Individual Stats on Single Game Outcome
Collin Travasos

In this project, I wanted to attempt to find a relationship from a single player's performance to that of the team. Furthermore, I wanted to predict the outcome of games based on a single player's statline. The thought process behind this was first, can a single player, especially those considered superstars, take over in a game in such a way that they single handedly win the game. Secondly, I wondered if not only can a single player's performance have such a drastic effect, but can it be used to accurately predict a game. If so, the applications for a model would be multifaceted, both useful for the teams and coaches as well as the fans and sports bettors.

A coach could use the model in a handful of ways. If they know that player X needs stat A to have a Y percent chance of victory compared to stat B, they would be able to install a game plan that encourages them to reach those numbers. For some that might be getting more assists, others it might mean shooting less three pointers, or getting to the free throw line more often. By playing with the model, the coach could find the best combinations of stats for each player on the roster and put them in the best spots to achieve those. While pregame planning is very useful, mid game adjustments might be even more beneficial for the teams. If they know player X needs stat Y by the ended of the game and at halftime they are only at .33Y and not .5Y then the coach can push them to be more aggressive and pursue that stat more in the second half, or conversely if they are at .66Y, a coach could tell the player to be more conservative and slow down as they get in to the second half. This could also help with lineup setting and substitutions.

From a sports betting perspective it is useful from both the perspective of that agency or book as well as the individual gambler. It is probably most useful for live betting. If player X is not on pace at halftime, the lines may not properly show what the user knows to be the team's chances of success if they do not reach their metrics. Similarly if player Y is on pace to reach their target metrics then the book can set the line properly even if the score does not reflect what we view the teams chances of success. This model would also be helpful in the creation of parlays. If a sports bettor is looking to bet on team Z to win and they know that when player X reaches stat Z the team's chances of victory increase significantly, it would be wise to parlay those metrics together seeing as the odds would likely not properly reflect the proper probabilities. Conversely the gambling agencies could use this information to set the lines accurately and entice those unaware of the models results. Depending on whose hands the model is in, it could be used to add another layer of useful information and ultimately increase earnings for the user.

Previous work done on this concept has been done with similar modeling strategies such as random forest and xg boost, however they used full team data to

predict the outcomes. This is obviously more accurate than just using a single player, however its applications are more limited. The results from a model that uses a whole team are more obvious and predictable. There are no doubt useful pieces of information involved but there is not as much room for coaching changes or gambling advantages.

While creating this model I first found team and player box scores for individual games. These had all the traditional stats for a basketball game that I needed. I cleaned up the data, added a few columns, and joined them together. Next I removed games that were blowouts, I determined that those decided by 25 points or more were outliers in the dataset and hurtful to the models. I also removed all player data points where they did not start the game and play at least one quarter, 12 minutes, of the contest. While this removes a lot of players, it puts a greater focus on the impact players on the court. While there are no doubt some very impactful players that come of the bench, think Finals MVP Andre Igoudala and other 6 Man of the Year award recipients, most of what we want in the model is in relation to the starters. That being said it would be very easy to add some or all players back in if a team wanted to do so. I played around with only using all stars, however there were not enough data points to properly conduct modeling and the applications were less than with all starters included.

I ran a handful of models. I first used random forest models on all players, first on points and then on points, rebounds, and assists. I then ran other models on a single player, opting for Lebron James as an example, using all of his game stats in logistics regression, gradient boosting, and k nearest neighbor models. I then ran single player random forest models using all game stats on a number of players with a spectrum of stars to role players. I compared their teams chances of victory when they had average games as well as games with one and two standard deviation increases and decreases – excellent, good, bad, and terrible – to determine how their performances affected the team.

My model had varying accuracy scores for each player, but overall was somewhat accurate. I was able to conclude a few themes from the variety of players I ran the model on. First and foremost was that star players can in fact take over games. When they have above average games their win chances can skyrocket. However doing too much can be detrimental, some had higher win rates when they had good games compared to excellent games. And unsurprisingly when they did not show up and had bad games the team was not as successful. It was also worth noting that the role players were vital for the team's success and that in coordination with their label, they needed to stay within their role. Various players had the highest chances of victory when they had average games. Other players I looked at were simply detrimental to the team the more they did. Rather concerningly their teams chances of success were the highest when they played the worst and did the least.

In the pursuit of a more accurate model, it would be beneficial to add a handful of facets. First and foremost the inclusion of defensive metrics would greatly improve the

model. While blocks and steals are helpful, there is much more to defensive than that. There is a rise in what have been coined “hustle stats” which include things like deflections, loose ball recoveries, charges drawn, and contested shots. Those could be useful for the model. There are various metrics that have been used for defense however it is a harder thing to quantify when compared to offensive output. What may be more important however, is the defense of the opponent and how well the player is being guarded. Incorporating the other team’s defensive metrics is crucial to improving the accuracy of this model going forward.

Additionally, I would like to alter specific stats from a player’s statline. In my model, I increased/decreased all of their stats by one standard deviation. What would create even more interesting results would be if you left stats A, B, and C at average levels but increased stats X and Y while decreasing Z. Can the user find the optimal statline for a player, and is that staline attainable? These are questions that I would love to be able to answer.

Furthermore, I would also like to look at multiple player stats to determine outcomes. If a team has a “big 3” what happens when players X, Y, and Z all have above average games? What about if A and B play well but C plays poorly. Looking at various combinations of players before looking at the whole team could prove interesting and valuable for the user.

Ultimately, basketball is a team sport. And it is important to remember that when using this model. The accuracy scores are not near 100% because there are nine other players on the floor at any given team, with another handful of players that are also included in the game that have impacts. This model should be a useful tool for both coaches and gamblers alike, but it should not be the only source for decision making. As much as those at their computers can try and quantify a basketball game, it is still a competition of individuals who are subject to all that comes with playing a game. With that being said, this is a successful way to predict the outcome of a single contest using one of the players performances.