

March Madness Basketball Turniment

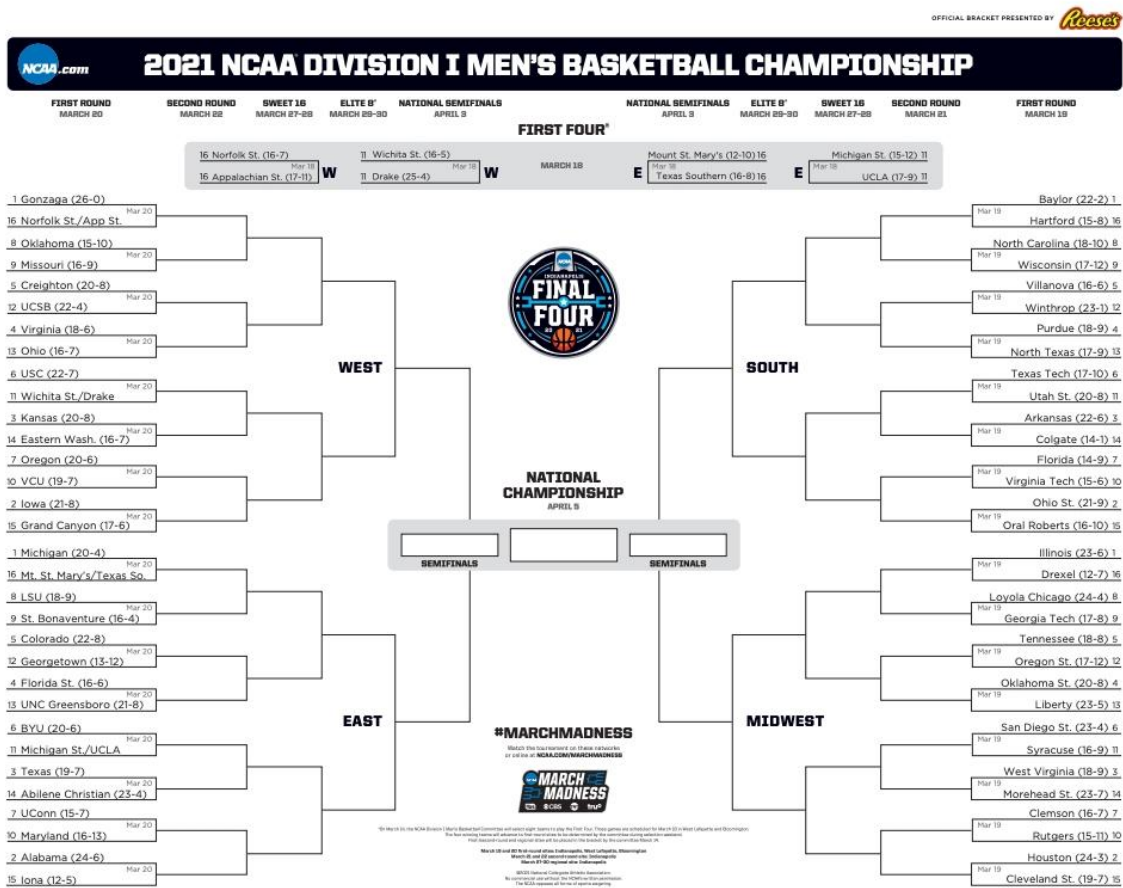
Sabrina Ball and Joe Despres

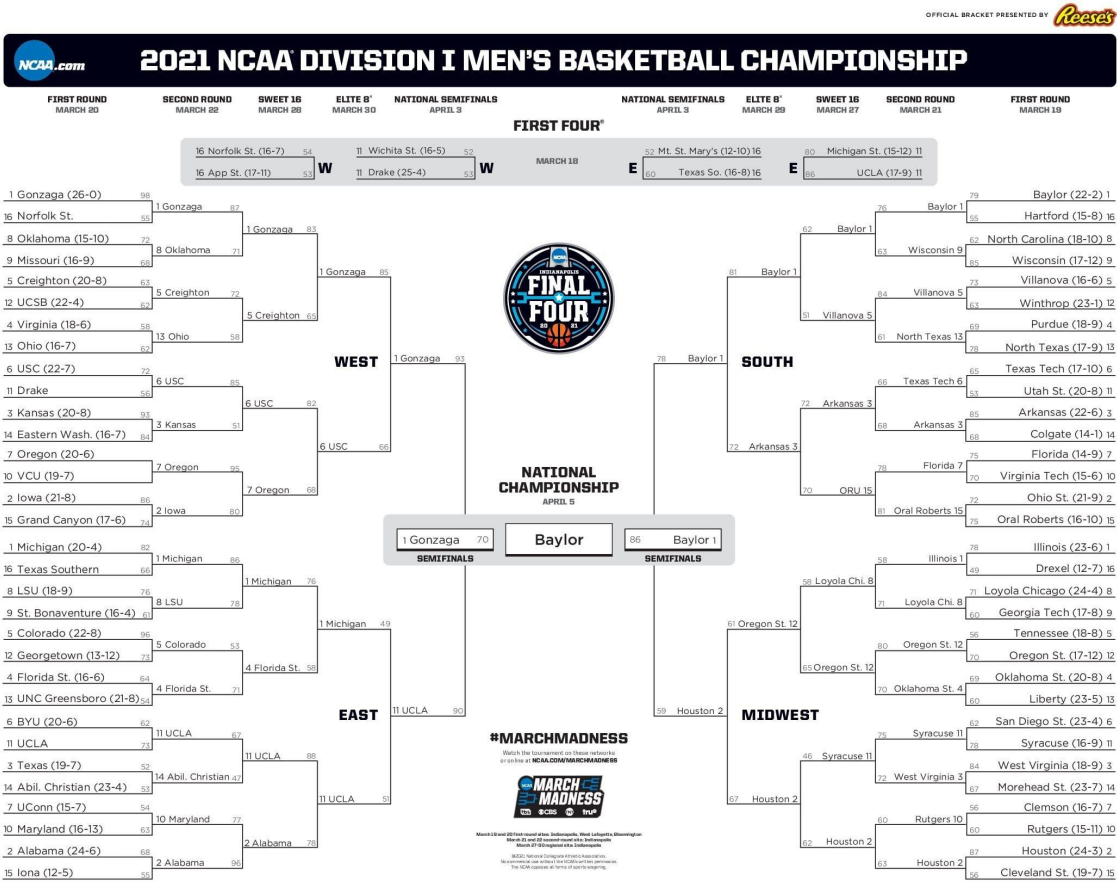
updated: 2021-04-15



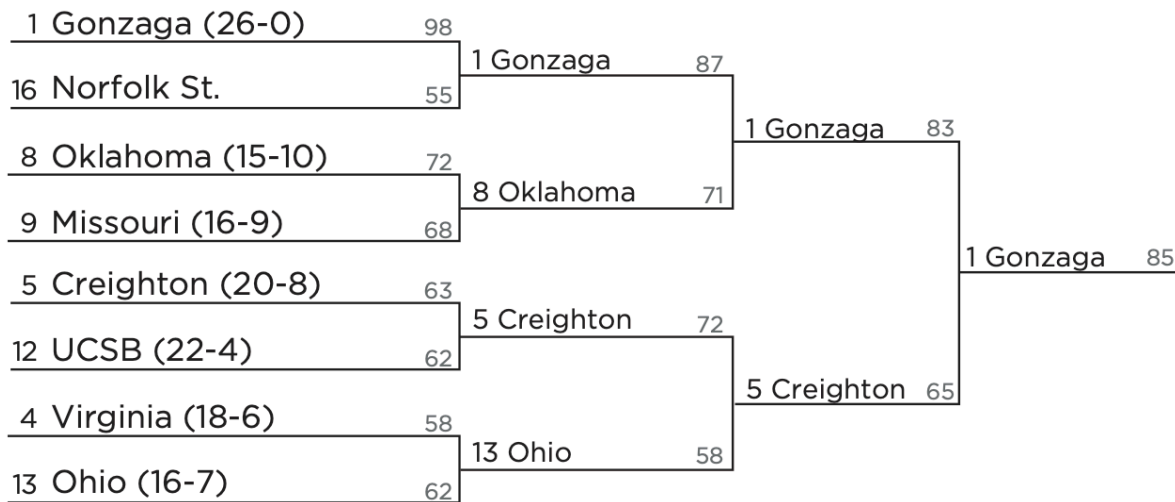
Background on March Madness Tournament

- First hosted in 1939 with 8 participating teams.
- Now, restricted to the top 64 qualifying NCAA division I teams.
- Currently 16.9 million viewers.
- Michigan State won in 1979 and 2000.
- Offices, organizations, and friend groups make and submit a total of 40 Million predictions each year.
- At the beginning small groups form and gamble on the outcomes by selecting winners and losers.
