Round 1 (the Round of 64)

For the first round, I checked for multicollinearity between my predictors before adjusting for league average. Some notable correlations were the following:

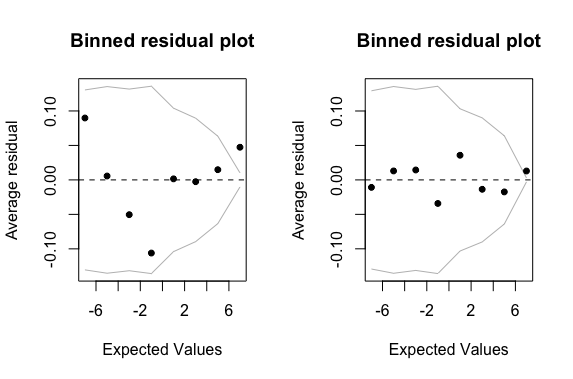
* effective field goal percentage (efg) and true shooting percentage (ts)
* efg with field goal percent, three-point percent, and points per possession (ppp)
* ts with field goal percent, three-point percent, and ppp
* points per game (ppg) with possessions per game, total points, field goals attempted (fga), and field goals made (fgm)

The same correlations mostly applied for the lower seeded team’s predictors. I did not check for multicollinearity in other rounds, since I made the assumption that the same correlations would mostly apply for other rounds. I also calculated a new predictor for the difference in seeds. I then adjusted the training data for the league average for each year.

I built my first model with a large amount of terms that I thought would be significant. I noticed that the p-values and coefficients of different terms would change by a lot each time I added a new predictor to the model. This was probably due to multicollinearity. I decided to use step-wise AIC to get a more simplistic model that would not over fit. This new model had four significant terms and an area under the ROC curve equal to 0.7746. The variables in this model were the lower seeded team’s defensive points per possession (opp\_dppp), ppp, ts, and seed difference.

I tried adding new predictors to this model and subsequently running Anova tests between both models. The predictor total possessions turned out to be pretty significant, so I added it to the model. I looked for significant interaction terms between seed difference and the other variables, but I was not able to find any significant interactions. I then looked at the binned residual plots for the predictors in the model, and I began to consider a quadratic transformation for seed difference and a cubic transformation for opp\_dppp. Both transformations increased the predicting power of the model, lowered the p-values of other variables, and fixed the binned residual plots for those variables. I decided to keep the transformations. The area under the ROC curve for this model increased to 0.8165.

The binned residual plot before and after the quadratic transformation for seed difference:



Until later rounds, I found it difficult to add other predictors due to multicollinearity. This was my final model for round 1:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 2.250e+00 2.755e-01 8.167 3.16e-16 \*\*\*

seed\_dif -1.963e-01 1.308e-01 -1.501 0.133474

seed\_dif2 2.569e-02 9.854e-03 2.607 0.009124 \*\*

opp\_dppp -1.336e+02 3.755e+01 -3.557 0.000375 \*\*\*

opp\_dppp3 4.936e+01 1.298e+01 3.803 0.000143 \*\*\*

ppp 1.425e+01 5.315e+00 2.681 0.007334 \*\*

ts -2.061e-01 8.936e-02 -2.306 0.021117 \*

total\_possessions -1.863e-03 1.053e-03 -1.769 0.076846 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 463.42 on 415 degrees of freedom

Residual deviance: 347.08 on 408 degrees of freedom

AIC: 363.08

I then predicted the winners of the first round and wrote the results to a csv file. After looking at the winners of the round, I created a schedule for round 2 according to where each team was on the bracket. I implemented a R function to create the matchups given the schedule. I ran this function on every new schedule for each round.

Round 2 (the Round of 32)

For round 2, I started with a model similar to my final round 1 model without the transformations. I discovered that many of the terms were not significant. I removed total possessions, ts, and ppp after running Anova tests and finding that these variables were not significant to the model. I looked at the binned residual plots of this altered model and considered a cubic transformation for seed difference and a quadratic transformation for opp\_dppp. These transformations improved the binned residual plots for both predictors. The area under the ROC curve for this model was 0.7547. I made a new model with opp\_dppp added to it. I ran an Anova test between the two models and found that this variable was significant. I decided to keep the variable. The area under my ROC curve increased to 0.7678.

This was my final model for round 2:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.342e+00 2.201e-01 6.099 1.07e-09 \*\*\*

seed\_dif 4.867e-01 2.104e-01 2.314 0.020689 \*

opp\_dppp -3.984e+02 1.029e+02 -3.872 0.000108 \*\*\*

opp\_dppp2 2.072e+02 5.232e+01 3.961 7.47e-05 \*\*\*

seed\_dif3 -4.999e-03 2.621e-03 -1.907 0.056494 .

opp\_ppp -9.273e+00 4.761e+00 -1.948 0.051418 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 251.70 on 207 degrees of freedom

Residual deviance: 207.65 on 202 degrees of freedom

AIC: 219.65

I did not run any interactions for this model, since I thought it would be difficult to interpret some of the transformed variables with interactions. I then predicted the matchups for round 2 and exported the winners and created the schedule for round 3.

Round 3 (the Sweet Sixteen)

For round 3, I decided to start with a very simplistic model, which included only seed difference and opp\_dppp. These predictors had proven to be significant for the other two rounds. The area under the ROC curve for this model was 0.7587.

