Using Python to Visualize the Impact of Covid 19 on the Game Performance of NBA Players

Md Rahman, undergraduate of the City College of New York

Hasibul Hassan, undergraduate of the City College of New York

Handell Vigniero, undergraduate of the City College of New York

City College of New York
May 21 2021

Using Python to Visualize the Impact of Covid 19 on the Game Performance of NBA Players

Md Rahman, undergraduate of the City College of New York
Hasibul Hassan, undergraduate of the City College of New York
Handell Vigniero, undergraduate of the City College of New York

Abstract

In this design analysis, we use various visualization tools provided by Python's vast library to interpret the significance of the novel Coronavirus(Covid 19) on the game performance of the athletes of the National Basketball Association(NBA). A diverse amount of visualization techniques are utilized to pinpoint where specific areas of performance were impacted, to determine which of these areas are the best indicators of performance as well as to evaluate changes in player behavior while on the court. Machine language classification techniques were utilized to compare the change in performance of teammates in different NBA teams during the affected season in comparison to their previous seasons. Regression techniques were used to create prediction models that predicted player statistics, player performance and team performance. The data that was used to perform these complex observations were provided by nba-api, an API Client package used to access APIs for NBA.com. It provides a plethora of information that includes player statistics(both regular season and postseason), league averages, team game logs and much more. This research shows that data visualization is a powerful tool that can explain phenomena that even the casual viewer cannot explain by the naked eye.

1 Introduction

On January 11, 2020 Chinese officials reported the death of an elderly Wuhan citizen that resulted from an unknown virus[1]. Months later this virus, now known as the Coronavirus(Covid 19), spread throughout the world infecting millions causing a global pandemic. During the early stages of the outbreak, the daily practices of various industries were put on a temporary halt to prevent possible infection. Eventually, some of these practices resumed during the quarantine procedure, although remotely. However, not all industries received the option of operating remotely, especially the sports industry. Many sports leagues such as Major Baseball League(MLB), Premier League(ELB), National Hockey League(NHL) and National Basketball Association (NBA) had issues adapting to the current climate of Covid 19. The most essential components of these industries, the athletes, were unable to participate in company activities since such activities required constant physical contact, which could potentially endanger both the players and other employees such as training staff, coaches, etc. League executives were pressured by owners to resume activity to reduce profit loss. Of these league executives, NBA commissioner Adam Silver and his associates were able to establish arguably the best approach to resumption that would guarantee employee safety.

On July 7, 2020, the NBA held its first game since league lockdown on March 11, 2020. Athletes played in an isolated "bubble" environment in a Florida facility that consistently tested players for Covid 19 and kept them under strict surveillance. All players were required to stay in the facility until the end of the season or until their teams were eliminated from playoff contention. As a result of the four month hiatus, some players experienced a significant change in performance. Some of them were able to exceed expectations, while others ultimately underachieved and were unable to overcome shortcomings. It is difficult to observe these drastic changes through casual observation since we are not able to pinpoint certain areas of performance change. Thankfully, the NBA has an incredibly vast amount of data that contains information

regarding player, team and league statistics. By properly, processing and analyzing this data we attempt to understand which areas of game performance and behavior were affected by the environment created by Covid 19.

2 Data Specifications

Data was collected from a multitude of sources. One of the most reliable sources was Nba_Api, which is a Python package that maps and analyzes various endpoints for the official NBA website: "www.NBA.com". This easy to access Python package was used to extract league and player statistics, which are in JSON format.The collected Data was filtered to include only active players since these are the only players that are directly effected by the Covid-19 pandemic. Pandas was utilized to manipulate the extracted data and convert the data into a data frame to be manipulated.For example, some essential data needed to created the Hex Shot Chart are the following: Loc_x(horizontal position of shot), Loc_y(vertical position of shot), shot_made (Boolean value indicating if shot was made), player name, player team(for the corresponding season) and league averages of shots made in specific areas of the court. When using Machine Language tools to classify player performance,the player statistics endpoint was used and filtered to include players of the same team. The PlayerCareerStats, PlayerEstimatedMetrics, TeamByYearStats and LeagueStandings endpoints of the Nba_Api were essential for creating the prediction models.

3 Background and Related Work

The examination of work related to our area of study provides us with an initial sense of direction in our approach to implementation and it also helps us anticipate our potential results. Vaquera [2] describes a technique used to differentiate between the significance of different performance indicators, in which some have more merit than others when used to indicate performance. McHill[3] is one of the few researchers to explicitly analyze the impact of Covid-19 on the athletic performance of NBA players with his focus being primarily on how the lack of traveling that occurred during the the NBA's temporary bubble environment during the 2019-2020 regular season impacted team statistics.

3.1 Determining Key Indicators of Performance

When searching for significant indicators of performance amongst NBA players, Vaquera implements the use of clusters consisting of data that contains information regarding advanced metrics such as ON, OFF, NET, Team wins, Pace, DRtg(defensive rating), ORtg(offensive rating), ORB%(offensive rebounding percentage), TOV%(turnover percentage) and eFg%(effective field goal percentage), which he refers to as Key Performance Indicators (KPI). Using such metrics, different groups of players can be categorized in these clusters to determine how they contribute to a team's win. For example, Vaquera mentions a cluster that contains players with strong MAX POS(maximum positive point difference). When these players played with players of a different cluster, those players had poor performance. In our implementation, we will implement a selection of advanced metrics to accomplish a comprehensive analysis of player production and how it is affected under varying circumstances such as the current pandemic.

3.2 Utilizing the National Basketball Association's COVID-19 restart "bubble" to uncover the impact of travel and circadian disruption on athletic performance

In his analysis, McHill claims that traveling done by athletes often disrupts their internal circadian clock, which has a direct correlation with their performance. This disruption is common amongst NBA players who must travel to the arena of the opposing team, but is avoided by those who have home court advantage, the benefit experienced by the team that has the opportunity to play in their home arena. During the 2019-2020 NBA season, the benefit of home court advantage was eliminated and using data extracted from http://www.basketball-reference.com/McHill observes how its absence

effected team statistics such as win percentage, effective field goal percentage, turnover percentage,offensive rebounding percentage,etc. Using linear mixed effect models, McHill was able to analyze the changes in win percentage. He also makes use of two tailed dependent t - tests to observe the correlation between game playing statistics and the location in which they are played.

Based on the results of McHill's analysis, prior to Covid-19, teams that traveled towards western time zones suffered a decrease in win percentage when compared to the win percentage experience in home games. During Covid-19, all games were played in the same arena where visual and auditory effects were used during the games to simulate home and away environments.According to McHill, the difference between home and away win percentages during this time were insignificant, with teams having a home win percentage of about 55.7% and away win percentage of 44.3%. Prior to Covid-19, the average difference in effective field goal percentage of teams playing at different time zones compared to playing at home was about 1.6%, while the average difference in turnover percentage was 1.5%. The effective field goal percentage and turnover percentage of teams in the bubble environment did not display any significant differences. The only team statics that portrayed significant changes during the NBA bubble were free throw rate, which increased by 2.6% and defensive rating, in which teams generally allowed 5.5 more points per 100 possessions. McHill's results give us an idea of the amount of change we should expect for certain team statistic, while also giving insight to the underlying causes of these changes.

