WHAT FACTORS PREDICT MEN'S COLLEGE BASKETBALL NCAA TOURNAMENT TEAMS?

Cole Jennings

TRACTABLE DATA

- The Kaggle datasets were originally scraped from “http://barttorvik.com/trank.php#”
- Some of the calculations and theory behind variables such as adjusted offensive efficiency come from kenpom.com, created by Ken Pomeroy
  - College basketball analytics and statistics

DATA RETRIEVAL

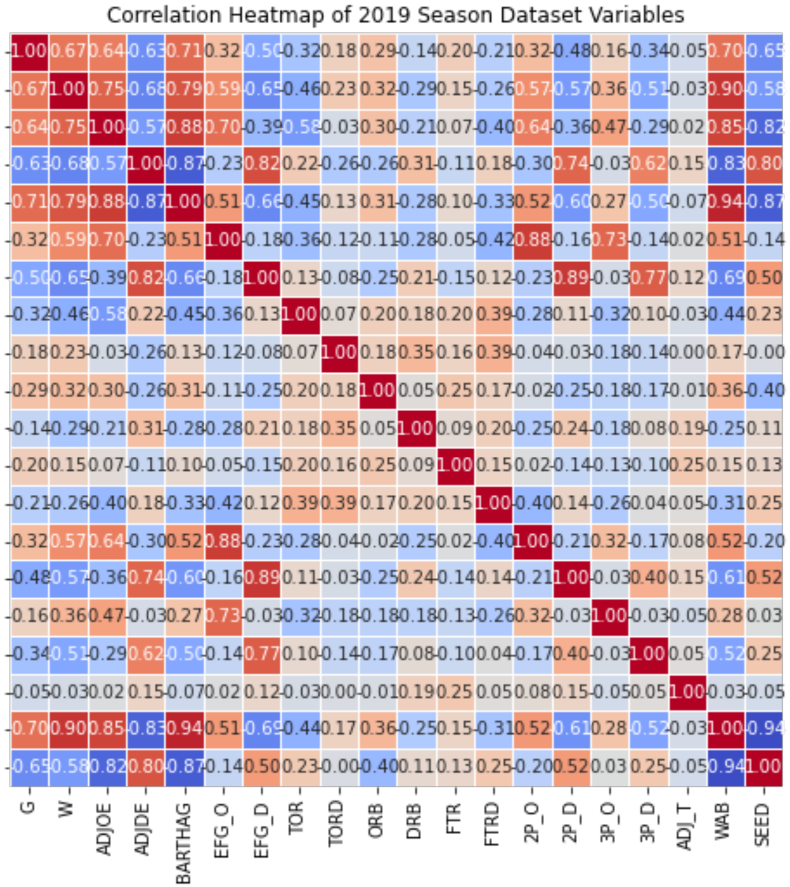
- Datasets obtained from Kaggle: 18-19, 21-22, 22-23 Men’s NCAA Division 1 College Basketball seasons
- Originally in csv files, read in using Python’s Pandas package
- The 3 different seasons’ data were downloaded separately and concatenated using Python

EXPLORATORY DATA ANALYSIS

- Variables include Team, Conference, Wins, Games Played, Postseason outcome, as well as various in game performance statistics
- I explore which variables are associated with making the NCAA Tournament, and build a model that can predict which teams will make it
- I added '2019' after each team in the 2019 dataset, and the same for 2022 and 2023 with their respective years
- I concatenated the three data frames so that they would be stacked on top of each other
- The new data frame has 1074 rows and 23 columns

VARIABLE CORRELATION

- Heatmap of all variables in the 2021 season dataset
- Sanity checks:
  - 2P_O (2-pt shooting %) and EFG_O (Effective Field Goal %) have a correlation coefficient of 0.88 (strong, positive relationship)
  - 2P_D and EFG_D have a 0.89 correlation coefficient

