Using Predictive Modeling to Determine Attendance at NBA Games

Elijah Sartin

[esartin@bellarmine.edu](mailto:esartin@bellarmine.edu)

14 January 2023

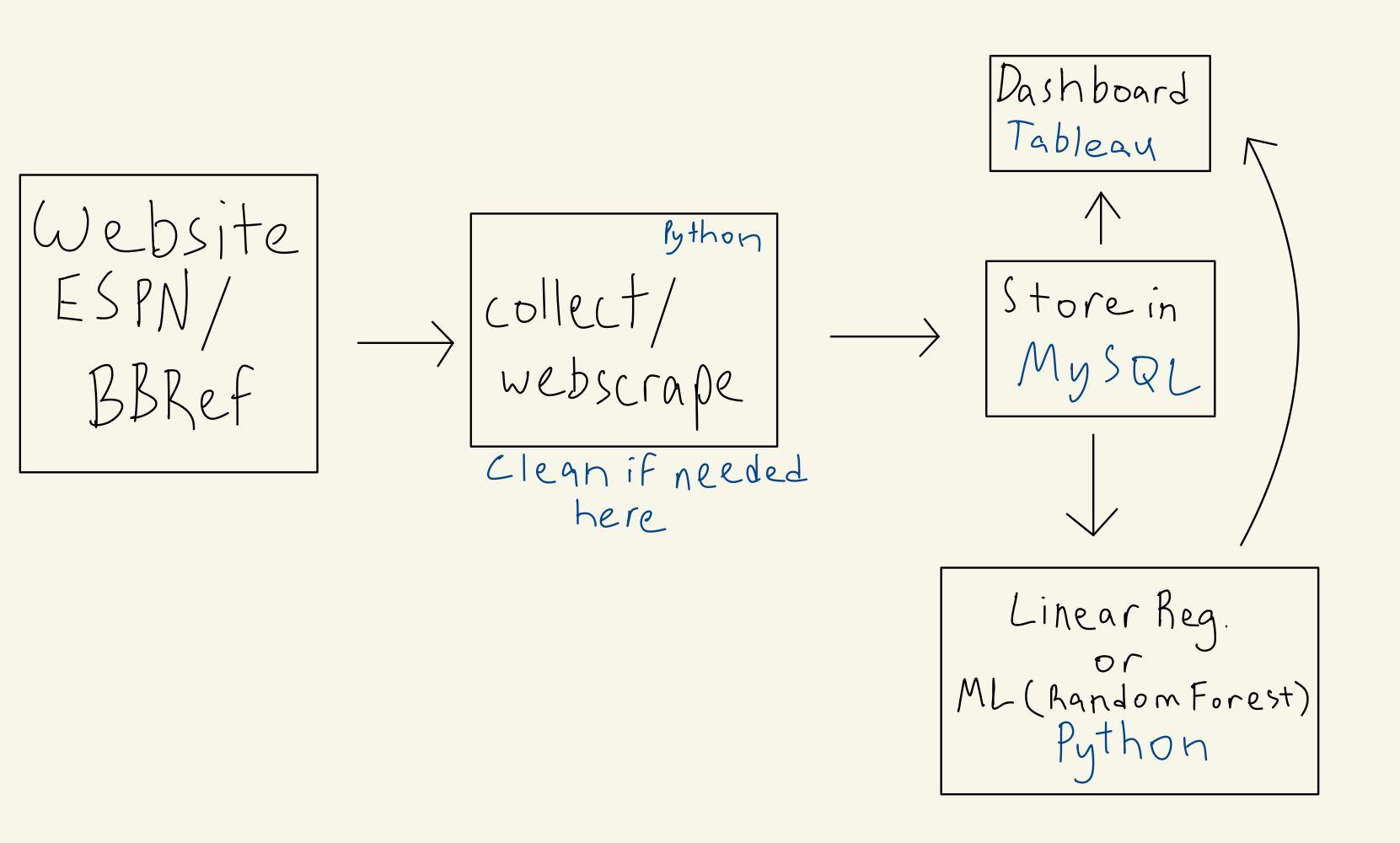
The following proposal contains the reasoning and methods for a model predicting attendance at NBA games. Accurately predicting this information would have many benefits to those involved and around the events. The data will be collected from the internet, stored, analyzed, and displayed. Python, MySQL, and Tableau will be used to obtain and manipulate the data. Python will be used to scrape, and data will be stored in MySQL, analyzed in Python, and displayed in Tableau. Ideally, the end result will be a model that is adaptable and applied to other sports aside from the NBA.

For this project, the goal is to be able to accurately predict attendance at any given NBA game. This will be done with a variety of factors that are determined before the game ever occurs. These statistics will include things such as start time, home and away teams, and the day of the week. This data will then be stored locally, put into a data dashboard, and used in a predictive model for attendance.