4 Methods

It is crucial to implement various data visualizations since it prevents us from interpreting and presenting our data in a one dimensional manner. By including this variety we can understand why we obtain particular results as well as the accuracy of the results of which we have obtained. These visualizations will serve as the blueprint for obtaining a valid conclusion based on the evidence that we have gathered. The visualizations that were implemented to conduct our analysis are the following: Scatter Plot, Bar Chart, Correlation Matrix, Dash Table Radar Plot, and Hex Shot Chart. Furthermore, when making comparison's amongst player performance, we must utilize difference classification techniques that rely on an abundance of machine language tools from Python. This will help ensure the accuracy of the classifications being made. The main classifications techniques that were used were Principal Component Analysis, Linear Discriminant Analysis and T-Distributed Stochastic Neighbor Embedding, all found under Python's Sklearn library. Also, our prediction of player and team performance is reliant on the underlying relationship between our set of feature variables and target variable(to be predicted), therefore it is necessary to utilize various Regression models. The Regression Models that were used to make our predictions were Logistic Regression,Support Vector Regress(SVR) and Support Vector Machines(SVM).

4.1 Visualization Techniques

Although data may be easily understood to those who have been thoroughly exposed to it, the average user will have difficultly understanding its technicalities. As a result, the programmer must make an effort to appropriately visualize the data in a manner that is easily digestible or broken down to the user. As a result, we have implemented various visualization methods that allow us to make accurate and detailed assessments of the impact Covid 19 had to player and team performance.

4.1.1 Radar Plot

After determining the variables that we want to go forward with emphasizing, it is necessary to convey how these numbers represent a player's performance. Using a radar plot, we can graph these attributes together to show how one player may excel in different aspects than another. Also, it will be interesting to see how a player stacks up against themselves when we compare their performance through multiple seasons, as well as their performance specifically inside of the bubble. Each player will have their potency in the statistical categories PER(Player Efficiency Rating), OWS(Offensive Win Shares), DWS(Defensive Win Shares), WS(Win Shares), BPM(Box Plus-Minus), and VORP(Value Above Replacement

Player). This provides us with insight on a player's strengths and weaknesses in these measurements for a particular season and their ability to outperform or under perform under strenuous conditions such as a pandemic. The area in the center of this plot will increase in the directions of the categories the player excelled in while staying closer to the origin if they were not as potent in that particular category.

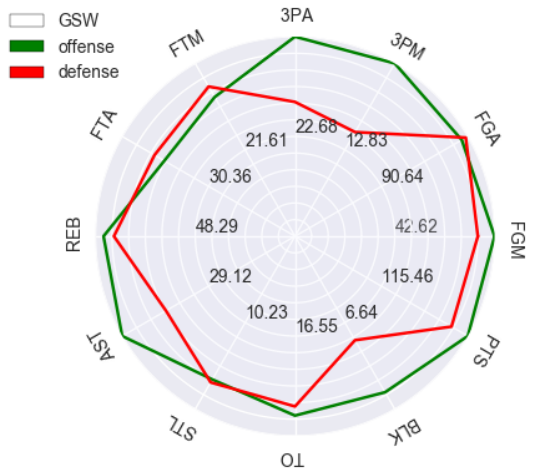

Figure 1: Radar Shot Chart

4.1.2 Hex Shot Chart

The common scatter plot is an inadequate form of visualization for large data since data points will tend to overlap and will obscure interpretation of the data. Hex plots are a remedy for this problem since this visualization method essentially allows us to map multiple samples to hex-bins and each of these bins have unique colors that indicate the density of points. Like scatter plot data points, hex-bins also have horizontal and vertical coordinates that allow them to be displayed visually. In our application, we map several shots to hex-bins. Raw shot chart data, which embeds scatter plots on top of a basketball court layout is difficult to interpret since points are densely packed as shown in figure 2. Players take thousands of shots and are bound to take shots in the same area, this causes many overlaps.

Rather than using the previous approach, we embed a hex plot into our basketball court layout(400 by 400), which was drawn using the Matplotlib library in Python. This hex-plot will group shots in certain areas to hex-bins of different sizes, since not all areas of the court have the same number of shots taken. We calculate the total amount of field goals attempted in each area and divide this value by the amount of field goals made (returns Boolean value of 1). We compare this value to the league average field goal percentage in said area. The colors used on the hex map shot chart are dark blue, light blue,yellow, orange and red, which respectively display an increasing success rate. A high success rate indicates that a player did well above league average, while the opposite indicates a player did not come close to meeting league standard. Furthermore, these colors are used to correspond with basketball terminology in which a player is "red hot" if he is shooting well or "ice cold" if he is under-performing.

Using the hex shot chart as shown in figure 3, we can describe a player's preferred shot selection relative to the areas on the basketball court and success rate in respective areas. The hex shot chart will allow us to make comparisons between the 2019-20 season that took place during Covid 19 with other seasons, in which the two seasons being compared can contain either Regular Season data or Playoff data. These comparisons will show us if certain players were able to obtain better performance due to an increased role in the offense, which is caused by key rotational players being unable to attend games due to contracting the virus or other personal reasons. It can also indicate which players were able to improve or gain new skills during the hiatus. This can be identified if players take more shots in an area during the Covid 19 season in which they did not take in previous seasons.

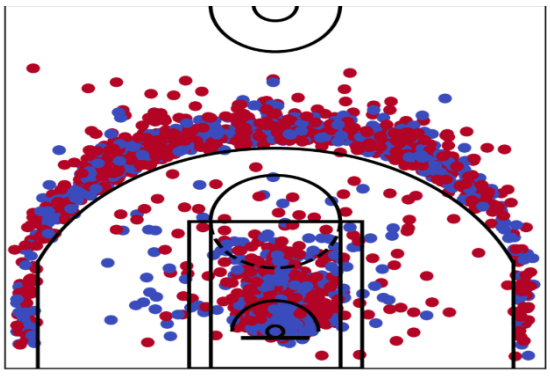

Figure 2: Raw shot chart data

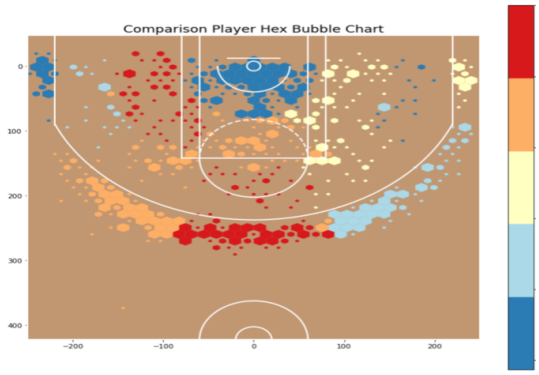

Figure 3: Hex Shot Chart

4.1.3 Bar Chart, Dash DataTable, & Scatter Plot

Python provides programmers with various data visualization libraries with MatPlotLib and Seaborn being the most popular. Despite their utility and popularity, these libraries often output static images that limit the amount of information available to the user. A remedy to this problem is to implement interactive visualizations that can be done using Python's Plotly and Dash libraries. Some interactions provided by Plotly include allowing users to hover over data points to view their associated values and mapped attributes, zooming into graphs to see the proximity between data points,etc. Users can also trigger sub graphs by hovering or clicking on data points, thus gaining additional insight on the data. Dash utilizes HTML, CSS and Javascript to produce user friendly web applications that allow users to specify parameters of interest that easily allows them to view data with certain specifications.

