Project C

```
# R studio API
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
```

Libraries

```
library(tidyverse)
library(caret)
library(careless)
library(xgboost)
```

Data import and cleaning

```
# Data import
c_tbl <- read_csv("../data/project c data.csv") %>%
  mutate(id     = factor(id))
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   Gender = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
# Data cleaning
# identify long strings
  # Personality scale
BFAS_longstring <- c_tbl %>%
  select(BFAS_A_1:BFAS_A_100) %>%
  longstring(., avg = T) %>%
  # add unique ID
  mutate(id               = c_tbl$id) %>%
  # rename longstr varaible
  rename(BFAS_longstr         = longstr) %>%
  select(-avgstr) %>%
  # reorder variables
  select(id, everything())

  # Self-efficacy scale
SE_longstring <- c_tbl %>%
  select(SE_1:SE_16) %>%
  longstring(., avg = T) %>%
  # add unique ID
```

```
  mutate(id              = c_tbl$id) %>%
  # rename longstr varaible
  rename(SE_longstr         = longstr) %>%
  select(-avgstr) %>%
  # reorder variables
  select(id, everything())

  # Test anxiety scale
anx_longstring <- c_tbl %>%
  select(anx_1:anx_10) %>%
  longstring(., avg = T) %>%
  # add unique ID
  mutate(id              = c_tbl$id) %>%
  # rename longstr varaible
  rename(anx_longstr         = longstr) %>%
  select(-avgstr) %>%
  # reorder variables
  select(id, everything())

# Cleaned dataset
c_tbl_cleaned <- c_tbl %>%
  # combine longstring variables to data
  full_join(BFAS_longstring, by = "id") %>%
  full_join(SE_longstring,   by = "id") %>%
  full_join(anx_longstring,  by = "id") %>%
  # drop missing values in dependent variable: Final exam
  drop_na(Final) %>%
  # convert variables to appropriate type
  mutate(Gender = factor(Gender)) %>%
  # remove id column
  select(-id) %>%
  # reorder to make DV as 1st column
  select(Final, everything())
```

Analysis

XGBoost

```
set.seed(511)
training_row <- sample(seq_len(nrow(c_tbl_cleaned)), size = floor(0.8*nrow(c_tbl_cleaned)))
# training dataset
training <- c_tbl_cleaned[training_row,]
# test dataset
test <- c_tbl_cleaned[-training_row,]


# Pre processing data
training_preproc <- preProcess(training[, 2:154],
                               method = c("medianImpute", "scale", "center"))
training_data <- predict(training_preproc, training)
```

```r
# Xgboost
xgb_mod <- train(Final ~ .,
                  data = training_data,
                  # XGBoost
                  method = "xgbLinear",
                  # treat missing values
                  na.action = na.pass,
                  # set cross-validation to be 10 fold
                  trControl = trainControl("cv", number = 10))
```

OLS

```r
lm_mod <- train(Final ~ .,
                  data = training_data,
                  # lm
                  method = "lm",
                  # treat missing values
                  na.action = na.pass,
                  # set cross-validation to be 10 fold
                  trControl = trainControl("cv", number = 10))
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```r
summary(lm_mod)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.870  -3.648   0.053   4.336  20.449
##
## Coefficients: (14 not defined because of singularities)
##                                              Estimate Std. Error t value
## (Intercept)                                 75.866470   0.491540 154.344
## Exam1                                        4.215631   0.617380   6.828
## Exam2                                        2.997457   0.664597   4.510
## Exam3                                        8.500612   0.643594  13.208
## HS_GPA                                       1.044435   0.408324   2.558
## COMP_ACT_SCORE                               0.694911   0.503792   1.379
## TRANS_TOT_CR                                 0.510570   0.406810   1.255
## BFAS_A_1                                     2.932787   2.253267   1.302
## BFAS_A_2                                    -0.103385   0.457177  -0.226
## BFAS_A_3                                    -0.216738   0.476626  -0.455
## BFAS_A_4                                     