Data: Tournament Simulation

Of particular interest in simulating the 2020 Tournament are datasets that encapsulate teams' summary statstics for any given season (2013-2021) and another that recounts the history of the Tournament by game. The first of these datasets contains a number of important summary statistics, many of which were discussed earlier in this section (e.g., ADJOE, ADJDE, ORB, TOR); each statistic is a representation of some aspect of the game of basketball, and higher/lower values of some statistics are often accompanied by higher win rates (see the Results section for additional commentary). Figure 1 provides a visualization of how this dataset is structured; note that, for 2020, the 'RESULT','SEED', and 'POSTSEASON' columns do not exist (these indicate how far a team went in the tournament, their seed that year, and whether or not they were invited, respectively).

	TEAM	CONF	G	W	ADJOE	ADJDE	BARTHAG	EFG_O	EFG_D	TOR
1	Michigan	B10	24	20	118.10	91.10	0.95	54.90	44.90	16.30
2	Baylor	B12	24	22	123.20	94.50	0.95	57.50	49.10	17.60
3	Illinois	B10	29	23	117.70	90.40	0.95	55.60	46.60	18.20
4	Gonzaga	WCC	26	26	125.40	89.80	0.98	61.00	47.50	16.10
5	Iowa	B10	29	21	123.50	95.70	0.95	54.60	48.30	13.30

Figure 1: Sample Summary Statistics from 2021 Season (some columns not shown due to cutoff)

Additionally, Figure 2 shows a small sample of games from the history of the NCAA Tournament (all games since 1985 were stored in this format).

Year	Seed1	Score1	Team1	Team2	Score2	Seed2	ScoreDiff	Upset
2013	1	79	Louisville	North Carolina A&T	48	16	31	0
2013	2	73	Duke	Albany	61	15	12	0
2013	3	65	Michigan St.	Valparaiso	54	14	11	0
2013	4	64	Saint Louis	New Mexico St.	44	13	20	0

Figure 2: Sample Data from History of NCAA Tournament Games

Sourcing Used Datasets

The following links or methods of obtaining the predominant resources used in the compilation of this report:

*Kaggle Repository containing dataset(s) depicted in Figure 1: Link

- ESPN Link to 2021 Tournament Data (merged with Kaggle Repository to include results from this year): Link
- Link to Warren Nolan's website containing historical BPI Data (merged with Kaggle Repository and ESPN data): Link
- Data.World Link containing datasets depicted in Figure 2: Link

Results: Simulating the 2020 Tournament

As the world grinded to a halt in March of 2020, college basketball, which had been preparing for conference tournaments and the "Big Dance," as the tournament to which the entire season culminates is coloquially referred, stopped alongside it. The cancellation brought about speculation of who would have comprised the elusive fields of 68, what "Cinderella stories" would have been, and who would have emerged the Champion.

Fortunately, advancements in machine learning and classification techniques, along with their practical applications in the blossoming field of data science, afford the ability to re-create college basketball's greatest tournament. The following sections will be dedicated to the discussion of how the 2020 NCAA D1 MBB Tournament can, and will, be simulated. The first section will be dedicated to the definition of the aforementioned "field of 68" teams, along with relevant terminology and slang phrases used to discuss the results of such tournaments. Next, the development of classification tools to structure the tournament field will be discussed, which include methods such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN) or K-Means Clustering algorithms, and the use of Artifical Neural Networks (ANN). Finally, the same methods will be appropriated to creating methods of determining the victor of a particular match-up. These two facets comprise how the tournament progresses. Results of each round, along with the eventual champion (no virtual banner for UNC, unforunately), will be presented as well.

