George Cole, Jack Crilly, Reuben Dayal, Brian Papiernik, David Sobek
Professor Barron
MSSA 60530: Human Performance Analytics
7 May 2024

NBA Injuries and Load Management Final Report

Introduction:

The primary focus of this project is to conduct analysis on the relationship between NBA player movement and play-by-play usage data with injury data in an attempt to develop an effective method of player workload management which can aid current injury prevention tactics. Our data is a collection of 39 Chicago Bulls games from the 2015-2016 NBA season, which contains GPS tracking data, injury data, play-by-play data, and player usage metrics. Additionally, we have created our own derived variables using the variables given to us in these datasets to create a more accurate model. These variables include velocity, acceleration, direction, and cumulative load (derived from box-score usage metrics). Our analysis aims to provide valuable insight into the world of high-level basketball by investigating how the in-game movement and workload management of players can contribute to injuries and be used to develop effective recovery methods.

Motivation:

Recently, there has been a big issue in the NBA when it comes to load management. Load management was originally designed for veteran players in the league to focus on the postseason by sitting out unimportant regular season games, but this system very quickly started being abused by league superstars. Perfectly healthy players began to sit out games on the tail end of a back-to-back, and superstar players have used the system to watch their minutes and purposely play less in regular season games. This can be a major issue for both on-court team performance and fan happiness, as many fans spend hundreds of dollars to see their favorite player live, only to get to the arena and watch them riding the bench the whole game. By developing a more effective load management system based on player movement and workload, we hope to increase on-court potential while still having star players perform at a high level. This will contribute to increased team performance and ensure that fans do not leave the arena disappointed after seeing their favorite players on the sidelines the one time they might get to see a player play live all season.

Problem: The project aims to investigate the relationship between injury prevention, specifically focusing on derivative variables created from the player tracking data, and various factors from the Play by Play and box score usage statistics that may contribute to a player's injury prevention.

Variables of Interest:

Understanding the impact of a player's tracking statistics, cumulative load from box-score usage metrics, and play by play metrics is crucial for both sports analytics and team roster management. This analysis could reveal insights into key workload indicators on a play-by-play basis that can lead to injuries and injury prevention. A description of derived variables (Velocity, Acceleration, Distance, Injured player, Injury event by game, Pace, PIE (Player Impact Estimate), usage rate, velocity thresholds, velocity threshold categories, and acceleration threshold categories) is provided on the attached Rmd file and below in the data description section.

Problem Framing:

The end goal of our project is to find a balance between managing player injuries and health while also making sure the NBA product as a whole doesn't suffer. Reducing the amount of injuries by improved injury detection methods and metrics would reduce the amount of games missed due to injury as well as the rest period following those injuries. To predict these injuries we will use the player's run speeds, accelerations, and high intensity efforts to determine which speeds can be detrimental to players lower body injuries. If a player gets hurt using a specific movement like a hard cut, we can use similar instances where the player has done that before to see whether it has an effect with multiple uses of the same movement. We will also utilize usage stats like PIE, pace, and usage rate from box score statistics over these 39 games to quantify the cumulative loads of different players throughout games and the season overall. While not all movements and high workloads are the responsibility of causing injuries, the ability to identify specific movements with respect to workload and usages can help a training staff teach the proper form of movements or give players rest days when they are needed in order to manage players workloads better.

The impact of healthy players extends beyond player performance by enhancing team dynamics and enabling better roster management. In turn, this impact leads to better allocation of resources which ultimately contributes to a more successful and sustainable season.

Proposed Solution: The project will utilize statistical methods and advanced machine learning techniques in R to analyze the dataset surrounding player movement metrics, game play-by-play usage metrics, and injury reports. A goal is to develop a predictive model that correlates specific player movements and usage metrics with the likelihood of causing injuries.

Related Work

Previous projects have looked into injuries over the course of NBA seasons and what factors contribute to players playing well and staying healthy. They have looked at tracking data, play-by-play data, and box-score data individually to predict performance, but these different types of data haven't been used in conjunction with each other to predict injury onset yet. These previous studies can help steer our project in the right direction by giving us ideas on how to use different variables and derived variables to solve the problem in question.

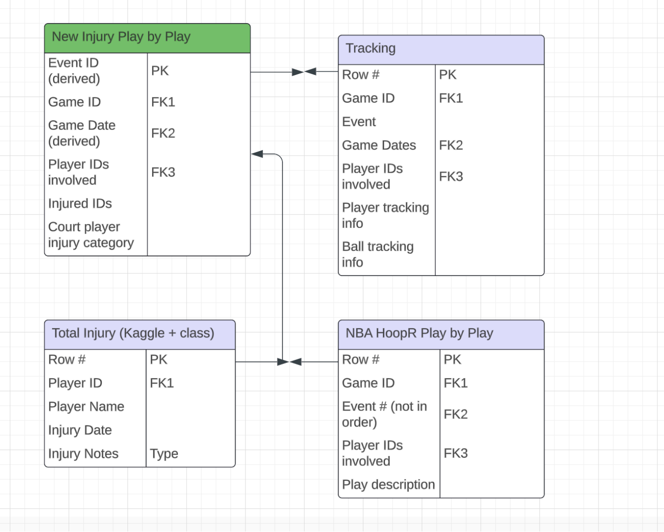
Other Studies:

Exploring Game Performance in the National Basketball Association Using Player Tracking Data by Jaime Sampaio, Tim McGarry, et al

The primary goal of this project is to use NBA tracking data from the 2013-2014 NBA season and see what on-the-court variables are significant in quantifying the performance of players. Some variables that were examined in this study include speed and distance, type of shot (i.e. drive, pull-up, catch and shoot, etc), passing variables, and shooting percentages. While this study was not really injury related and more performance related, it provided us with extra derived variable ideas that we applied to our own project.

