Crafting the Perfect March Madness Bracket

By: Joe Coyne

The goals for my project remained relatively the same throughout the semester. The one slight shift is that I focused less on the individual matchups in the tournament, but rather the overall trends of the tournament. For instance, what teams advance further in the tournament and ultimately win, and what differentiates these teams?

Research Questions

- Central research question: How can we best predict the outcomes and top teams of an NCAA Tournament?
  - Are offensively-minded teams better suited for the tournament over defensively-minded teams?
  - Does conference matter? How strong of predictors are SOS (Strength of Schedule) and SRS (Simple Rating System)?
  - How important are the newer and more advanced statistics to predicting postseason wins?
- If the 2019-2020 season were to have run uninterrupted, how would the tournament unfold and who would end up as the champion?

Research questions remained pretty much the same as well
- Focused mostly on overall trends and general predictability

Data

All data is from 2002-2003 season through the 2019-2020 season

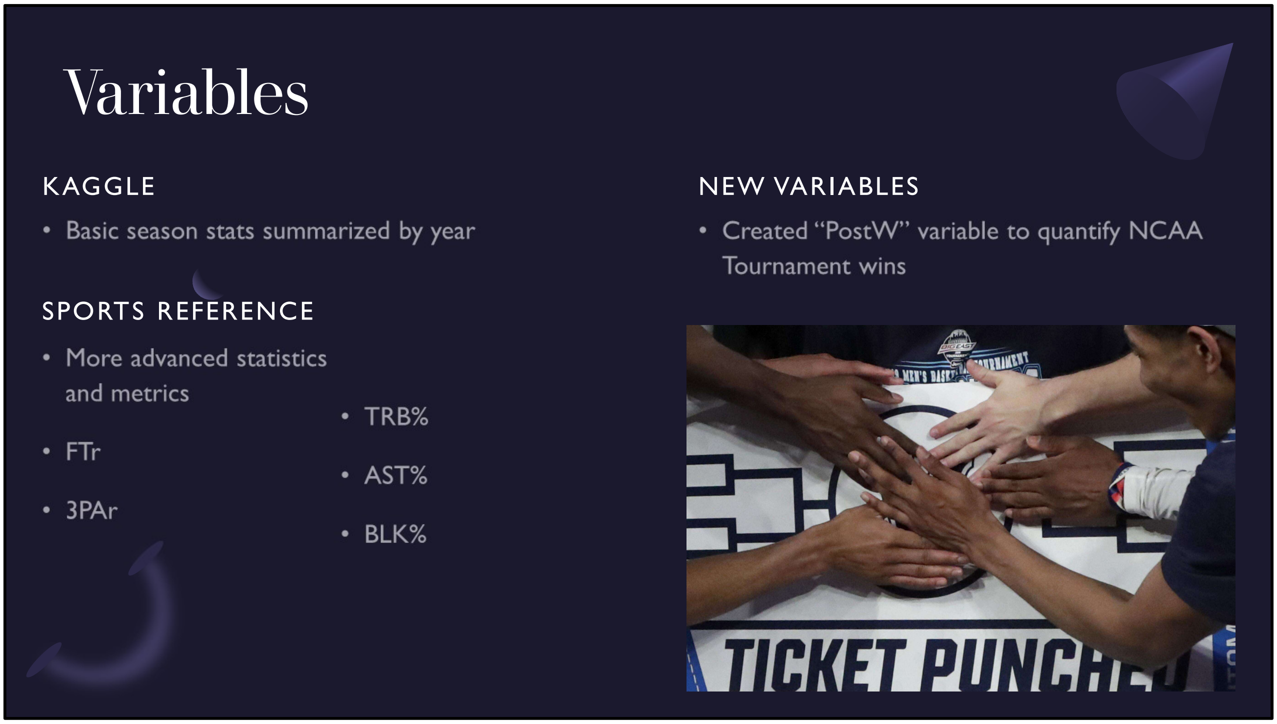
Kaggle (from competition: "March Machine Learning Mania 2021 - NCAAM – Spread")

- Basic statistics
    - Tracked on a game by game basis

Sports Reference (https://www.sports-reference.com/cbb/)

