Noah Blum, Ryan Kops, Joseph Re, Kara Conrad, Sichan Kim

Professor Gillett

STAT 451

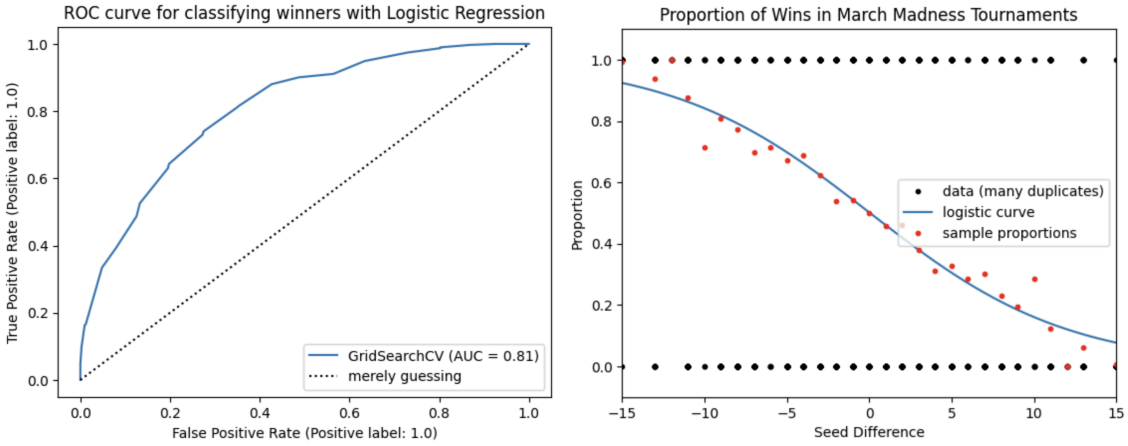
December 10th, 2023

# March Madness Bracket Prediction Report

March Madness is an annual NCAA basketball tournament that begins with the 64 teams, and narrows down to the best college basketball team. For the past few decades there has been a widespread competition to create “the perfect bracket”, which is almost impossible. The odds of one filling out a perfect bracket each year are around 1 in 9.2 quintillion (USA Today)[1]. Our group has chosen to analyze which basketball game variables are most useful in predicting the winners of NCAA Men’s March Madness basketball games using logistic regression.

The dataset we used for our model was a cleaned up version of a Sports Reference set from Akkio[2]. It included every bracket game dating back to 1985, comparing Team1 vs Team2. There were two rows for each game, switching Team1 and Team2 with each other, which was made to eliminate bias. For each team, their seed in the bracket was included, along with statistics for that year’s season (games played, points, rebounds, assists, and more). We narrowed down these statistical columns to about 28 that we believed were most relevant. We created a column called ‘Seed Difference’, which was Team1’s Seed subtracted by Team2’s Seed. Further, OneHotEncoding was also used on the ‘winner’ column to turn the values into binary values, with 1.0 representing the winner, and 0.0 representing the loser. Logistic regression, utilizing GridSearchCV(), was run on these variables to try and predict which variable was most useful in predicting each winner.

Initially, logistic regression was performed using only ‘Seed Difference’ as a predictor. The model achieved an accuracy of 72.9% on the test data, with the best parameters found to be C: 0.01, max_iter: 100, multi_class: ‘auto’, penalty: l1, and solver: 'liblinear'. The ROC curve results in an AUC of 0.81, also showing good performance. The logistic regression graph showed a logistic curve fitting the data, indicating the relationship between seed difference and the proportion of wins. The lower the seed difference, the higher the win proportion. This makes sense as a 1 seed, for example, that defeats a 16 seed, has a seed difference of -15. As seed differences get closer together, the team’s skills become more equal, resulting in a more equal chance of winning for both teams, which we see below.



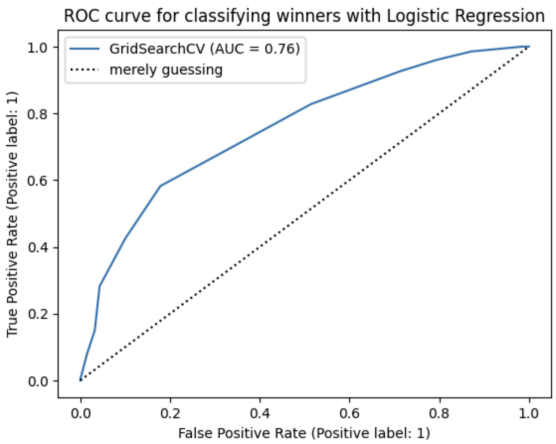

Next, logistic regression was performed for each team statistic (28 variables) to assess their individual predictive power. The accuracy scores range from 48.3% to 70.0% (seen below).