The main Plotly graphs used to visualize the impact of Covid 19 on NBA performance were scatter plots, bar charts and Dash DataTables. Scatter Plots allowed data with distinguishing attributes to be plotted on the same graph. The size of these data points were dependent on attributes like player age and were color coded based on player name. Each of these scatter plot data points contained hover information containing information regarding player name, player age, the current season and performance. An example of a Plotly scatter plot can be seen in figure 4. Bar charts were utilized in a variety of analyses, in which the heights of horizontally aligned rectangular bars were used to compare categorical data. Like the scatter plots, the Plotly bar charts were also color coded, some being based on shot area on the court and others being based on player name. An example of a Plotly bar chart can be seen in figure 5. Dash DataTables, implemented from React.js, are tables that formatted in a similar style to a spreadsheet and contain meaningful information that is presented in an organized manner. They contained interactive columns that can perform a multitude of tasks such as triggering another sub graph or table. An example of a Dash DataTable can be seen in figure 6. Dash objects such as radiobuttons and dropdowns played a critical role in allowing users to dictate the data that they wished to analyze.

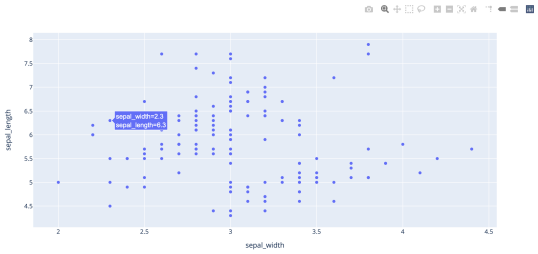

Figure 4: Example of Plotly Scatter Plot

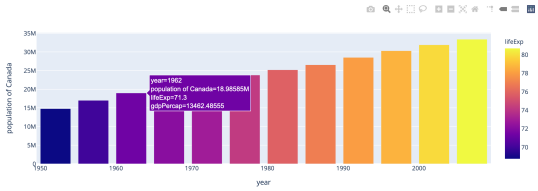

Figure 5: Plotly Bar Chart

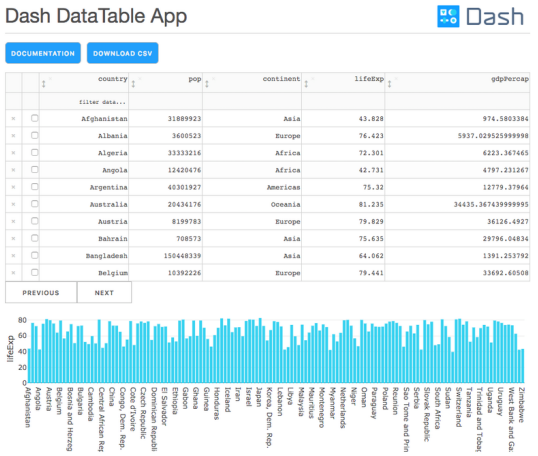

Figure 6: Example of Dash DataTable

4.2 Classification Techniques

NBA Players generate a plethora of statistical categories for each game played, which include points per game, steals per game, field goal percentage, rebounds per game, etc. The pitfall of comparing these features to determine a player's level of performance is that they do not always share the same scale. Nevertheless, there are techniques available that allow us visualize data with numerous attributes which would

be otherwise messy and confusing. These attributes may be reduced to a smaller number of components for more significant comparison and results. In particular, we utilized the following classification techniques: Principal Component Analysis, Linear Discriminant Analysis, and T-Distributed Stochastic Neighbor Embedding, which were all available in the sklearn Python library.

4.2.1 Principal Component Analysis

Principal Component Analysis, or PCA, is a classification technique that prioritizes reducing dimensionality while still maintaining the variation within the original data set. The data set is transformed and distributed amongst new sets of variables known as Principal Components. In order to maximize variation, each component, a one-dimensional set of attributes, can be observed to eliminate variables that do not display sufficient variation. To better establish how much variance each feature presents, the values must share the same scale. Therefore, each column of the principal components are standardized to numerical data of the same order of magnitude. Python provides a built in function to proceed with the next coupe steps to complete a principal component analysis which include computing the co-variance matrix, eigenvalues and eigenvectors. The final principal component values are derived from the dot product of eigenvector and the standardized columns.

The specific features under consideration for this analysis were minutes played, field goal percentage, turnovers, three-point field goal percentage, points, rebounds, assists, steals, and blocks. As stated before, plotting so many different variables on one graph would create an incoherent cluster. Also, the scale at which these statistics are measured vary greatly. While minutes played is just an accumulation of all the time a player has spent playing, the rest are averages taken on a per-game basis. Taking into account that only two of these statistics are percentages difference in magnitude is very evident. Using Python's built in functions to standardize scale and fit these values, we were able to reduce the dimensions of our data to three principal components. This simplifies class separability greatly as our independent variable, the Players' unique player IDs, can be plotted concurrently to evaluate their performance. Each class can be differentiated by position, color and size. The position on the plot will relate to the specific principal components, each unique player id and name is represented by a different color, and the size of the point represents the player's age. Points that stray away from the cluster can be identified as outliers and an in-depth look of their measurements can be observed by hovering over the point.

4.2.2 Linear Discriminant Analysis

Linear Discriminant Analysis(LDA) is a supervised(unlike PCA) classification technique in which high dimensional data set is projected into a lower dimensional space with optimal class separability, thus decreasing the computational task of classification and steering clear of overfitting. Prior to this dimensionality reduction, a standard scalar is required to pre-process the data set to ensure that it is normally distributed. As its name suggest, LDA makes the assumption that classes are linearly separable and are distinguishable through the inclusion of multiple hyperplanes, whose main function is to maximize space between data of different classes and minimize their variation. The distance between classes can be found by calculating the difference between their means. Note that finding this difference will be more complicated depending on the amount of classes that are being separated, in which distances between the means will rely on their distances from a central point in each category. Furthermore, we must ensure that after maximizing the distance between the central point and different classes, we must retain proximity amongst similar classes and reduce scatter.

LDA played a critical role in classifying players based on performance, in which base statistics such as minutes played, field goal percentage, three point field goal percentage,turnovers,points,rebounds,assists, steals,and blocks served as independent variables.Certain base statistics like total points and assists had values that significantly larger compared to other stats like field goal percentage(always floating point values less than 1), therefore a standardscaler was used to normalize our feature variables. Class separability was solely dependent on unique player ids. After successfully fitting our dataset using sklearn's LDA function

and reducing our nine feature variables to two LDA components, we can classify players based on performance, in which high class separability signifies a large gap in performance and impact. Each class is color coded based on player name and points that deviate from their class can be categorized as an outlier. It is important to note that players who have more seasonal data are more likely to have better class separation as opposed to those who are relatively new to the league.

4.2.3 T-Distributed Stochastic Neighbor Embedding

T-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised classification technique applied to a high dimensional data. It takes a high dimensional dataset and reduces it to a low dimensional graph that still contains the same relevant information as the high dimensional graph. When transforming the dataset into a low dimensional graph, t-SNE attempts to preserve the clustering from the high dimension. Although t-SNE and PCA may seem similar, t-SNE differs from PCA because t-SNE is concerned with retaining small local clusters while PCA is concerned with large clusters in the dataset. This process can be broken into simple steps. Initially, points are measured to find neighboring points with similarities. A probability distribution is created to measure that similarity. Then, a dataset of points from the same target dimension are used to create a probability distribution for all the points. Lastly, gradient descent (algorithm to find the minimum of a function) is used to minimize the overall error so that the joint probability distribution of the new low dimension is similar to the high dimension. Overall, t-SNE can help determine if a dataset is separable or find a specific structure to the data. Visualizing a t-SNE aids this process as well. t-SNE was used in this project to determine some identifiable structure about NBA data. We observed player performance in different NBA teams using t-SNE.