0.180124   0.519963   0.346
## BFAS_A_5                                    -0.532785   0.488227  -1.091
## BFAS_A_6                                     0.635110   0.538998   1.178
## BFAS_A_7                                     0.080832   0.467412   0.173
## BFAS_A_8                                    -0.017610   0.456882  -0.039
## BFAS_A_9                                     0.436678   0.549782   0.794
## BFAS_A_10                                    0.273037   0.431320   0.633
## BFAS_A_11                                   -2.764270   2.182637  -1.266
## BFAS_A_12                                    0.127435   0.520473   0.245
## BFAS_A_13                                    0.945562   0.530412   1.783
## BFAS_A_14                                    0.270178   0.506129   0.534
## BFAS_A_15                                   -0.306098   0.466574  -0.656
## BFAS_A_16                                    0.466785   0.548794   0.851
## BFAS_A_17                                   -0.546214   0.470563  -1.161
## BFAS_A_18                                    0.505520   0.508093   0.995
## BFAS_A_19                                    0.511996   0.497364   1.029
## BFAS_A_20                                   -0.361061   0.535605  -0.674
## BFAS_A_21                                    2.385145   1.860463   1.282
## BFAS_A_22                                    0.451083   0.499589   0.903
## BFAS_A_23                                   -0.392410   0.537500  -0.730
## BFAS_A_24                                   -0.355601   0.477901  -0.744
## BFAS_A_25                                   -0.137632   0.507179  -0.271
## BFAS_A_26                                    0.782219   0.541618   1.444
## BFAS_A_27                                   -0.295686   0.414439  -0.713
## BFAS_A_28                                   -0.414004   0.514628  -0.804
## BFAS_A_29                                   -0.404503   0.477652  -0.847
## BFAS_A_30                                    0.355571   0.429888   0.827
## BFAS_A_31                                   -3.151001   2.027961  -1.554
## BFAS_A_32                                    0.709984   0.466622   1.522
## BFAS_A_33                                   -0.458714   0.468226  -0.980
## BFAS_A_34                                    0.262689   0.508852   0.516
## BFAS_A_35                                    0.517100   0.470442   1.099
## BFAS_A_36                                    0.336879   0.496665   0.678
## BFAS_A_37                                   -0.664902   0.485470  -1.370
```

```
## BFAS_A_38  -0.052027  0.533889 -0.097
## BFAS_A_39  -0.361264  0.508074 -0.711
## BFAS_A_40   0.194333  0.448810  0.433
## BFAS_A_41   3.299097  2.400076  1.375
## BFAS_A_42   0.341514  0.587679  0.581
## BFAS_A_43  -0.059866  0.519882 -0.115
## BFAS_A_44  -0.760806  0.483261 -1.574
## BFAS_A_45   0.291558  0.437824  0.666
## BFAS_A_46  -0.738640  0.564113 -1.309
## BFAS_A_47  -0.004057  0.442685 -0.009
## BFAS_A_48   0.441416  0.492919  0.896
## BFAS_A_49   0.337607  0.561618  0.601
## BFAS_A_50  -0.917309  0.480070 -1.911
## BFAS_A_51  -3.563735  2.011085 -1.772
## BFAS_A_52  -0.200247  0.516635 -0.388
## BFAS_A_53   0.713140  0.492214  1.449
## BFAS_A_54  -0.639630  0.448465 -1.426
## BFAS_A_55   0.392177  0.477451  0.821
## BFAS_A_56   0.192479  0.475552  0.405
## BFAS_A_57  -0.516177  0.455241 -1.134
## BFAS_A_58   0.035542  0.478124  0.074
## BFAS_A_59   0.163394  0.549761  0.297
## BFAS_A_60  -0.850998  0.455176 -1.870
## BFAS_A_61  -2.176133  2.106125 -1.033
## BFAS_A_62   0.212199  0.479145  0.443
## BFAS_A_63   0.362414  0.472247  0.767
## BFAS_A_64  -0.352013  0.490595 -0.718
## BFAS_A_65   0.557750  0.496202  1.124
## BFAS_A_66   0.344678  0.622629  0.554
## BFAS_A_67   0.758817  0.477040  1.591
## BFAS_A_68  -0.299286  0.469655 -0.637
## BFAS_A_69   0.004460  0.455963  0.010
## BFAS_A_70   0.202976  0.474020  0.428
## BFAS_A_71   2.883904  2.189006  1.317
## BFAS_A_72   0.004506  0.508686  0.009
## BFAS_A_73   0.655486  0.473535  1.384
## BFAS_A_74   0.464689  0.475588  0.977
## BFAS_A_75  -0.690007  0.495016 -1.394
## BFAS_A_76  -0.778898  0.514672 -1.513
## BFAS_A_77   0.580344  0.469625  1.236
## BFAS_A_78  -0.246796  0.454377 -0.543
## BFAS_A_79   0.501116  0.488258  1.026
## BFAS_A_80  -0.540058  0.488489 -1.106
## BFAS_A_81  -2.667568  2.097710 -1.272
## BFAS_A_82  -0.031430  0.449834 -0.070
## BFAS_A_83   0.502235  0.469797  1.069
## BFAS_A_84  -0.086651  0.514070 -0.169
## BFAS_A_85  -0.296168  0.508433 -0.583
## BFAS_A_86  -0.026612  0.552301 -0.048
## BFAS_A_87  -0.255934  0.494509 -0.518
## BFAS_A_88  -0.505397  0.468690 -1.078
## BFAS_A_89   0.578696  0.491739  1.177
## BFAS_A_90   0.709477  0.482119  1.472
## BFAS_A_91  -3.053313  2.118226 -1.441
```

```
## BFAS_A_92                                        -0.056678   0.466177  -0.122
## BFAS_A_93                                        -0.297854   0.468344  -0.636
## BFAS_A_94                                        -0.152065   0.501058  -0.303
## BFAS_A_95                                        -0.495248   0.464416  -1.066
## BFAS_A_96                                        -0.222362   0.504156  -0.441
## BFAS_A_97                                        -0.412823   0.439437  -0.939
## BFAS_A_98                                        -0.340200   0.467943  -0.727
## BFAS_A_99                                         0.207038   0.488916   0.423
## BFAS_A_100                                       -1.153003   0.494458  -2.332
## BFAS_Withdrawal                                  18.242964  13.109926   1.392
## BFAS_Volatility                                         NA         NA      NA
## BFAS_5_N                                                NA         NA      NA
## BFAS_Compassion                                         NA         NA      NA
## BFAS_Politeness                                         NA         NA      NA
## BFAS_5_A                                                NA         NA      NA
## BFAS_Industriousness                                    NA         NA      NA
## BFAS_Orderliness                                        NA         NA      NA
## BFAS_5_C                                                NA         NA      NA
## BFAS_Enthusiasm                                         NA         NA      NA
## BFAS_Assertiveness                                      NA         NA      NA
## BFAS_5_E                                                NA         NA      NA
## BFAS_Intellect                                          NA         NA      NA
## BFAS_Openness                                           NA         NA      NA
## BFAS_5_O                                                NA         NA      NA
## SE_1                                              0.285941   0.599894   0.477
## SE_2                                              0.477737   0.513070   0.931
## SE_3                                              0.492383   0.590333   0.834
## SE_4                                              0.296404   0.556910   0.532
## SE_5                                              1.090473   0.608917   1.791
## SE_6                                             -1.001878   0.603675  -1.660
## SE_7                                             -0.047313   0.518281  -0.091
## SE_8                                             -0.038050   0.534501  -0.071
## SE_9                                              0.588384   0.617487   0.953
## SE_10                                             0.700095   0.710953   0.985
## SE_11                                            -0.218111   0.559421  -0.390
## SE_12                                             0.501176   0.755256   0.664
## SE_13                                             0.345520   0.687469   0.503
## SE_14                                             1.352819   0.615581   2.198
## SE_15                                            -0.655073   0.639307  -1.025
## SE_16                                            -0.356446   0.699707  -0.509
## SE_total                                         -1.571482   2.720277  -0.578
## anx_1                                            -0.987686   0.663915  -1.488
## anx_2                                            -0.696662   0.733947  -0.949