- Nobody has ever made a perfect bracket prediction and there are 2^{63} possible outcomes.





Zoomed in



We use basketball season game scores and statistics to predict this year's tournament outcome.

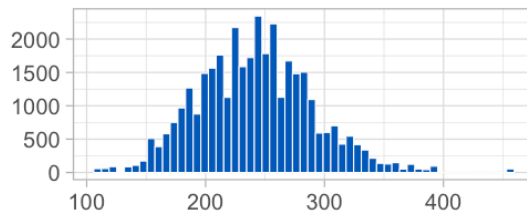
We need to derive a scoring metric.

Evaluate and tune the model.

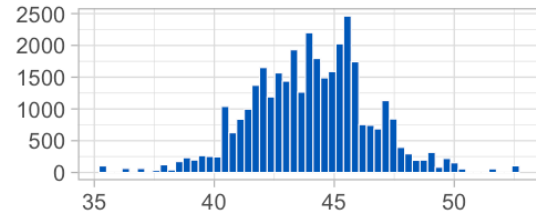
Compare the model's performance.

We made several models, in the interest of time we are going to discuss the logistic because its the most straight-forward.

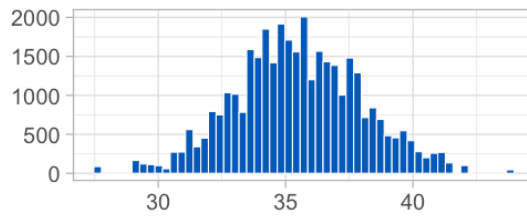
Three-Points Goals



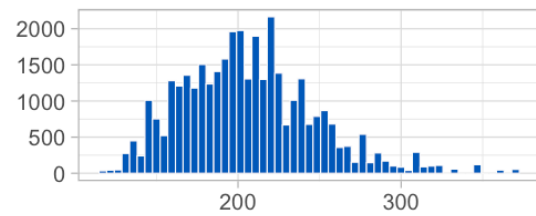
Percentage of Goals Scored



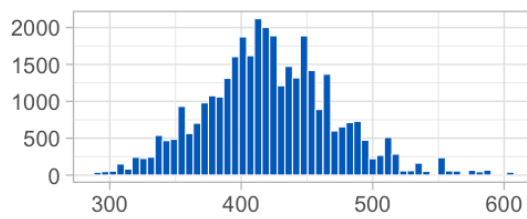
Rebounds Per Game



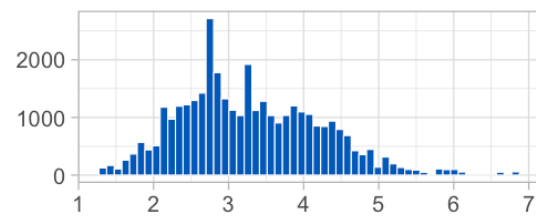
Steals Per Game



Turnovers Per Game



Blocks Per Game



Real Quick Descriptive Statistics

	Mean	Median	Std. Dev	Min	Max
Field Goal Percentage	44.03	44.10	2.46	35.4	52.60
Free Throw Percentage	70.65	70.60	3.73	58.0	80.20
Rebounds Per Game	35.48	35.35	2.49	27.6	43.81
Steals	206.34	202.00	40.59	116.0	369.00
Turnovers	421.84	419.00	49.01	290.0	604.00
Blocked Shots Per Game	3.30	3.20	0.94	1.3	6.80
	Games Played	Teams	Tournament		
	35248	343	64		

- Normal-ish distributions.
- No catastrophic skews or heavy tails.
- Ideally we would have had steals and turn-overs per game rather than season total. Not all played the same game.