4.3 Prediction Techniques

Regression models are extremely powerful tools that allow one to make accurate predictions using appropriate feature variables. They make use of training data that allows them to pinpoint correlations amongst feature variables as well as their relationship with the target variable, the variable to be predicted. Testing data, which comes from the same data set of the training data, is used to evaluate the accuracy of a particular regression model, often returned as a percentile score. Note that it is important to pre-process the data to ensure the feature variables are normalized as well as filtering out any redundant or misleading feature variables that could cause potential bias in the prediction. Various regression models were observed and tested based on the accuracy and effectiveness. The regression models that produced the most favorable outcomes were ultimately retained for further use, which includes the following models: Logistic Regression, Support Vector Machine, and Support Vector Regression, which were all available in the sklearn Python library.

4.3.1 Logistic Regression

Linear Regression is a supervised machine learning algorithm used to predict a set of values (dependent variable) based on the independent variables in a dataset. Similar to linear regression, logistic regression assumes relationships between data are linear. However, it differs because it models the data in an "s" shape using the sigmoid function. Logistic regression is known to be used to obtain binary outcomes. The raw data was first obtained from nba-api (API client for www.nba.com). Data was combined using different endpoints in the api to obtain NBA team data between the years 1983 to 2020. Data was then prepped into training and testing data where it was first filtered with either the eastern or western conference. Then, a sample year is selected . The training data is then filtered to be all team data for all previous years from the sample year selected. The testing data would be team data for the sample year that was selected. For example, in order to predict NBA teams playoff selection from the 2019 eastern conference, we filter the training data to be only eastern conference teams and only teams for years before 2019. The testing data is filtered to be only 2019 eastern conference teams. The features variables used for prediction range from seasonal field goals made to seasonal points scored. The target variable is a boolean which indicates if the team made it to the playoffs. The

outcome of the prediction is a probability value between 0 and 1 and a boolean value for each team indicating if they make it to the playoffs. This data is then visualized on a dash table.

4.3.2 Support Vector Machine

Support vector machine(SVM) is a supervised machine learning algorithm more commonly used for classification but is also used for regression. Support vector machine is similar to linear regression in that it involves generating a line based on the data points. More specifically, SVM works by finding a hyperplane that divides the dataset into two classes that you can visualize. The term support vectors from the name are the data points near the hyperplane that if removed changes the position of the hyperplane that divides it. The goal is to choose a hyperplane with the greatest possible space or margin between the hyperplane and the data points. The purpose of this is to be able to classify new data points as accurately as possible. Support vector classifier was used to make playoff predictions for NBA teams. Support vector classifier is just a different implementation of support vector machine with the same principle. The raw data needed for this prediction was first obtained from nba-api (API client for www.nba.com). Data was combined using different endpoints in the api to obtain NBA team data between the years 1983 to 2020. Then target and predicted variables were chosen. Lastly, data was then prepped for prediction by dividing data into training data and testing data for the desired prediction. The outcome of the prediction is a probability value between 0 and 1 and a boolean value for each team indicating if they make it to the playoffs. This data is then visualized on a dash table.

4.3.3 Support Vector Regression

Support Vector Regression(SVR) is an extension of SVM that prioritizes regression rather than classification. Its a supervised learning algorithm that uses principles from SVM to predict continuous variables. This prediction is heavily reliant on the training sample provided by a portion of a particular data set. Decision boundaries are created on top of the data of interest and hyper-plane containing the greatest number of points is used to specify the line of best fit. Compared to other regression models, SVR is highly sensitive to outliers, however, tolerances can be adjusted to allow the decision boundaries to have a few outliers in order to fine tune the model. Additionally, SVR to inner data points since it utilizes a kernel. These kernels project low dimensional input vectors into a higher dimensional feature space that makes previously inseparable data separable, making SVR one of the most flexible forms of Regression.

Using the SVR functions found in Python's sklearn library we were able to formulate predictions on the total number of points and assists a player would have each season and compare these results to the actual amount of points and assists that were accumulated during the corresponding season. Pre Processing was mandatory to scale the appropriate feature variables using a standard scaler and target variables and reverse engineering was used to bring the predicted variables back to their original scales. Cross Validation was used to train the regression model by providing the model with various sets of training data and was done iteratively to ensure the optimal training set was supplied, making the prediction as accurate as possible. A Dash web application was implemented allowing users to chose a specific prediction kernel as well as the amount of folds to be used in the cross validation. The kernels that were provided are the following: linear, poly,radial basis function(rbf),sigmoid and pre-computed. The number of folds ranged three to ten folds, however, it is important to note that the number of splits must always be greater than the number of samples when using cross validation. Thus,when choosing the amount of folds one must take into account the number of seasons played by a particular player, since choosing a high number of folds may not be applicable to those who have not been active in the league for a long time.

5 Observations

Various observations on Covid 19's impact on NBA performance using the multitude of visualizations, prediction models and classification models that we have implemented. Such observations will pertain to both the 2019-20 and 2020-21 NBA seasons, since these seasons were significantly impacted by the pandemic. Using our regression models that predict player statis-

tics using a particular set of feature variables, we will be able to identify the best performance indicators. A Hex Shot Chart Dash web application will be utilized to describe the change in on court shooting behavior as well as efficiency. Using a Dash web application that employs various classification methods and another Dash application that features radar plot, we will be able to compare amongst players the impact Covid 19 had on their base statistics. Finally, using a Dash web application that predicts playoff eligibility, we will analyze how player's contributed to their teams overall success.

5.1 Best Performance Indicators

NBA players generate numerous statistics of various complexities as they complete games, which can be further analyzed through visualization techniques to better understand a player's impact on the game. Observing both base game statistics as well as advanced metrics are essential to determine if a player regressed, maintained, or excelled their usual performance considering circumstances created by Covid 19. Using SVR, we were able to create accurate prediction on the total number of points and assists a player accumulated each season as well as a players efficiency rating for the same year. Minutes, field goal attempts, field goal percentage, free throws attempted, free throw Percentage and effective usage rating were the features variables used to predict accumulated points per season, while turnovers, minutes, effective assist ratio and effective offensive rating were used to predict accumulated assists per season. Points per game, three point field goal percentage, two point field goal percentage, total rebounds per game, blocks per game, turn overs per game and personal fouls per game were used to predict player efficiency rating. Other data like age and games played decreased the accuracy of the model due to their lack of correlation with the target variable.

Our predictions were plotted on a scatter plot indicating if a correlation exists between our feature variables and target variables. As seen in figure 7, the feature variables were useful in predicting points of superstar player Lebron James, since the predicted and actual values plotted on the graph have close proximity when using a linear kernel and six folds for cross validation. This same proximity is displayed in figure 8 when predicting the accumulated number of assists per season for Chris Paul. This is done using an rbf kernel and six folds for cross validation. Unlike the previous plots, the prediction of PER, shown in figure 9, plots the actual values with respect to the predicted values. Note how the data points follow a linear line of best fit, with minimal outliers indicating that our prediction was accurate. Furthermore, each of the three regression models mentioned previously produced high accuracy scores(all of which were over 90%), which were evaluated using sklearn's score function, which utilizes test data to measure the accuracy of any of its regression model. Accuracy can also be evaluated by finding the mean square error,the average of a set of errors that monitors the distances of various data points with respect to the regression line. Note that although we are aware of which statistics are the best performance indicators, we can extend our analysis in the future to make internal comparisons amongst the different indicators to order each indicator based on its contribution towards the prediction.