Data Description

To build our predictive model, we required linking multiple different datasets together that we found through an NBA HoopR package (NBA Box Score Player Usage Statistics and Play by Play datasets), an injury dataset found on Kaggle with injury reports dating back to 2008, and game tracking data for the 39 Bulls Games for the 2015-2016 NBA Season.The complete clean dataset that would be used to predict injuries required the linkage of multiple datasets which is shown in the visual to the right. The final merged dataset consisted of 55 variables reflecting a comprehensive mix of game performance usage metrics, player movements, historical injury data, and unique GameIDs, PlayerIDs, and EventIDs. These variables include player velocity, acceleration, distance covered on a play, time on the court, and detailed injury history which are essential for understanding the dynamic between player activities and potential injury risk. The final merged dataset spanned 83,000 rows which represents detailed play-by-play data for each of the five Bulls players on the court during the 39 games of the season. Each row captures a snapshot of a specific game event that occurs every 0.04 seconds.



New Injury Play by Play	
Event ID (derived)	PK
Game ID	FK1
Game Date (derived)	FK2
Player IDs involved	FK3
Injured IDs	
Court player injury category	

Tracking	
Row #	PK
Game ID	FK1
Event	
Game Dates	FK2
Player IDs involved	FK3
Player tracking info	
Ball tracking info	

Total Injury (Kaggle + class)	
Row #	PK
Player ID	FK1
Player Name	
Injury Date	
Injury Notes	Type

NBA HoopR Play by Play	
Row #	PK
Game ID	FK1
Event # (not in order)	FK2
Player IDs involved	FK3
Play description	

Data Sources:

NBA Box Score description: The Game ID column found in this dataset was used to join this dataset to the other ones used in the project. The main statistics that were extracted from this dataset were minutes played, usage rate/percentage, pace, and PIE (player impact estimate). Minutes played was used to quantify how long a player played in each game during this 39 game stretch. Usage rate quantifies a player's presence on the offensive end of the floor. This metric takes into account statistics like Field Goals Attempted (FGA), Free Throws Attempted (FTA), Turnovers (TOV), Minutes Played (MP) in a comprehensive formula to output a player's usage on offense. Pace is the amount of possessions a team has over the course of a 48 minute period (full NBA game). Since this was an individual statistics dataset, the pace for each player corresponded to the time that they were on the court. Finally, PIE was used to quantify the overall impact of a player on both ends of the floor. This metric took into account statistics like Points, FGA, FTM, Steals, Blocks, and Rebounds (offensive and defensive). The formulas for Usage Rate and PIE are attached in the visuals below.

Usage Rate formula:

$$USG\% = \frac{100\,(0.33\,AST + FGA + 0.44\,FTA + TO)}{POSS}$$

PIE formula:

PIE Formula=(PTS + FGM + FTM – FGA – FTA + Deff.REB + Off.REB/2 + AST + STL + BLK/2 – PF – TO) / (Game.PTS + Game.FGM + Game.FTM – Game.FGA – Game.FTA + Game.Deff.REB + Game.Off.REB/2 + Game.AST + Game.STL + Game.BLK/2 – Game.PF – Game.TO)

NBA Tracking Data: The tracking dataset consisted of 6.8 millions rows where each row is 0.04 seconds of the game clock for a specific event of a specific game. The Tracking Dataset serves as a critical component in our analysis of the likelihood of injury based on the player movement derivative variables. This dataset needed a lot of cleaning but it contained the X and Y coordinates for all 10 players on the court at any given moment. The coordinates provided spatial analysis insights for player movements.. Basically, our group was able to create these derivatives based on lag differences between the two rows. These derivative variables were created based on functions that are provided in our Rcode. Many of the derivative variables were based on physics metrics that were calculated between the two rows. Also, each row has a unique PlayerID and TeamID indicator for all 10 players on the court. These were used linking the tracking data and our derivative variables with our play-by-play usage metrics. The GameID, EventID, and PlayerID are the unique keys that were used to join this dataset to the other datasets used in this project. Each row of data represents a snapshot taken every 0.04 seconds of game time which offers a granular view of the game's progression. This time-stamping is vital for synchronization of events within the game like shots, passes, and defensive plays, and which team has possession. Below are the derivative variables that were created from the tracking data.

- Velocity: This variable measures the speed of a player at a specific moment, providing insights into their movement dynamics during the game.
- Acceleration: Reflects the change in speed of a player at a given point in time, crucial for analyzing how quickly a player can adjust their pace during play.
- Velocity Threshold Category: Represents the count of consecutive data points where a player's speed exceeds 10 ft/sec, useful for identifying sustained high-speed activities.
- Direction: Indicates the direction in which a player moves from one point to another, from coordinates (x1, y1) to (x2, y2), essential for understanding movement patterns.
- Velocity Category: Categorizes a player's speed into one of six groups based on the overall distribution of velocity data, which helps in comparing different levels of player activity.
- Acceleration Category: Similar to velocity category, this variable divides acceleration into six distinct groups based on how acceleration data is distributed among players.
- Possession Team: Identifies which team is in possession of the ball at any given time, derived from player positioning and ball proximity.
- Ball Handler: Specifies the player who is closest to the ball, effectively indicating who is controlling the ball at that moment.
- Time with Ball: Measures the duration a player holds the ball during a particular event, offering insights into player control and influence on the game.
- Time on Offense/Time on Defense: Quantifies the amount of time spent by a player or team in offensive or defensive play during a particular event, providing a clear picture of game strategy and player roles.

NBA Injury Dataset: We found players on the injury report and, within the last game that they appeared in prior to injury, pinpointed which play was their last of the game, assuming this is the play on which play the injury may have occurred. This method of injury-play flagging was done in conjunction with the play-by-play dataset.