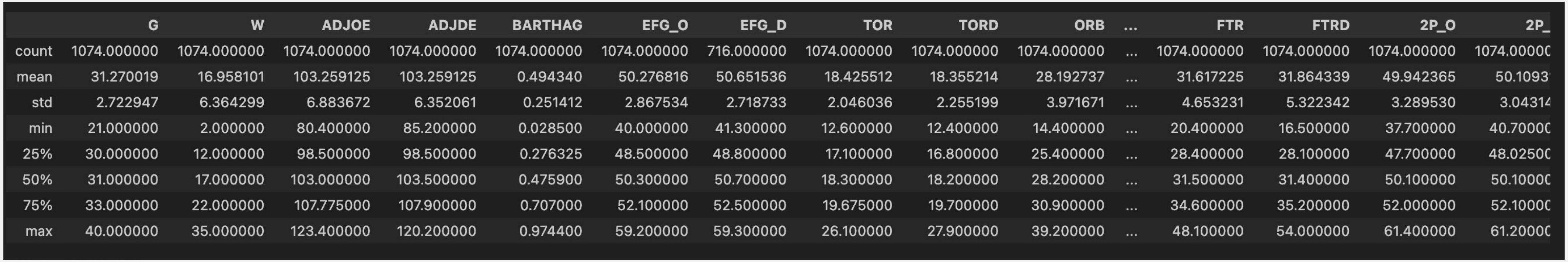


Correlation Heatmap of 2019 Season Dataset Variables

G	W	ADJOE	ADJDE	BARTHAG	EFG_O	EFG_D	TOR	TORD	ORB	DRB	FTR	FTRD	2P_O	2P_D	3P_O	3P_D	ADJ_T	WAB	SEED
1.00	0.67	0.64	-0.63	0.71	0.32	-0.50	-0.32	0.18	0.29	-0.14	0.20	-0.21	0.32	-0.48	0.16	-0.34	0.05	0.70	-0.65
0.67	1.00	0.75	-0.68	0.79	0.59	-0.65	-0.46	0.23	0.32	-0.29	0.15	-0.26	0.57	-0.57	0.36	-0.51	-0.03	0.90	-0.58
0.64	0.75	1.00	-0.57	0.88	0.70	-0.39	-0.58	-0.03	0.30	-0.21	0.07	-0.40	0.64	-0.36	0.47	-0.29	0.02	0.85	-0.82
-0.63	-0.68	-0.57	1.00	-0.87	-0.23	0.82	0.22	-0.26	-0.26	0.31	-0.11	0.18	-0.30	0.74	-0.03	0.62	0.15	-0.83	0.80
0.71	0.79	0.88	-0.87	1.00	0.51	-0.66	-0.45	0.13	0.31	-0.28	0.10	-0.33	0.52	-0.60	0.27	-0.50	-0.07	0.94	-0.87
0.32	0.59	0.70	-0.23	0.51	1.00	-0.18	-0.36	-0.12	-0.11	-0.28	-0.05	-0.42	0.88	-0.16	0.73	-0.14	0.02	0.51	-0.14
-0.50	-0.65	-0.39	0.82	-0.66	-0.18	1.00	0.13	-0.08	-0.25	0.21	-0.15	0.12	-0.23	0.89	-0.03	0.77	0.12	-0.69	0.50
-0.32	-0.46	-0.58	0.22	-0.45	-0.36	0.13	1.00	0.07	0.20	0.18	0.20	0.39	-0.28	0.11	-0.32	0.10	-0.03	-0.44	0.23
0.18	0.23	-0.03	-0.26	0.13	-0.12	-0.08	0.07	1.00	0.18	0.35	0.16	0.39	-0.04	-0.03	-0.18	-0.14	0.00	0.17	-0.00
0.29	0.32	0.30	-0.26	0.31	-0.11	-0.25	0.20	0.18	1.00	0.05	0.25	0.17	-0.02	-0.25	-0.18	-0.17	-0.01	0.36	-0.40
-0.14	-0.29	-0.21	0.31	-0.28	-0.28	0.21	0.18	0.35	0.05	1.00	0.09	0.20	-0.25	0.24	-0.18	0.08	0.19	-0.25	0.11
0.20	0.15	0.07	-0.11	0.10	-0.05	-0.15	0.20	0.16	0.25	0.09	1.00	0.15	0.02	-0.14	-0.13	-0.10	0.25	0.15	0.13
-0.21	-0.26	-0.40	0.18	-0.33	-0.42	0.12	0.39	0.39	0.17	0.20	0.15	1.00	-0.40	0.14	-0.26	0.04	0.05	-0.31	0.25
0.32	0.57	0.64	-0.30	0.52	0.88	-0.23	-0.28	-0.04	-0.02	-0.25	0.02	-0.40	1.00	-0.21	0.32	-0.17	0.08	0.52	-0.20
-0.48	-0.57	-0.36	0.74	-0.60	-0.16	0.89	0.11	-0.03	-0.25	0.24	-0.14	0.14	-0.21	1.00	-0.03	0.40	0.15	-0.61	0.52
0.16	0.36	0.47	-0.03	0.27	0.73	-0.03	-0.32	-0.18	-0.18	-0.18	-0.13	-0.26	0.32	-0.03	1.00	-0.03	-0.05	0.28	0.03
-0.34	-0.51	-0.29	0.62	-0.50	-0.14	0.77	0.10	-0.14	-0.17	0.08	-0.10	0.04	-0.17	0.40	-0.03	1.00	0.05	-0.52	0.25
0.05	-0.03	0.02	0.15	-0.07	0.02	0.12	-0.03	0.00	-0.01	0.19	0.25	0.05	0.08	0.15	-0.05	0.05	1.00	-0.03	-0.05
0.70	0.90	0.85	-0.83	0.94	0.51	-0.69	-0.44	0.17	0.36	-0.25	0.15	-0.31	0.52	-0.61	0.28	-0.52	-0.03	1.00	-0.94
-0.65	-0.58	-0.82	0.80	-0.87	-0.14	0.50	0.23	-0.00	-0.40	0.11	0.13	0.25	-0.20	0.52	0.03	0.25	-0.05	-0.94	1.00