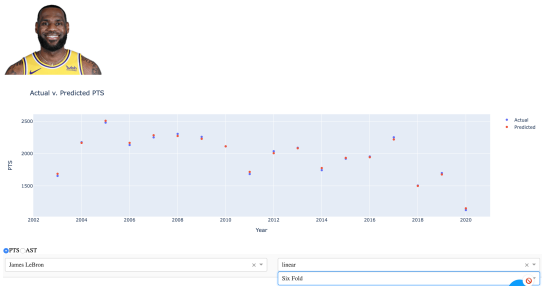

Figure 7: Support Vector Regression Model: Prediction of PTS vs Year (with actual data on predicted on same graph)

Figure 8: Support Vector Regression Model: Prediction of ASTS vs Year (with actual data on predicted on same graph)

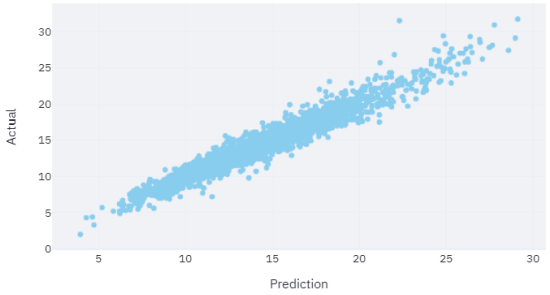

Figure 9: Support Vector Regression Model: Prediction of PER vs Actual

5.2 Change in on court shooting behavior and shooting efficiency due to Covid 19

Due to the circumstances created by Covid 19, disruptions occurred among daily activities of players including team practice, athletic conditioning, etc, since the number of interactions amongst players, coaches and training staff were minimized. This disruption has created some changes in players' on court behavior,specifically their shooting.Using a hex shot chart visualization we can directly analyze, which players were directly affected by the disruption caused by Covid 19 as well as the changes made to players that have contracted the virus. Bar charts will also be used to numerically analyze, the accuracy of players shots in different positions as well as the frequency. Comparisons will be made between prior seasons and the 2019-2020 and 2020-21 season(only effected seasons). When analyzing changes in shooting behavior during the post season, only the 2019-20 season will be taken into account, since the 2020-21 post season is currently progressing.

Analysis done on the 2019-20 season indicates that certain players who were given more offensive responsibilities were able to take the challenge head on and display remarkable performance. A majority of such players were younger players with little playoff experience. An example of a player with a drastic spike in performance was Nugget's point guard Jamal Murray. Based on comparisons between Murray's Covid 19 playoff shot chart and his shot chart from the 2018-19 season shown in figure 10 and figure 11, it is apparent that he was able to improve in a multitude of areas such as center of the three point arc, left corner of three point arc, right wing of three point arc as well as 8 feet within the basket.These changes can also be seen when comparing the accuracy charts shown in figures 12 and 13. These charts portray a 2.1% increase in accuracy in the right side center(24 feet+), a 4.2% increase in left side center(24 feet+), an 11.4% increase in the center(8 feet and below) as well as many other areas. One can also observe that Jamal Murray took much more shots than he did in the previous season which benefited him. He took more shots in the center of the three point arc as well as midrange areas on the right side of the court. When comparing figures 14 and 15, one can see that there was an increase in frequency of shots in the left center(16-24 feet) position, in which there was a 14.2% increase as well as a 29.8% increase in shots attempted in the right center(6-24 feet).It is probable to assume that Murrays preparation during the hiatus potentially increased his confidence, leading to him taking more shots and being effective in the areas in which these shots are taken. Along with Murray other young stars like Donovan Mitchel, Bam Adebayo and Luka Doncic also experienced more success due to their increased involvement during Covid 19.

The 2020-21 NBA season, which also occurred during the pandemic, provided more data than the previous season since players were effected during the entire regular and post season as opposed to only a small portion of the regular season and entirety of the playoffs. During this time period some players were able to be vaccinated, however, others were unfortunate and fell victim to the disease. The after effects of Covid 19 symptoms had an impact on some players such as Jayson Tatum, who claims that the disease had led him to experiencing breathing problems. When making comparisons between figures 16 and 17, one can see that Jayson Tatum's accuracy took a hit on a majority of the areas around the three point line as well as a few areas in the midrange. This changes can also be seen numerically when making comparisons between figures 18 and 19. For example, Tatum's accuracy in the Left Side Center(24+) fell by about 12% and his accuracy in the Right Side Center(16-24),his most efficient spot on the court, fell by 10.4%. Tatum's drops in performance indicate that those previously infected by Covid 19 are susceptible will suffer from decreased shooting efficiency. Stamina and breathing rate play a critical role in a players performance and drastic alterations to these factors

will naturally lead to issues in athletic performance. Future research revolved around player health after Covid 19 will be able to provide insight into why such decreases in performance occur, however such information is often confidential.

