Final Project ML Model Creation

Jackson Gasperack & Sawan Pandita

2025-11-26

Table of contents


```
load(file = "EDA.RData")
```

Front Matter

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.2
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- 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 becom
Loading required package: lattice


Attaching package: 'caret'
```

```
The following object is masked from 'package:purrr':

    lift

Loading required package: Matrix

Attaching package: 'Matrix'

The following objects are masked from 'package:tidyr':

    expand, pack, unpack

Loaded glmnet 4.1-10

randomForest 4.7-1.2

Type rfNews() to see new features/changes/bug fixes.

Attaching package: 'randomForest'

The following object is masked from 'package:dplyr':

    combine

The following object is masked from 'package:ggplot2':

    margin

Type 'citation("pROC")' for a citation.

Attaching package: 'pROC'
```

```
The following objects are masked from 'package:stats':

    cov, smooth, var


Loading required package: rlang


Attaching package: 'rlang'


The following objects are masked from 'package:purrr':

    %@%, flatten, flatten_chr, flatten_dbl, flatten_int, flatten_lgl,
    flatten_raw, invoke, splice



Attaching package: 'xgboost'


The following object is masked from 'package:dplyr':

    slice



Attaching package: 'e1071'


The following object is masked from 'package:ggplot2':

    element
```

Mentioned in class:

Models: Simple/Multiple Regression, Logistic Regression, Binomial Regression, k-Nearest Neighbors (Regression/Classification), Naïve Bayes, SVM, Spline Regression, Neural Networks

Model Evaluations: RMSE, k-fold Cross Validation, Confusion Matrix (Accuracy, Recall, …), ROC Curves, AUC, R-squared, Subset Selection (Forward, Backward), Shrinkage (LASSO, Ridge), AIC, BIC, MAE, Bootstrap Validation

NOT Mentioned in class:

Models: RandomForest, XGBoost, Poisson/Gamma Regression, t-SNE, PCA, k-Means

Model Evaluations: Precision-Recall Curve, F1 Score, MAPE

Question

Can we use the stats of the last 12 years to predict who will be in the final four for March Madness?

New variable

Find best feature set

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Forward model is the one we'll move forward with, lowest AIC. X = W, ADJOE, TOR, FTR, 2P_D, 2P_O, and BARTHAG but since it's not statistically significant we will try and replace it with some other variables

```
Call:
glm(formula = isFinalFour ~ W + ADJOE + TOR + FTR + `2P_D` +
    `2P_O` + ADJDE + TORD, family = "binomial", data = cbb)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -25.96295    9.68782  -2.680  0.00736 **
W             0.40978    0.07573   5.411 6.27e-08 ***
ADJOE         0.28057    0.05482   5.118 3.09e-07 ***
TOR           0.30914    0.14199   2.177  0.02947 *
FTR          -0.10084    0.04266  -2.364  0.01809 *
`2P_D`        0.24083    0.10260   2.347  0.01891 *
`2P_O`       -0.27311    0.08755  -3.119  0.00181 **
ADJDE        -0.18132    0.08122  -2.232  0.02559 *
TORD         -0.11019    0.09345  -1.179  0.23836
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 481.80  on 3884  degrees of freedom
Residual deviance: 191.16  on 3876  degrees of freedom
AIC: 209.16

Number of Fisher Scoring iterations: 10
```

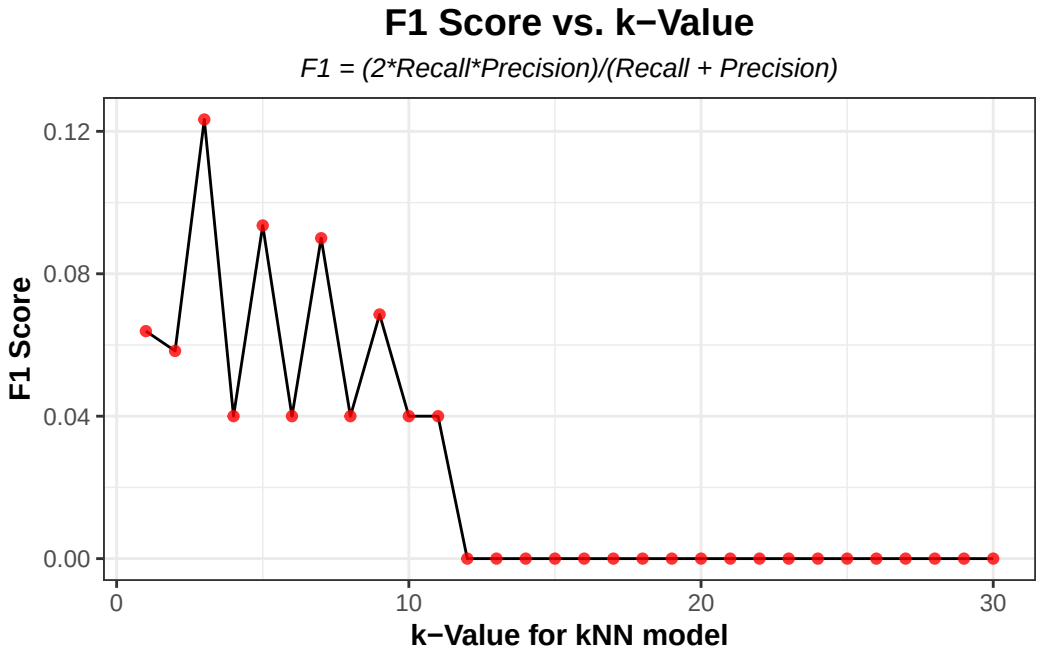
Found model1 adds two more predictors to reduce underfitting while also decreasing AIC

kNN