Relevant Terminology

Before discussing how any classification methods work or their respective performances, a framework for understanding them must be established, especially for audiences who are not acclimated nor familiar with the college basketball environment in recent years (i.e since 2011, when the 'First Four' games were included to broaden the field from 64 to 68 teams). Therefore, let the following list serve as a guide to the words and phrases repeatedly used throughout this analysis:

- *Field*: This is a coloquial term referencing the entire scope of teams who receive an invitation to the tournament.
- *Automatic [Bid]*: This term refers to teams who are conference champions within their contemporaneous and respective tournaments. Conference champions receive an invitation to the tournament, no matter from where they originate.
- *At-Large*: This term refers to teams who received an invitation to the tournament but were not automatically entered; a committee is responsible for determining these teams.
- *Seed*: A team's ranking within their region/pool, ranging from 1 (Best) to 16 (Worst).
- *Region*: The tournament is divided into four regions, and thus four of each seed enter the first round, dubbed the 'Round of 64.' (Six '11-seeds' and '16-seeds' are chosen, but four of each enter the 'First Four' games, likely determined by a committee.)
- *Upset*: Such refers to a game in which the result is a significantly lower-seeded team defeating a high-seeded team (e.g., see Duke-Lehigh in 2012, Duke-Mercer in 2014, Virginia-UMBC in 2019, Ohio State-Oral Roberts in 2021).

Other terms, such as the names given to different rounds of the tournament, should be self-explanatory; that is, the number in any respective name (aside of the Championship Round) indicates the number of remaining teams.

The Field of 68

Since 2011, in which the tournament has been held included 68 teams who shared one common goal: "Win or Go Home." But before such goals can be realized, they first are made aware of their initial opponent. A committee is appointed (as to who exactly comprises said group, the authors are unsure) to carry out the selection process, which involves the assignment of seeds 1-16 to teams based on a number of statistics (e.g., Strength of Schedule, or SOS, which is a ranking given to identify the team(s) with the most difficult slate of games over the entire season). Four teams are allotted each seed, with the aforementioned explanation of two additional seeds being offered to fill out the first games that are played the day before the tournament is scheduled to begin. As a result, "1-seed" teams play "16-seed" opponents, "2-seed" teams play "15-seed" opponents, and so on. Teams are then split into four regions, where each region is a microcosm consisting of a single team seeded 1-16. The winners of each region arrive at the 'Final Four' round, and the winners of those games compete for the title.

Methods

Accurately structuring any tournament field can be difficult, and March Madness is no exception. Bracketologists (such as Joe Lunardi), individuals who employ analysis of historical records and current performances to predict fields, labor throughout the season to communicate to the public their opinions on who will reach the postseason. This section allows the authors to assume this role (and, when the time comes, hopefully win their next office or family bracket competitions) with the help of aforementioned classification techniques.

The goal of this subsection is to structure the field of 68 teams that would have hypothetically competed for the championship. Such is a two-step process: first, the field will be generated by determining worthiness for a postseason berth. Afterwards, teams will be seeded according to a separate set of classifiers. Finally, regression techniques will be used to simulate matchups designed in a traditional tournament format (see Appendix for the full results)!

It is worth noting that all of the described methods were tested on data *independent* from data on the 2020 season beforehand to ensure validity of the process (e.g., prevent overfitting and poor classification amongst test data) and become acclimated with the programming necessary to achieve an end-goal. Some of the included output will be from randomly selected test data, others exclusively the 2020 season.

Postseason Berth

SVM

Simply put, the term "postseason berth" refers to any given team being included in the tournament, whether in the First Four or Round of 64. In order to simulate this for the 2020 season, classifiers were trained on data from the 2013-2021 tournaments (excluding 2020). No weighting nor mutation of the data was conducted as a result of how far a team progressed in the tournament, but instead only their presence was noted.

A train-test split ratio of 80/20 (that is, trained on 80% of the data, tested on 20%) was used in the generation of an SVM classifier. Hyperparameters, along with their respective results of Accuracy ("Acc"), Precision ("Prec"), and Sensitivity ("Sens"), common metrics used to describe SVMs and Confusion Matrices, were used to tune the classifier; the best set of aforementioned metrics belonged to the parameters $C = 1, \gamma = 1$, where C is the 'cost' parameter and γ the 'gamma' parameter.

Additionally, when tested, this SVM performed relatively well, as indicated by the confusion matrix

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} = \begin{bmatrix} 437 & 15 \\ 28 & 80 \end{bmatrix}$$

where TP signifies "True Positive," FN "False Negative," and the like.