NBA Play by Play: The Play-by-Play dataset is for every game of the Chicago Bulls 39 games in question from the 2015-2016 season. The GameID and EventID, and PlayerID are the unique keys that were used to join this dataset to the other datasets used in the project. The interesting part of this dataset is that it shows which 5 PlayerIDs are on the court for each team during each play. Additionally, each event is able to show which team scored and other result variables like the score margin for each specific event. From here, the play-by-play data was further lengthened (then filtered) so that each row contained a Bull’s player’s data during a specific play. The main purpose of this dataset is to see which 5 Bulls Players were on the court and link this play-by-play dataset to the usage box score statistics through the gameID. After, we were created new variables to increase the level of granularity that took the usage variables (Usage Rate and PIE) and multiplied them by the cumulative seconds of that player in order to capture the players’ impact on the court progressively throughout the game (PACE was also used, but divided by cumulative seconds instead in order to create the cumulative standard).

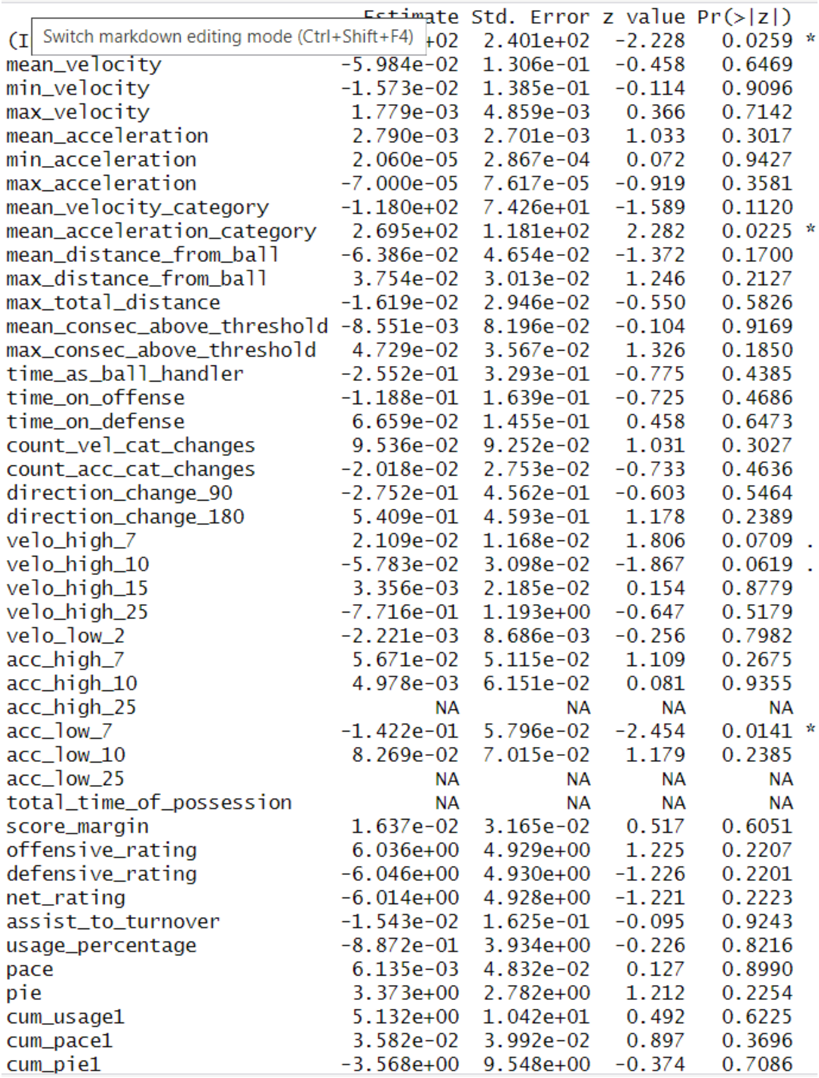
Data Split:

To train and evaluate our injury predictive models, we divided the dataset into distinct subsets, typically following the standard practice of splitting data into training, and test sets. This separation allowed us to assess the model's performance on unseen data, ensuring that it could make accurate injury predictions for nba players beyond the data it was trained on. The split was 80/20… meaning that 80% of data was included in the train data while the other 20% was our test data.

Methods

Logistic Regression: The first model that we will use is the **generalized linear model (glm)** and we will use the logistic regression function. Generalized linear models are used to understand the relationship between a binary outcome variable and multiple input variables. It estimates the probability that the outcome variable will be "1" given the input variables. A strength of using generalized linear models is that they are easy to implement and interpret. We can see which basketball variables have a significant effect on contributing to our injury prediction. A weakness of these models is that they assume that predictor variables are independent of each other, and this is not the case in a real world scenario. When trying to predict injuries, there are multiple different variables that affect the prediction. Logistic Regression provides valuable insights into how changes in these variables above impact the likelihood of increasing injury risk, allowing us to quantify the influence of various factors on the injury outcome.

XGBoost: The next model run is a **XGBoost model** where the decision trees are built sequentially in order to reduce error from the previous trees, updating the residual errors. This model is used to find relationships within complex datasets as we have here with multiple datasets merged together. Effectively, a XGBoost model uses weak learners where the bias is high and the predictive power is only a small amount higher than random guessing. By sequentially combining these weak learners using boosting; creates a strong learner that has low bias and variance. A XGBoost model is perfect for this data as it is extremely complex with multiple different variables and it is hard to predict injury in basketball without a multitude of inputs. Another reason this model is great to use is that XGBoost is easy to tune though it can take awhile. Also, it provides a ranking of variables. With the high number of variables this dataset has, the variable ranking will help those creating insights understand which ones to focus on. So that we can properly give advice on how game and training plans should be altered.



