Bayesian Model to Predict the Outcome of NCAA March Madness Games

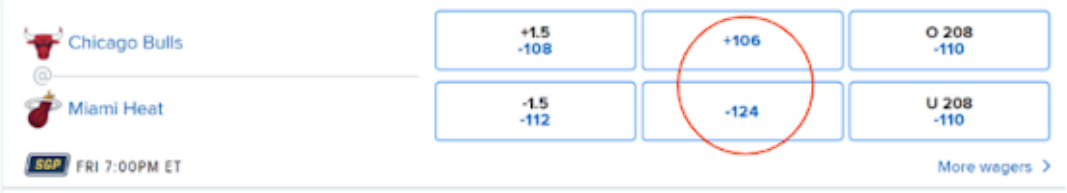
By: Brock Etzel

Introduction

Sports gambling has become more and more prevalent in today's society. In 2023 U.S. bettors wagered nearly $120 billion with around $10 billion of those being profits according to the American Gaming Association. But what if it was possible to beat the odds? This is obviously the goal of every sports bettor but it is a much more complicated question to answer. For starters oddsmakers use lots of data to determine the odds they set, especially for bigger games such as March Madness. Secondly odds are purposefully mispriced to give the casino an advantage. This is where the phrase "the house always wins" comes from, while it is possible to make money on a smaller scale in the long run the casino is going to be profitable due to mispricing. This project will attempt to create a logistic regression model that predicts the probability a given team will beat another team in March Madness.

Sports Gambling and Odds

A brief aside on how sports gambling actually works is necessary to understand how exactly oddsmakers misprice games and how it could be possible to find underpriced games. Gambling odds are very similar to the odds of any probability however scaled up 100 times. They are also altered to not include decimals, if a team is favored to win the odds would be calculated based on their winning probability then multiplied by -100 and if a team is favored to lose the odds would be calculated based on their losing probability then multiplied by 100. In theory these values should be direct opposites of one another because the probability of team A to win equals the probability of team B to lose. However this is never the case.



For example the Miami Heat moneyline (betting on them to win the game outright) for their Friday play-in tournament game is -124. This means that in order to win $100 a bettor would have to risk losing $124. These odds imply that the Heat would have a 55.36% chance of winning this game. On the other hand the Atlanta Hawks moneyline is +106 meaning a bet of $100 would earn $106 if correct. These odds imply a 48.54% chance of winning this game for the Hawks. Anyone with any experience in probability can immediately see the issue here; the probability of the two possible outcomes do not add to one but rather 1.039. By boosting the implied total probability above one oddsmakers can virtually ensure they profit no matter which team wins.

The Data

The data used for this study will attempt to use advanced analytics taken from Ken Pomperoy. Pomeroy is the creator of the KenPom rating system which attempts to rank every single Division I College Basketball team based on a number of metrics. He has worked alongside

Daryl Morey with the Houston Rockets who is well known for his use of analytics when constructing rosters. Pomeroy's website kenpom.com has a relatively comprehensive list of all the data he takes into account when making his rankings. This project will focus on four non-advanced measures and then a further five more advanced ones.

The non-advanced measures will be Conference, Wins, Losses, and Tournament Rank. The latter three are very easy to understand, how many wins and losses a given team has and what seed they are for the NCAA Tournament ranked one through sixteen. Conference however is a bit more complicated as it is a qualitative variable that can take 32 different values. There is also no base conference where deviations from such a base could be studied. However it is clear that the conference a team plays is important as it's well known that teams from the SEC tend to be better than teams from the Ivy League. In order to correct for this as well as make the regression simpler, teams that are apart of a Power 7 Conference in 2024 (SEC, ACC, Pac-12, Big Ten, Big 12, Big East, and Mountain West) will be coded with a 1 and teams in any other mid-major conference will be coded with a 0.