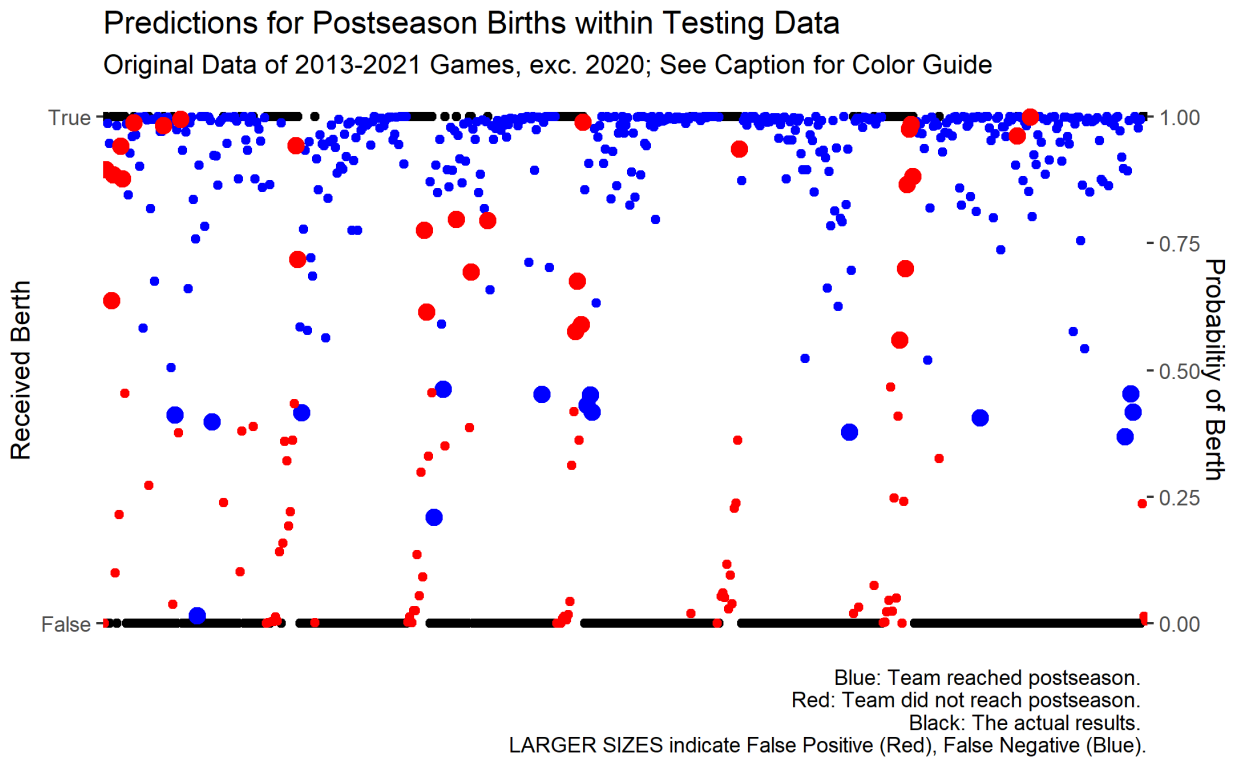

Figure 3: Probabilities of Postseason Berths, 2013-2019 and 2021)

As evidenced by the above plot and confusion matrix, the SVM was confident in many of its classifications of teams' eligibility for postseason appearances, so much so that it failed to identify an adequate-sized field when validated on data from 2020. To combat this shortcoming, any results marking a team eligible for an appearance were coupled with the necessary number of *low-confidence* predictions of a team *not* making it into the postseason (e.g., as 46 teams were automatically selected, 22 were chosen among those initially not selected).

KNN

In addition to an SVM, a K-Means Clustering approach known as "K-Nearest Neighbors" was introduced. To tune the hyperparameter k for this algorithm, classifiers tested on data from 2020 were compared with respect to $\mu_{pv!=av}$. Note that 'pv' and 'av' represent the predicted and actual values (e.g., the predictions and test dataset, respectively). The plot on the following page shows that a value of $k = 6$ minimized this function and was thus used in the determination of postseason berths.