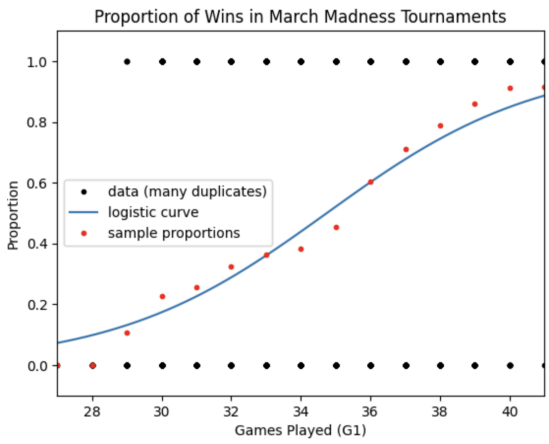
```
key:

{'G_': 'Games Played',
 'PTS_': 'Points',
 'TOV_': 'Turnovers',
 'FG%_': 'Field Goal Percentage',
 '2P%_': '2 Point Percentage Made',
 '3P%_': '3 Point Percentage Made',
 'AST_': 'Assists',
 'TRB_': 'Total Rebounds',
 'BLK_': 'Blocks',
 'DRB_': 'Defensive Rebound Percentage',
 'ORB_': 'Offensive Rebound Percentage',
 'PF_': 'Personal Fouls',
 'FT%_': 'Free Throw Percentage Made',
 'STL_': 'Steals'}
```

```
{'G1': 0.7,
 'G2': 0.624,
 'PTS2': 0.579,
 'TOV1': 0.577,
 'PTS1': 0.571,
 'FG%2': 0.562,
 '2P%2': 0.562,
 '3P%2': 0.562,
 'AST2': 0.557,
 'TRB2': 0.553,
 'BLK1': 0.546,
 'DRB1': 0.544,
 'BLK2': 0.544,
 'ORB2': 0.541,
```

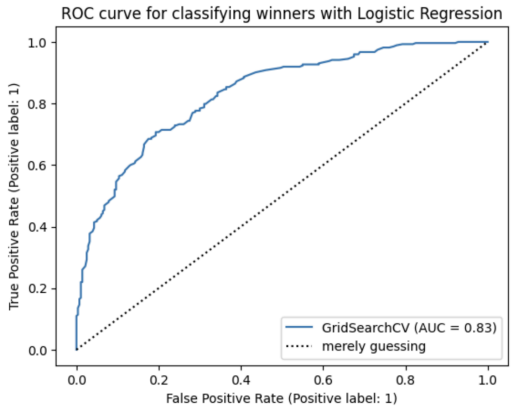
```
 'TRB1': 0.537,
 'PF2': 0.535,
 'FT%2': 0.532,
 'DRB2': 0.532,
 'STL2': 0.528,
 'TOV2': 0.528,
 'FG%1': 0.524,
 'FT%1': 0.521,
 '2P%1': 0.519,
 'AST1': 0.515,
 '3P%1': 0.506,
 'PF1': 0.505,
 'ORB1': 0.505,
 'STL1': 0.483}
```

The highest accuracy was obtained for the 'G1' variable (Team1 Games Played), which did not outperform 'Seed Difference'. We still ran logistic regression on 'G1' to compare to 'Seed Difference.' Analysis of games played is useful as the number of games can vary per team due to conference tournaments. Some teams will play more games if they do better in these conference tournaments, which also results in conference champions receiving better seeding. The ROC curve shows good performance with an AUC of 0.76 (below) and accuracy of 70%. From the logistic regression curve, we see the sample proportions have an oscillating curve, which suggests the data is overfitting.

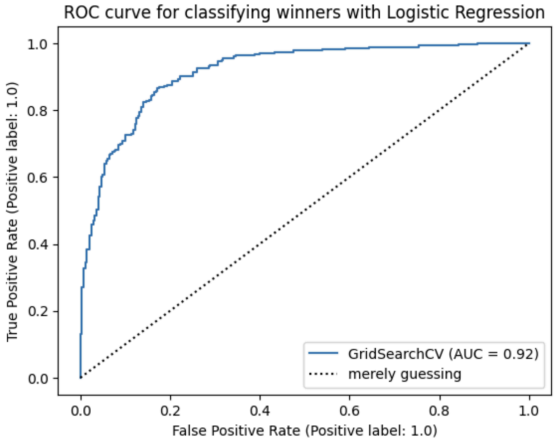


A subsequent analysis was conducted by training on ‘Seed Difference’ and 'G1' to see if the two best individual scores would be an improvement together. It did, as accuracy improved to 73.6%, suggesting that the combination of these two variables provides better predictions than individually. The AUC is also 0.83, showing good performance.




Finally, logistic regression was performed using all available statistics as predictors. The model achieved an accuracy of 84.3%, indicating that additional team statistics improved the predictive performance compared to using ‘Seed Difference’ and 'G1'.




We tested our logistic regression for ‘Seed Difference’ on the most recent 2023 Men’s bracket, which would have resulted in a 78% correct through the first round. This equates to choosing about 25 out of the 32 games correctly. This helps to show that the seedings are effective, and choosing the higher seed results in a better percentage than randomly guessing at

50/50. It also proves how upsets are common and unpredictable, otherwise our model would be 100% correct.

In conclusion, this analysis demonstrates how using training and testing logistic regression with use of GridSearchCV on both teams' statistics can be used to predict the winners of NCAA March Madness basketball games with an accuracy of 84.3%. These findings can potentially be valuable for basketball enthusiasts, analysts, and gamblers looking to make informed predictions during the tournament.

## Citations

[1] https://www.usatoday.com/story/gameon/2013/03/19/ncaa-tournament-perfect-bracket-odds-quintillion/1999795/

[2] https://www.akkio.com/post/crushing-your-march-madness-bracket-with-ai