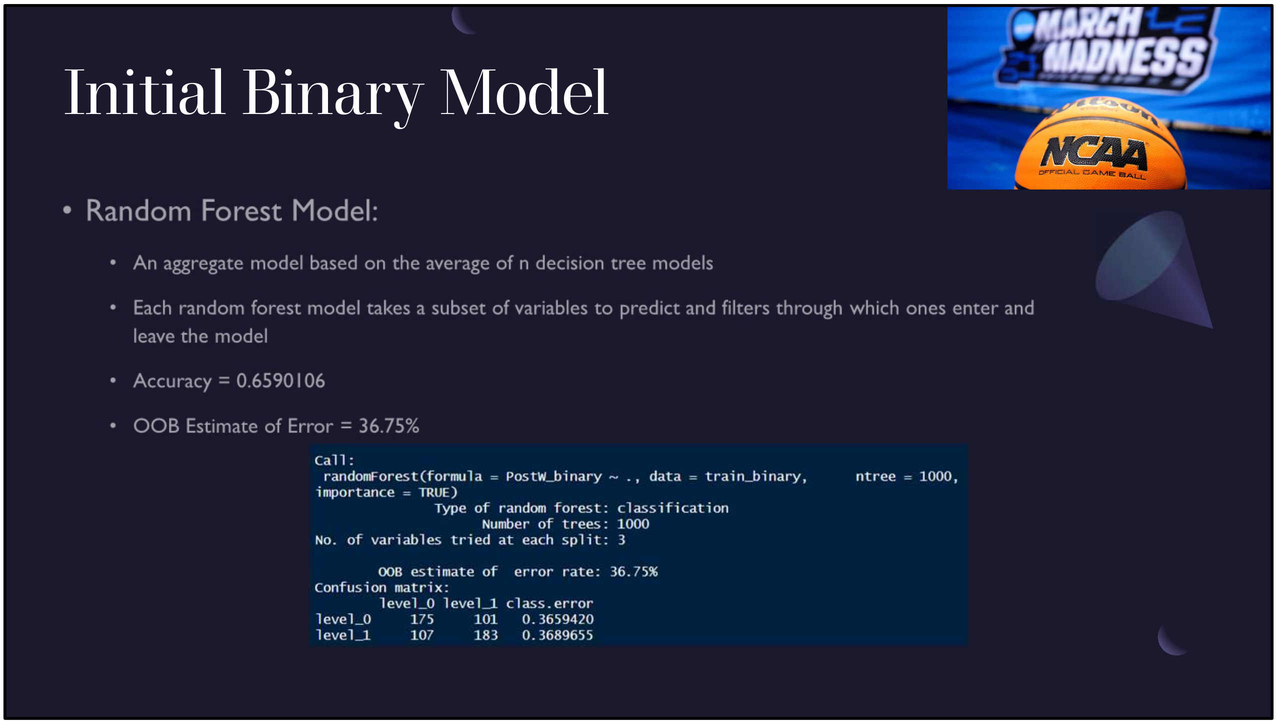
- More advanced metrics
    - Season statistics for each team