Logistic Model

$$\text{win} \sim \text{Bernoulli} \left(\text{prob}_{\text{win}=1} = \hat{P} \right)$$

$$\log \left[\frac{\hat{P}}{1 - \hat{P}} \right] = \beta_0 + \beta_1(\text{x3fg}) + \beta_2(\text{opposingx3fg}) + \beta_3(\text{fg_percent}) + \beta_4(\text{opposingfg_percent}) + \beta_5(\text{ft_percent}) + \beta_6(\text{opposingft_percent}) + \beta_7(\text{rpg}) + \beta_8(\text{opposingrpg}) + \beta_9(\text{st}) + \beta_{10}(\text{opposingst}) + \beta_{11}(\text{to}) + \beta_{12}(\text{opposingto}) + \beta_{13}(\text{opposingbkpg}) + \beta_{14}(\text{bkpg})$$

Summary

Term	Estimate	Std. Error	T-Stat	P-Value
Intercept	-0.333	0.523	-0.636	0.52498
Three Pointers	0.002	0.000	8.832	0.00000
O* Three Pointers	-0.003	0.000	-9.161	0.00000
Field Goal Percentage	0.183	0.005	33.439	0.00000
O* Field Goal Percentage	-0.176	0.005	-32.748	0.00000
Free-throw percentage	0.032	0.004	9.106	0.00000
O* Free-throw percentage	-0.031	0.004	-8.718	0.00000
Rebounds	0.143	0.005	26.359	0.00000
O* Rebounds	-0.145	0.005	-26.736	0.00000
Steals	0.007	0.000	19.885	0.00000
O* Steals	-0.006	0.000	-19.700	0.00000
Turnovers	-0.006	0.000	-21.856	0.00000
O* Turnovers	0.006	0.000	21.558	0.00000
Blocks	-0.143	0.014	-10.188	0.00000
O* Blocks	0.165	0.014	11.652	0.00000

Where O* is the Mertric associated with their opposing team

	LRT	DEViance
Stat	8194.126	40659.49
P Value	0.000	0.00

Summary

Term	Estimate	Std. Error	T-Stat	P-Value
Intercept	-0.333	0.523	-0.636	0.52498
Three Pointers	0.002	0.000	8.832	0.00000
O* Three Pointers	-0.003	0.000	-9.161	0.00000
Field Goal Percentage	0.183	0.005	33.439	0.00000
O* Field Goal Percentage	-0.176	0.005	-32.748	0.00000
Free-throw percentage	0.032	0.004	9.106	0.00000
O* Free-throw percentage	-0.031	0.004	-8.718	0.00000
Rebounds	0.143	0.005	26.359	0.00000
O* Rebounds	-0.145	0.005	-26.736	0.00000
Steals	0.007	0.000	19.885	0.00000
O* Steals	-0.006	0.000	-19.700	0.00000
Turnovers	-0.006	0.000	-21.856	0.00000
O* Turnovers	0.006	0.000	21.558	0.00000
Blocks	-0.143	0.014	-10.188	0.00000
O* Blocks	0.165	0.014	11.652	0.00000

