SIADS 591 Report: Dribble Science

Contributors: Aseem Sachdeva, Nick Essner

Section I: Motivation

The introduction of data science techniques into the lexicon of sports analysis has provided additional opportunities for teams to generate a competitive advantage for themselves in novel ways. However, the utilization of advanced analytics techniques in professional sports raises more nebulous questions of application, namely: can different physical and cognitive traits serve as reliable indicators for performance? It is this intersection of practicality and complexity that drew us to this field.

This project aims to analyze historical NBA, NCAA, and AIQ(Athletic Intelligence Quotient) data, with a focused lens on three questions:

1. Is there a correlation between standing reach (height + wingspan) and defensive plus/minus[1],change in field goal percentage, block rate, and steal rate, amongst NBA players? And does the strength of the correlation vary positionally?
    a. Defensive Plus/Minus:

       Null Hypothesis: As standing reach increases, defensive plus/minus will not increase or decrease at statistically significant levels.

       Alternative Hypothesis: As standing reach increases, defensive plus/minus will increase or decrease at statistically significant levels.

    b. Effect on Field Goal Percentage

       Null Hypothesis: As standing reach increases, the effect on field goal percentage will not increase nor decrease at statistically significant levels.

       Alternative Hypothesis: As standing reach increases, the effect on field goal percentage will increase or decrease at statistically significant levels.

[1] Defensive plus/minus is the player's average impact in terms of net point differential per 100 offensive and defensive possessions.

c. Block Rate

Null Hypothesis: As standing reach increases, block rate will not increase nor decrease at statistically significant levels.

Alternative Hypothesis: As standing reach increases, block rate will increase or decrease at statistically significant levels.

d. Steal Rate

Null Hypothesis: As standing reach increases, steal rate will not increase nor decrease at statistically significant levels.

Alternative Hypothesis: As standing reach increases, steal rate will increase or decrease at statistically significant levels.

2. Do defensive and offensive win shares from the NCAA correlate with NBA defensive and offensive win shares? We will consider players who have played 5 years in the NBA to normalize.

a. Defensive Win Shares[2]:

Null Hypothesis: As defensive win shares increase within the NCAA player data, defensive win shares within the NBA player data will not increase nor decrease at statistically significant levels.

Alternative Hypothesis: As defensive win shares increase within the NCAA player data, defensive win shares within the NBA player data will increase or decrease at statistically significant levels.

b. Offensive Win Shares[3]

Null Hypothesis: As offensive win shares increase within the NCAA player data, offensive win shares within the NBA player data will not increase nor decrease at statistically significant levels.

[2] Defensive Win Share is a metric that estimates the number of wins a player produces for his team due to his defensive ability

[3] Offensive Win Shares is a metric that estimates the number of wins a player produces for his team due to his offensive ability

Alternative Hypothesis: As offensive win shares increase within the NCAA player data, offensive win shares within the NBA player data will increase or decrease at statistically significant levels.

3. Is there a correlation between Athletic Intelligence Quotient scores(target comparison[4], learning efficiency[5], full scale score[6], decision making score[7]) and several player performance metrics in the NBA? And how strong are the correlations?

    a. Target Comparison (potentially look more broadly at Decision Making)

    Null Hypothesis: As the Target Comparison and/or Reaction Time scores increase, steal and block rates will not increase nor decrease at statistically significant levels

    Alternative Hypothesis: As the Target Comparison and/or Reaction Time scores increase, steal and block rates will increase or decrease at statistically significant levels

    b. Learning Efficiency compared to change in Player Efficiency Rating (PER)[8] over time

    Null Hypothesis: As the Learning Efficiency score increases, the change in Player Efficiency Rating (PER) over time will not increase nor decrease at statistically significant levels.

    Alternative Hypothesis: As the Learning Efficiency score increases, the change in Player Efficiency Rating (PER) over time will increase or decrease at statistically significant levels.

    c. Full Scale Score correlation with Player Efficiency Rating (PER) for players' fifth year in the league

[4] Target comparison assesses the ability to quickly compare information in a visual field. A strength in this area may allow a player to quickly decide what to do next, based on the actions of opposing players.