Only change from my last presentation is that I added in Sports Reference data



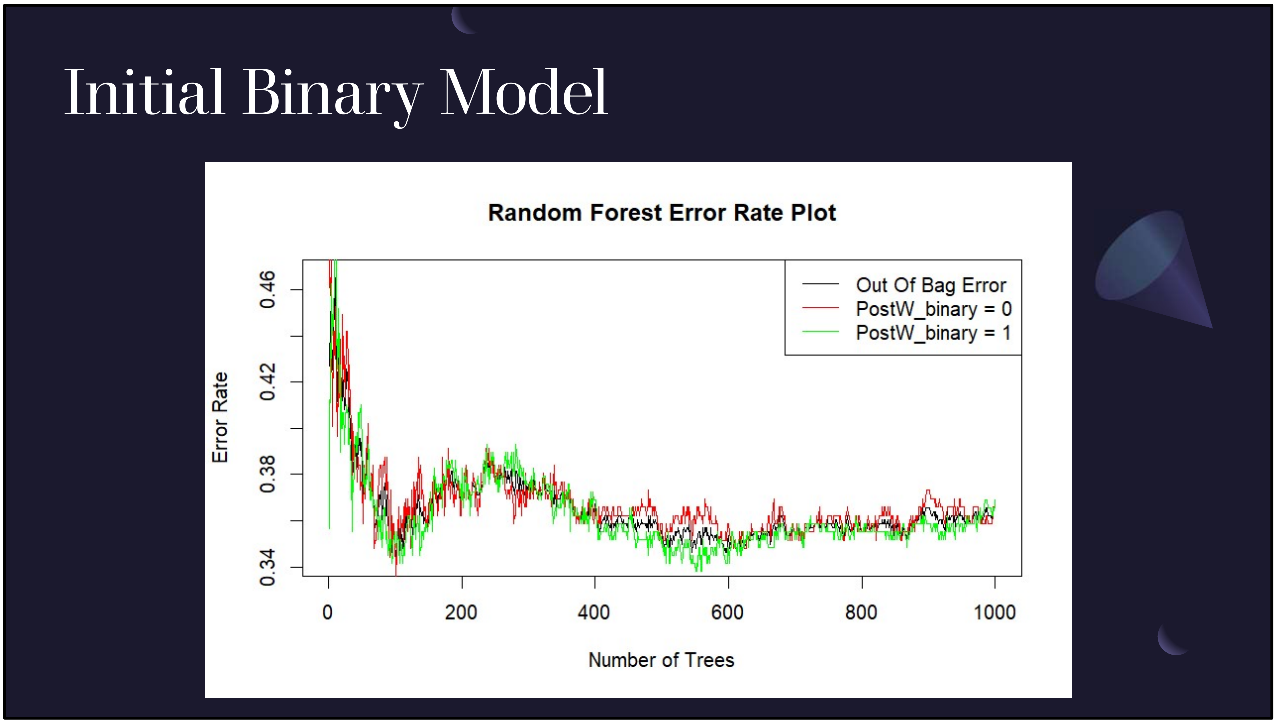
To create the models, I only took teams that made that year's tournament to reduce the class imbalance (there are a lot more teams each year that win 0 postseason games (about 300 teams) vs teams that win at least 1)

Advanced stats that I used:
- FTr – Free Throw Rate (Number of FT Attempts per FG Attempt)
- 3PAr – 3-Point Attempt Rate (% of FG attempts from 3-point range)
- TRB% - Total Rebound Percentage (estimate of % of available rebounds a player grabbed while on the floor)
- AST% - Assist Percentage (estimate of % of teammate field goals a player assisted while on the floor)
- BLK % - Block Percentage (estimate of % of opponent two-point field goal attempts blocked by the player while on the floor)



In order to further address the class imbalance issue, an initial binary model was created (by creating a new variable called PostW_binary, where if the team lost in the first round, it would be denoted with a 0, and all others that won in the first round got a 1)

- The Random Forest Model had the highest accuracy of all the binary models, with an accuracy of 0.6590106
(aggregate model based on the average of n decision tree models; each random forest model takes a subset of variables to predict and filters through which ones enter and leave the model)
- Out of Bag Estimate of 36.75% -> acts as a test or validation data set; this represents the variables that were left out of the specific iteration of the model
- The other binary models I ran (jackknife, stepwise selection) can be found in my paper



X-axis -> number of decision trees built in the random forest model (graph shows how the misclassification / error rate changes depending on the number of decision trees built into the model

