# STAT 172 CBB Project

## Data

https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset/data

## Model Specification

Since our Y variable is binary (it can only take either a 0 or a 1), we need to use a Bernoulli random component. Since we want interpretable results, specifically odds ratios, we chose to use a logit link function. Probit regression and complementary log-log link functions would not provide us with an odds interpretation, which is why using a logit link is better for our model.

Let:

$y_i$ = {1 if team i made the tournament,
0 otherwise

Let:

$G_i$ represent the number of games played by team i
$W_i$ represent the number of games won by team i
$ADJOE_i$ represent team i's adjusted offensive efficiency
$ADJDE_i$ represent team i's adjusted defensive efficiency
$BARTHAG_i$ represent team i's power rating (chance of beating an average D1 team)
$EFG\_O_i$ represent team i's effective field goal percentage shot
$EFG\_D_i$ represent team i's effective field goal percentage allowed
$ADJ\_T_i$ represent team i's effective adjusted tempo
$LM_i$ = 1 if team i is part of the LM conference group, and 0 otherwise
$MM_i$ = 1 if team i is part of the MM conference group, and 0 otherwise

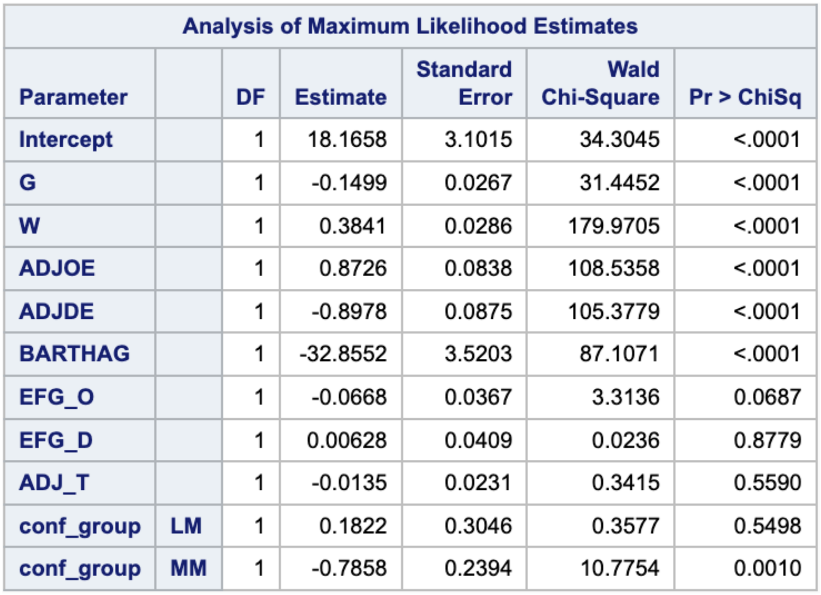
Random Component:

$Y_i \sim \text{Bernoulli}(\pi_i)$

Systematic Component:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1(G_i) + \beta_2(W_i) + \beta_3(ADJOE_i) + \beta_4(ADJDE_i) + \beta_5(BARTHAG_i) + \beta_6(EFG\_O_i) + \beta_7(EFG\_D_i) + \beta_8(ADJ\_T_i) + \beta_9(LM_i) + \beta_{10}(MM_i)$$

Model Results



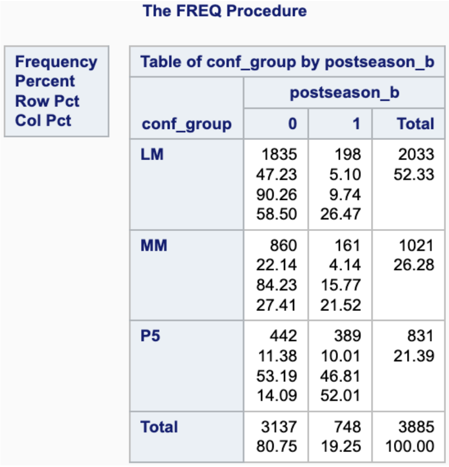
Analysis of Maximum Likelihood Estimates