Where O* is the Mertric associated with their opposing team

	LRT	DEViance
Stat	8194.126	40659.49
P Value	0.000	0.00

Summary

Term	Estimate	Std. Error	T-Stat	P-Value
Intercept	-0.333	0.523	-0.636	0.52498
Three Pointers	0.002	0.000	8.832	0.00000
O* Three Pointers	-0.003	0.000	-9.161	0.00000
Field Goal Percentage	0.183	0.005	33.439	0.00000
O* Field Goal Percentage	-0.176	0.005	-32.748	0.00000
Free-throw percentage	0.032	0.004	9.106	0.00000
O* Free-throw percentage	-0.031	0.004	-8.718	0.00000
Rebounds	0.143	0.005	26.359	0.00000
O* Rebounds	-0.145	0.005	-26.736	0.00000
Steals	0.007	0.000	19.885	0.00000
O* Steals	-0.006	0.000	-19.700	0.00000
Turnovers	-0.006	0.000	-21.856	0.00000
O* Turnovers	0.006	0.000	21.558	0.00000
Blocks	-0.143	0.014	-10.188	0.00000
O* Blocks	0.165	0.014	11.652	0.00000

Where O* is the Mertric associated with their opposing team

	LRT	DEViance
Stat	8194.126	40659.49
P Value	0.000	0.00

Summary

Term	Estimate	Std. Error	T-Stat	P-Value
Intercept	-0.333	0.523	-0.636	0.52498
Three Pointers	0.002	0.000	8.832	0.00000
O* Three Pointers	-0.003	0.000	-9.161	0.00000
Field Goal Percentage	0.183	0.005	33.439	0.00000
O* Field Goal Percentage	-0.176	0.005	-32.748	0.00000
Free-throw percentage	0.032	0.004	9.106	0.00000
O* Free-throw percentage	-0.031	0.004	-8.718	0.00000
Rebounds	0.143	0.005	26.359	0.00000
O* Rebounds	-0.145	0.005	-26.736	0.00000
Steals	0.007	0.000	19.885	0.00000
O* Steals	-0.006	0.000	-19.700	0.00000
Turnovers	-0.006	0.000	-21.856	0.00000
O* Turnovers	0.006	0.000	21.558	0.00000
Blocks	-0.143	0.014	-10.188	0.00000
O* Blocks	0.165	0.014	11.652	0.00000

Where O* is the Mertric associated with their opposing team

	LRT	DEViance
Stat	8194.126	40659.49
P Value	0.000	0.00

Summary

Term	Estimate	Std. Error	T-Stat	P-Value
Intercept	-0.333	0.523	-0.636	0.52498
Three Pointers	0.002	0.000	8.832	0.00000
O* Three Pointers	-0.003	0.000	-9.161	0.00000
Field Goal Percentage	0.183	0.005	33.439	0.00000
O* Field Goal Percentage	-0.176	0.005	-32.748	0.00000
Free-throw percentage	0.032	0.004	9.106	0.00000
O* Free-throw percentage	-0.031	0.004	-8.718	0.00000
Rebounds	0.143	0.005	26.359	0.00000
O* Rebounds	-0.145	0.005	-26.736	0.00000
Steals	0.007	0.000	19.885	0.00000
O* Steals	-0.006	0.000	-19.700	0.00000
Turnovers	-0.006	0.000	-21.856	0.00000
O* Turnovers	0.006	0.000	21.558	0.00000
Blocks	-0.143	0.014	-10.188	0.00000
O* Blocks	0.165	0.014	11.652	0.00000

Where O* is the Mertric associated with their opposing team

	LRT	DEViance
Stat	8194.126	40659.49
P Value	0.000	0.00

Summary

Term	Estimate	Std. Error	T-Stat	P-Value
Intercept	-0.333	0.523	-0.636	0.52498
Three Pointers	0.002	0.000	8.832	0.00000
O* Three Pointers	-0.003	0.000	-9.161	0.00000
Field Goal Percentage	0.183	0.005	33.439	0.00000
O* Field Goal Percentage	-0.176	0.005	-32.748	0.00000
Free-throw percentage	0.032	0.004	9.106	0.00000
O* Free-throw percentage	-0.031	0.004	-8.718	0.00000
Rebounds	0.143	0.005	26.359	0.00000
O* Rebounds	-0.145	0.005	-26.736	0.00000
Steals	0.007	0.000	19.885	0.00000
O* Steals	-0.006	0.000	-19.700	0.00000
Turnovers	-0.006	0.000	-21.856	0.00000
O* Turnovers	0.006	0.000	21.558	0.00000
Blocks	-0.143	0.014	-10.188	0.00000
O* Blocks	0.165	0.014	11.652	0.00000