[5] Learning efficiency measures the ability to store information into long-term memory and then retrieve that information later.

[6] Based on all ten subtest scores, the full scale score is considered the best overall estimate of athletic intelligence.

[7] Decision making score measures the speed and accuracy of decision making over time.

[8] Player Efficiency Rating is the overall rating of a player's per-minute statistical production.

Null Hypothesis: As the Full Scale Score increases, Player Efficiency Rating (PER) will not increase nor decrease at statistically significant levels

Alternative Hypothesis: As the Full Scale Score increases, Player Efficiency Rating (PER) will increase or decrease at statistically significant levels

d. Decision Making correlation with Turnover Ratio[9]

Null Hypothesis: As the Decision Making score increases, Turnover Ratio will not increase nor decrease at statistically significant levels

Alternative Hypothesis: As the Decision Making score increases, Turnover Ratio will increase or decrease at statistically significant levels

Section II: Data Sources

Name: NBA API

Short Description: The NBA API contains data from https://stats.nba.com/. It contains simple and advanced stats for both teams and players across numerous years.

Size: At present, the NBA dataset contains records for seasons ranging from 1946 to the present NBA season. Outputting data for a single season(2019) returns about 100,000 records. Although we will not be accessing data for every single NBA season, it is reasonable to assume that the NBA dataset contains upwards of seven million records(100,000*(2020-1946))

Location: https://pypi.org/project/nba-api/

Format: Data scraped from one of the NBA's publicly available endpoints through the nba_api python library is returned in raw json format. Given this fact, it will be necessary for us to massage this return output into a format that is more accessible(i.e. Into a pandas dataframe).

Access method: Much of the data from nba.com utilized within this project will be captured using some of the many publicly available endpoints on nba.com

Name: Sports Reference API

Short Description: The Sports Reference API contains data from https://www.sports-reference.com/. It contains simple and advanced stats for both teams and players across numerous years.

[9] Turnover ratio is the percentage of a player's possessions that end in a turnover.

Size: The Sports Reference dataset contains All-League, All-Season data from the NCAA, NHL, NFL, NBA, and MLB leagues. Although we are not accessing the full sports reference dataset, it is reasonable to assume that the amount of data available through the sportsreference api is several magnitudes larger than the data available through the NBA API.

Location: https://pypi.org/project/sportsreference/

Format: While the output returned from hitting any of the sportsreference endpoints is in raw json format, the sportsreference python library contains many functions that work under the hood capable of transforming this raw json into a pandas dataframe, which eliminates a lot of intermediary work for us.

Access method: Much of the data from sportsreference.com utilized within this project will be captured using some of the many publicly available endpoints on sportsreference.com

Name: Athletic Intelligence Quotient Data

Short Description: The Athletic Intelligence Quotient is an exam administered by Athletic Intelligence Measures LLC, a private organization affiliated with the NBA responsible for administering athletic intelligence measure tests to all incoming NBA players.

Size: Since this data is not available to the general public without express permission from Athletic Intelligence Measures LLC, current dataset size is not readily known to us.

Location: https://athleticintel.com/

Format: Format is unknown to us at the present time, though we anticipate data will be formulated in a tabular form.

Access Method: Data will be obtained by contacting personnel employed at Athletic Intelligence Measures, LLC. A preliminary meeting was held between our team, Athletic Intelligence Measures, and a Georgia Tech research team to assess our data needs and potential for collaboration.

Section III: Data Manipulation Methods

Standing Reach x Defensive Statistics (DPM, Effect on field goal percentage, steal rate, block rate)

1) We will have to pull NBA data from the nba-api and use get_data_frames() to convert it into a dataframe
2) We will merge shot defense data with player “anthro” data on player_id to combine standing reach data with other metrics from relevant seasons(2016-2020 for shot defense data, 2013-2017 for anthro data). The data is offset by 3 years following the draft, in order to ensure all players in the sample have had equivalent tenure in the NBA. Naturally, players who have been tenured in the NBA for less time will exemplify worse performance on key performance metrics as an expected consequence.
3) A correlation between standing reach (height + wingspan) and defensive plus/minus, effect on field goal percentage, block rate, and steal rate will be evaluated after combining defensive data and anthro data.

