1 Introduction

College basketball has always been one of the most popular sports in the United States, with the NCAA Tournament being one of the most anticipated sporting events every year. During the NCAA Tournament or “March Madness”, everyone tries to predict the winners of all of the games in order to get the elusive perfect bracket. However, no one ever talks about the score differential of those games. If there was a way to predict the score differential in a matchup, then maybe that helps sway your mind on a toss up game. If someone is trying to go back and forth on whether or not they want to pick an upset, but you find out it’s most likely going to be a 20 point differential, then you are probably not going to pick that upset. But on the flip side, if that same model tells you it is going to be a two point game, you are probably going to take a chance and pick that upset.

A lot of studies have been done trying to predict the winner of athletic events, but rarely does anyone pay attention to the score and try to predict the differential of the game. The purpose of this study is to build a model that will accurately predict the differential of a basketball game by using box score statistics from the two teams. A model that can accurately predict score differential could be used not only by fans, but also by coaches of basketball teams. If the model predicts accurately, then the factors put into the model obviously matter a lot to winning. A coach could use that information to focus on those particular variables or to recruit a certain way. As a fan, using it to predict March Madness games is an example, or using it to bet on a game if you are betting the spread. If the model says the point differential is going to be 2, but the spread is 11, then a fan can feel very confident in the bet that would be made. There is a place for this kind of prediction and it could be of use for a lot of people.

2 Literature Review

Perhaps the most important aspect to this research is figuring out exactly what statistics matter the most to the outcome of games. A study found "rebounding, assists, and 3-point shooting efficiency were determinant to win balanced and unbalanced games" (Giovanini 1). Rebounding, taking care of the ball, and shooting the three are all pretty standard as far as thinking of the most important statistics in basketball. But, there are several other statistics that are very important when looking at predicting the outcome of basketball games. Of all statistics in basketball, no matter what level, three point shooting is going to be the most talked about. Sports Illustrated wrote in 2014 "The NBA average from three-point range this season was 36 percent, compared with 23.8 percent in 1983-83" (Jenkins 2). As three-point shooting continues to grow, it has to be one of the variables in the model. Whether it is the number of threes made, or percentage of threes made, it is vital that three-point shooting be represented.

Another aspect that should be studied is actually predicting basketball games in general. West used "regression to model the number of NCAA Tournament games a team would win. His predictors were the teams proportion of games won, cumulative points scored minus points allowed, Sangarin strength-of-schedule metric, and number of wins against teams with top-30 Sangarin rating" (Gupta 3). There have been a lot of studies and models created to try and predict the winners of the game, but this study is trying to predict the score differential of the games as opposed to the winner of the game. However, this model will be similar and can provide direction on how a model should be formulated. Different metrics will be used as the outcome variable is different, but the formatting will be similar in a regression style.

Lastly, looking at how data is used overall in basketball is necessary to understand how the data works. Analytics and data are becoming extremely popular in basketball over the last decade. Demenius wrote about how data and analytics is changing basketball and said “Advanced game statistics analysis gives a possibility to make better evaluations of everything that happens in a basketball game” (Demenius 2). Everything in basketball is now analyzed more than it was before, from the amount of points on the scoreboard, to the distance players run, to how many times a guy high fives his teammates on the bench. Since everything is measured, it can be decided what the most valuable measurements are to actually winning basketball games. That is where this model is so important, finding out exactly how to accurately measure what matters in the final score differential.

3 Data

Data in this study was found online on Kaggle and downloaded so that it could be used. This data contained several different files that contained several years of college basketball information from regular and postseason. The files that were needed for this study were detailed game-by-game results, team names, and NCAA Tournament seedings. The detailed results contained winning team ID, losing team ID, location, and box score stats for both teams and each row or observation was a game. The team names data matched the name of the school with the team ID that was used for that specific team. NCAA Tournament seeding data had the seeds of the tournament and the team ID that was matched with that seed for the specific year.

For this study, the data was filtered to only data in the year 2019. This is because the 2021 data was not in the dataset and 2020 did not have an NCAA Tournament due to the Covid-19 pandemic. Data cleaning was done so that the name of the team was merged into the detailed results so that there was a team name attached to the result instead of just a team ID. Columns were also created to represent the differential between the winning team and losing team of games of important statistics. So, taking the number of steals by the winning team and subtracting the number of steals of the losing team gives you the differential column. For example, if the winning team had 11 steals and the losing team had 5 steals, then the differential column is 6 and that was done for each game.

4 Methods

5 Findings

6 Conclusion