SUMMARY STATISTICS

- The average # of wins in a season is 16.958
- The median adjusted offensive efficiency rating is 103.26
- The maximum effective field goal percentage observed is 59.2%

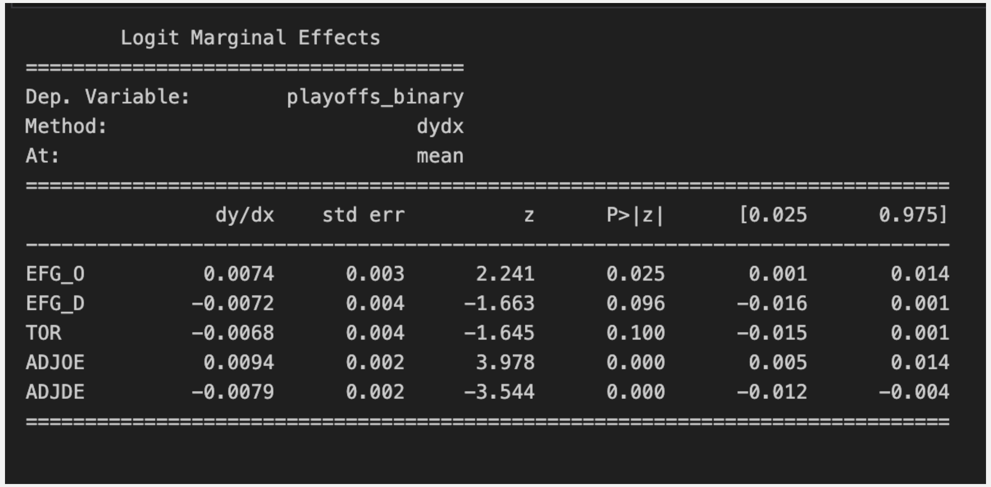


	G	W	ADJOE	ADJDE	BARTHAG	EFG_O	EFG_D	TOR	TORD	ORB	...	FTR	FTRD	2P_O	2P_
count	1074.000000	1074.000000	1074.000000	1074.000000	1074.000000	1074.000000	716.000000	1074.000000	1074.000000	1074.000000	...	1074.000000	1074.000000	1074.000000	1074.00000
mean	31.270019	16.958101	103.259125	103.259125	0.494340	50.276816	50.651536	18.425512	18.355214	28.192737	...	31.617225	31.864339	49.942365	50.1093
std	2.722947	6.364299	6.883672	6.352061	0.251412	2.867534	2.718733	2.046036	2.255199	3.971671	...	4.653231	5.322342	3.289530	3.04314
min	21.000000	2.000000	80.400000	85.200000	0.028500	40.000000	41.300000	12.600000	12.400000	14.400000	...	20.400000	16.500000	37.700000	40.70000
25%	30.000000	12.000000	98.500000	98.500000	0.276325	48.500000	48.800000	17.100000	16.800000	25.400000	...	28.400000	28.100000	47.700000	48.02500
50%	31.000000	17.000000	103.000000	103.500000	0.475900	50.300000	50.700000	18.300000	18.200000	28.200000	...	31.500000	31.400000	50.100000	50.10000
75%	33.000000	22.000000	107.775000	107.900000	0.707000	52.100000	52.500000	19.675000	19.700000	30.900000	...	34.600000	35.200000	52.000000	52.10000
max	40.000000	35.000000	123.400000	120.200000	0.974400	59.200000	59.300000	26.100000	27.900000	39.200000	...	48.100000	54.000000	61.400000	61.20000