## anx_3                                            -0.068077   0.760319  -0.090
## anx_4                                            -0.490145   0.760537  -0.644
## anx_5                                            -0.291324   0.717409  -0.406
## anx_6                                            -0.837412   0.795892  -1.052
## anx_7                                             0.510422   0.775104   0.659
## anx_8                                            -0.544927   0.790532  -0.689
## anx_9                                            -0.599267   0.778561  -0.770
## anx_10                                           -0.793429   0.766022  -1.036
## anxiety_total                                     3.022281   3.941469   0.767
## `GenderI prefer to not answer this question.` -8.789992   3.495701  -2.515
## GenderMale                                       -0.905185   0.920823  -0.983
```

```
## `GenderTransgender/non-binary/gender fluid`    4.125398   3.902009   1.057
## BFAS_longstr                                   0.442921   0.390283   1.135
## SE_longstr                                    -0.220002   0.441985  -0.498
## anx_longstr                                   -0.760786   0.383238  -1.985
##                                               Pr(>|t|)
## (Intercept)                                    < 2e-16 ***
## Exam1                                         2.43e-11 ***
## Exam2                                         8.03e-06 ***
## Exam3                                          < 2e-16 ***
## HS_GPA                                          0.0108 *
## COMP_ACT_SCORE                                  0.1684
## TRANS_TOT_CR                                    0.2100
## BFAS_A_1                                        0.1936
## BFAS_A_2                                        0.8212
## BFAS_A_3                                        0.6495
## BFAS_A_4                                        0.7292
## BFAS_A_5                                        0.2757
## BFAS_A_6                                        0.2392
## BFAS_A_7                                        0.8628
## BFAS_A_8                                        0.9693
## BFAS_A_9                                        0.4274
## BFAS_A_10                                       0.5270
## BFAS_A_11                                       0.2059
## BFAS_A_12                                       0.8067
## BFAS_A_13                                       0.0752 .
## BFAS_A_14                                       0.5937
## BFAS_A_15                                       0.5121
## BFAS_A_16                                       0.3954
## BFAS_A_17                                       0.2463
## BFAS_A_18                                       0.3202
## BFAS_A_19                                       0.3038
## BFAS_A_20                                       0.5005
## BFAS_A_21                                       0.2004
## BFAS_A_22                                       0.3670
## BFAS_A_23                                       0.4657
## BFAS_A_24                                       0.4572
## BFAS_A_25                                       0.7862
## BFAS_A_26                                       0.1493
## BFAS_A_27                                       0.4759
## BFAS_A_28                                       0.4215
## BFAS_A_29                                       0.3975
## BFAS_A_30                                       0.4086
## BFAS_A_31                                       0.1209
## BFAS_A_32                                       0.1287
## BFAS_A_33                                       0.3277
## BFAS_A_34                                       0.6059
## BFAS_A_35                                       0.2722
## BFAS_A_36                                       0.4979
## BFAS_A_37                                       0.1714
## BFAS_A_38                                       0.9224
## BFAS_A_39                                       0.4774
## BFAS_A_40                                       0.6652
## BFAS_A_41                                       0.1699
## BFAS_A_42                                       0.5614
```