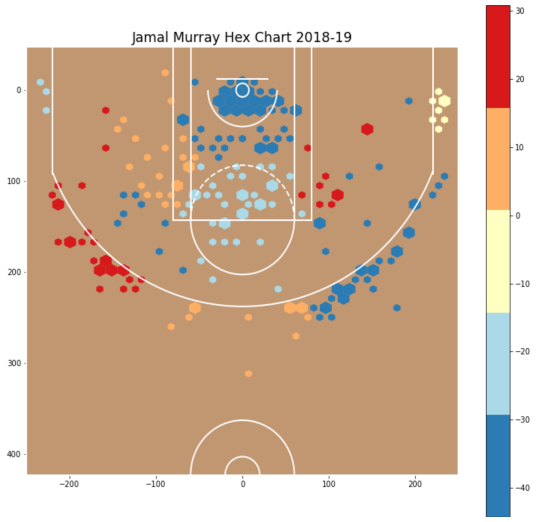

Figure 10: Jamal Murray Hex Chart for 2018-19 Playoffs

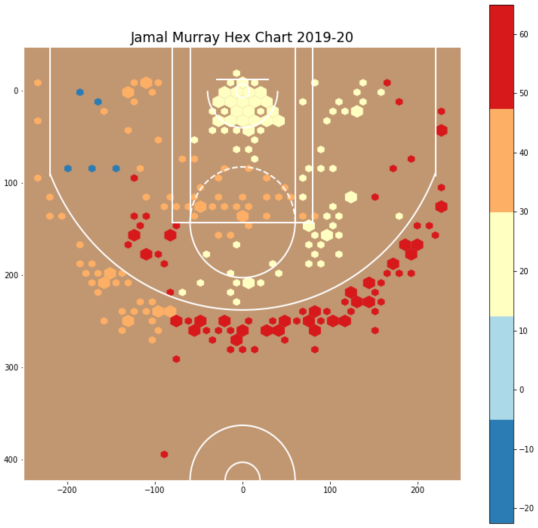

Figure 11: Jamal Murray Hex Chart for 2019-20 Playoffs

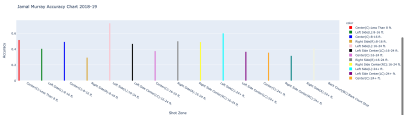

Figure 12: Jamal Murray Accuracy Chart for 2018-19 Playoffs

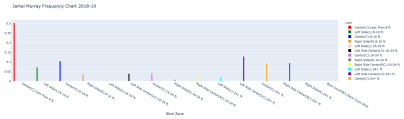

Figure 13: Jamal Murray Frequency Chart for 2018-19 Playoffs

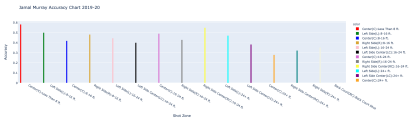

Figure 14: Jamal Murray Accuracy Chart for 2019-20 Playoffs

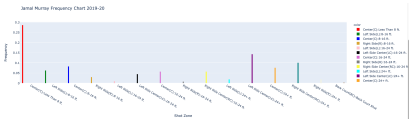

Figure 15: Jamal Murray Frequency Chart for 2019-20 Playoffs

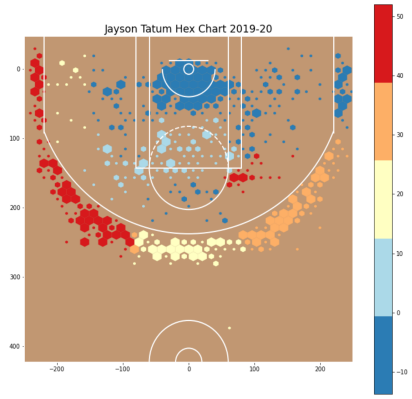

Figure 16: Jayson Tatum Hex Chart for 2019-20 Regular Season

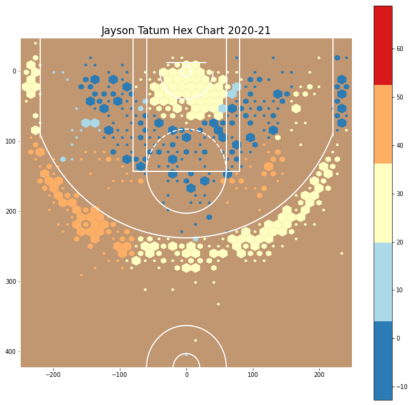

Figure 17: Jayson Tatum Hex Chart for 2019-21 Regular Season

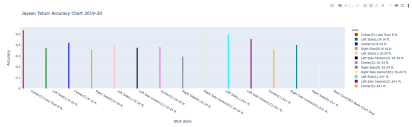

Figure 18: Jayson Tatum Accuracy Chart for 2019-20 Regular Season

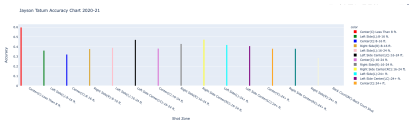

Figure 19: Jayson Tatum Accuracy Chart for 2020-21 Regular Season

5.3 Comparing the change in base statistics during affected seasons amongst NBA Players

After identifying various significant statistical variables, we can utilize these variables to compare players based on performance, to analyze which players were most effected by Covid 19. This is done through the radar plot visualization tool, which makes direct comparisons between two players of choice. Comparisons are made based on the following statistics: win shares(WS), defensive win shares(DWS), offensive win shares(OWS), Box Plus Minus(BPM) and Value over Replacement Player(VORP). We will also use classification techniques provided by Python's machine learning tools to make comparisons in performance amongst players of the same team. The data shown using these techniques will include every season played by these players to observe an deviations to their expected performance due to Covid 19. The classification techniques to be used to compare performances are as follows: Principal Component Analysis, Linear Discriminant Analysis, and T Distributed Stochastic Analysis.

5.3.1 Observations made from Radar Plot

Based on the analysis obtained from using the Radar plot visualization on various athletes, role players did not excel in the new bubble environment, while veterans performed on par with their average performance. However, there were some cases where younger players blossomed in the new setting portrayed in figure 20. In this Radar Plot, we make comparisons between 17 year veteran Carmelo Anthony and Bam Adebayo, a young player that was recently drafted in 2017. Adebayo, depicted by the orange radar, outperformed Carmelo Anthony in every category, thus resulting in the area of his radar far exceeding that of Anthony's. This indicates that the bubble did not hinder the development of some young players. Rather, the increased amount of involvement in the offense has been a benefit to their growth. However, average role players showed poor results, which can be attributed to the notion that role players thrive off of crowd involvement. Without the constant criticism on the court, players were much more loose with the ball resulting in inefficient play. The best performers in the bubble environment were star players who continued to display their normal dominance as well as several outliers including young players(Devin Booker, Jamal Murray, and T.J. Warren) who increased their production tremendously. These occurrences can be attributed to multiple factors that the shortened season 2019-2020 put a strain on. These include stress, health, team chemistry, motivation, and training. As well as disrupting the schedule and "normal" feel of games with lack of fan attendance, a long break mid-season, inability to practice with your team, and more back-to-back games upon the restart.

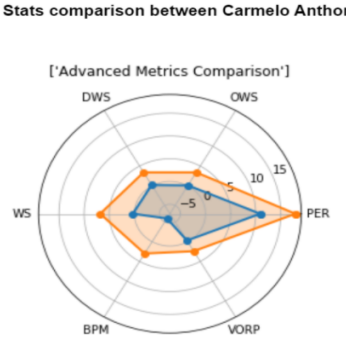

Figure 20: Radar Plot Comparison: Carmelo Anthony vs Bam Adebayo

5.3.2 Observations made from Principal Component Analysis

Before making any conclusions based on the PCA graphs, it is important to understand the feature contribution to each principal component. Because we are only graphing two at a time, different features that carry more weight in each component will significantly impact player positioning. These weights are displayed in figures 21-23. Principal component relies has considerably even distribution excluding strong representation for the field goal percentages. On the other hand, the second principal component receives much more contribution from both percentages as well as blocks. Lastly, principal component 3 is dominated by three-point field goal percentage. Figures 24 and 25 portray the analysis for the 2019-2021 season for the teams that played in the NBA Championship game. We can ascertain that regardless of the components, LeBron James is always very prominent in the graph. This is due to the fact that he has consistently piled up some of the best statistics in the league. This analysis further compliments

that idea showing that his success compared to the rest of his team excluding his star teammate Anthony Davis. Likewise, the Heat are led by players such as Jimmy Butler, a veteran, and Bam Adebayo, a young player previously mentioned for his amazing performance in the playoffs. This trend continues for the rest of the teams in the league. In other words, this analysis supports the idea that veterans and stars who have been consistent throughout their careers remain at the top of the league regardless of the bubble scenario, as well as a few outliers such as Bam Adebayo and other young players that elevated their game immensely despite the circumstances.