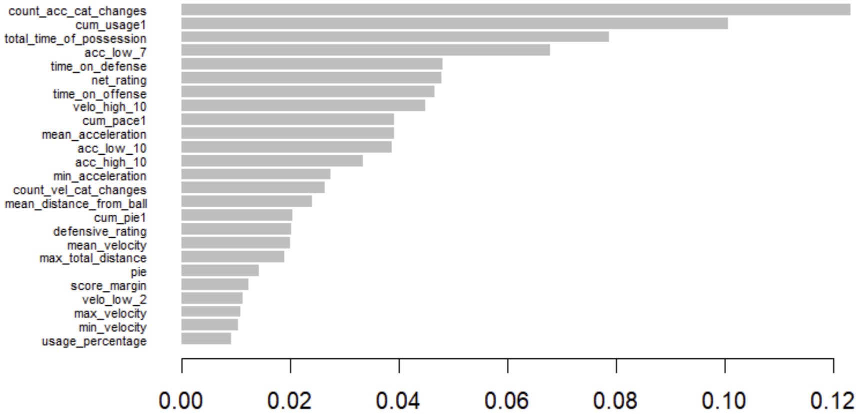
```
                           Estimate Std. Error z value Pr(>|z|)
(I Switch markdown editing mode (Ctrl+Shift+F4) +02  2.401e+02  -2.228   0.0259 *
mean_velocity              -5.984e-02  1.306e-01  -0.458   0.6469
min_velocity               -1.573e-02  1.385e-01  -0.114   0.9096
max_velocity                1.779e-03  4.859e-03   0.366   0.7142
mean_acceleration           2.790e-03  2.701e-03   1.033   0.3017
min_acceleration            2.060e-05  2.867e-04   0.072   0.9427
max_acceleration           -7.000e-05  7.617e-05  -0.919   0.3581
mean_velocity_category     -1.180e+02  7.426e+01  -1.589   0.1120
mean_acceleration_category  2.695e+02  1.181e+02   2.282   0.0225 *
mean_distance_from_ball    -6.386e-02  4.654e-02  -1.372   0.1700
max_distance_from_ball      3.754e-02  3.013e-02   1.246   0.2127
max_total_distance         -1.619e-02  2.946e-02  -0.550   0.5826
mean_consec_above_threshold -8.551e-03 8.196e-02  -0.104   0.9169
max_consec_above_threshold  4.729e-02  3.567e-02   1.326   0.1850
time_as_ball_handler       -2.552e-01  3.293e-01  -0.775   0.4385
time_on_offense            -1.188e-01  1.639e-01  -0.725   0.4686
time_on_defense             6.659e-02  1.455e-01   0.458   0.6473
count_vel_cat_changes       9.536e-02  9.252e-02   1.031   0.3027
count_acc_cat_changes      -2.018e-02  2.753e-02  -0.733   0.4636
direction_change_90        -2.752e-01  4.562e-01  -0.603   0.5464
direction_change_180        5.409e-01  4.593e-01   1.178   0.2389
velo_high_7                 2.109e-02  1.168e-02   1.806   0.0709 .
velo_high_10               -5.783e-02  3.098e-02  -1.867   0.0619 .
velo_high_15                3.356e-03  2.185e-02   0.154   0.8779
velo_high_25               -7.716e-01  1.193e+00  -0.647   0.5179
velo_low_2                 -2.221e-03  8.686e-03  -0.256   0.7982
acc_high_7                  5.671e-02  5.115e-02   1.109   0.2675
acc_high_10                 4.978e-03  6.151e-02   0.081   0.9355
acc_high_25                        NA         NA      NA       NA
acc_low_7                  -1.422e-01  5.796e-02  -2.454   0.0141 *
acc_low_10                  8.269e-02  7.015e-02   1.179   0.2385
acc_low_25                         NA         NA      NA       NA
total_time_of_possession           NA         NA      NA       NA
score_margin                1.637e-02  3.165e-02   0.517   0.6051
offensive_rating            6.036e+00  4.929e+00   1.225   0.2207
defensive_rating           -6.046e+00  4.930e+00  -1.226   0.2201
net_rating                 -6.014e+00  4.928e+00  -1.221   0.2223
assist_to_turnover         -1.543e-02  1.625e-01  -0.095   0.9243
usage_percentage           -8.872e-01  3.934e+00  -0.226   0.8216
pace                        6.135e-03  4.832e-02   0.127   0.8990
pie                         3.373e+00  2.782e+00   1.212   0.2254
cum_usage1                  5.132e+00  1.042e+01   0.492   0.6225
cum_pace1                   3.582e-02  3.992e-02   0.897   0.3696
cum_pie1                   -3.568e+00  9.548e+00  -0.374   0.7086
```

Results

In this project, we employed two distinct machine learning methods – Logistic Regression and XGBoost – that we had previously learned about in our machine learning classes. The methods that we utilized to facilitate our analysis of the relationship between the "injury binary variable" and the various mix of game performance usage metrics and derived player movements variables. Our logistic regression regression model was fitted using a regression dataset (nba_regression_train) that we created when splitting the dataset which focused on specific derivative player movement metrics and play-by-play usage metrics for the different players for each .04 seconds within a game event within a specific game as described earlier. For the XGBoost model, we used the training dataset described in the Data Split section. It's worth noting that XGBoost required an additional preprocessing step due to its utilization of a specialized data type called DMatrix, which efficiently stores sparse matrices, a particularly valuable feature when dealing with datasets containing numerous zero values.

