**Predicting Win Shares:  
A case study for the Boston Celtics**

By: Joe Dinan

**Abstract**

After their worst season since 2014, the Boston Celtics underwent a significant makeover. Over ½ the roster was shipped out via trades or free agency. In addition, the President of Basketball Operations stepped down and was replaced by the Head Coach. After all this turmoil and these roster changes, the Head of Basketball operations wants to understand if the team is better or not? He has tasked me to build a model that predicts the win shares of the players on the roster to get a sense if the team has improved or not. I built a linear regression model predicting win-shares based on previous year statistics.

**Design**

The data analyzed was based on 26 advanced statistics scraped from basketball-reference.com. The data was scraped over the time period 2014-2021 (which coincided with Brad Stevens reign as Celtics head coach). From there I followed a three-step process: 1.) Determine which variables best predict the following year’s win share. 2.) Build a model based on those variables – train, validate and test that model 3.) Use that model to predict next year’s win share for the current Celtics roster. As part of that process, it was important to identify the ways that the model was successful in predicting win share and areas where there might be gaps. I highlighted two specific case studies to explain the values and the drawbacks of this particular model.

**Data**

The dataset contains 5,163 rows with 26 advanced statistics yearly over the years 2014-2021. Data included: Age, Games Played, Minutes Played, Player Efficiency Rating, True Shooting %, 3-Point Attempt Rate, Free Throw Attempt Rate, Off. Reb %, Def Reb %, Total Reb %, Assist %, Steal %, Block %, Turnover %, Usage %, Offensive Win Share, Defensive Win Share, Win Share, Win Share per 48 minutes, Off. Box Plus Minus, Defensive Box, Plus Minus, Box Plus Minus, Value over Replace Player.

**Analysis**

1. Scrape the data from basketball-reference.com
2. Clean the data
3. Identify which features are correlated with win share
4. Iterate over multiple versions of a linear regression model to determine which model is best.
5. Train, validate and test that model
6. Predict win shares
7. Examine predictive win share data to draw conclusions for next season

**Tools**

* Beautiful Soup for Web Scraping
* NumPy and Pandas for data manipulation
* Statsmodels and SciKitLearn for linear regression/modeling
* Matplotlib and Seaborn for plotting