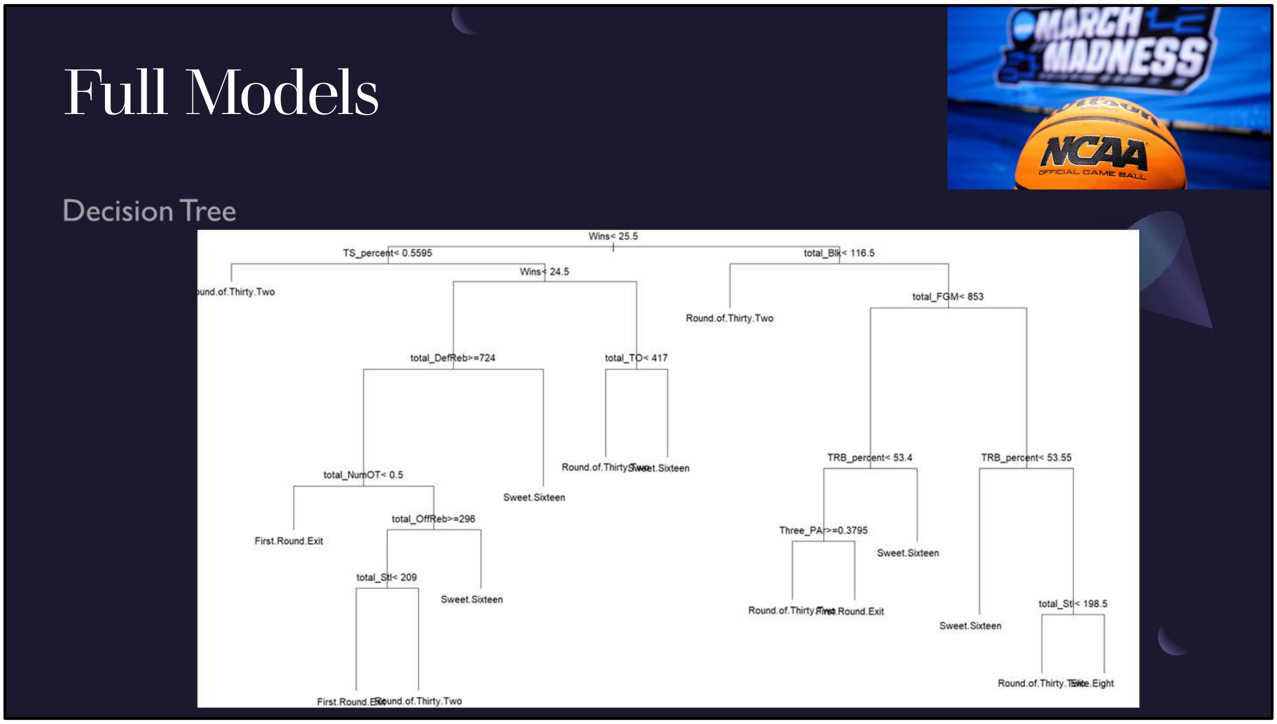
Once again, Out of Bag Error is the baseline / validation error

Applied the binary predictions to the full dataset and filters only teams that were predicted to have at least 1 win

Before creating the full models, some variables were removed for multicollinearity, such as Losses, total_pts, total_FGA, TOV_percent

Similar to K fold cross validation, the jackknife and leave one out cross validation resample by systematically leaving one observation out of the dataset at a time and then calculating the estimate of interest for each subset. So in a dataset with N observations, you would create N subsets, each containing N-1 observations.

Area Under ROC Curve = 0.6318
We use area under the curve instead of accuracy as our performance metric for all the full models, because even though we partially dealt with the class imbalance issue, when all the seasons are pooled together, there is still an overwhelming number of observations in the Round of 32 value. The model is bound to predict some of these wrong, but since it's at a larger scale, these incorrect predictions will skew the accuracy, whereas it won't skew the area under the ROC curve



Interpretation: If you are predicted to have at least 1 postseason win and have more than 25.5 wins, but have less than 116.5 blocks, you will be eliminated in the Round of 32

Full Models

Decision Tree

```
Call:
multiclass.roc.default(response = test$PostW, predictor = pred_prob_tree)

Data: pred_prob_tree with 7 levels of test$PostW: First.Round.Exit, Round.of.Thirty.Two, Sweet.Sixteen, Elite.Eight, Final.Four, Runner.Up, Champion.
Multi-class area under the curve: 0.625
```

```
Confusion Matrix and Statistics

                     Reference
Prediction            First.Round.Exit Round.of.Thirty.Two Sweet.Sixteen Elite.Eight Final.Four Runner.Up Champion
  First.Round.Exit                   2                   3             2           0          1         0        0
  Round.of.Thirty.Two               39                  93            44          23         10         5        2
  Sweet.Sixteen                      2                  18             9           6          3         2        3
  Elite.Eight                        0                   0             2           6          1         0        3
  Final.Four                         0                   0             0           0          0         0        0
  Runner.Up                          0                   0             0           0          0         0        0
  Champion                           0                   0             0           0          0         0        0

Overall Statistics

               Accuracy : 0.3943
                 95% CI : (0.3365, 0.4543)
    No Information Rate : 0.4086
    P-Value [Acc > NIR] : 0.7072

                  Kappa : 0.057

 Mcnemar's Test P-Value : NA
```