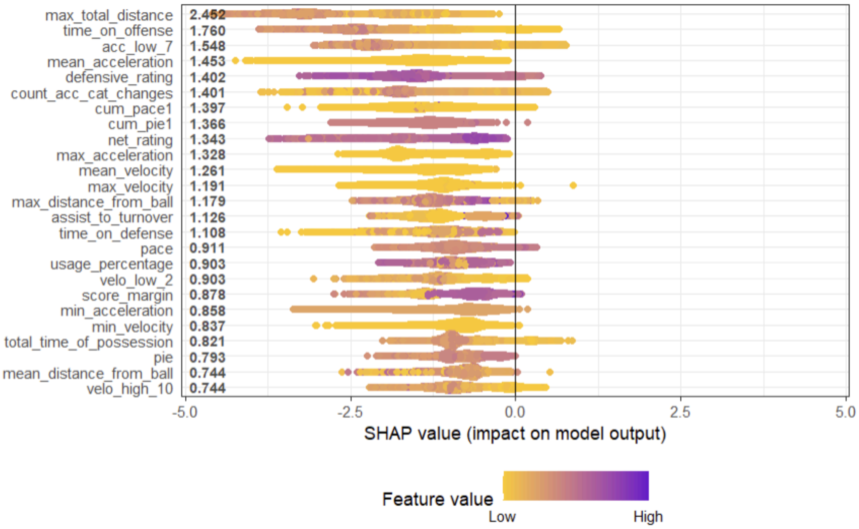
Logistic Regression: For our logistic regression, we created a logistic regression model for the relationship between our dependent variable (injury likelihood "player_inj") and multiple independent variables (various player tracking metrics and play-by-play usage performance metrics). Unlike our XGBoost model that requires complex hyperparameter tuning, our logistic regression offers a more straightforward statistical approach. The logistic regression output provides coefficients for each independent variable, which indicates the direction and strength of influence on predicting the likelihood of injury. In this specific model, we observed that certain variables had statistically significant effects on the likelihood of injury. Notably, our deceleration count greater than -7 ft2/sec, our mean acceleration category, and velocity count greater than 7 ft/sec and 10 ft/sec were proven to be statistically significant. In order, the p-value for these variables were 0.0141, 0.0225, 0.0709, and 0.0619. The two variables with positive coefficients that increased the likelihood of injury that were statistically significant were the mean acceleration category and the velocity count greater than 7 ft/sec. This insight means that maintaining a velocity above 7 ft/sec can increase the likelihood of injury and sporadic sprinting in order to maintain high acceleration numbers will increase the likelihood of injury The two variables with negative coefficients that decreased the likelihood of injury were our deceleration count greater than -7 ft/sec and our velocity count greater than 10 ft/sec. This insight might indicate that stopping your movements with a deceleration lower than -7 ft2/sec decreases the likelihood of injury. Deceleration might have less wear on the body than acceleration. Also, constantly sprinting above 10 ft/sec also decreases the likelihood of injury which might indicate that these players had better fitness than other injured players. It is important to not that the logistic regression model at an AIC value of 341.4 which suggests the model's effectiveness of explaining the variability of the response data while penalizing for the number of predictors.

XGBoost: The XGBoost method was used to predict the likelihood of injury binary variable based on some complicated relationships between the play-by-play usage performance metrics and our various player tracking metrics. For our XGBoost, the hyperparameters that required tuning included learning rate (0.3), max depth (15), minimum child weight (15), gamma (0.1), subsample (0.6), and column sample by tree (0.6). These hyperparameters were tuned based on the highest Area Under Curve (AUC) for each hyperparameter. The output for our XGBoost model creates a probability of a binary variable (likelihood of injury) based on the hyperparameters above and the play-by-play usage performance metrics and our various player tracking metrics. The most important variables are in the visualization to the right. The top 10 most important variables in order are counting the number of acceleration category changes between 0.04 seconds time difference (6 to 1, 6 to 2, 5 to 1, 1 to 5, 2 to 6, 1 to 6), cumulative usage, total time of possession, deceleration greater than -7 ft2/sec, time on defense, net efficiency rating, time on offense, velocity count greater than 10 ft/sec, cumulative PACE, and mean acceleration. The AIC when tuning the model was around 0.80. The model's ability to adjust the probability of injury based on these metrics allows for a unique understanding of



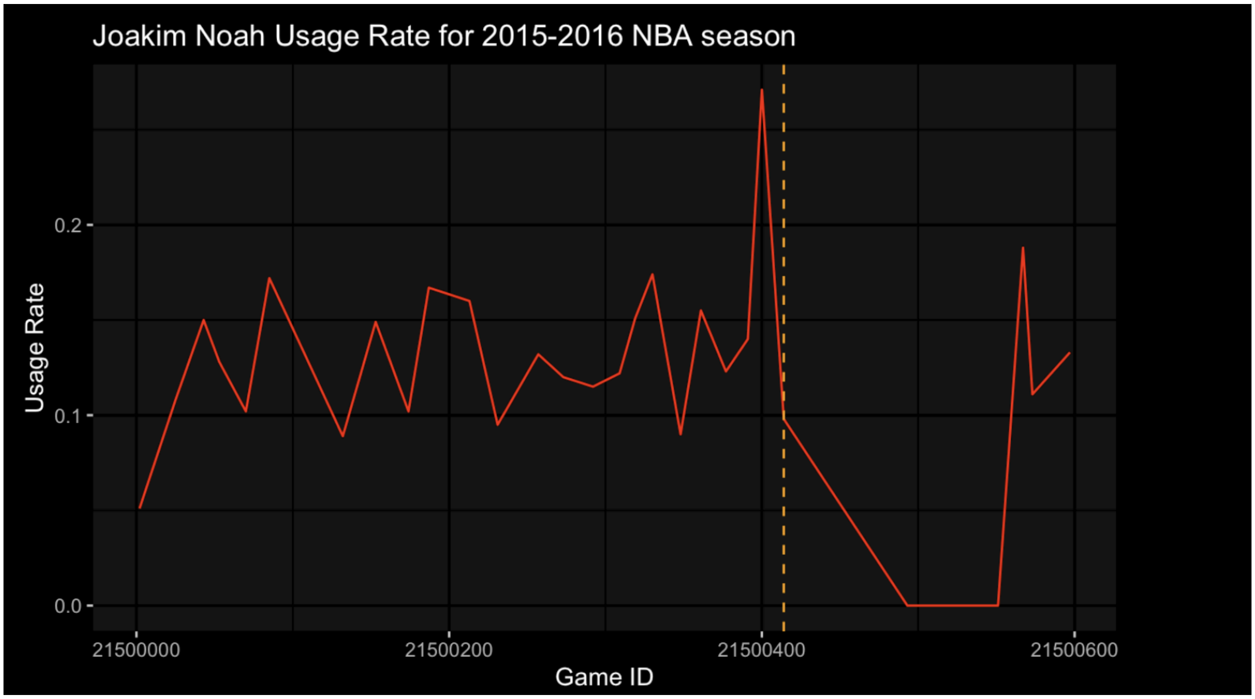
injury risk factors specific to NBA players. This can be instrumental in developing strategies for injury prevention and player load management. This modeling approach not only identifies at-risk individuals, but also helps in planning the workload and recovery protocols, thereby optimizing player performance and their career longevity.