```
[1] 3
```

```
[1] 0.1233333
```

```
[1] 0.9922879
```



Neural Networks

```
# weights:  51
initial  value 2250.650309
iter  10 value 101.817811
iter  20 value 79.847059
iter  30 value 78.172895
iter  40 value 75.805273
iter  50 value 72.640532
iter  60 value 66.737695
iter  70 value 58.584165
iter  80 value 52.304031
iter  90 value 49.773489
iter 100 value 46.391530
iter 110 value 42.592233
iter 120 value 41.398456
iter 130 value 37.755815
iter 140 value 35.409370
iter 150 value 33.897085
iter 160 value 32.233844
iter 170 value 31.582054
iter 180 value 31.066695
iter 190 value 30.497692
```

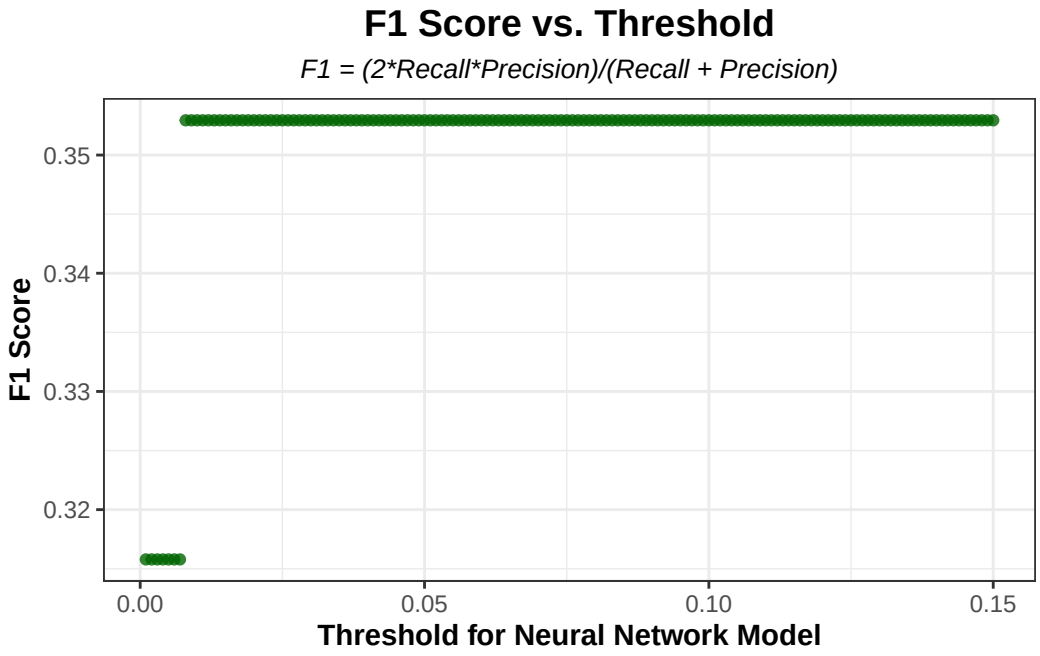
```
iter 200 value 29.259468
iter 210 value 27.584972
iter 220 value 25.109052
iter 230 value 24.602170
iter 240 value 22.845850
iter 250 value 22.062019
iter 260 value 20.612603
iter 270 value 20.218157
iter 280 value 20.044892
iter 290 value 19.770097
iter 300 value 19.308538
iter 310 value 18.990695
iter 320 value 18.912976
iter 330 value 18.806523
iter 340 value 18.582917
iter 350 value 18.543215
iter 360 value 18.508755
iter 370 value 18.489926
iter 380 value 18.395439
iter 390 value 18.350940
iter 400 value 18.338235
iter 410 value 18.323064
iter 420 value 18.305248
iter 430 value 18.275686
iter 440 value 18.272927
iter 450 value 18.271068
iter 460 value 18.269752
iter 470 value 18.269321
iter 480 value 18.269059
iter 490 value 18.268581
iter 500 value 18.268207
iter 510 value 18.267604
iter 520 value 18.266935
iter 530 value 18.266477
iter 540 value 18.266158
iter 550 value 18.265859
iter 560 value 18.265699
iter 570 value 18.265394
iter 580 value 18.265047
iter 590 value 18.264788
iter 600 value 18.263636
iter 610 value 18.262794
iter 620 value 18.260648
```