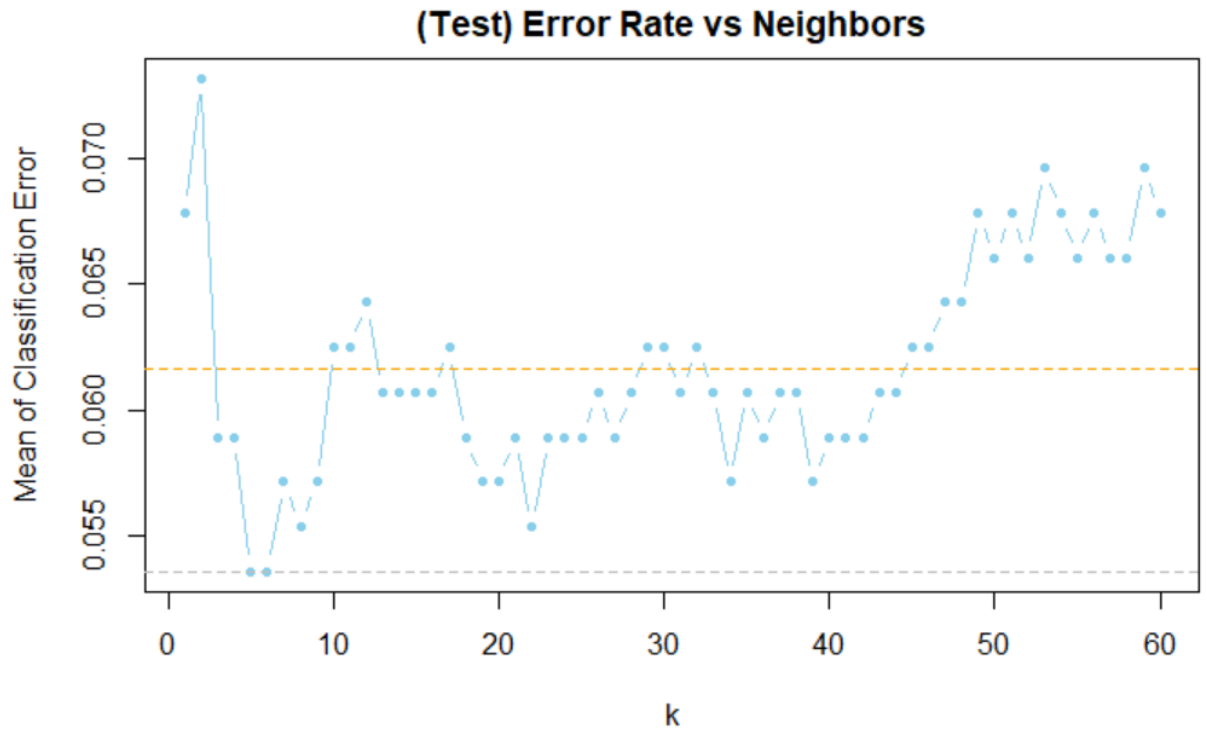

Figure 4: Plot of Hyperparameter Tuning for Postseason-Berth KNN Classifier)

When validated using data from 2020, a similar scenario occurred to that when using the SVM classifier; as a result, 49 teams received automatic berths, and 19 were chosen among those initially not selected.

ANN

Finally, an Artificial Neural Network was included for determining postseason eligibility. The model was trained and tested on the same data sources as the SVM and KNN before also being validated with data from 2020. Surprisingly, this network performed extremely well, boasting a "train-accuracy" of ~0.925 and "test-accuracy" of ~0.916.

The model structure is defined below (taken from output of the .py file from where the model originates).