Area under the ROC curve = 0.625, which is slightly worse than the jackknife model

Full Models

Stepwise Selection

```
Call:
multiclass.roc.default(response = test$PostW, predictor = rf_pred)

Data: rf_pred with 7 levels of test$PostW: 0, 1, 2, 3, 4, 5, 6.
Multi-class area under the curve: 0.6265
```

```
reg.step_pred         First.Round.Exit Round.of.Thirty.Two Sweet.Sixteen Elite.Eight Final.Four Runner.Up Champion
  First.Round.Exit                   0                   2             0           0          1         0        0
  Round.of.Thirty.Two               42                 107            54          29         11         5        5
  Sweet.Sixteen                      1                   3             3           3          2         2        0
  Elite.Eight                        0                   2             0           3          1         0        3
  Final.Four                         0                   0             0           0          0         0        0
  Runner.Up                          0                   0             0           0          0         0        0
  Champion                           0                   0             0           0          0         0        0
Accuracy: 0.4050179
```

Had to use polr( ) function since target variable PostW is an ordinal variable
- polr = Ordered Logistic or Probit Regression, which fits a logistic or probit regression model to an ordered factor response
- Area under the ROC curve = 0.6265, which is slightly worse than the jackknife model as well

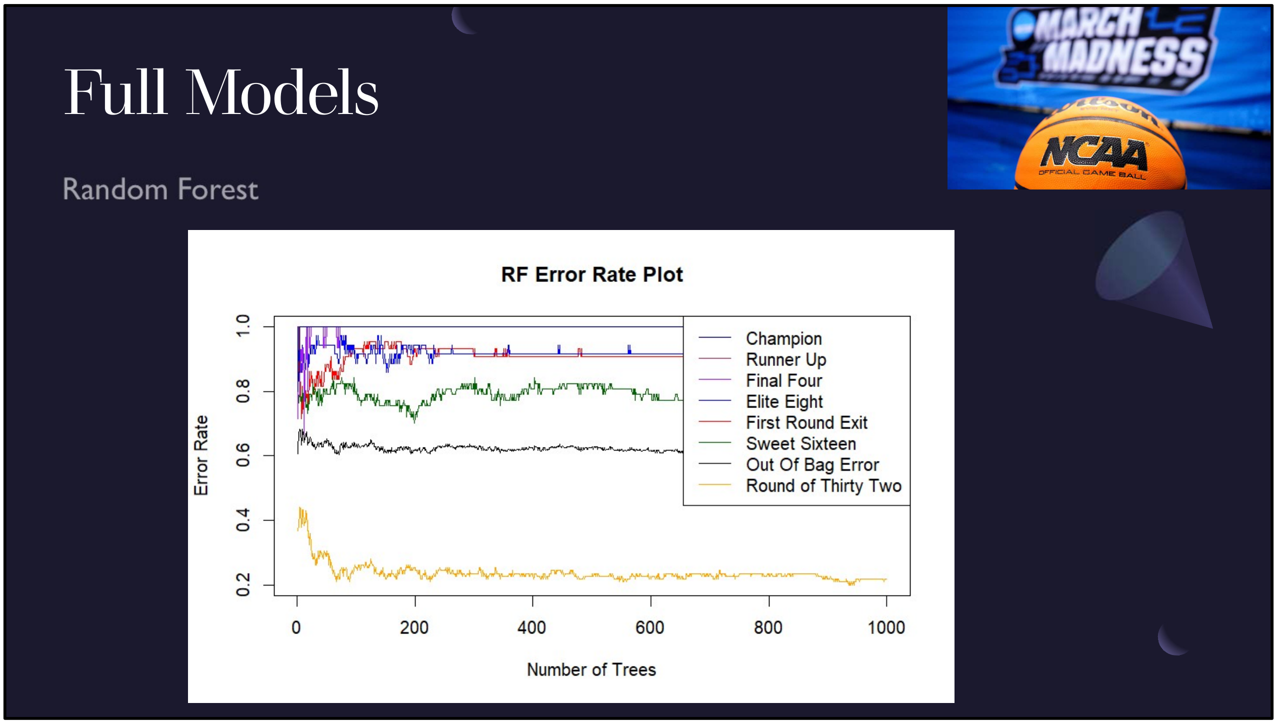
Full Models

Random Forest

```
Call:
 randomForest(formula = PostW ~ ., data = train, ntree = 1000,      importance = TRUE)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 4

        OOB estimate of  error rate: 62.5%
Confusion matrix:
  0  1  2 3 4 5 6 class.error
0 4 34  5 0 0 0 0   0.9069767
1 5 89 18 2 0 0 0   0.2192982
2 7 37 10 3 0 0 0   0.8245614
3 3 18 11 2 0 0 1   0.9428571
4 0 11  2 2 0 0 0   1.0000000
5 1  3  1 2 0 0 0   1.0000000
6 1  0  7 1 0 0 0   1.0000000
```

OOB Error = Out of Bag Error



As you can see, Round of 32 has the lowest misclassification rate. This is most likely because there are so many more observations of this value of PostW than any others

Full Models

Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1  2  3  4  5  6
         0  2  4  2  1  0  0  0
         1 38 91 41 21 10  5  2
         2  3 19 12  7  4  2  3
         3  0  0  2  6  1  0  3
         4  0  0  0  0  0  0  0
         5  0  0  0  0  0  0  0
         6  0  0  0  0  0  0  0

Overall Statistics

               Accuracy : 0.3978
                 95% CI : (0.34, 0.4579)
    No Information Rate : 0.4086
    P-Value [Acc > NIR] : 0.6639

                  Kappa : 0.0713
```

```
Call:
multiclass.roc.default(response = test$PostW, predictor = rf_pred)