```
## BFAS_A_43                                0.9084
## BFAS_A_44                                0.1160
## BFAS_A_45                                0.5058
## BFAS_A_46                                0.1910
## BFAS_A_47                                0.9927
## BFAS_A_48                                0.3709
## BFAS_A_49                                0.5480
## BFAS_A_50                                0.0566 .
## BFAS_A_51                                0.0770 .
## BFAS_A_52                                0.6985
## BFAS_A_53                                0.1480
## BFAS_A_54                                0.1544
## BFAS_A_55                                0.4118
## BFAS_A_56                                0.6858
## BFAS_A_57                                0.2574
## BFAS_A_58                                0.9408
## BFAS_A_59                                0.7664
## BFAS_A_60                                0.0621 .
## BFAS_A_61                                0.3020
## BFAS_A_62                                0.6580
## BFAS_A_63                                0.4432
## BFAS_A_64                                0.4734
## BFAS_A_65                                0.2615
## BFAS_A_66                                0.5801
## BFAS_A_67                                0.1123
## BFAS_A_68                                0.5242
## BFAS_A_69                                0.9922
## BFAS_A_70                                0.6687
## BFAS_A_71                                0.1883
## BFAS_A_72                                0.9929
## BFAS_A_73                                0.1669
## BFAS_A_74                                0.3290
## BFAS_A_75                                0.1639
## BFAS_A_76                                0.1308
## BFAS_A_77                                0.2171
## BFAS_A_78                                0.5873
## BFAS_A_79                                0.3052
## BFAS_A_80                                0.2694
## BFAS_A_81                                0.2041
## BFAS_A_82                                0.9443
## BFAS_A_83                                0.2855
## BFAS_A_84                                0.8662
## BFAS_A_85                                0.5605
## BFAS_A_86                                0.9616
## BFAS_A_87                                0.6050
## BFAS_A_88                                0.2814
## BFAS_A_89                                0.2398
## BFAS_A_90                                0.1417
## BFAS_A_91                                0.1501
## BFAS_A_92                                0.9033
## BFAS_A_93                                0.5251
## BFAS_A_94                                0.7616
## BFAS_A_95                                0.2867
## BFAS_A_96                                0.6594
```

```