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_3 (Dense)              (None, 64)                1280
_________________________________________________________________
dense_4 (Dense)              (None, 32)                2080
_________________________________________________________________
dense_5 (Dense)              (None, 2)                 66
=================================================================
Total params: 3,426
Trainable params: 3,426
Non-trainable params: 0
_________________________________________________________________
```

Figure 5: Table of Postseason ANN Structure)

Field Generation

When each of the aforementioned classifiers were validated using data from 2020, their predictions were used to each determine a set of teams to label with seeds. The code (under the 'Field Generation' section of the main script) for the following section employs filtered datasets from the validation data for each set of predictions. Rather than re-hash descriptions of each classifier for this different purpose (because accuracies cannot be measured, as there are no values to use as comparison), the results of each classifier will be presented, along with an explanation of necessary methods of imputation and analysis to arrive at a finalized tournament field.

For *each* classifier, its prediction indices for postseason berths were cross-referenced with team-names from the validation data to obtain a list of 68 teams. As mentioned before, if a classifier failed to predict the necessary number of berths, then additional values were imputed based on having the lowest remaining probabilities of *not* being given an invitation. Each list was then trained on a separate set of classifiers (of the same techniques as above) that labeled each team with a seed. As a result, each seed classifier yielded a *68x2* table represent its predicted field and seeds for each field.

The three tables were then joined, allowing for missing values, to arrive at a *105x3* table. Automatic bids were determined by selecting teams who had predicted values for all three classifiers (before imputation), ranking their average seeds, and seeding in chunks of 4 teams per increment. Missing values were imputed with the value 16 as any classifier who desired *not* to include them in predictions communicated that something about such teams' statistics did not warrant a postseason appearance and, as 16-seeds often have an early exodus (only *1* in the entire tournament history has won a game), such value acts as a penalty to ultimately match such teams with higher-seeds who are, from a historical perspective, more likely to win.

At-large bids were selected under the following process: first, the average seed of all teams not receiving automatic bids were computed and ranked in ascending order. Then, the highest n rankings were selected; in practice, 25 teams were selected after 43 received automatic bids. It should be admitted that **this process is different than how selection committees normally proceed**, with conference champions being given automatic seeds and all other teams competing for at-large bids; the authors felt that the process used in this report incorporated a rigorous use of classification techniques and that such process would be better suited without the premature filtration of teams that were not conference champions in 2020.

The following table, beginning on the next page, shows the rankings of the 43 automatic bids and 25 at-large bids (the 'Rank' column starts anew because the two groups were compiled and ranked separately):

TEAM	KNN	SVM	ANN	means	Rank	Seed
Gonzaga	1	1	1	1	1	1
Kansas	2	1	1	1.333333	2	1
Baylor	2	3	1	2	3	1
Dayton	2	3	1	2	4	1
San Diego St.	5	2	1	2.666667	5	2
Florida St.	3	4	1	2.666667	6	2
Duke	3	3	3	3	7	2
Creighton	5	3	1	3	8	2
Villanova	5	3	1	3	9	3
Ohio St.	2	5	3	3.333333	10	3
Michigan St.	7	3	3	4.333333	11	3
Louisville	5	5	3	4.333333	12	3