The more advanced measures will be KenPom Rank, AdjO, AdjD, Luck, and AdjEm SoS. The KenPom Rank takes into account all of these metrics as well as others to determine a team's final rank. Thus it will be of interest to see whether certain metrics or even the KenPom Rank as a whole are significant in the regression since they are already accounted for. AdjO and AdjD represent points scored and points allowed per 100 possessions adjusted for a team's opponents. Luck is a team's deviation from expected wins (South Carolina was the fourth luckiest team in Division I and the second luckiest team in the Tournament this season). Finally AdjEm SoS is a measure of strength of schedule which finds the average difference of AdjO and AdjD for all the opponents of a given team.

The sample of games will be all 63 NCAA 2024 Tournament games sorted as higher seed vs. lower seed so as to not double count games and ensure there are both wins and losses in the data. The previously discussed variables will be in use for both teams so there is a possibility of using up to 18 variables however, presumably a number of them will be redundant and removed from the model.

Prior Elicitation

Obviously fitting normal priors for all the variables would be easiest and make sense. The model is set up so that it predicts the probability of the higher seeded team winning the game, not just any team winning any game (which would have a known probability of 0.5). Because of that a "typical" higher seed could reasonably be expected to have a win probability between 0.55 and 0.75. The log-odds of these values would be 0.087 and 0.477 which would be centered at 0.282 with a standard deviation of 0.0975.

KenPom Rank: A one unit increase in KenPom Rank would probably lead to a change by a factor of 0.9 to 1 in odds since lower ranked teams are worse. Leading to a Normal(-0.0229, 0.0114)

AdjO/D: It is reasonable to believe that a one point increase in offense and a one point decrease in defense would have similar effects on the outcome of a game thus they will be discussed

together. Such changes would provide between a 1.03 and 1.15 factor increase in win probability leading to a Normal(0.0367, 0.0119) for offense and a Normal(-0.0419, 0.0143) for defense.
Luck: Luck is more a measure of variation rather than the mean outcome so because of that both Luck and Opponent Luck will use non-informative priors with means of 1.
AdjEm SoS: Strength of Schedule is determinant on the previous opponents of a team and therefore it is reasonable to believe that the effect of SoS may not have a huge impact on a current game so a non-informative prior with mean of 1 will be used.
Conference: If the team is in a Power 7 conference it is reasonable to believe their odds of winning increase by a factor of between 1.1 and 2 leading to a Normal(0.1712, 0.0655) when compared to if they were in a non-power conference.
Wins: An increase in one win would most likely increase win odds by a factor between 1 and 1.05 leading to a Normal(0.0105, 0.0053)
Losses: Similarly a one loss increase to a teams record would likely decrease win odds by a factor between 0.95 and 1 leading to a Normal(-0.0111, 0.0056)
Seed: An increase of one seed rank would likely have an increase in win odds by a factor between 1.05 to 1.2 leading to a Normal(0.0502, 0.0145)
Opponent KenPom Rank: A one unit increase in the opponent's KenPom Rank would most likely lead to a change in win odds by a factor between 1 and 1.1 as discussed above leading to a Normal(0.0207, 0.0103)
Opponent AdjO/D: Similar to AdjO and AdjD a one unit decrease in the opponent's AdjO or increase in AdjD would lead to changes in win odds by a factor of 1.03 and 1.15 which corresponds to a Normal(-0.0419, 0.0143) for offense and a Normal(0.0367, 0.0119) for defense.
Opponent Conference: If the opponent is in a power conference it is reasonable to assume that odds decrease by a factor of 0.5 to 0.9 leading to a Normal(-0.1734, 0.0638)
Opponent Wins: An increase in one win for the opponent would likely lead to a change in odds by a factor between 0.95 and 1, providing a Normal(-0.0111, 0.0056)
Opponent Losses: An increase in one loss for the opponent would lead to a change in odds by a factor between 1 and 1.05, providing a Normal(0.0105, 0.0053)
Opponent Seed: An increase of the opponent's seed by one rank would likely lead to a decrease in win odds by a factor between 0.8 and 0.95 leading to a Normal(-0.0596, 0.0187)

