_car_14      ps_car_15
##              1              0              0          63805              0
##     ps_calc_01     ps_calc_02     ps_calc_03     ps_calc_04     ps_calc_05
##              0              0              0              0              0
##     ps_calc_06     ps_calc_07     ps_calc_08     ps_calc_09     ps_calc_10
##              0              0              0              0              0
##     ps_calc_11     ps_calc_12     ps_calc_13     ps_calc_14 ps_calc_15_bin
##              0              0              0              0              0
## ps_calc_16_bin ps_calc_17_bin ps_calc_18_bin ps_calc_19_bin ps_calc_20_bin
##              0              0              0              0              0
```

Imputing Missing Values

```
Mode = function (x, na.rm) {
  xtab = table(x)
  xmode = names(which(xtab == max(xtab)))
  if (length(xmode) > 1) xmode = ">1 mode"
  return(xmode)
}
```

```
train$ps_ind_02_cat[is.na(train$ps_ind_02_cat)] = Mode(train$ps_ind_02_cat)
train$ps_ind_04_cat[is.na(train$ps_ind_04_cat)] = Mode(train$ps_ind_04_cat)
train$ps_ind_05_cat[is.na(train$ps_ind_05_cat)] = Mode(train$ps_ind_05_cat)
train$ps_car_01_cat[is.na(train$ps_car_01_cat)] = Mode(train$ps_car_01_cat)
train$ps_car_02_cat[is.na(train$ps_car_02_cat)] = Mode(train$ps_car_02_cat)
train$ps_car_07_cat[is.na(train$ps_car_07_cat)] = Mode(train$ps_car_07_cat)
train$ps_car_09_cat[is.na(train$ps_car_09_cat)] = Mode(train$ps_car_09_cat)
train$ps_car_11[is.na(train$ps_car_11)] = Mode(train$ps_car_11)
train$ps_car_12[is.na(train$ps_car_12)] = mean(train$ps_car_12,na.rm=T)
train$ps_car_14[is.na(train$ps_car_14)] = mean(train$ps_car_14,na.rm=T)
```

```
test$ps_ind_02_cat[is.na(test$ps_ind_02_cat)] = Mode(test$ps_ind_02_cat)
test$ps_ind_04_cat[is.na(test$ps_ind_04_cat)] = Mode(test$ps_ind_04_cat)
test$ps_ind_05_cat[is.na(test$ps_ind_05_cat)] = Mode(test$ps_ind_05_cat)
test$ps_car_01_cat[is.na(test$ps_car_01_cat)] = Mode(test$ps_car_01_cat)
test$ps_car_02_cat[is.na(test$ps_car_02_cat)] = Mode(test$ps_car_02_cat)
test$ps_car_07_cat[is.na(test$ps_car_07_cat)] = Mode(test$ps_car_07_cat)
test$ps_car_09_cat[is.na(test$ps_car_09_cat)] = Mode(test$ps_car_09_cat)
test$ps_car_11[is.na(test$ps_car_11)] = Mode(test$ps_car_11)
test$ps_car_12[is.na(test$ps_car_12)] = mean(test$ps_car_12,na.rm=T)
test$ps_car_14[is.na(test$ps_car_14)] = mean(test$ps_car_14,na.rm=T)
```

```
sum(is.na(train))
```

```
## [1] 0
```

```
sum(is.na(test))
```

```
## [1] 0
```

Sampling Data

```
zero = train[train$target==0,]


one = train[train$target==1,]
len = nrow(one)

zero_sample = sample(1:nrow(zero),len)
length(zero_sample)
```

```
## [1] 21694
```

```
train_new = train[c(zero_sample,row.names(one)),]
nrow(train_new)
```

```
## [1] 43388
```

XG Boost

```
train_mat = model.matrix(~.+0,train_new %>% select(-target))
dmat_train = xgb.DMatrix(train_mat,label=as.numeric(as.character(train_new$target)))

tst_mat = model.matrix(~.+0,test)
dmat_tst = xgb.DMatrix(tst_mat)

param = list(colsample_bytree = 0.8,
             subsample_bytree = 0.7,
             booster="gbtree",
             objective="binary:logistic",
             eta=.2,
             gamma=5,
             max_depth=5,
             eval_metric = "auc",
             nthread = 1)

xg_mod = xgb.train(params = param,
                   data = dmat_train,
                   nrounds = 45)


xg_predict = predict(xg_mod,dmat_tst)
predic = ifelse(xg_predict>0.5,1,0)

output1 = data.frame(id=test$id,target=xg_predict)

write.csv(output1,file="E:\\porto pred\\output.csv",row.names = F)
```

Confusion Matrix