```
iter 630 value 18.260175
iter 640 value 18.259470
iter 650 value 18.259241
iter 660 value 18.258959
iter 670 value 18.258752
iter 680 value 18.258689
iter 690 value 18.258585
iter 700 value 18.258424
iter 710 value 18.258139
iter 720 value 18.257867
final  value 18.257806
converged
```

```
[1] 0.008
```

```
[1] 0.3529412
```

```
[1] 0.9811321
```



Random Forest

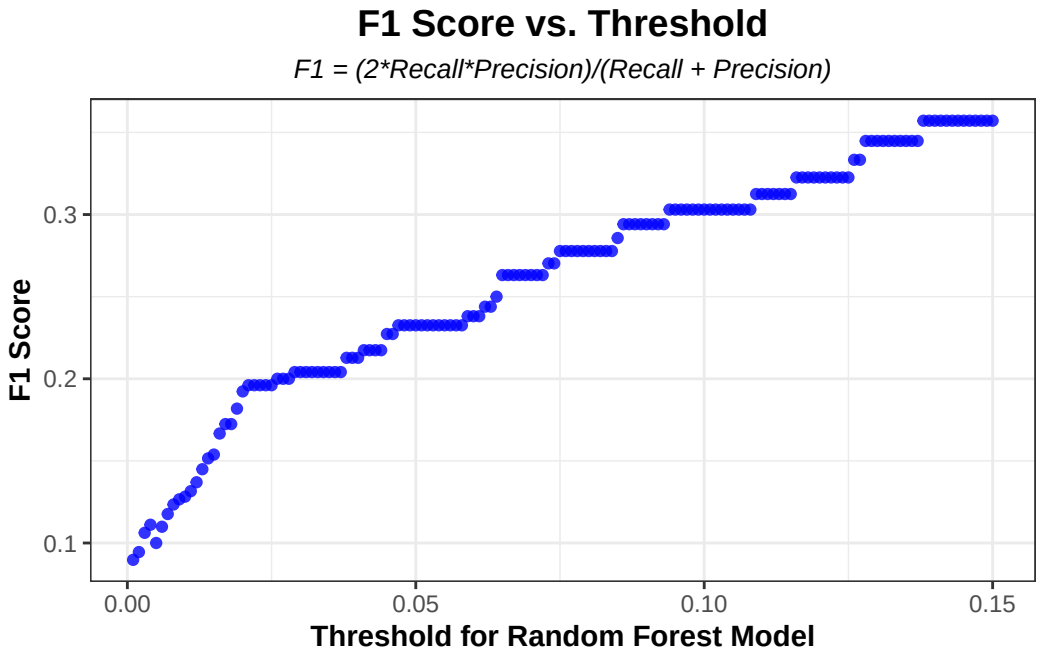
```
Call:
 randomForest(formula = isFinalFour ~ TPR + TPRD + W + ADJOE +      ADJDE + TOR + TORD + FTR
               Type of random forest: classification
                     Number of trees: 5000
No. of variables tried at each split: 2

        OOB estimate of  error rate: 1%
Confusion matrix:
     0 1  class.error
0 3264 1 0.0003062787
1   32 5 0.8648648649
```

```
[1] 0.138
```

```
[1] 0.3571429
```

```
[1] 0.9691252
```

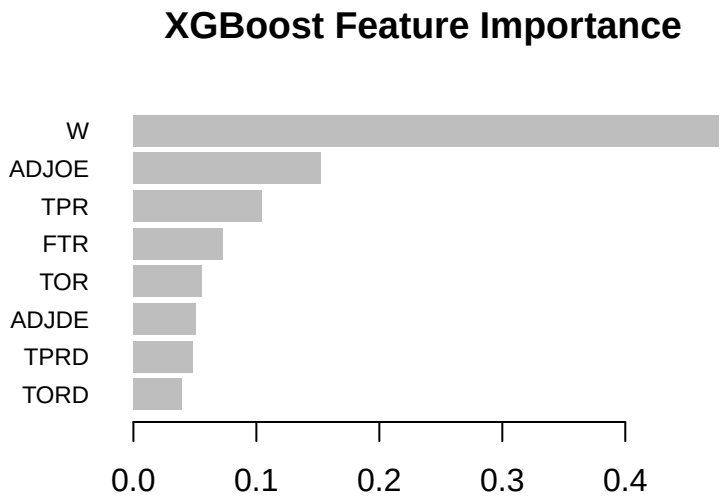



```
[1] "My Best XGBoost Parameters:"
```

```
       objective      eval_metric        max_depth              eta
"binary:logistic"            "auc"              "6"            "0.1"
```