Where O* is the Mertric associated with their opposing team

	LRT	DEViance
Stat	8194.126	40659.49
P Value	0.000	0.00

Summary

Term	Estimate	Std. Error	T-Stat	P-Value
Intercept	-0.333	0.523	-0.636	0.52498
Three Pointers	0.002	0.000	8.832	0.00000
O* Three Pointers	-0.003	0.000	-9.161	0.00000
Field Goal Percentage	0.183	0.005	33.439	0.00000
O* Field Goal Percentage	-0.176	0.005	-32.748	0.00000
Free-throw percentage	0.032	0.004	9.106	0.00000
O* Free-throw percentage	-0.031	0.004	-8.718	0.00000
Rebounds	0.143	0.005	26.359	0.00000
O* Rebounds	-0.145	0.005	-26.736	0.00000
Steals	0.007	0.000	19.885	0.00000
O* Steals	-0.006	0.000	-19.700	0.00000
Turnovers	-0.006	0.000	-21.856	0.00000
O* Turnovers	0.006	0.000	21.558	0.00000
Blocks	-0.143	0.014	-10.188	0.00000
O* Blocks	0.165	0.014	11.652	0.00000

Where O* is the Mertric associated with their opposing team

	LRT	DEViance
Stat	8194.126	40659.49
P Value	0.000	0.00

Summary

Term	Estimate	Std. Error	T-Stat	P-Value
Intercept	-0.333	0.523	-0.636	0.52498
Three Pointers	0.002	0.000	8.832	0.00000
O* Three Pointers	-0.003	0.000	-9.161	0.00000
Field Goal Percentage	0.183	0.005	33.439	0.00000
O* Field Goal Percentage	-0.176	0.005	-32.748	0.00000
Free-throw percentage	0.032	0.004	9.106	0.00000
O* Free-throw percentage	-0.031	0.004	-8.718	0.00000
Rebounds	0.143	0.005	26.359	0.00000
O* Rebounds	-0.145	0.005	-26.736	0.00000
Steals	0.007	0.000	19.885	0.00000
O* Steals	-0.006	0.000	-19.700	0.00000
Turnovers	-0.006	0.000	-21.856	0.00000
O* Turnovers	0.006	0.000	21.558	0.00000
Blocks	-0.143	0.014	-10.188	0.00000
O* Blocks	0.165	0.014	11.652	0.00000

Where O* is the Mertric associated with their opposing team

	LRT	DEViance
Stat	8194.126	40659.49
P Value	0.000	0.00

Scoring

$$\sum_{i=1}^n [\mathbf{1}_{\{\text{correct prediction}\}} \hat{p}_i - \mathbf{1}_{\{\text{incorrect perdition}\}} \hat{p}_i] = \text{Model Score}$$

- Basketball games are often quite close and odds in betting markets are often close to 1:1.
- Our goal is to weight correct prediction higher when it assigns a higher probability. Conversely, a lower score to an incorrect prediction.
- If our predictions are no better than randomly guessing $E(\text{Model Score}) = 0$.

Scoring

$$\sum_{i=1}^n [\mathbf{1}_{\{\text{correct prediction}\}} \hat{p}_i - \mathbf{1}_{\{\text{incorrect prediction}\}} \hat{p}_i] = \text{Model Score}$$