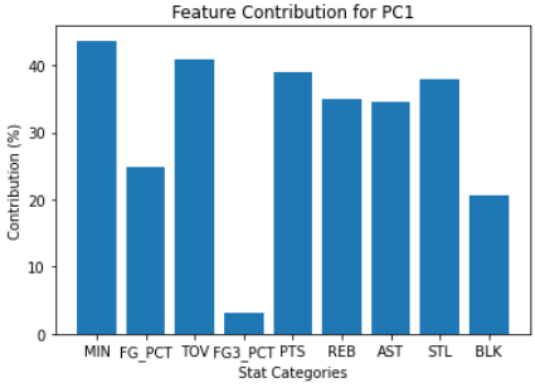

Figure 21: Feature Contribution for PC1

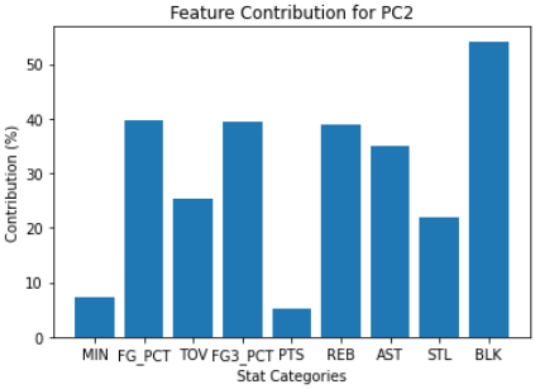

Figure 22: Feature Contribution for PC2

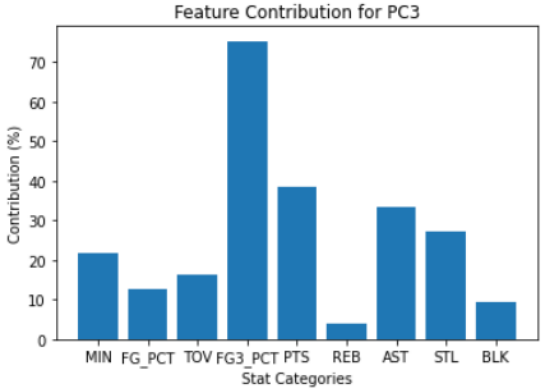

Figure 23: Feature Contribution for PC3

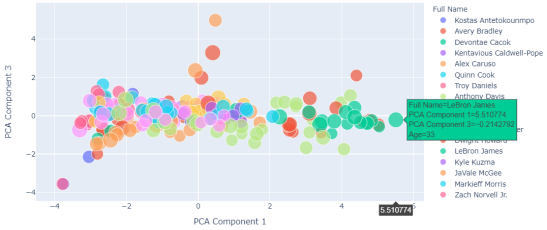

Figure 24: Los Angeles Lakers PC1 vs PC2

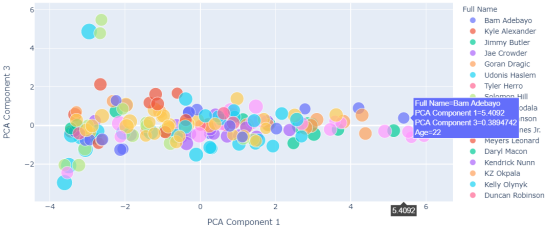

Figure 25: Miami Heat PC1 vs PC2

5.3.3 Observations made from Linear Discriminant Analysis

Using Linear Discriminant Analysis we will be able to classify players based on performance and the time period observed will range from the 2001-02 season to the current 2020-21 season. Players with low class separability are assumed to have similar performance to the classes of which they are closest to. Based on the observations made using the LDA classification method, it is apparent that certain players deviated from their expected performance. This can be seen in figure 26, in which players like LeBron James(shown in orange) experienced slight changes in performance. The seasons in which he played during Covid 19 are indicated by the black arrows shown in black. His 2019-20 season appears to deviate slightly from his normal performance, however, their is more significant deviation in his 2020-21 season. His All Star teammate Anthony Davis(shown in black) experience more significant deviation in his 2020-21 season since he has low class separability during that season when compared skilled role players like Marc Gasol(shown in yellow) and Markieff Morris (shown in gold). However, his performance in the season 2019-20 was up to par with his regular performance. It is important to note that, players who average more minutes per game are more susceptible to performance changes than others. Role players like Jared Dudley(shown in grey), Ben Mclemore(shown in red), etc. are less likely to portray performance changes and are more likely to have low class separability due to their low contribution to their teams. Furthermore, players who have played a greater number

of season's are more likely to have more class separability and few outliers than those who are relatively new to the league.

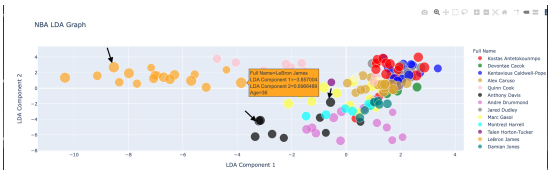

Figure 26: LDA depicting the performance of Players in the Los Angeles Lakers

Figure 27: TSNE classification used for the Los Angeles Lakers 2020-21 Team

5.3.4 Observations made from TSNE

The t-SNE classification graph shows clusters of player career statistics of a selected team. The graph above shows the t-SNE classification graph for the Lakers 2020-21 season roster. It takes the list of all players in the current season and collects season statistics for each year the player was active in the NBA. For example, LeBron James is a member of the Los Angeles Lakers. Since James has played the most seasons in the NBA, the t-SNE graph will have the most data points for him. Each circular data point represents a player's season statistics. Each player is represented using different colors. The t-SNE classifies the clusters of player data really well. It is easy to see in the visualization how t-SNE is able to find clusters of the same players data despite their stats being different every year. It is also interesting to view how the best players are located at the corner of the t-SNE graph. LeBron James who is a phenomenal player is located at the upper and right corner. Player statistics located on the bottom left indicate those may be the worst player statistics. There are definitely some noticeable correlation between position of data points on the t-SNE graph and player performance. Players who score the most points in games are usually located higher. This is why LeBron is at the highest position. It is also important to note that players on the right most side of the graph are all centers. Being on the right side may correlate to players with higher rebounds and better defensive statistics. According to this example, it seems higher position correlates to higher offensive statistics and right positions indicated higher defensive statistics.

5.4 Player Contribution to Playoff Success

This visualization predicts the NBA teams that will make the playoffs based on team stats from previous season. The visualization also includes graphs to show player statistics from each team help understand how player contribution affects a teams ability to make the playoffs. In order to view a playoff prediction, you have to select 2 options from the drop-down menu. The left drop-down selects the conference (eastern conference/ western conference). The right drop-down selects the season or season year and there are three different seasons to choose from. By default, the drop-down is set to western conference for the 2020-21 season year. This is the most recent season. The accuracy for the logistic regression model is 73% while the accuracy for support vector machine is 80%. Clearly the support vector machine model predicted the playoff teams better. The columns with the probabilities for logistic regression model and SVM model are styled using a heat map color coding. Teams with the higher probabilities according to the model are shaded in with darker shades of blue while teams with smaller probabilities have lighter shades of blue as the cell background. The accurate playoff selections are under the column "playoffsBool". This column indicates a boolean value (1 means the team qualified to be in the playoffs and 0 means they were not qualified for NBA playoffs). The predicted boolean values are the columns to the far right. There are also bar graphs below the dash table that aid the visualization. The first graph shows player statistics (offensive rating, defensive rating, net rating, and usage percentage).

The second graph shows changes in player statistics from the previous year (player had to have played for same team both years).

The Phoenix Suns are an interesting team to analyze with this visualization. The Phoenix Suns did not make the playoffs for the 2019-20 season. However, both logistic regression model and SVM model predicted with a probability of 70% that the Suns would make the playoffs. During this season, teams were affected by the pandemic and had to play in the NBA bubble. Diving deeper into player statistics, it's easy to see that three players had the highest net ratings and usage percentages (Devin Booker, Chris Paul, Cameron Payne and Dario Saric). These players also had the highest usage percentages with Devin Booker clearly distinguished at the top with a usage percentage of 32.6%. When you change the option for the season from 2019-20 to 2020-21, you can see the Suns actually made the playoffs for this season. Now when you check the graph for changes in player statistics, you can see improvement in certain players, which is most likely attributed to the change in usage rating caused by many players being suspended from games for a period of time due to Covid protocol. Due to this, some players have had to step up big time to fill these roles. According to the bar graph, Dario Saric has an increase in net ratings of 11.9. Some other players who saw improvement in net ratings are Mikal Bridges(3.4), Cameron Johnson(4.4) and more. Dario Saric also saw an improvement in usage percentages by 4.7% followed by Devin Booker(2.7%) who saw the second more improved usage percentage. These changes in player statistics attributed to the Suns making the 2020-21 playoffs.