Data: rf_pred with 7 levels of test$PostW: 0, 1, 2, 3, 4, 5, 6.
Multi-class area under the curve: 0.6265
```

Confusion matrix is slightly different because this is after the random forest model is applied to the test data

Area under the ROC curve = 0.6265

Jackknife model was the best performing, with an AUC of 0.6318

Jackknife Model Application

Final Predictions (2003 Season)

	PostW <ord>	predicted_PostW <ord>
Kansas	Runner.Up	Champion
Syracuse	Champion	Runner Up
Arizona	Elite.Eight	Final Four
Kentucky	Elite.Eight	Final Four
Butler	Sweet.Sixteen	Elite Eight
Duke	Sweet.Sixteen	Elite Eight
Pittsburgh	Sweet.Sixteen	Elite Eight
Florida	Round.of.Thirty.Two	Elite Eight
Texas	Final.Four	Sweet Sixteen
Oklahoma	Elite.Eight	Sweet Sixteen
Notre Dame	Sweet.Sixteen	Sweet Sixteen
Wisconsin	Sweet.Sixteen	Sweet Sixteen
Gonzaga	Round.of.Thirty.Two	Sweet Sixteen
Louisville	Round.of.Thirty.Two	Sweet Sixteen
Tulsa	Round.of.Thirty.Two	Sweet Sixteen
Xavier	Round.of.Thirty.Two	Sweet Sixteen
Connecticut	Sweet.Sixteen	Round of Thirty Two
California	Round.of.Thirty.Two	Round of Thirty Two
Indiana	Round.of.Thirty.Two	Round of Thirty Two
Missouri	Round.of.Thirty.Two	Round of Thirty Two
Stanford	Round.of.Thirty.Two	Round of Thirty Two
UNC Asheville	Round.of.Thirty.Two	Round of Thirty Two
Utah	Round.of.Thirty.Two	Round of Thirty Two

- I predicted Kansas to beat Syracuse in the title game, but Syracuse ended up winning (not a huge discrepancy)

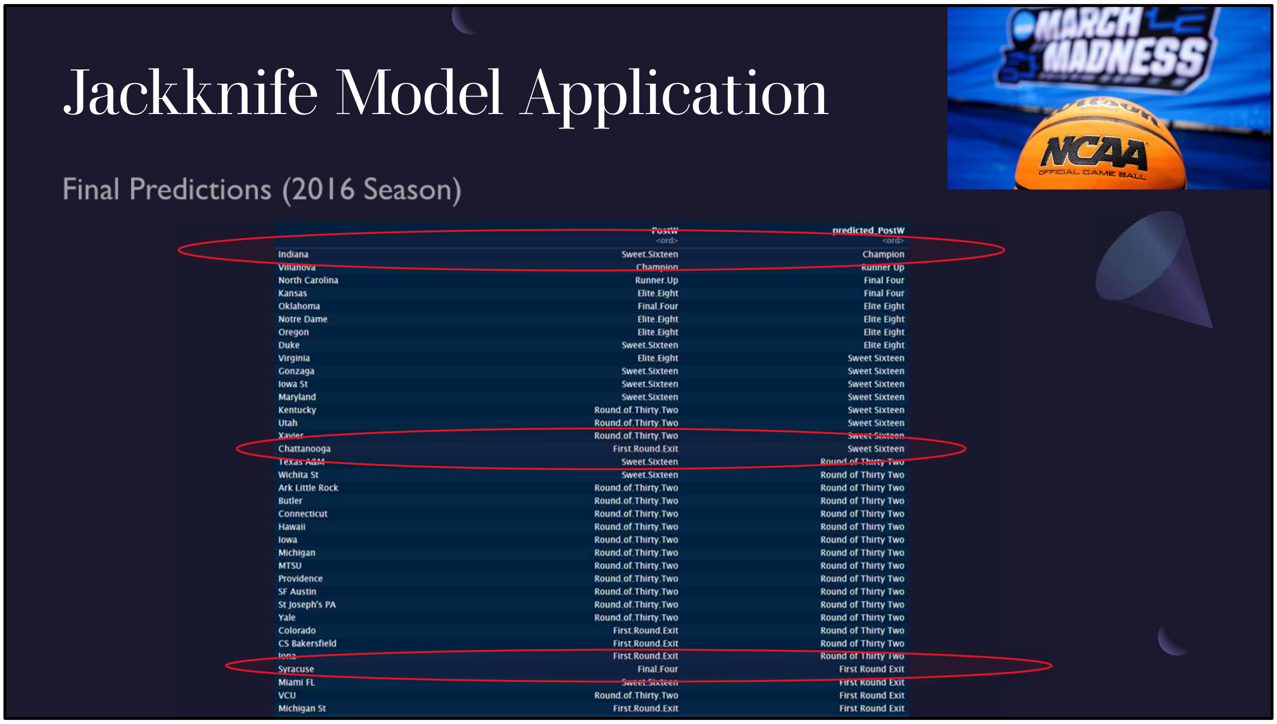
Two biggest incorrect predictions:
- Florida was eliminated in the 2nd round but I predicted them to make it to the Elite 8
- Texas made the Final 4, but I predicted them to lose in the Sweet 16

Jackknife Model Application

Final Predictions (2008 Season)

	PostW <ord>	predicted_PostW <ord>
North Carolina	Final.Four	Champion
Kansas	Champion	Runner Up