| Parameter | | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|
| Intercept | | 1 | 18.1658 | 3.1015 | 34.3045 | <.0001 |
| G | | 1 | -0.1499 | 0.0267 | 31.4452 | <.0001 |
| W | | 1 | 0.3841 | 0.0286 | 179.9705 | <.0001 |
| ADJOE | | 1 | 0.8726 | 0.0838 | 108.5358 | <.0001 |
| ADJDE | | 1 | -0.8978 | 0.0875 | 105.3779 | <.0001 |
| BARTHAG | | 1 | -32.8552 | 3.5203 | 87.1071 | <.0001 |
| EFG_O | | 1 | -0.0668 | 0.0367 | 3.3136 | 0.0687 |
| EFG_D | | 1 | 0.00628 | 0.0409 | 0.0236 | 0.8779 |
| ADJ_T | | 1 | -0.0135 | 0.0231 | 0.3415 | 0.5590 |
| conf_group | LM | 1 | 0.1822 | 0.3046 | 0.3577 | 0.5498 |
| conf_group | MM | 1 | -0.7858 | 0.2394 | 10.7754 | 0.0010 |

Note: After completing our analysis, we further examined the variables used in our model and found evidence of multicollinearity, specifically with BARTHAG, which uses ADJOE and ADJDE as inputs. While we would certainly address this issue in the real world, likely by removing BARTHAG (or both ADJOE and ADJDE), we were unable to adequately update our model and the resulting change in coefficients in the timeframe given. As such, we have intentionally decided to avoid interpreting BARTHAG.

Complete Separation

Pre Analysis: There are no zeros observed in the cross table between conf_group (our only categorical variable) and postseason_b (our response variable), indicating we aren't worried about complete separation.



The FREQ Procedure

Frequency<br>Percent<br>Row Pct<br>Col Pct

Table of conf_group by postseason_b

| | postseason_b | | |
|---|---|---|---|
| conf_group | 0 | 1 | Total |
| LM | 1835<br>47.23<br>90.26<br>58.50 | 198<br>5.10<br>9.74<br>26.47 | 2033<br>52.33 |
| MM | 860<br>22.14<br>84.23<br>27.41 | 161<br>4.14<br>15.77<br>21.52 | 1021<br>26.28 |
| P5 | 442<br>11.38<br>53.19<br>14.09 | 389<br>10.01<br>46.81<br>52.01 | 831<br>21.39 |
| Total | 3137<br>80.75 | 748<br>19.25 | 3885<br>100.00 |

Post Analysis: Additionally, we observed no standard errors greater than 5, further indicating no concern for complete separation.

Interpretations, Confidence Intervals

**Interpretation for ADJOE:**
ADJOE: Adjusted Offensive Efficiency
β3=ADJOE: Holding all other factors constant, the odds of making the tournament increase by a factor of $e^{.8726} = 2.393$ for each additional point scored in ADJOE. This means that for each additional point in a team's adjusted offensive efficiency, the odds of making the tournament more than double, suggesting that stronger offensive performance significantly boosts a team's chances of making the tournament.

What happens to the odds of making the tournament when adjusted offensive efficiency points scored per 100 possessions increase by 5?

Odds Ratio = $\frac{Odds\ at\ x+5}{Odds\ at\ x}$

$$\frac{e^{\beta 0+\beta 1G+\beta 2W+\beta 3ADJOE(x+5)+\beta 4ADJDE+\beta 5BARTHAG+\beta 6EFGO+\beta 7EFGD+\beta 8ADJT}}{e^{\beta 0+\beta 1G+\beta 2W+\beta 3ADJOE(x)+\beta 4ADJDE+\beta 5BARTHAG+\beta 6EFGO+\beta 7EFGD+\beta 8ADJT}}=\frac{e^{\beta 3(x+5)}}{e^{\beta 3(x)}}=e^{\beta 3(x+5)-\beta 3(x)}=e^{5*\beta 3}=e^{5*(.8726)}=78.49$$

Holding all else constant, the odds of making the NCAA tournament increase by a factor of e^5*.8726=78.49 times with every increase of 5 in points scored per 100 possessions in adjusted offensive efficiency.

**Confidence Interval for ADJOE:**
A 95% likelihood-based confidence interval.

Holding all else constant, we are 95% confident that the odds of a team making the NCAA Tournament increase by a factor of $e^{.7135} = 2.04112$ and $e^{1.042} = 2.83488$ for each additional point scored per 100 possessions. We are 95% confident that the odds increase anywhere from 104.112% to 183.488% holding other factors constant.

A 95% likelihood-based confidence interval for the odds ratio with an increase of 5 points scored per 100 possessions.

Holding all else constant, we are 95% confident that the odds of a team making the NCAA Tournament increase by a factor of $e^{5*.7135} = 32.4279$ and $e^{5*1.042} = 183.094$ for an increase of 5 points scored per 100 possessions. We are 95% confident that the odds increase anywhere from 3142.79% to 18209.4% holding other factors constant.