```
        subsample
            "0.8"
```

```
[1] "My Best XGBoost F1 Score: 0.344827586206897"
```




```
[1] "My Best SVM F1 Score: 0.769230769230769"
```

Model Comparison

```
Setting levels: control = 0, case = 1
```

```
Setting direction: controls < cases
```

```
Setting levels: control = 0, case = 1
```

```
Setting direction: controls < cases
```

```
Setting levels: control = 0, case = 1
```

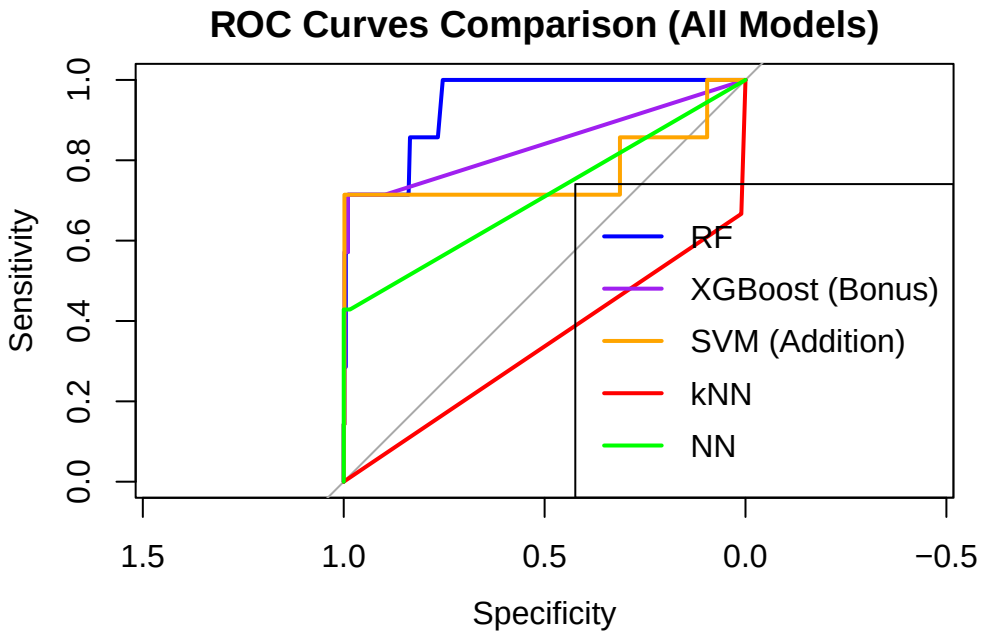
```
Setting direction: controls < cases
```

```
Setting levels: control = 0, case = 1
```

```
Setting direction: controls < cases
```

```
Setting levels: control = 0, case = 1

Setting direction: controls < cases
```




```
[1] "AUC of K-Nearest Neighbors: 0.339"

[1] "AUC of Neural Network: 0.71"

[1] "AUC of Random Forest: 0.939"

[1] "AUC of XGBoost (Mine): 0.84"

[1] "AUC of SVM (Mine): 0.772"

[1] "--- Random Forest ---"

$`F1 Score`
[1] 0.357

$Accuracy
[1] 0.969
```

```
$`Area Under Curve`
[1] 0.939

$`Balanced Accuracy`
[1] 0.843

[1] "--- XGBoost (Bonus) ---"

$`F1 Score`
[1] 0.345

$Accuracy
[1] 0.967

$`Area Under Curve`
[1] 0.84

$`Balanced Accuracy`
[1] 0.842

[1] "--- SVM ---"

$`F1 Score`
[1] 0.769

$Accuracy
[1] 0.995

$`Area Under Curve`
[1] 0.772

$`Balanced Accuracy`
[1] 0.856

[1] "--- kNN ---"

$`F1 Score`
[1] 0.123

$Accuracy
```

```
[1] 0.992

$`Area Under Curve`
[1] 0.339

$`Balanced Accuracy`
[1] 0.5


[1] "--- Neural Network ---"


$`F1 Score`
[1] 0.353

$Accuracy
[1] 0.981

$`Area Under Curve`
[1] 0.71

$`Balanced Accuracy`
[1] 0.708
```