Memphis	Runner.Up	Final Four
UCLA	Final.Four	Final Four
Texas	Elite.Eight	Elite Eight
Duke	Round.of.Thirty.Two	Elite Eight
Siena	Round.of.Thirty.Two	Elite Eight
Drake	First.Round.Exit	Elite Eight
Davidson	Elite.Eight	Sweet Sixteen
Louisville	Elite.Eight	Sweet Sixteen
Tennessee	Sweet.Sixteen	Sweet Sixteen
WKU	Sweet.Sixteen	Sweet Sixteen
Wisconsin	Sweet.Sixteen	Sweet Sixteen
Georgetown	Round.of.Thirty.Two	Sweet Sixteen
Marquette	Round.of.Thirty.Two	Sweet Sixteen
Pittsburgh	Round.of.Thirty.Two	Sweet Sixteen
Xavier	Elite.Eight	Round of Thirty Two
Michigan St	Sweet.Sixteen	Round of Thirty Two
Stanford	Sweet.Sixteen	Round of Thirty Two
Villanova	Sweet.Sixteen	Round of Thirty Two
West Virginia	Sweet.Sixteen	Round of Thirty Two
Arkansas	Round.of.Thirty.Two	Round of Thirty Two
Butler	Round.of.Thirty.Two	Round of Thirty Two
Kansas St	Round.of.Thirty.Two	Round of Thirty Two
Miami FL	Round.of.Thirty.Two	Round of Thirty Two
Mississippi St	Round.of.Thirty.Two	Round of Thirty Two
Mt St Mary's	Round.of.Thirty.Two	Round of Thirty Two
Notre Dame	Round.of.Thirty.Two	Round of Thirty Two
Texas A&M	Round.of.Thirty.Two	Round of Thirty Two
BYU	First.Round.Exit	Round of Thirty Two
Oral Roberts	First.Round.Exit	Round of Thirty Two
UMBC	First.Round.Exit	Round of Thirty Two

Three biggest incorrect predictions:
- Duke and Siena were eliminated in the Round of 32, but I predicted them to make the Elite 8
- Xavier made the Elite 8, but I predicted them to lose in the Round of 32



	PostW <ord>	predicted_PostW <ord>
Indiana	Sweet.Sixteen	Champion
Villanova	Champion	Runner Up
North Carolina	Runner.Up	Final Four
Kansas	Elite.Eight	Final Four
Oklahoma	Final.Four	Elite Eight
Notre Dame	Elite.Eight	Elite Eight
Oregon	Elite.Eight	Elite Eight
Duke	Sweet.Sixteen	Elite Eight
Virginia	Elite.Eight	Sweet Sixteen
Gonzaga	Sweet.Sixteen	Sweet Sixteen