I added both the win percent for the higher seeded team and the win percent for the lower seeded team, and through Anova tests, I discovered that neither was significant to the model. I looked at the binned residual plots for both seed difference and opp\_dppp. I then considered a log or quadratic transformation for seed difference and a quadratic transformation for opp\_dppp. The log transformation of seed difference lowered the p-values for all the terms and fixed the binned residual plot. However, a quadratic transformation for opp\_dppp was not significant and did not increase predicting power.

This was the final model for round 3:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.0724 0.2570 4.173 3.01e-05 \*\*\*

log\_seed\_dif 1.2265 0.3781 3.244 0.00118 \*\*

opp\_dppp 9.3888 4.7541 1.975 0.04828 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 126.72 on 103 degrees of freedom

Residual deviance: 109.07 on 101 degrees of freedom

AIC: 115.07

This model was much smaller than the models that I created in other rounds. I contributed this to the size decrease in my training data. Every new round had less historical matchups, since less teams got to those rounds. Additionally, I did not want to over fit my model, so I was hesitant about adding terms unless the predictors increased the predicting power and had significant p-values. Moreover, I predicted the matchups for round 3 and exported the winners and created the schedule for round 4.

Round 4 (the Elite Eight)

For round 4, I started with my final model from round 3. None of the variables in this model were significant. I then switched to a very large model and used stepwise AIC to get a shortened model, which only had the predictor points allowed from the lower seeded team. I concluded then that many of the predictors I included in the models from my earlier rounds were not significant in round 4. This is probably because those predictors are very general, such as points per possession and defensive points per possession, which try to summarize the offenses and defenses of the teams. However, at this point in the tournament, the higher seeded teams and lower seeded teams probably have very similar offenses and defenses or else they would not have been able to make it this far in the tournament. This small model had an area under the ROC curve equal to 0.6859.

I decided to include very specific predictors into my model, such as assists per game, free throw percentage, and offensive rebounds. The Anova test concluded that this group of variables were significant. The addition of these variables increased the ROC of the model to 0.7852. I determined that free throw percentage was only significant for the higher seeded team, offensive rebounds were only significant for the higher seeded team, and assists per game were not significant to the model. Removing the predictor assists per game only decreased my predicting accuracy to 0.7659. After looking at the binned residual plots, I considered a quadratic transformation on most of the predictors. None of these transformations, however, were able to add any predicting power or fix the binned residual plots.

This was my final model for round 4:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.05205 0.31933 -0.163 0.8705

opp\_points\_allowed 0.35780 0.14357 2.492 0.0127 \*

ft\_percent 0.26455 0.11708 2.260 0.0239 \*

or 0.60169 0.30211 1.992 0.0464 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 72.010 on 51 degrees of freedom

Residual deviance: 58.115 on 48 degrees of freedom

AIC: 66.115

I predicted the matchups for round 4 and exported the winners and created the schedule for round 5.

Round 5 (the Final Four)

The training data for round 5 was very small with only 26 observations. I initially thought that a model for this round would be similar to my final model in round 4, thus I created a model with the same predictors. However, the only predictor that had a significant p-value was offensive rebounds for the higher seeded team. I decided to create a simpler model with only offensive rebounds. After running an Anova test between my models, I concluded that the group of variables besides offensive rebounds were not significant. The simpler model returned an area under the ROC curve equal to 0.6845.

I decided to add personal fouls to the model, and it was significant for the lower seeded team. The addition of this variable increased my predicting power to 0.7738. I decided to keep the variable. The residuals in the binned residual plot for offensive rebounds looked periodical and had a clear trend of going up and down like a trigonometric function. I think that this could be due to serial correlation, since my training data is ordered in time, and there is a small number of matchups for each year in this round. The binned residual plot for personal fouls looked random. I tried different transformations on offensive rebounds, but these only inflated the p-values for other terms to such an extent that they were no longer significant. I stuck with the model without any transformations.

This was my final model for round 5:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.2047 0.4548 0.450 0.6527

or 0.8060 0.4249 1.897 0.0578 .

opp\_pf 0.8841 0.4990 1.772 0.0764 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 35.890 on 25 degrees of freedom

Residual deviance: 28.833 on 23 degrees of freedom

AIC: 34.833

I predicted the matchups for round 5 and exported the winners and created the schedule for round 6.

Round 6 (the Championship)

For round 6, the training data for this round only had 13 observations. In order to not over fit my model, I decided that I should only include one or two predictors. The only term that had a significant p-value in my model was opp\_ppp, which was the points per possession for the lower seeded team. This surprised me, since my models for the later rounds had very specific variables and points per possession is a very general description of the offense. The binned residual plot for this predictor looked random, so I did not consider any transformations. The area under the ROC curve was equal to 0.8095. This was also very surprising, since I only had one predictor in my model. I tried adding the variable opp\_total\_possessions to my model, since this lowered the p-value for opp\_ppp. However, the variable did not have a significant p-value and an Anova test concluded that it was not significant to my model.

This was my final model for round 6:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.1286 0.6649 0.193 0.8466

opp\_ppp -36.6469 22.1039 -1.658 0.0973 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 17.945 on 12 degrees of freedom

Residual deviance: 13.656 on 11 degrees of freedom

AIC: 17.656