```
tran=train[sample(1:nrow(train),0.7*nrow(train)),]

tst=train[sample(1:nrow(train),0.3*nrow(train)),]


zer = tran[tran$target==0,]
on = tran[tran$target==1,]
leng = nrow(on)

set.seed(1001)
ze = sample(x = row.names(zer),3*leng)



trn = tran[c(ze,row.names(on)),]

trn_mat <- model.matrix(~.+0,trn %>% select(-target))
test_mat <- model.matrix(~.+0,tst %>% select(-target))
dmat_train <- xgb.DMatrix(trn_mat,label=trn$target)
dmat_test <- xgb.DMatrix(test_mat,label=tst$target)

param <- list(colsample_bytree = 0.8,
              subsample_bytree = 0.7,
              booster="gbtree",
              objective="binary:logistic",
              eta=.2,gamma=5,
              max_depth=15,
              eval_metric = "auc")

xg_mod <- xgb.train(params = param,
                    data = dmat_train,
                    nrounds = 50)

xg_predict <- predict(xg_mod,dmat_test)
predic<- ifelse(xg_predict>0.5,1,0)
tst$target = as.factor(tst$target)
tst$predi = as.factor(predic)
```

Kappa and Accuracy

```
cu = confusionMatrix(tst$predi,tst$target,positive = "1")
cu$overall[1]
```

```
##  Accuracy
## 0.9432021
```

```
cu$overall[2]
```

```
##     Kappa
## 0.2845605
```

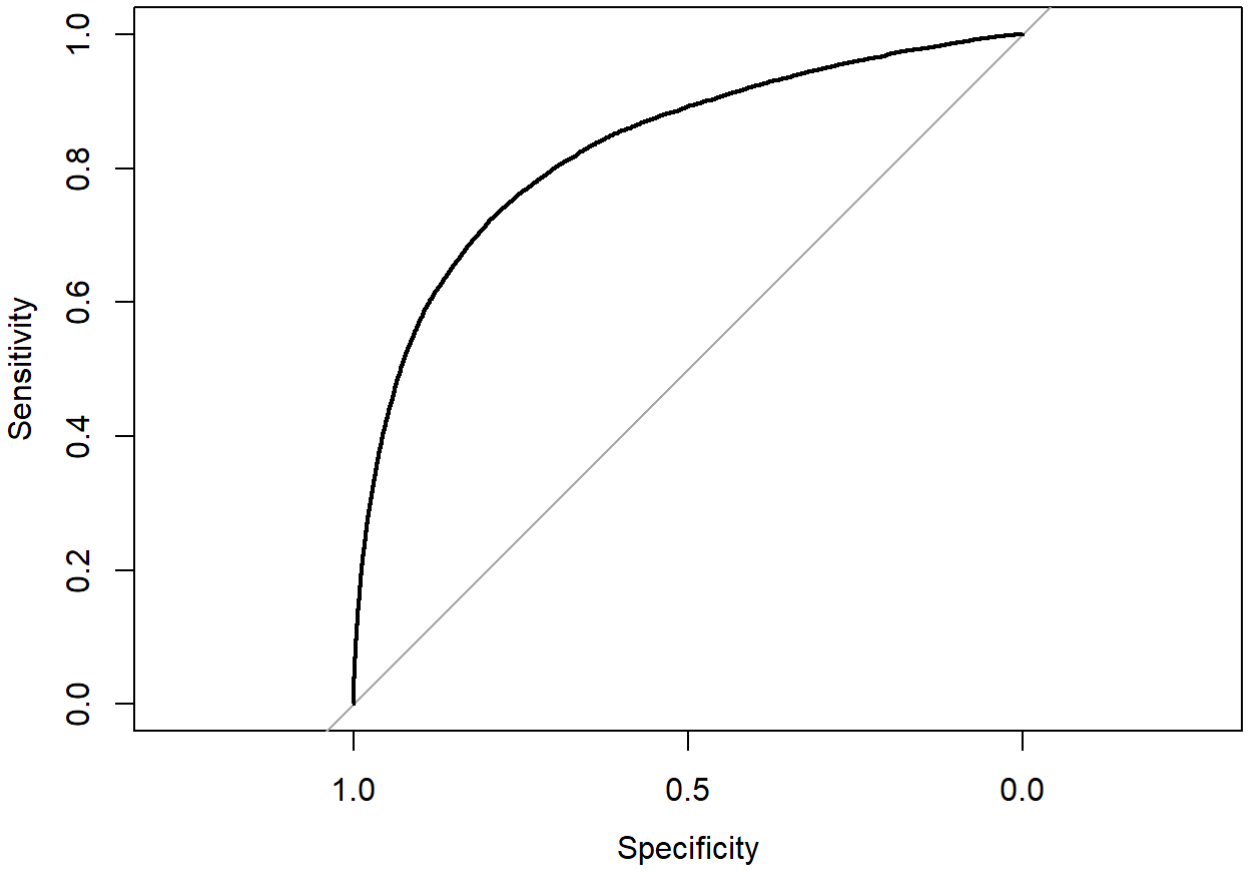
```
#F1 Score = 2*(Recall * Precision) / (Recall + Precision)

precisi = cu$byClass[5]
recall = cu$byClass[6]
f = 2*(recall * precisi) / (recall + precisi)
names(f) = "F1"
f
```

```
##        F1
## 0.3138024
```

ROC

```
x = roc(predictor = xg_predict,response = tst$target)
plot(x)
```



AUC

```
x$auc
```

```
## Area under the curve: 0.8288
```