## BFAS_A_97                                    0.3479
## BFAS_A_98                                    0.4675
## BFAS_A_99                                    0.6721
## BFAS_A_100                                   0.0201 *
## BFAS_Withdrawal                              0.1647
## BFAS_Volatility                                  NA
## BFAS_5_N                                         NA
## BFAS_Compassion                                  NA
## BFAS_Politeness                                  NA
## BFAS_5_A                                         NA
## BFAS_Industriousness                             NA
## BFAS_Orderliness                                 NA
## BFAS_5_C                                         NA
## BFAS_Enthusiasm                                  NA
## BFAS_Assertiveness                               NA
## BFAS_5_E                                         NA
## BFAS_Intellect                                   NA
## BFAS_Openness                                    NA
## BFAS_5_O                                         NA
## SE_1                                         0.6338
## SE_2                                         0.3522
## SE_3                                         0.4046
## SE_4                                         0.5948
## SE_5                                         0.0739 .
## SE_6                                         0.0976 .
## SE_7                                         0.9273
## SE_8                                         0.9433
## SE_9                                         0.3411
## SE_10                                        0.3252
## SE_11                                        0.6968
## SE_12                                        0.5073
## SE_13                                        0.6155
## SE_14                                        0.0284 *
## SE_15                                        0.3060
## SE_16                                        0.6107
## SE_total                                     0.5637
## anx_1                                        0.1374
## anx_2                                        0.3430
## anx_3                                        0.9287
## anx_4                                        0.5196
## anx_5                                        0.6849
## anx_6                                        0.2932
## anx_7                                        0.5105
## anx_8                                        0.4909
## anx_9                                        0.4418
## anx_10                                       0.3008
## anxiety_total                                0.4436
## `GenderI prefer to not answer this question.`  0.0122 *
## GenderMale                                   0.3261
## `GenderTransgender/non-binary/gender fluid`  0.2909
## BFAS_longstr                                 0.2570
## SE_longstr                                   0.6189
## anx_longstr                                  0.0477 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.1 on 515 degrees of freedom
##   (52 observations deleted due to missingness)
## Multiple R-squared:  0.8066, Adjusted R-squared:  0.7537
## F-statistic: 15.24 on 141 and 515 DF,  p-value: < 2.2e-16
```