Minimal Example

Game	Predicted Winner	Winner	\hat{p}	Score
Abilenechristian Vs. Texas	Abilenechristian	Abilenechristian	0.52	0.52
Abilenechristian Vs. Ucla	Abilenechristian	Ucla	0.39	-0.61
Arkansas Vs. Baylor	Baylor	Baylor	0.53	0.53
Baylor Vs. Gonzaga	Gonzaga	Baylor	0.30	-0.70
Clevelandst Vs. Houston	Houston	Houston	0.87	0.87
Colgate Vs. Arkansas	Colgate	Arkansas	0.28	-0.72
Colorado Vs. Floridast	Floridast	Floridast	0.63	0.63
Creighton Vs. Gonzaga	Gonzaga	Gonzaga	0.80	0.80
Drexel Vs. Illinois	Illinois	Illinois	0.71	0.71
Easternwash Vs. Kansas	Easternwash	Kansas	0.43	-0.57

Scoring

$$\sum_{i=1}^n [\mathbf{1}_{\{\text{correct prediction}\}} \hat{p}_i - \mathbf{1}_{\{\text{incorrect prediction}\}} \hat{p}_i] = \text{Model Score}$$

Correct Predictions

Game	Predicted Winner	Winner	\hat{p}	Score
Abilenechristian Vs. Texas	Abilenechristian	Abilenechristian	0.52	0.52
Abilenechristian Vs. Ucla	Abilenechristian	Ucla	0.39	-0.61
Arkansas Vs. Baylor	Baylor	Baylor	0.53	0.53
Baylor Vs. Gonzaga	Gonzaga	Baylor	0.30	-0.70
Clevelandst Vs. Houston	Houston	Houston	0.87	0.87
Colgate Vs. Arkansas	Colgate	Arkansas	0.28	-0.72
Colorado Vs. Floridast	Floridast	Floridast	0.63	0.63
Creighton Vs. Gonzaga	Gonzaga	Gonzaga	0.80	0.80
Drexel Vs. Illinois	Illinois	Illinois	0.71	0.71
Easternwash Vs. Kansas	Easternwash	Kansas	0.43	-0.57

Scoring

$$\sum_{i=1}^n [\mathbf{1}_{\{\text{correct prediction}\}} \hat{p}_i - \mathbf{1}_{\{\text{incorrect prediction}\}} \hat{p}_i] = \text{Model Score}$$

Incorrect Predictions

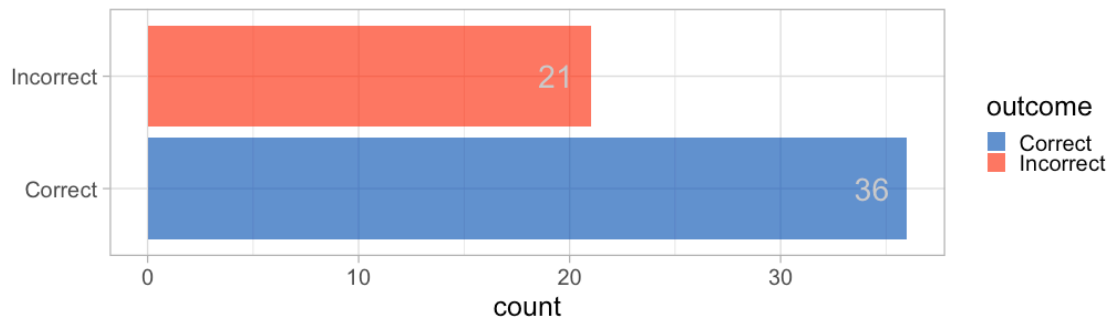
Game	Predicted Winner	Winner	\hat{p}	Score
Abilenechristian Vs. Texas	Abilenechristian	Abilenechristian	0.52	0.52
Abilenechristian Vs. Ucla	Abilenechristian	Ucla	0.39	-0.61
Arkansas Vs. Baylor	Baylor	Baylor	0.53	0.53
Baylor Vs. Gonzaga	Gonzaga	Baylor	0.30	-0.70
Clevelandst Vs. Houston	Houston	Houston	0.87	0.87
Colgate Vs. Arkansas	Colgate	Arkansas	0.28	-0.72
Colorado Vs. Floridast	Floridast	Floridast	0.63	0.63
Creighton Vs. Gonzaga	Gonzaga	Gonzaga	0.80	0.80
Drexel Vs. Illinois	Illinois	Illinois	0.71	0.71
Easternwash Vs. Kansas	Easternwash	Kansas	0.43	-0.57