Penn St.	3	5	5	4.333333	13	4
Oregon	5	5	3	4.333333	14	4
Kentucky	6	4	3	4.333333	15	4
Virginia	5	5	3	4.333333	16	4
Houston	2	7	5	4.666667	17	5
Maryland	10	3	1	4.666667	18	5
Wisconsin	5	7	3	5	19	5
Seton Hall	7	5	3	5	20	5
Butler	5	7	3	5	21	6
West Virginia	5	6	5	5.333333	22	6
Arizona	3	5	11	6.333333	23	6
Michigan	6	8	5	6.333333	24	6
Illinois	6	8	5	6.333333	25	7
Auburn	11	6	3	6.666667	26	7
Iowa	8	8	5	7	27	7
Oklahoma	10	7	6	7.666667	28	7
Wichita St.	5	7	11	7.666667	29	8
BYU	11	7	6	8	30	8
Saint Mary's	11	7	6	8	31	8
USC	5	8	11	8	32	8
LSU	6	8	11	8.333333	33	9
Cincinnati	7	8	11	8.666667	34	9
Rutgers	9	8	11	9.333333	35	9
Marquette	7	10	11	9.333333	36	9
Richmond	10	7	11	9.333333	37	10
Providence	7	10	11	9.333333	38	10
Texas Tech	7	9	13	9.666667	39	10
Florida	7	11	11	9.666667	40	10
Colorado	11	8	11	10	41	11
Indiana	11	11	11	11	42	11
Utah St.	13	11	11	11.66667	43	11
Purdue	7	10	16	11	1	11
Georgia Tech	5	16	16	12.33333	2	11
Saint Louis	9	16	13	12.66667	3	11
Northern Iowa	11	16	11	12.66667	4	12
East Tennessee St.	12	16	11	13	5	12
Liberty	10	16	14	13.33333	6	12
Texas	13	16	11	13.33333	7	12
Washington	9	16	16	13.66667	8	13
Tennessee	9	16	16	13.66667	9	13

TEAM	KNN	SVM	ANN	means	Rank	Seed
Arkansas	11	16	15	14	10	13
Oklahoma St.	10	16	16	14	11	13
St. John's	10	16	16	14	12	14
San Francisco	10	16	16	14	13	14
Northern Colorado	11	16	16	14.33333	14	14
Arizona St.	16	16	11	14.33333	15	14
Rhode Island	16	16	11	14.33333	16	15
Stephen F. Austin	16	16	11	14.33333	17	15
Davidson	13	16	16	15	18	15
SMU	13	16	16	15	19	15
Iowa St.	13	16	16	15	20	16
Memphis	16	16	13	15	21	16
Yale	16	16	13	15	22	16
South Carolina	14	16	16	15.33333	23	16
VCU	15	16	16	15.66667	24	16
Louisiana Tech	16	16	15	15.66667	25	16

Figure 6: Rankings and Seeds of Simulated 2020 Tournament Teams

Tournament Simulation

In order to simulate the tournament with the above list of teams, a method to determine a victor for any given contest must first be derived. Using the "TEAM", "Team1", and "Team2" columns from the datasets shown in Figures 1 and 2 as primary keys, the following process was carried out:

- For all games from 2013-2019, compute the differences of all quantitative statistics (except 'SEED' and other categorical variables) between the two teams that competed in each contest (Figure 7).
- Create a regression model that predicts the difference in score (along with its sign, as it indicates which team emerged victorious).
- For teams in Figure 6, create and simulate games according to whether or not the regression model returned a positive or negative value (the magnitude represents a quasi-"betting line", how many points one team is expected to win by). Repeat this step for each region of the tournament and record the victor(s).

The regression model initially contained all of the aforementioned differences, and was minimized using stepwise regression under the *BIC* criterion. Ultimately, the following regression equation was used to simulate the 2020 Tournament: $ScoreDiff \sim ADJOE\_DIFF + ADJDE\_DIFF + TORD\_DIFF + WAB\_DIFF$.

Figure 7 includes sample differences from the games depicted in Figure 1 (the figures are snapshots of the same dataset).

ADJOE_DIFF	ADJDE_DIFF	BARTHAG_DIFF	EFG_O_DIFF	EFG_D_DIFF
25.3	-11.2	0.628	4.9	-0.4
17.2	-7.2	0.378	4.3	-2.8
3.6	-10.1	0.188	-5.1	-1.7
5.5	-7.8	0.219	1.1	0.1
5.5	-7.8	0.219	1.1	0.1

Figure 7: Differenced Statistics for Games from Figure 1 (most columns not depicted due to cutoff).

The difference statistics were created and passed to the regression model for each simulated game. Figure 8 shows the results of simulating each round of the tournament, including the victor and the difference in score (negative values indicate an "upset" victory).

ID	DIFF	TEAM	ROUND
Saint Louis-Louisiana Tech	1.6	Saint Louis	First Four
Georgia Tech-Yale	5.2	Georgia Tech	First Four
Gonzaga-Iowa St.	10.3	Gonzaga	East - Round of 64
San Diego St.-Rhode Island	8.2	San Diego St.	East - Round of 64