Figure 28: Machine Learning Predictions in Dash Table

Figure 29: Player Statistic Bar Charts

6 Conclusion

Based on the results obtained from our observations, we conclude that the bubble environment created by the circumstances of Covid 19 benefited the performance of younger players. The claim is supported by analyses performed using the Radar Plot and Hex Shot Chart. They exhibited the greatest positive change in performance. This is partly due to the fact that they have lower expectations and workload due to inexperience. As our observations have displayed, some rise above this adversity to produce career higher stats in a tumultuous time. The observations from the hex shot chart show that when younger players were given an increased opportunity on offense, a majority of them were able to step up to the challenge and were able to portray significant efficiency. A prime example of a player that was able to portray this phenomena was Jamal Murray who increased his proficiency at three point range and midrange when comparing his success rate from a previous season. This claim is reinforced in the Radar Plot in which a younger player like Bam Adebayo reaped more benefits from playing in the bubble than an older veteran such as Carmelo Anthony, since he exceeded Anthony in every advanced metric as shown in the plot. Utilizing various key features such as usage rating, three point field goal percentage, minutes played and more we were able to formulate accurate regression models that predict both base and advanced statistics. Additionally, we conclude that players that contracted the virus were likely to experience a hit in performance as seen when comparing the hex shot charts and accuracy bar charts of Jayson Tatum during the 2019-2020 and 2020-2021 seasons. Veteran reliability can be observed through our classification models, as the better positioned points were mostly made up of vet-

erans and star players in the Principal Component Analysis. This speaks to their experience and ability to remain prepared regardless of unexpected adversity. The middle of the pack was made up mostly of young players and consistent veterans while the worse positions consisted of below average players that played as poorly as they usually do and some stars that far under performed their high expectations. Thus, this reinforces the idea that star players are most likely to perform at the highest level compared to the rest of the league along with certain young players that flourished in their expanded roles. Using the LDA classification, we can further assess that experienced players are more sensitive to performance changes in Covid 19 indicated by data points of LeBron James and Anthony Davis slightly deviating from their respective classes and the performance of the majority of role players remaining relatively unchanged. The playoff prediction visualization showed the changes in player statistics which contributed to teams making the playoffs. Dario Saric is an NBA player who plays for the Phoenix Suns. Saric's usage percentage and net ratings increased during the pandemic. Saric and many other NBA players had to step up during the pandemic because other players were often suspended because of Covid related restrictions. Many players missed games because they were affected by Covid. As a result, bench players got more time to play. Saric is a good example of a player whose net ratings increased with his usage percentage. Saric was able to maximize his output as result of Covid. This might have been one of the reasons Suns made the playoffs. t-SNE classification enabled us to view clusters of player statistics. We are able to observe how players from different NBA teams play throughout their career. We are also able to observe how talented the players are in each team. We can see consistent players like LeBron James who has all his data points in one area. His data points are usually far away from the other members of his team so show how he is NBA on a whole different level. Then, you can view players like Demarcus Cousins who didn't quite have such a consistent career. You can see two clusters of data for Cousins. You have a cluster of data points which represents his productive early career. This cluster was further away in one corner of the visualization which signified how he played really well or really bad. In this case, those were his best years. However, there is a small cluster in the middle of the visualization which represents how he hasn't been playing his best. One overall observation is that great players and not so great players are usually on opposite sides of the graph.

7 Limitations

The underlying nature of our research poses a few limitations. For example, despite having access to an enormous quantity of statistical data, we do not have any access to data that may concern non statistical factors such as time spent training/exercising during the Covid 19 hiatus, Body Mass Index(BMI), or Psychological Well Being(PWB). The scarcity of such data is most likely due to the fact that this information remains confidential for concern over player privacy. This forces us to rely heavily on statistical game data that does not capture the full effect of Covid on an NBA player's life. Another limitation that occurs is that our results can be skewed if we have a small sample size. For example, according to figure 30, the results of our Hex Shot Chart Graph show that Houston Rockets guard, Russell Westbrook, had a poor playoff performance during the 2019-2021 post season. However, Westbrook missed seven playoffs games due to contracting the virus and only played four. Had he played all or most of the games, we would have a more accurate interpretation of the impact Covid 19 had on his playoff performance. Additionally, it's difficult to forecast our predictions, since many of our feature variables like usage rating and minutes played do not exhibit a steady rate of change. The only statistical category the exhibits a predictable rate of change in our data is player age, however this statistic is detrimental to the accuracy of our regression model.

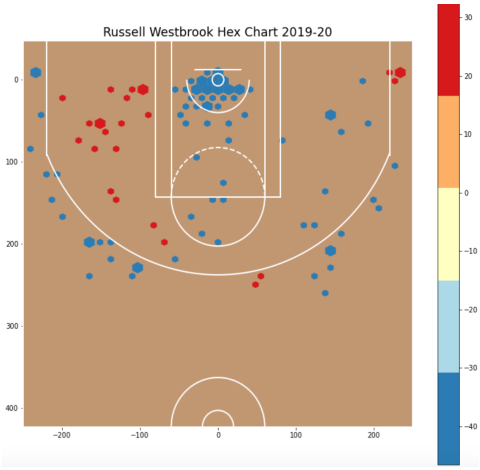

Figure 30: Russel Westbrook Playoff ShotChart 2019-2020

References

[1] A. Qin and J. C. Hernández, "China Reports First Death From New Virus," The New York Times, 11-Jan-2020. [Online]. Available: https://www.nytimes.com/2020/01/10/world/asia/china-virus-wuhan-death.html. [Accessed: 15-Dec-2020].

[2] McHill, Andrew W, and Evan D Chinoy. "Utilizing the National Basketball Association's COVID-19 restart "bubble" to uncover the impact of travel and circadian disruption on athletic performance." Scientific reports vol. 10,1 21827. 11 Dec. 2020, doi:10.1038/s41598-020-78901-2

[3] A. Vaquera, " Key game indicators in NBA players' performance profiles," ResearchGate. [Online]. Available: https://www.researchgate.net/publication/332032047_Key_game_indicators_in_NBA_players'_performance_profiles. [Accessed: 17-Dec-2020].

[4] J. P. Hwang, "NBA shot data analytics & visualization with Python, Pandas and Matplotlib Part 2 - Grouping data by area," The Visual in the Noise, 29-Oct-2020. [Online]. Available: https://www.jphwang.com/nba-shot-data-analytics-visualization-with-python-pandas-and-matplotlib-part-2-grouping-data-by-area/. [Accessed: 18-Dec-2020].

[5] A. Kapri, "PCA vs LDA vs T-SNELet's Understand the difference between them!," Medium, 21-May-2020. [Online]. Available: https://medium.com/analytics-vidhya/pca-vs-lda-vs-t-sne-lets-understand-the-difference-between-them-22fa6b9be9d0.

[6] M. Beckler and H. Wang, NBA Oracle. [Online]. Available: https://www.mbeckler.org/coursework/2008-2009/10701_report.pdf. [Accessed: 10-May-2021].

[7] "NBA player performance prediction," Google Sites. [Online]. Available: https://sites.google.com/view/kutouxiyiji/home/data-science-projects/nba-player-performance-prediction. [Accessed: 25-May-2021].