Attendance at a sports event has many different implications for the stadium that hosts it, as well as the surrounding businesses. Higher attendance at sporting events leads to more workers being needed at concession stands, nearby restaurants, and more room in parking garages. Teams would be able to react to low predictions with discounts on tickets or events in the stadium to attempt to draw in larger crowds. In turn, this means that knowing the attendance of a given sports event is invaluable information for the home team, as well as many others. The NBA is currently in season and this will allow a constant stream of fresh data to ensure the accuracy and continue to build the model. This is a project that has been undertaken before in 2017 by Barry King, a Butler University Associate Professor of Operations Management, who used many years of past data to form a model to predict NBA attendance. King’s model was able to predict the attendance of NBA games from 2015-2017, on average, within 5%. These results are very compelling and show promise for the success of this project. It will be interesting to see how covid may have affected attendance numbers and the impact of data from the past couple of seasons being unreliable. The goal is to eventually be able to apply this new model to a myriad of different professional sports.

The data for this project will be scraped from ESPN or basketballreference. This will be based on ease of access to acquire the data. Both of these sites are updated daily with information on games as well as the teams playing in them. Some external variables that do not vary from game to game will be taken from other sources but will not require constant scraping. Basic information about a game should have a huge impact on attendance. Start time, day of the week, and date of the game all provide valuable information about attendance. Night games will almost always have more attendance than day games. Games that happen on a Monday are less likely to have a large crowd than a game on Friday. The date is also important due to the massive effect that holidays should have on attendance at games. The winning percentage of the home team, as well as the away team, also should have an effect on the attendance of games. People are more likely to attend basketball when the teams involved are better. In his paper “Predicting National Basketball Association Game Attendance Using Random Forests” (2017) King lays out a few other important variables that should be taken into account. These variables include match type, previous season winning percentage, last home game attendance, two home games ago attendance, attendance last time hosting team, venue capacity, the city’s personal income, conference, and the number of professional sports teams in the city. While only trial and error will be able to tell how significant these variables will be in the model, King makes solid arguments in favor of the inclusion of each of them. A game being either for the playoffs or for the regular season has an obvious impact on attendance, more people should come to playoff games. The other lagged variables that King introduced should help the model to make more informed decisions based on recent data. Information regarding the conference as well as the city the team is located in should account for popularity bias as well as incorporate some factors that impact the local fan base of a team.

The execution of the project will begin with the collection of data using web scraping. Data will be taken from either ESPN or basketballreference depending on the ease of collection. Basketballreference is seeming more likely due to the large amount of historical data readily available. Neither of these sites has a built-in API so either a new scraping tool will be built in Python using BeautifulSoup for this project or an existing kit will be used. This premade kit may be one made specifically for basketballreference or may be as simple as a chrome extension. The data will be cleaned using either Python or R and then stored in MySQL. Additional cleaning may occur once here but the goal is for the data to be largely prepared by the time it is stored. A dashboard will be created using Tableau or Power Bi in order to give an easy-to-understand visual representation of the data. The data will be modeled using Python or potentially R and the results will also be displayed in the data dashboard.



The collection and display of the data acquired will not be significant until meaningful results are produced and then incorporated. Linear regression will be the first method of predictive analysis used. Surely there is a method that will be in some way superior to linear regression, but in the case that significant results can be produced then there are a few benefits to using it. Simplicity and understanding of the exact impacts of variables are some of the best features of linear regression. Even if satisfaction is achieved with linear regression, the data will also be examined using machine learning methods. This is the method that King was able to find the best results with and it would not make sense to at least attempt to apply the same. Having a large amount of past data to train makes random forests rather alluring even without the previous context. A weakness of the random forest is needing a lot of data to train, but when this is avoided, the consistency of the results should surpass that of a basic linear regression. This modeling will all be done through either Python or R.

Making any sort of predictions in sports can be difficult but determining attendance at games should be a reasonable undertaking. Using a myriad of variables including previous attendance and other team factors it should be possible to accurately predict the attendance at any given NBA game. Data acquired from basketballreference and ESPN will be stored, displayed, and modeled. These results will provide valuable information for the venues hosting these events, nearby restaurants, and more. Tools such as Python, Tableau, and MySQL will be used in the collection and analysis of the information. The goal is to create an adaptable model that can be used for other sports or similar events as well. The applications for a system like this would be expansive.

Works Cited

By: Rachel Stern | February 5, 2019. “Butler Professor Uses Past to Predict Sports Attendance.” *Stories.Butler*, 5 Feb. 2019, https://stories.butler.edu/butler-professor-uses-past-to-predict-sports-attendance/.

King, Barry E. “Predicting National Basketball Association Game Attendance Using Random Forests.” *Journal of Computer Science and Information Technology*, 2017, https://doi.org/10.15640/jcsit.v5n1a1.