**Interpretation of Conf_Group Variable:** Conf_group represents the classification of the conference team i is in (either P5, MM, or LM). With the initial observation of P5 being accounted for in the intercept estimate, the predicted low major $\beta_9$-hat is 0.1822. The predicted mid-major $\beta_{10}$-hat is -0.7858.

Holding all other variables in the equation constant, the following odds are attributed to each conference classification and a team's odds of making the NCAA Tournament:

Low Major: $e^{0.1822} = 1.1999$
Mid Major: $e^{-0.7858} = 0.4558$

This means that in comparison to the Power 5 conference teams, a team in a low major conference has roughly 20% greater odds of making the NCAA Tournament, given they have the same statistics.

A team in a mid-major conference has only 0.45 odds to make the NCAA Tournament relative to a team in a Power 5 conference, given all statistics are held constant

The following intervals demonstrate the range to which we are 95% confident the true odds of making the NCAA Tournament per conference group relative to being in the Power 5 lie:

Given the following for low major:
$e^{-0.4140} = 0.660$
$e^{0.7806} = 2.180$
We are 95% confident that given a team is in a low-major conference, the odds of them making the NCAA Tournament compared to a team in a Power 5 conference are between (0.660,2.180)

Given the following for mid-major:
$e^{-1.2553} = 0.285$
$e^{-0.3161} = 0.729$
We are 95% confident that given a team is in a mid-major conference, the true odds of them making the NCAA Tournament compared to a team in a Power 5 Conference are between (0.285,0.729). This means I can say with 95% confidence that a team in a Power 5 conference has better odds of making the NCAA Tournament than a mid-major team.

**Interpretation of Wins Variable ($W_i$):**
$W_i$ represents the number of games won by team i. It corresponds to $\beta_2$ in our model. We observed a predicted $\beta_2$-hat of 0.3841.

Holding all other factors constant, the odds of a team making the NCAA tournament increase by a factor of $e^{0.3841} = 1.4682922$ for each one additional win a team has. In other words, the odds of making the tournament increase by about 46.83%, holding all other factors constant.

Holding all other factors constant, the odds of a team making the NCAA tournament increase by a factor of $e^{(2*0.3841)} = 2.155882$ for every two additional wins a team has. In other words, the odds of making the tournament increase by about 115.59%, holding all other factors constant.

**Confidence Interval for Wins Variable ($W_i$):**

We are 95% confident that the odds of a team making the NCAA tournament increase by a factor between $e^{0.3288}$ = 1.389300 and $e^{0.4411}$ = 1.554416 for each additional win a team has, holding all other factors constant. In other words, we are 95% confident the odds increase anywhere from 38.93% to 55.44%, holding all other factors constant.

We are 95% confident that the odds of a team making the NCAA tournament increase by a factor between $e^{(2*0.3288)}$ = 1.930154 and $e^{(2*0.4411)}$ = 2.416210 for every two additional wins a team has, holding all other factors constant. In other words, we are 95% confident the odds increase anywhere from 93.02% to 141.62%, holding all other factors constant.

## Hypothesis Testing

**Hypothesis Test of Model:**

$H_0$: $\beta 1 = \beta 2 = \beta 3 = \ldots = 0$
$H_a$: $At\ least\ one\ \beta \neq 0$
Test statistic: 34.3045
Null Distribution: $\chi^2(10)$
P-Value: <.0001
Conclusion: We reject $H_0$ in favor of $H_a$ at all reasonable levels of alpha.

**Hypothesis Test of categorical variable Conf_group.**

$H_0$: $\beta 9 = \beta 10 = 0$
Ha: $At\ least\ one\ \beta \neq 0$
Test statistic: 30.9934
Null Distribution: $\chi^2(2)$
P-Value: <.0001
Conclusion: We reject $H_0$ in favor of $H_a$ at all reasonable levels of alpha.

**Hypothesis Test of Numeric variable ADJOE.**

Ho:$\beta 3 = 0$
Ha:$\beta 3 \neq 0$
Test statistic: 108.5358
Null Distribution: $^{2(1)}$
P-Value: <.0001
Conclusion: We reject $H_0$ in favor of $H_a$ at all reasonable levels of alpha. We have significant evidence to suggest that the coefficient for adjusted offensive efficiency is different from zero.