Villanova-St. John's	1.6	Villanova	East - Round of 64
Penn St.-Washington	3	Penn St.	East - Round of 64
Houston-Northern Iowa	8.5	Houston	East - Round of 64
Butler-Colorado	2	Butler	East - Round of 64
Illinois-Richmond	2.4	Illinois	East - Round of 64
Wichita St.-LSU	0.4	Wichita St.	East - Round of 64
Gonzaga-Wichita St.	9.4	Gonzaga	East - Round of 32
San Diego St.-Illinois	4.5	San Diego St.	East - Round of 32
Villanova-Butler	1.7	Villanova	East - Round of 32
Penn St.-Houston	0.6	Penn St.	East - Round of 32
Gonzaga-Penn St.	6.2	Gonzaga	East - Sweet 16
San Diego St.-Villanova	3.5	San Diego St.	East - Sweet 16
Gonzaga-San Diego St.	5.8	Gonzaga	East - Elite 8
Dayton-VCU	8.2	Dayton	West - Round of 64
Creighton-SMU	9.7	Creighton	West - Round of 64
Louisville-Arizona St.	7	Louisville	West - Round of 64
Virginia-Oklahoma St.	-0.2	Oklahoma St.	West - Round of 64
Seton Hall-Texas	6.8	Seton Hall	West - Round of 64
Michigan-Purdue	-0.3	Purdue	West - Round of 64
Oklahoma-Florida	-2.9	Florida	West - Round of 64
USC-Marquette	-3.3	Marquette	West - Round of 64
Dayton-Marquette	8.4	Dayton	West - Round of 32
Creighton-Florida	4.6	Creighton	West - Round of 32
Louisville-Purdue	0.5	Louisville	West - Round of 32
Oklahoma St.-Seton Hall	-1.6	Seton Hall	West - Round of 32
Dayton-Seton Hall	7.1	Dayton	West - Sweet 16
Creighton-Louisville	3.4	Creighton	West - Sweet 16
Dayton-Creighton	3	Dayton	West - Elite 8
Baylor-South Carolina	9.2	Baylor	Midwest - Round of 64
Duke-Davidson	7.4	Duke	Midwest - Round of 64
Michigan St.-Northern Colorado	10.1	Michigan St.	Midwest - Round of 64
Kentucky-Arkansas	0.3	Kentucky	Midwest - Round of 64
Wisconsin-Liberty	6.7	Wisconsin	Midwest - Round of 64
Arizona-Utah St.	7.3	Arizona	Midwest - Round of 64
Iowa-Texas Tech	-3.2	Texas Tech	Midwest - Round of 64
Saint Mary's-Rutgers	-0.7	Rutgers	Midwest - Round of 64
Baylor-Rutgers	7.7	Baylor	Midwest - Round of 32
Duke-Texas Tech	2.1	Duke	Midwest - Round of 32
Michigan St.-Arizona	0.7	Michigan St.	Midwest - Round of 32
Kentucky-Wisconsin	-0.2	Wisconsin	Midwest - Round of 32
Baylor-Wisconsin	7.6	Baylor	Midwest - Sweet 16
Duke-Michigan St.	2.8	Duke	Midwest - Sweet 16
Baylor-Duke	2.5	Baylor	Midwest - Elite 8
Kansas-Memphis	11.3	Kansas	South - Round of 64

ID	DIFF	TEAM	ROUND
Florida St.-Stephen F. Austin	8.4	Florida St.	South - Round of 64
Ohio St.-San Francisco	8	Ohio St.	South - Round of 64
Oregon-Tennessee	5.5	Oregon	South - Round of 64
Maryland-East Tennessee St.	4.9	Maryland	South - Round of 64
West Virginia-Indiana	5.7	West Virginia	South - Round of 64
Auburn-Providence	-4.6	Providence	South - Round of 64
BYU-Cincinnati	5.4	BYU	South - Round of 64
Kansas-BYU	4.3	Kansas	South - Round of 32
Florida St.-Providence	-0.9	Providence	South - Round of 32
Ohio St.-West Virginia	3.1	Ohio St.	South - Round of 32
Oregon-Maryland	5.9	Oregon	South - Round of 32
Kansas-Oregon	4.7	Kansas	South - Sweet 16
Providence-Ohio St.	0	Providence	South - Sweet 16
Kansas-Providence	4.4	Kansas	South - Elite 8
Gonzaga-Dayton	1.6	Gonzaga	Final Four (East vs. West)
Kansas-Baylor	0.7	Kansas	Final Four (South vs. Midwest)
Gonzaga-Kansas	2.2	Gonzaga	Championship Game

Figure 8: Results of Simulating 2020 Tournament

Of particular interest are the games in which the regression model predicted an upset victory; there are a number of them occurring in the First Round (the "Round of 64"), one of emphasis being No. 13 Oklahoma St.'s victory over No. 5 Virginia in the West Region. Over time, however, *higher seeds* were predicted to prevail, with all four one-seeds reaching the Final Four games and Gonzaga emerging as the Champion. The authors would like to acknowledge this conclusion as an interesting one, considering that Gonzaga has reached the title game twice in the last 4 tournaments, losing most recently in 2021 to Baylor, and four years prior to North Carolina [in the Tar Heels' "Redemption Season", as many newspapers soon nicknamed it].