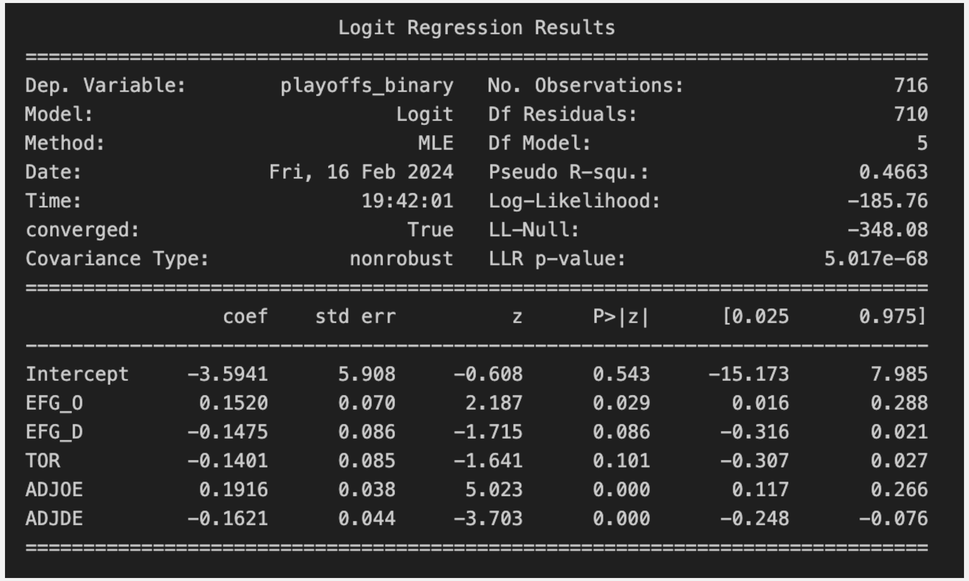
LOGISTIC REGRESSION

- Created the variable “playoffs_binary”
  - 1 if a team made the NCAA tournament, 0 if not
- Explanatory (X) Variables:
  - EFG_O
  - EFG_D
  - TOR
    - Estimated per 100 plays
  - ADJOE
  - ADJDE

RESULTS AND MARGINAL EFFECTS



```
        Logit Marginal Effects       
=====================================
Dep. Variable:        playoffs_binary
Method:                          dydx
At:                              mean
==============================================================================
                dy/dx    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
EFG_O          0.0074      0.003      2.241      0.025       0.001       0.014
EFG_D         -0.0072      0.004     -1.663      0.096      -0.016       0.001
TOR           -0.0068      0.004     -1.645      0.100      -0.015       0.001
ADJOE          0.0094      0.002      3.978      0.000       0.005       0.014
ADJDE         -0.0079      0.002     -3.544      0.000      -0.012      -0.004
==============================================================================
```



```
                           Logit Regression Results                           
==============================================================================
Dep. Variable:        playoffs_binary   No. Observations:                  716
Model:                          Logit   Df Residuals:                      710
Method:                           MLE   Df Model:                            5
Date:                Fri, 16 Feb 2024   Pseudo R-squ.:                  0.4663
Time:                        19:42:01   Log-Likelihood:                -185.76
converged:                       True   LL-Null:                       -348.08
Covariance Type:            nonrobust   LLR p-value:                 5.017e-68
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.5941      5.908     -0.608      0.543     -15.173       7.985
EFG_O          0.1520      0.070      2.187      0.029       0.016       0.288
EFG_D         -0.1475      0.086     -1.715      0.086      -0.316       0.021
TOR           -0.1401      0.085     -1.641      0.101      -0.307       0.027
ADJOE          0.1916      0.038      5.023      0.000       0.117       0.266
ADJDE         -0.1621      0.044     -3.703      0.000      -0.248      -0.076
==============================================================================
```

Calculating interpretations:

(dy/dx)/mean of y variable = interpretation

A 1 percentage point increase in EFG_O is associated with 3.9% increase in the likelihood of a team making the NCAA tournament

Each additional turnover per 100 plays is associated with a 3.6% lower likelihood of a team making the NCAA tournament

IMPLICATIONS FOR STAKEHOLDERS

- College athletics generate large sums of revenue for their respective institutions
- Monetary incentives make anyone invested in a university interested in their team making the NCAA tournament
- Illuminating which factors contribute to making the NCAA tournament might change how teams recruit, and institutions invest
  - Facilities
  - Coaches
  - NIL Deals

ETHICAL, LEGAL, SOCIETAL IMPLICATIONS

- Ethical implications
  - Is using a model to determine personnel for a program ethical? What about potential biases?
  - Coaches/recruits overlooked because of model bias
- Legal implications
  - Should universities be allowed to use models like this without disclosing it to the NCAA?
  - Should all models be open-source?
- Societal implications
  - College sports fans will want their favorite teams/institutions to be at the cutting-edge
  - Institutions will have to invest in this practice if others find success with it