SHAP Visual based on XGBoost Model:

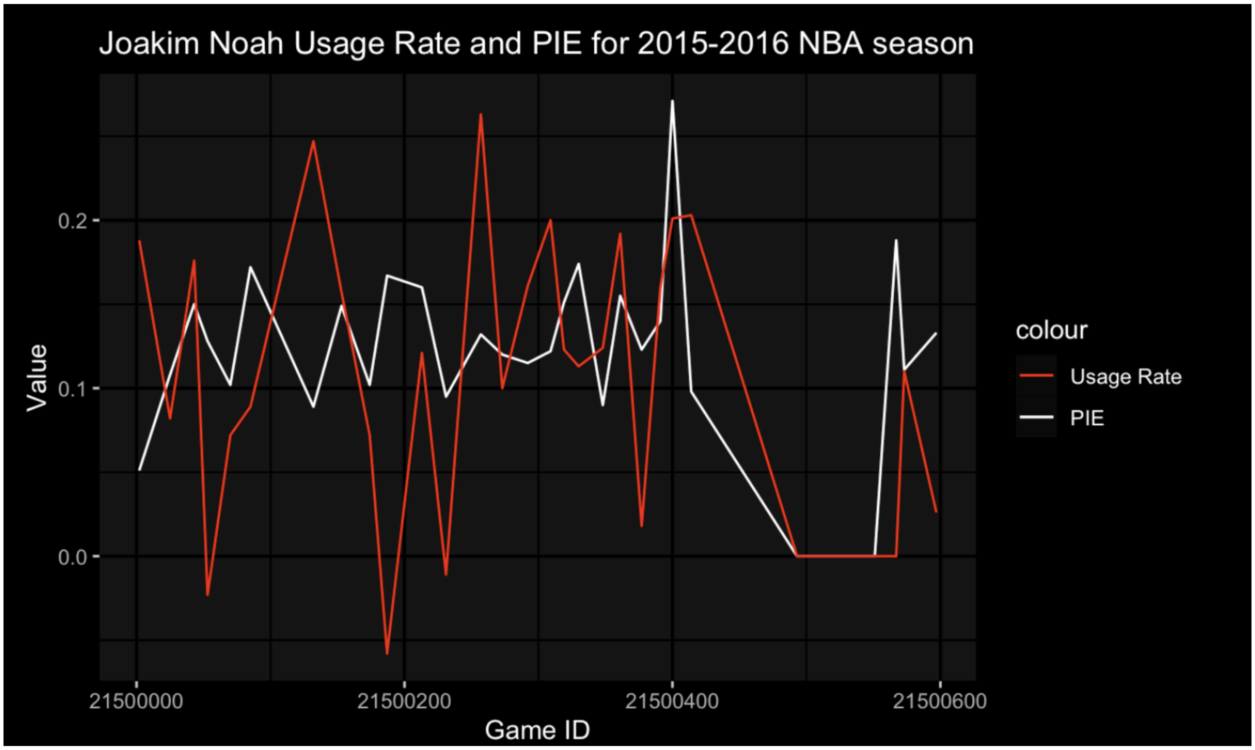


The SHAP visual that is provided on the right is based on the XGBoost. The SHAP visual is color coded where purple indicates high values and yellow indicates low values. It is interesting to see the distribution of how high and low values for these specific variables affect the likelihood of injury. Dots to the right of the center increase the likelihood of injury where dots to the left of the center decrease the likelihood of injury. The top influencers for variables affecting injury are max total difference, time on offense, deceleration count greater than -7 ft2/sec, and mean acceleration. Max Total Distance has a surprising impact on the model's predictions. This is suggesting that greater distances covered may actually decrease the likelihood of injury. We believe that this insight indicates that players conditioned to endure more extensive physical activity might be less prone to injuries, Similarly, longer durations spent on offensive plays also appear to reduce the risk of injury which can mean that defensive plays cause more collisions for your players which can impact the likelihood of injuries. This insight means that offensive gameplay, which has controlled strategic movements, causes fewer risks than more chaotic defensive play. Acceleration metrics such as Acc Low 7 and Mean Acceleration show that lower rates of acceleration are associated with reduced injury risks. This supports the notion that less explosive and sudden movements are generally safer for athletes. Such a trend might reflect a strategic approach to gameplay and training against these sudden movements where steady and controlled movements are emphasized over quick movements that have higher potential to cause bodily harm. This understanding could influence training and gameplay strategies, promoting techniques that prioritize sustainability over short-term performance bursts. This can potentially enhancing player's career length and reducing injury incidents.