Evaluation

```
test_preproc <- preProcess(test[, 2:154],
                           method = c("medianImpute", "scale", "center"))
test_data <- predict(test_preproc, test)
# XGB predict
xgb_predict <- predict(xgb_mod, test_data, na.action = na.pass)
# lm predict
lm_predict <- predict(lm_mod, test_data, na.action = na.pass)
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient fit
## may be misleading
```

```
# correlation between predicted y and y in test set
xgb_cor <- cor(xgb_predict, test_data$Final, use = "complete.obs")
lm_cor <- cor(lm_predict, test_data$Final, use = "complete.obs")
(cor <- list(OLS = round(lm_cor,2),
             XGB = round(xgb_cor,2)))
```

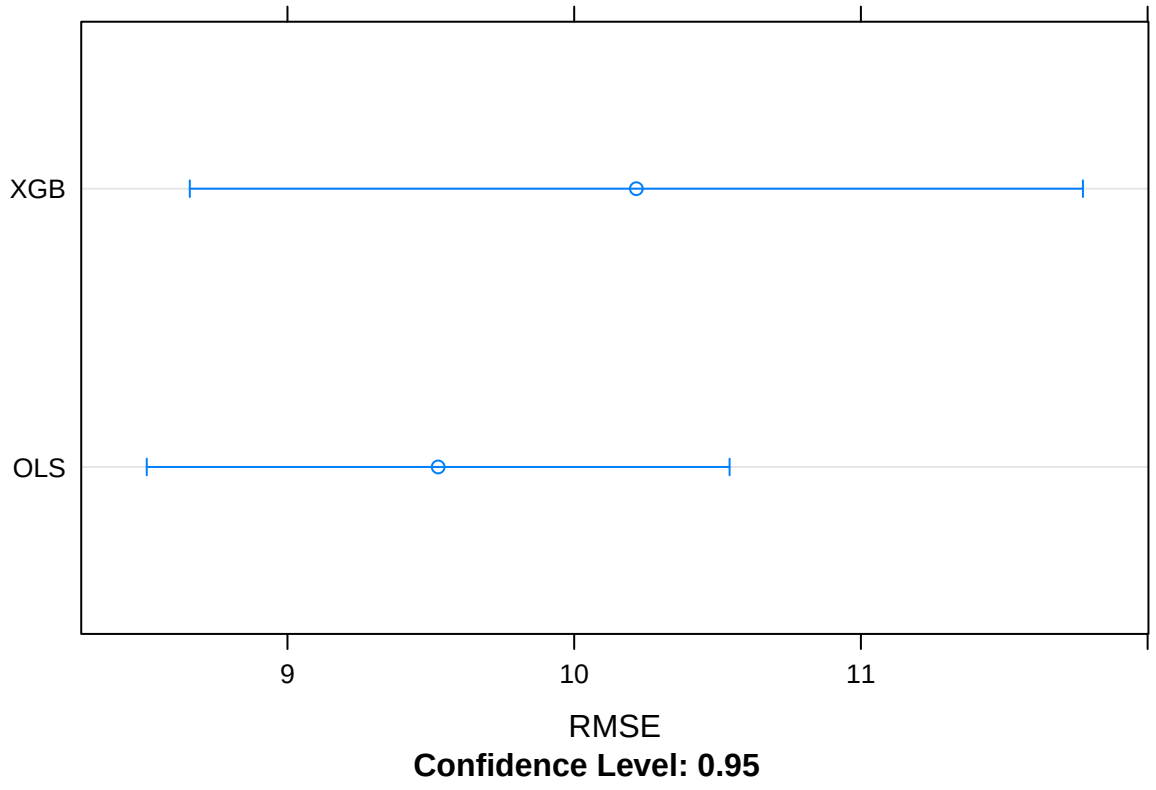
```
## $OLS
## [1] 0.83
##
## $XGB
## [1] 0.87
```

```
# Comparing RMSE and R-squared
summary(resamples(list(OLS = lm_mod,
                       XGB = xgb_mod)))
```

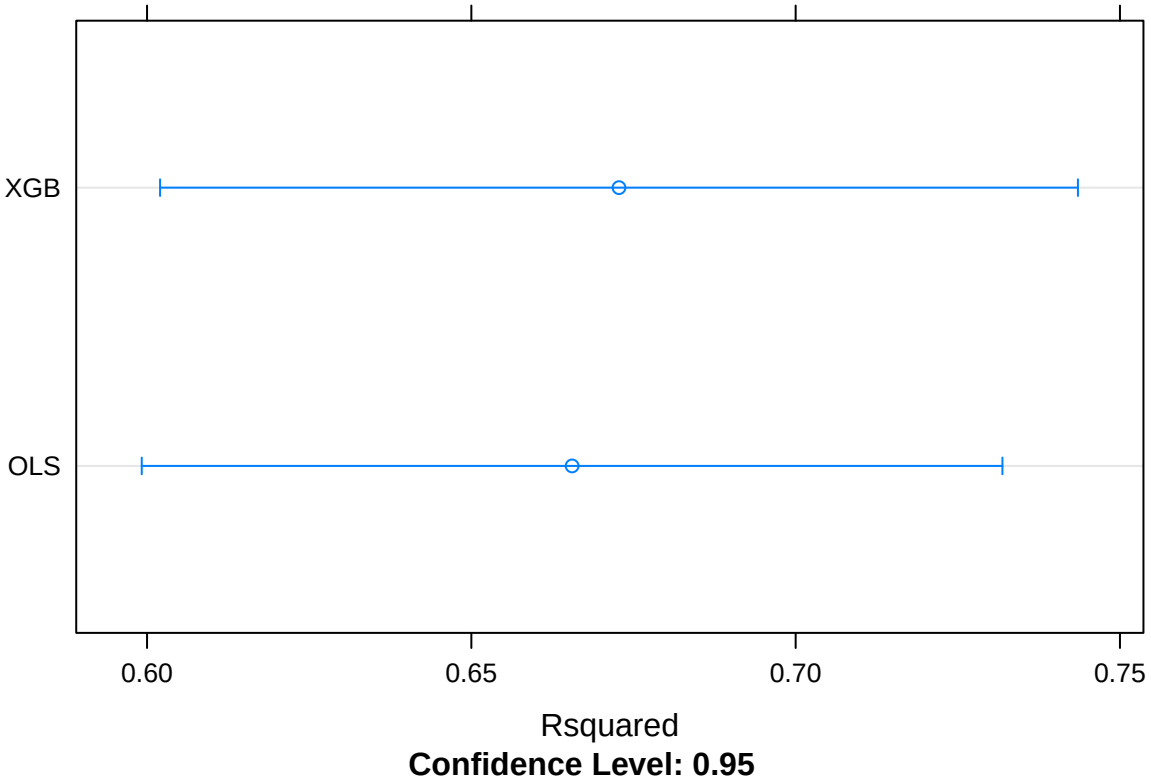
```
##
## Call:
## summary.resamples(object = resamples(list(OLS = lm_mod, XGB = xgb_mod)))
##
## Models: OLS, XGB
## Number of resamples: 10
##
## MAE
##         Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## OLS 5.641801 6.397241 6.884397 6.796746 7.177508 7.548654    0
## XGB 5.513959 6.529132 7.260011 7.133643 7.491283 8.690355    0
##
## RMSE
##         Min.  1st Qu.   Median      Mean   3rd Qu.     Max. NA's
## OLS 7.119486 8.890665 9.477239  9.525647  9.916739 12.00183    0
## XGB 7.711214 8.420654 9.869469 10.216810 11.280339 14.12245    0
##
## Rsquared
```

```
##          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## OLS 0.4900355 0.6198587 0.6743851 0.6655303 0.7118295 0.8410604    0
## XGB 0.5123270 0.6116018 0.6780337 0.6727609 0.7348700 0.8098053    0
```

```
dotplot(resamples(list(OLS = lm_mod,
                       XGB = xgb_mod)), metric="RMSE")
```




```
dotplot(resamples(list(OLS = lm_mod,
                       XGB = xgb_mod)), metric="Rsquared")
```



Data and intepretation

The dataset includes academic performance of college students enrolled in Introduction to Psychology course. Their exam scores, HSGPA, ACT scores, number of transfer credits, and gender are predictors of Final exam score. Some individual difference variables are also available: item-level and scale-level personality (coded as BFAS), self-efficacy (coded as SE), and test anxiety (coded as anx). Since both item- and scale-level data are in the dataset, using OLS would return rank-deficient results since scale-level predictors are linear transformation of item-level predictors. Thus, using machine learning is more beneficial.

The question of interest: What predicts performance in students' final exam?

I include all variables as predictors, but no interaction among predictors to reduce the complexity of the analysis. Previous exams are good predictors of performance on final exam. Individual differences can also predict academic performance.