Other Meaningful Insights

**Meaningful Insight 1**

Comparing two non-baseline levels of Conf_group odds of making the NCAA Tournament.

$$\frac{odds\ of\ LM\ making\ the\ NCAA\ Tournament}{odd\ of\ MM\ making\ the\ NCAA\ Tournament}$$

$e^{\beta 9 - \beta 10}$=$e^{.1822 - -.7858}$=2.632

All else constant, the odds that a Low Major makes the NCAA Tournament are 2.632 times that a Mid Major makes the NCAA Tournament. That means the odds of a Low Major team making the NCAA Tournament are 163.2% higher than the odds of a Mid Major Team making the NCAA Tournament.

**Meaningful Insight 2:** How does a team's probability of making the NCAA tournament change when adjusting 1 variable, wins?

To demonstrate our model, we will calculate the probability of a team with certain characteristics making the NCAA Tournament. Consider a team with the following statistics:

$G_i$ = 30
$W_i$ = 22
$ADJOE_i$ = 110
$ADJDE_i$ = 105
$BARTHAG_i$ = 0.75
$EFG\_O_i$ = 51
$EFG\_D_i$ = 53
$ADJ\_T_i$ = 67
$LM_i$ = 0
$MM_i$ = 0

Note, the team above is a member of the P5 conference group ($LM_i$ and $MM_i$ are set to 0).

$\log(\frac{\pi_i}{1-\pi_i})$ = 18.1658 − 0.1499(30) + 0.3841(22) + 0.8726(110) − 0.8978(105) − 32.8552(0.75) − 0.0668(51) + 0.00628(53) − 0.0135(67) + 0.1822(0) − 0.7858(0) = −4.78386
$\pi_i = e^{-4.78386} / (1 + e^{-4.78386})$
$\pi_i$ = 0.00829428 = 0.829428%

The probability of a team with the above characteristics making the NCAA tournament is 0.829428%.

Now, let's find the probability of a team making the playoffs if they win one additional game, holding all other factors above constant. In other words, let $wins_i$ = 23, while holding the rest of the variables specified above constant.

$\log(\frac{\pi_i}{1-\pi_i})$ = 18.1658 − 0.1499(30) + 0.3841(23) + 0.8726(110) − 0.8978(105) − 32.8552(0.75) − 0.0668(51) + 0.00628(53) − 0.0135(67) + 0.1822(0) − 0.7858(0) = −4.39976
$\pi_i = e^{-4.39976} / (1 + e^{-4.39976})$
$\pi_i$ = 0.0121313 = 1.21313%

Holding all other variables constant, increasing the number of wins for a team by 1 increases the team's probability of making the NCAA tournament from 0.829428% to 1.21313%

Now, instead of increasing the number of wins by 1, let's consider the impact of increasing the number of wins by 5, holding all other factors constant. In other words, let $wins_i$ = 27, while holding the rest of the variables specified above constant.

$\log(\frac{\pi_i}{1-\pi_i})$ = 18.1658 − 0.1499(30) + 0.3841(27) + 0.8726(110) − 0.8978(105) − 32.8552(0.75) − 0.0668(51) + 0.00628(53) − 0.0135(67) + 0.1822(0) − 0.7858(0) = −2.86336
$\pi_i = e^{-2.86336} / (1 + e^{-2.86336})$
$\pi_i$ = 0.0539948 = 5.39948%

Holding all other variables constant, increasing the number of wins for a team by 5 (from 22 to 27) increases the team's probability of making the NCAA tournament from 0.829428% to 5.39948%

**Meaningful Insight 3**
Given a Power 5 team has the following resume:

$G_i$ = 30
$W_i$ = 23
$ADJOE_i$ = x
$ADJDE_i$ = 105
$BARTHAG_i$ = 0.75
$EFG\_O_i$ = 51
$EFG\_D_i$ = 53
$ADJ\_T_i$ = 67
$LM_i$ = 0
$MM_i$ = 0

I would like to know what ADJOE would give this team a 95% probability of making the NCAA Tournament. Using the equation below, we can solve for x:

$$0.95 = e^z/1 + e^z$$

Where z = 18.1658 − 0.1499(30) + 0.3841(23) + 0.8726(x) − 0.8978(105) − 32.8552(0.75) − 0.0668(51) + 0.00628(53) − 0.0135(67) + 0.1822(0) − 0.7858(0)

0.95 = e^(-110.39 + 0.8726(x)) / 1+e^(-110.39 + 0.8726(x))
X = 118.44

Solving for x, we determine that when a team has the above resume and is in a Power 5 conference, an adjusted offensive efficiency score of 118.44 gives them a 95% probability of making the postseason.