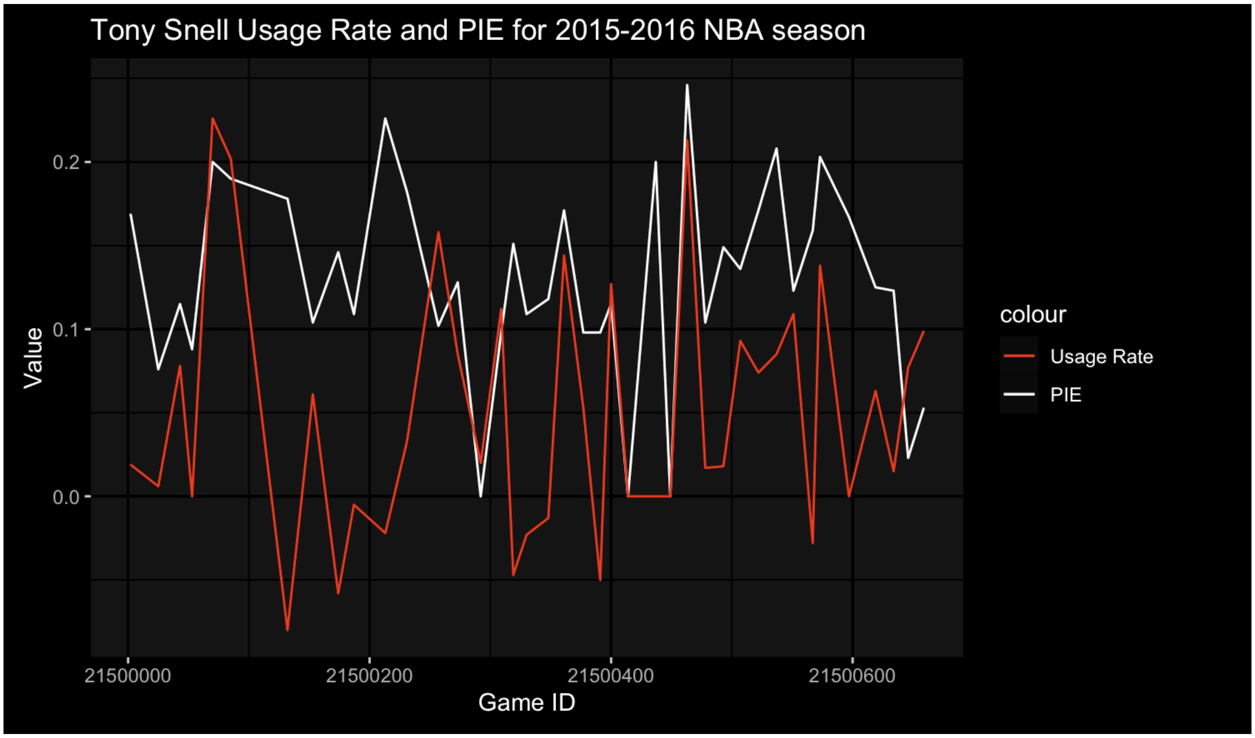
Actions



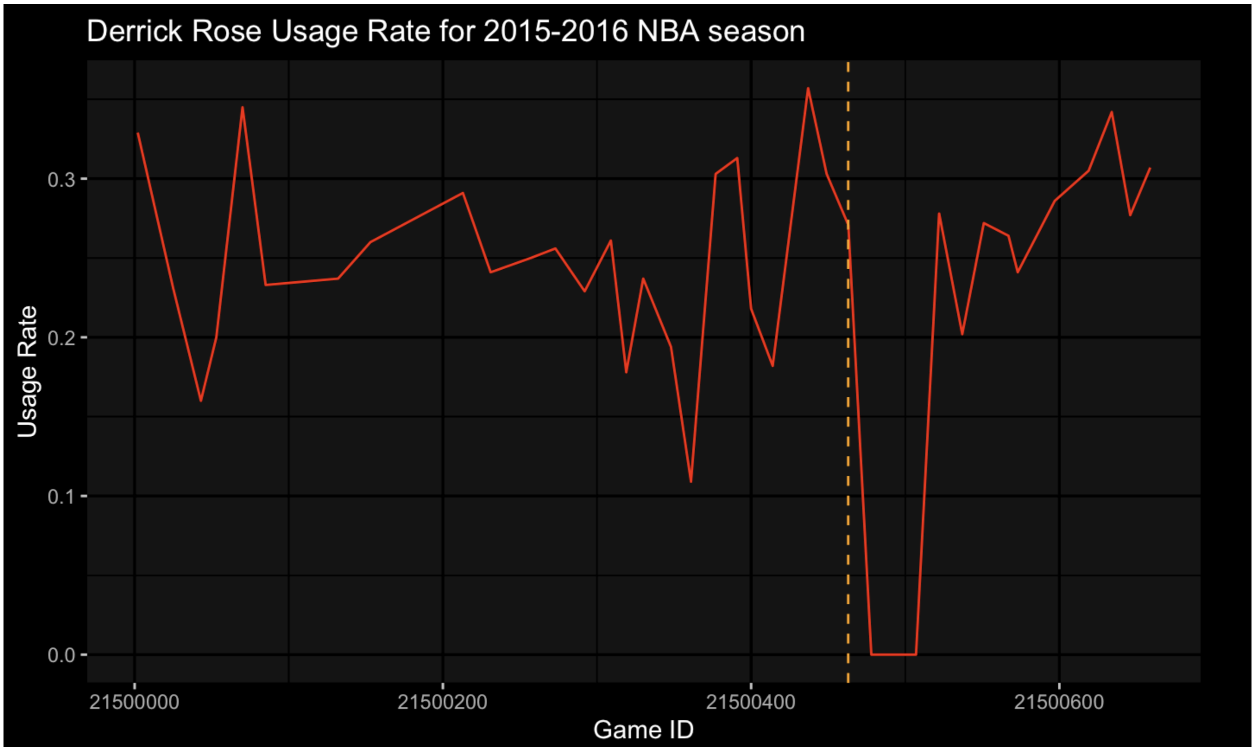
Joakim Noah's usage rate was pretty stable and was in a relatively uniform range of values for the first half of the data points. However, shortly after the halfway point, he had a huge spike in usage rate, which was almost immediately followed by an injury. This is just one case study, but it is interesting to see that an injury came right after a significant change from his usual usage rate range.