Scoring

$$\sum_{i=1}^n [\mathbf{1}_{\{\text{correct prediction}\}} \hat{p}_i - \mathbf{1}_{\{\text{incorrect prediction}\}} \hat{p}_i] = \text{Model Score}$$

Score

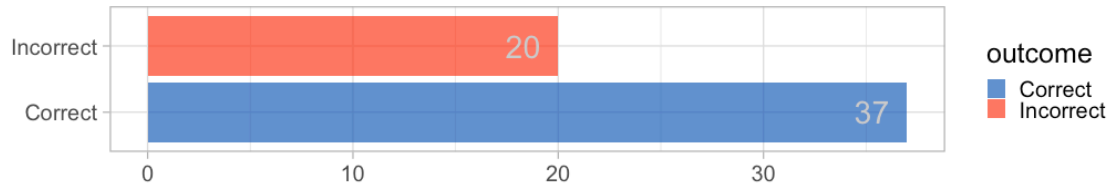
Game	Predicted Winner	Winner	\hat{p}	Score
Abilenechristian Vs. Texas	Abilenechristian	Abilenechristian	0.52	0.52
Abilenechristian Vs. Ucla	Abilenechristian	Ucla	0.39	-0.61
Arkansas Vs. Baylor	Baylor	Baylor	0.53	0.53
Baylor Vs. Gonzaga	Gonzaga	Baylor	0.30	-0.70
Clevelandst Vs. Houston	Houston	Houston	0.87	0.87
Colgate Vs. Arkansas	Colgate	Arkansas	0.28	-0.72
Colorado Vs. Floridast	Floridast	Floridast	0.63	0.63
Creighton Vs. Gonzaga	Gonzaga	Gonzaga	0.80	0.80
Drexel Vs. Illinois	Illinois	Illinois	0.71	0.71
Easternwash Vs. Kansas	Easternwash	Kansas	0.43	-0.57
$\sum_i \text{Score}$				
Model Score = 1.46				



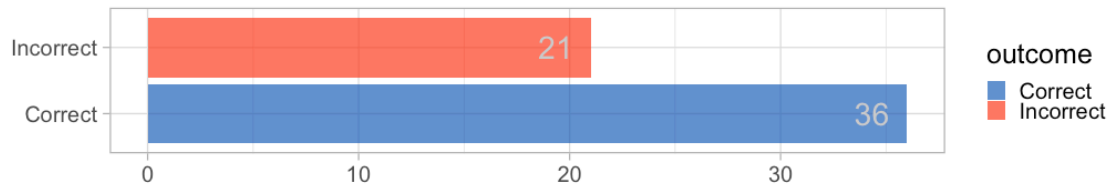
Result	Score	Total Score	11.24
Correct	24.71	Standard Error	0.65
Incorrect	-13.47	T-Statistic	17.17
Total	11.24	P-value	0.00

Model Performance

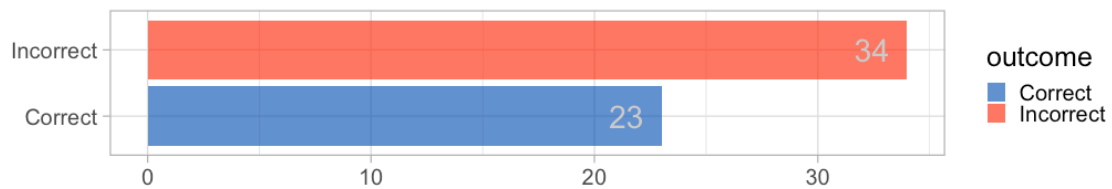
#1 Poisson Model



#2 Logistic Model



#3 Multinomeal Model



Conclusion

Both Logistic and Poisson Regression perform well.

However, we are scoring the model based on independent predictions. While real-world predictions are made dependent on the previous round.

These predictions are not wildly different than betting markets and seeds because they use similar statistics to derive their predictions.