Model

The full model with all possible variables in the model would be:

$$Y_i|\beta_0, \beta_1, \dots, \beta_{18} \sim Bernoulli(\pi_i) \text{ with } log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_{18} X_{i18}$$

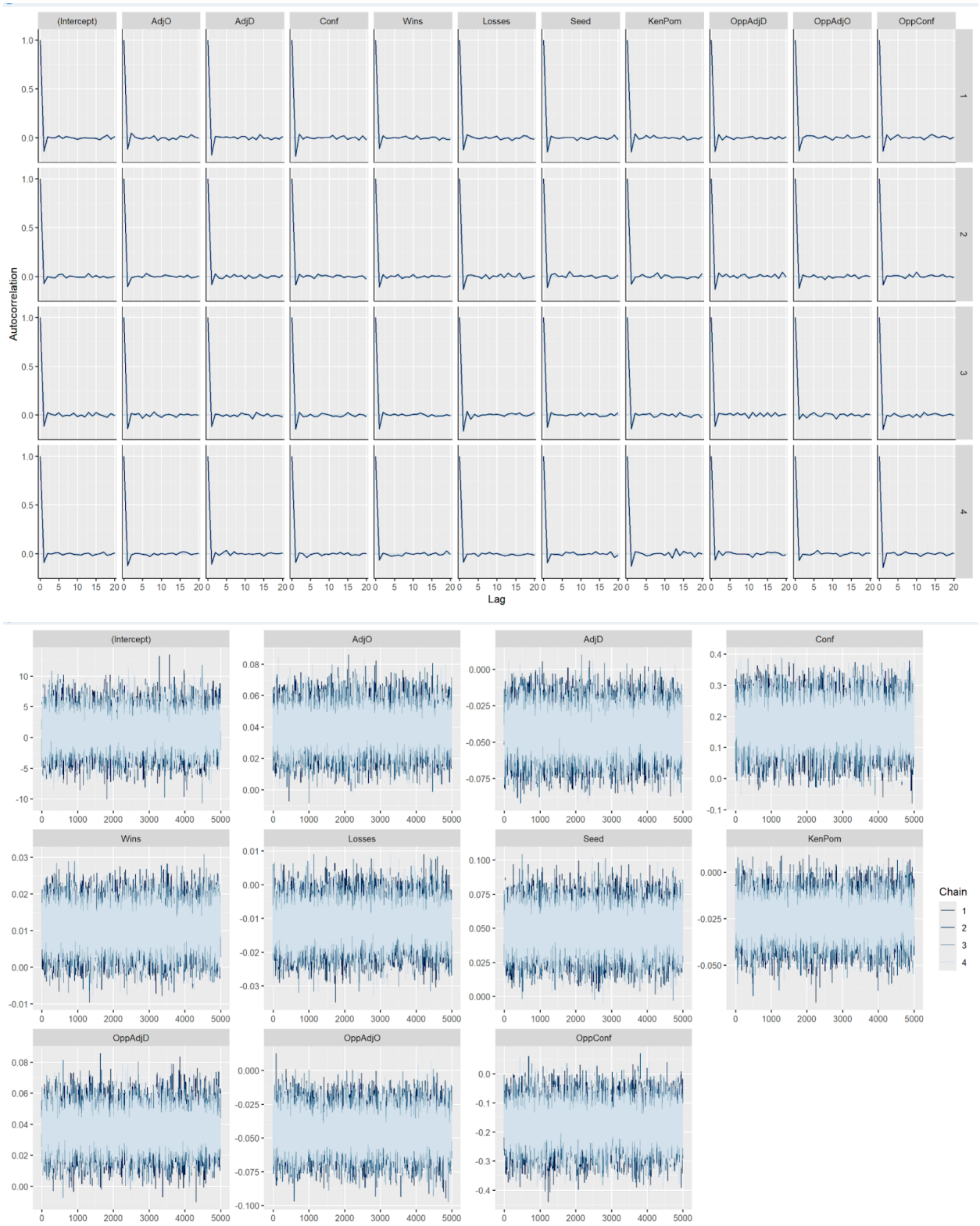
When fitting the full model 10 of the variables seem to be significant. These 10 are; AdjO, AdjD, Conference, Wins, Losses, Seed, KenPom Rank, Opponent AdjO, Opponent AdjD, and Opponent Conference.

```
# A tibble: 19 × 5
   term           estimate std.error  conf.low conf.high
   <chr>             <dbl>     <dbl>     <dbl>     <dbl>
 1 (Intercept)  0.758        3.12     -5.39       6.91
 2 AdjO         0.0387       0.0117    0.0159     0.0612
 3 AdjD        -0.0421       0.0141   -0.0692    -0.0146
 4 Conf         0.171        0.0650    0.0445     0.298
 5 Wins         0.0109       0.00527   0.000407   0.0214
 6 Losses      -0.0115       0.00559  -0.0224    -0.000392
 7 Seed         0.0496       0.0145    0.0211     0.0777
 8 KenPom      -0.0245       0.0103   -0.0446    -0.00457
 9 Luck        -0.00000399   0.00999  -0.0197     0.0198
10 SoS          0.000659     0.0101   -0.0191     0.0203
11 OppAdjD      0.0353       0.0116    0.0123     0.0584
12 OppAdjO     -0.0427       0.0141   -0.0700    -0.0146
13 OppConf     -0.173        0.0627   -0.296     -0.0480
14 OppKenPom    0.0123       0.00691  -0.000813   0.0265
15 OppLosses    0.0101       0.00529  -0.000476   0.0206
16 OppLuck     -0.0000221    0.0101   -0.0198     0.0200
17 OppSeed     -0.104        0.0964   -0.291      0.0839
18 OppSoS      -0.000727     0.0101   -0.0201     0.0186
19 OppWins      0.0111       0.0482   -0.0834     0.106
```

