How can we best model game outcomes in sports? Earlier in the class, we posed a similar question in the vein of how to best predict the results of future games. The uncertainty of sports plays a huge part in its success in society, with millions of people tuning in on a daily basis to find out the outcome of a particular matchup. A couple weeks back, we attempted to model these matchups using a Bradley-Terry model. Based primarily on score differential, the Bradley-Terry model computes one-number estimates for team strength, which can then be used in comparison with each other to predict a particular matchup. However, this approach is not without its flaws. For one, this one-number estimate assumes that team strength can be measured in only one aspect, and that score differential (or win probability) can be properly estimated using this sole estimate. In many sports, however, this is not a fair assumption. Take basketball, for example. In the 2019-20 season, the Washington Wizards were what many fans today would consider "mid". A pretty average team by record (34-38) and point differential, they were not a great team by any stretch of the imagination, but they weren't bad either. Most statistical models, such as Bradley-Terry, would rate them as average-to-slightly-below-average, and no one would bat an eye. However, in reality, the way the Wizards played that season was anything but average. Boasting a dynamic offense led by Russell Westbrook and Bradley Beal, the Wizards finished 3rd in the NBA in points per game at 116.6. Sadly, their performance on the other side of the court was Jekyll-and-Hyde-esque, as they also finished dead last in the NBA in points allowed per game at 118.5. Models that solely look at point differential would miss this key difference in the Wizard's performance from one side of the floor to the other. The solution: enter the Rasch Model. The Rasch model differentiates itself from the Bradley-Terry and other models because it realizes that team strength does not necessarily remain constant when the matchup is flipped (as in Wizards defense vs Rockets offense is a very different matchup than Wizards offense vs Rockets defense). In this report, I used the regularized ridge regression Rasch model to estimate NCAA men's basketball offensive and defensive strengths when it comes to scoring for individual teams. The data for this model was taken from the hoopR package from the SportsDataverse, and it contains for every matchup in all of college basketball this season, the home team, away team, and points scored for each team (as well as other box score statistics not included in the model). As March Madness rapidly approaches, I wanted to more accurately gauge team strengths when it comes to scoring to more accurately inform my bracket predictions. My model attempts to estimate the points scored for a given team by including: the offensive team, defensive team, and whether the offense has home-court advantage or not. Given that the NCAA Tournament is all played on neutral courts, I wanted to eliminate the chances that teams became overvalued in my model by scoring outlandishly well at home.

RESULTS

The intercept, or in this case the amount of points an exactly average offense would expect to score against an exactly average defense on a neutral floor, was **69.1.** Home-court advantage ended up being a very significant factor in my model, with an estimated extra **4** points scored and **4 less** points allowed for the home team against an identical opponent, good for about an **8 point** advantage. Below are the top 5 and bottom 5 offenses as estimated by the model, with the

number representing the estimated points scored above or below an average offense when facing an identical defense.

Top 5 Offenses	**Bottom 5 Offense**
1) Alabama Crimson Tide (2.50)	1) Mississippi Valley State Delta Devils (-2.52)
2) Arizona Wildcats (2.22)	2) Army Black Knights (-1.91)
3) Kentucky Wildcats (2.12)	3) Siena Saints (-1.68)
4) Samford Bulldogs (2.04)	4) Coppin State Eagles (-1.66)
5) Wright State Raiders (1.98)	5) Virginia Cavaliers (-1.56)

The results from the model seem to match the common intuition pretty well, as Alabama currently ranks first on Ken Pomeroy(kenpom.com)'s adjusted offensive efficiency metric, while Mississippi Valley State also ranks last and have stumbled their way to a 1-27 record so far this season. One area the model could perhaps improve upon is adjusting for strength of schedule, Samford and Wright State both rank in the top 5 in the country in points per game, however they face extremely weak defenses on a consistent basis playing in sub-mid-major conferences.

Below is the model's rankings for defensive teams:

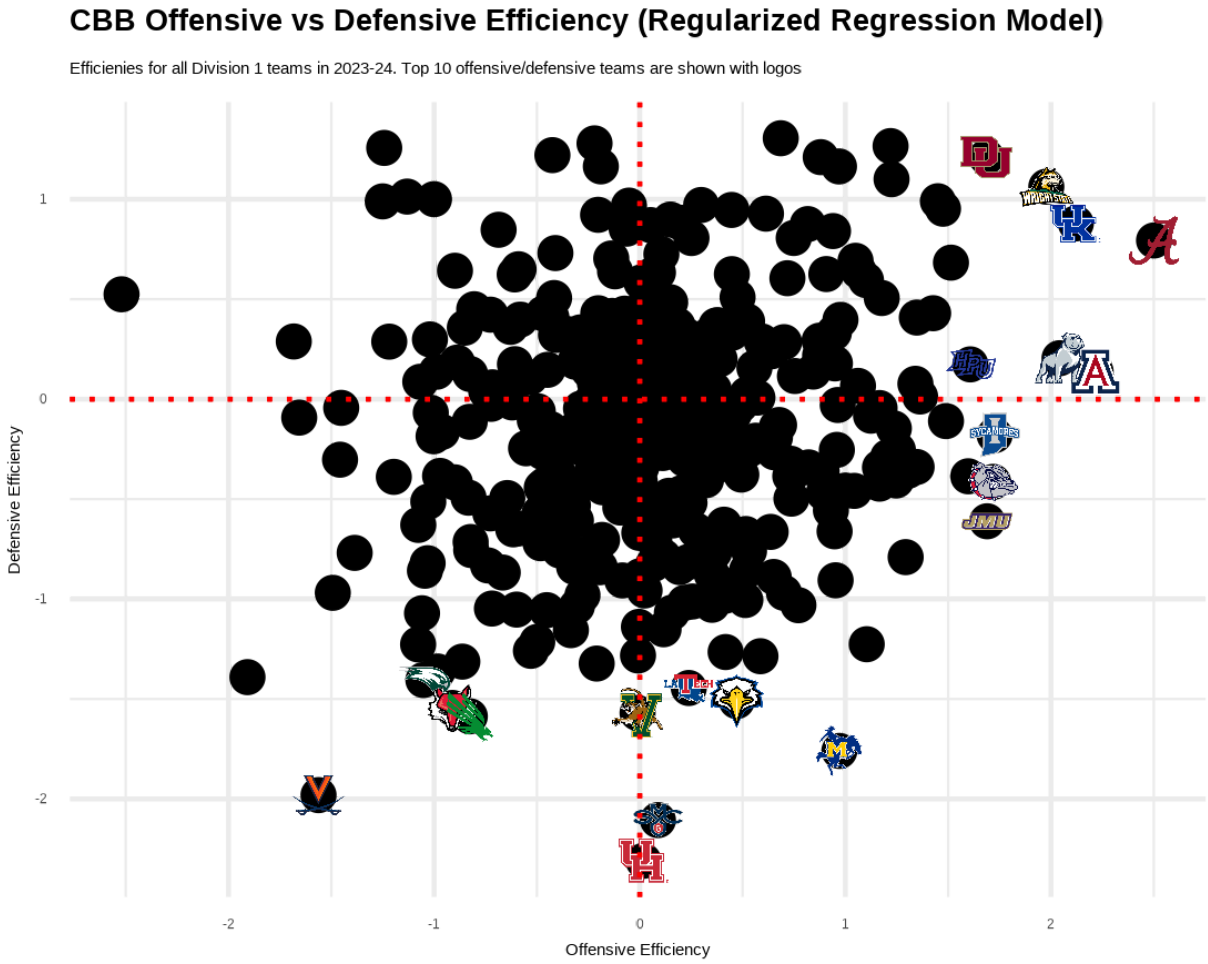
Top 5 Defenses	**Bottom 5 Defenses**
1) Houston Cougars (-2.31)	1) Virginia Lynchburg Dragons (1.75)
2) Saint Mary's Gaels (-2.11)	2) Bethesda University Flames (1.73)
3) Virginia Cavaliers (-1.98)	3) Champion Christian Tigers (1.57)
4) McNeese Cowboys (-1.76)	4) UTSA Roadrunners (1.31)
5) North Texas Mean Green (-1.59)	5) Houston Christian Huskies (1.28)

Once again, the results seem to match common intuition, as Houston ranks number 1 in KenPom's defensive efficiency metric and teams like Virginia and Saint Mary's have had strong defenses for years.

DATA VISUALIZATION

The following plot shows the relationship between Offensive and Defensive Efficiency as measured by our Regularized Regression model. Teams in the bottom right quadrant are above-average in both metrics, and top 10 teams in each metric are represented with their team logos. We can see that most teams either specialize in offense or defense, as there is no overlap between the top 10's of both metrics. Teams like Gonzaga, Saint Mary's Indiana State, and James

Madison appear to be potential sleepers in the NCAA tourney as they are above-average in both metrics while being top 10 in one.



ANALYSIS

The famous British statistician George P. Box once famously quoted, “All models are wrong, but some are useful.” This phrase holds especially true in the sports analytics realm. By definition, any statistical representation we build of game outcomes is exactly that, an *approximation* of reality. We can improve upon our assumptions and make more accurate models by attempting to mimic reality as best we can. The Rasch model implemented here is certainly an upgrade over the Bradley-Terry model, but it is by no means the be-all end-all in predicting games. The sorting of teams into offensive and defensive strengths is also not quite 100% accurate. A team like Houston, top 10 in the country in turnovers created, may get an artificial “boost” to their offensive scores in our model because they are shooting layups in transition. The model also does not account for possessions (my attempts to do so ended up crashing my RStudio memory), so teams who play slower (i.e. Virginia) may seem to have better defense than they actually do, and vice-versa for offense. Overall, this model, if applied with context, can be a great starting

point for your bracket challenge this year, but one must be careful not to get lost in the numbers and acknowledge that the name “March Madness” was coined for a reason.