Iowa St	Sweet.Sixteen	Sweet Sixteen
Maryland	Sweet.Sixteen	Sweet Sixteen
Kentucky	Round.of.Thirty.Two	Sweet Sixteen
Utah	Round.of.Thirty.Two	Sweet Sixteen
Xavier	Round.of.Thirty.Two	Sweet Sixteen
Chattanooga	First.Round.Exit	Sweet Sixteen
Texas A&M	Sweet.Sixteen	Round of Thirty Two
Wichita St	Sweet.Sixteen	Round of Thirty Two
Ark Little Rock	Round.of.Thirty.Two	Round of Thirty Two
Butler	Round.of.Thirty.Two	Round of Thirty Two
Connecticut	Round.of.Thirty.Two	Round of Thirty Two
Hawaii	Round.of.Thirty.Two	Round of Thirty Two
Iowa	Round.of.Thirty.Two	Round of Thirty Two
Michigan	Round.of.Thirty.Two	Round of Thirty Two
MTSU	Round.of.Thirty.Two	Round of Thirty Two
Providence	Round.of.Thirty.Two	Round of Thirty Two
SF Austin	Round.of.Thirty.Two	Round of Thirty Two
St Joseph's PA	Round.of.Thirty.Two	Round of Thirty Two
Yale	Round.of.Thirty.Two	Round of Thirty Two
Colorado	First.Round.Exit	Round of Thirty Two
CS Bakersfield	First.Round.Exit	Round of Thirty Two
Iona	First.Round.Exit	Round of Thirty Two
Syracuse	Final.Four	First Round Exit
Miami FL	Sweet.Sixteen	First Round Exit
VCU	Round.of.Thirty.Two	First Round Exit
Michigan St	First.Round.Exit	First Round Exit

Three biggest incorrect predictions:
- Indiana was eliminated in the Sweet 16, but I predicted them to win the entire tournament
- Chattanooga was eliminated in the first round, but I predicted them to make the Sweet 16
- Syracuse made the Final 4, but I predicted them to lose in the first round

With more time, I would like to actually put these predictions into a bracket and score it like any march madness bracket would be scored

There is a more robust way of comparing multi class models using Area Under ROC Curve called Multiclass ROC Index, where you can compare micro and macro average ROC curves, which takes into account all of AUC's from each value of the target variable