Refitting the model with only the significant predictors results in:

```
# A tibble: 11 × 5
   term         estimate std.error  conf.low conf.high
   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
 1 (Intercept)   1.07      2.73     -4.27       6.48
 2 AdjO          0.0389    0.0115    0.0164     0.0614
 3 AdjD         -0.0424    0.0141   -0.0695    -0.0151
 4 Conf          0.172     0.0652    0.0446     0.300
 5 Wins          0.0110    0.00545   0.000300   0.0213
 6 Losses       -0.0115    0.00556  -0.0224    -0.000605
 7 Seed          0.0493    0.0147    0.0211     0.0772
 8 KenPom       -0.0256    0.00997  -0.0455    -0.00615
 9 OppAdjD       0.0367    0.0118    0.0137     0.0598
10 OppAdjO      -0.0451    0.0134   -0.0715    -0.0187
11 OppConf      -0.177     0.0634   -0.303     -0.0519
```

Wins and losses both also seem to be barely significant at the 95% confidence level but will remain in the model.



Both the trace plot and the autocorrelation function of the MCMC chain look good indicating this model fits the data well and can be used for future study of games.

Overall the final model with the 10 variables deemed to be important to determining outcome of a game would be:

$$\log(\text{odds}) = 1.07 + 0.0389 \cdot \text{AdjO} - 0.0424 \cdot \text{AdjD} + 0.172 \cdot \text{Conference} + 0.011 \cdot \text{Wins} - 0.115 \cdot \text{Losses} + 0.0493 \cdot \text{Seed} - 0.0256 \cdot \text{KenPom} - 0.0451 \cdot \text{OppAdjO} + 0.0367 \cdot \text{OppAdjD} - 0.177 \cdot \text{OppConference}$$

Use of the Model

This model would not necessarily be used to predict whether or not a team is going to win because the higher seeded team is almost always expected to win, that is why they are the higher seeded team. Thus looking at confusion matrices or setting cutoff values for every game as a whole seems trivial. Instead the model can be used in comparison to betting odds. If a sportsbook has the implied probability of winning or losing lower than the probability of winning or losing given by the model then betting on the line that the sportsbook seemingly has mispriced too low may be a “sharp” play.