Correlation Between NCAA Win Shares and NBA Win Shares

1) We will have to format player names into specific conventions in order to access data corresponding to them, for the NCAA and NBA data. For example, NBA players are identified according to a “(first five letters of last name)(first two letters of first name)(01)” convention. If player names are not in the correct format, then the sportsreference/nba api packages will not be able to POST the necessary JSON bodies correctly to the relevant endpoints, and will thus be unable to pull the necessary data.
2) We will merge men’s NCAA basketball data from the sports-reference api and NBA data from the nba-api. We will join on the NBA players’ ‘player_id’ to filter out players from the NCAA that did not play in the NBA.
3) Combining the two data sources will allow us to identify the correlation between NCAA win shares and NBA win shares by player.

Correlation Between AIQ Scores and Various NBA Performance Metrics (steal rate, block rate, Player Efficiency Rating (PER), turnover ratio)

1) We do not know what format we will receive the AIQ data in. But we may need to adjust player naming conventions to allow merging with other data sets.
2) The AIQ dataset will be merged with NBA API and Sports Reference API datasets on player name or the player_id after configured for the AIQ dataset.
3) We will be able to assess the correlation between various AIQ scores and various NBA performance metrics for players thereafter.

Section IV: Analysis And Visualization

Our analysis will employ different visualizations of varying complexity. In order to initially discern the relationship between standing reach and the various performance metrics we are examining, and evaluate our null hypothesis, we intend to visualize these variables in the form of a scatter plot, with standing reach encoded on the x-axis, player performance metrics encoded on the y axis, and position encoded in color. We will create separate scatter plots for each performance metric. For us to accomplish this, it will be necessary for us to query the NBA API for two different sets of data(shot defense data and player anthro data) and munge the raw json into a single, concatenated pandas dataframe. For initial exploration purposes, we will utilize the matplotlib standard library, and can migrate these to a more advanced visualization framework like altair, taking care to utilize interactive capabilities where it will aid the 'effectiveness' of our visualizations. In order to examine differences across positions in our comparison between standing reach and the four comparison metrics we have chosen, we can also employ histograms, grouping positionally, segmenting the correlation evaluations appropriately.

We will merge data from the NCAA and NBA for NBA players to look for correlations between offensive win shares per 40 minutes and defensive win shares per 40 minutes to assess the value of each as predictors for NBA success. Once we've gotten our data in a workable form, we can visualize the difference between win shares in the NCAA and NBA using an interactive altair scatterplot(so the user can hover over individual data points), taking care to limit our analysis to data from each players' fifth year in the NBA and each player's final year in the NCAA. (***If sample size is exceptionally small after implementing these filtering parameters, we can revisit our time in league constraint).**

We will evaluate the correlation between various AIQ scores and various NBA statistics to assess whether certain cognitive traits play a significant role in performance on the court. We will visualize this data using a correlation matrix heatmap to highlight strong and weak correlations across scores and statistics. For example, we could see a strong positive correlation between target comparison and steals, and we could represent that in a dark red. Conversely, we could see a weak negative correlation between full scale score and player efficiency rating, which could be represented as a light blue. With tone indicating strength of correlation and color indicating positivity.

Section V: Statement of Work

We intend to implement a divide and conquer working strategy, splitting the analysis/visualization/written requirements equally. Additionally, we intend to have semi-frequent working sessions via Slack audio call, in order to make work more collaborative and to eliminate chances of producing duplicate work.

Source URLs

https://www.sportingcharts.com/dictionary/nba/defensive-win-shares-dws.aspx#:~:text=A%20metric%20that%20estimates%20the,calculated%20using%20full%2Dseason%20statistics.

http://insider.espn.com/nba/hollinger/statistics/_/sort/turnoverRatio/order/true

https://athleticintel.com/wp-content/uploads/2020/05/MLB-CF-De-Indentified-AIQ-Report.pdf