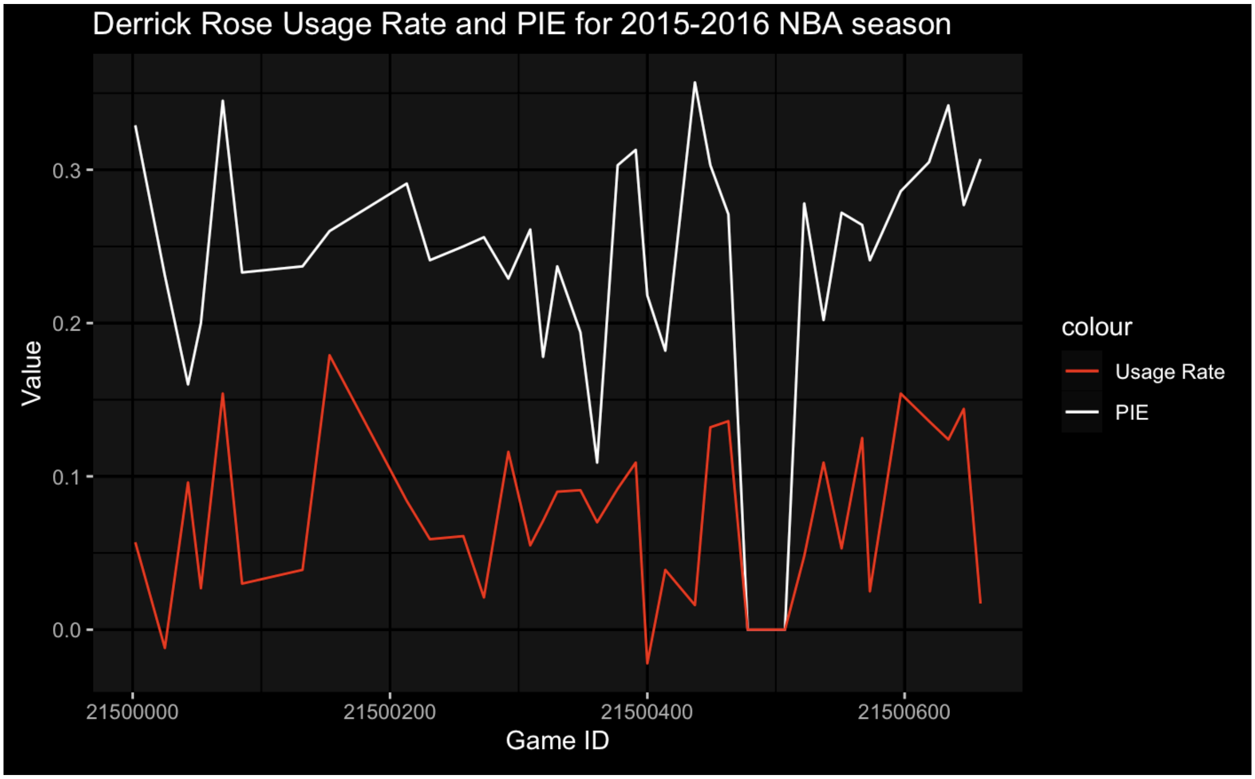
The graphic shows that Joakim Noah's usage rate and PIE are usually peaking at different times. In fact, there seems to be a lag between the two variables. One would typically expect both of these stats to peak and valley together, as PIE has all of the same variables going into its formula, but with defensive metrics added on as well. This suggests that his defensive impact does not necessarily happen at the same time as his offensive impact. Since Joakim Noah was an exceptional defender, he may exhibit different tendencies in these two metrics from offense minded players.



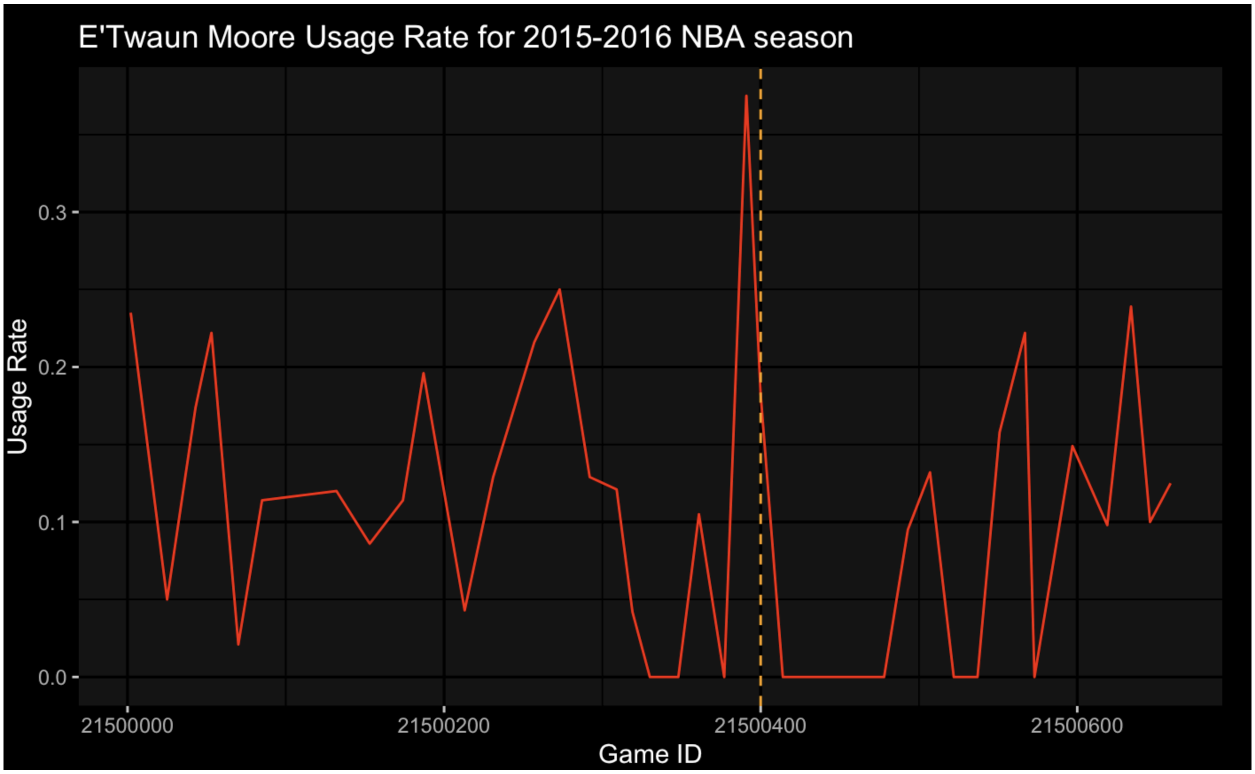
Tony Snell's usage rate and PIE diagram shows more of what a typical player's metrics would look like. For the most part, he had his peaks and valleys together in these two metrics.



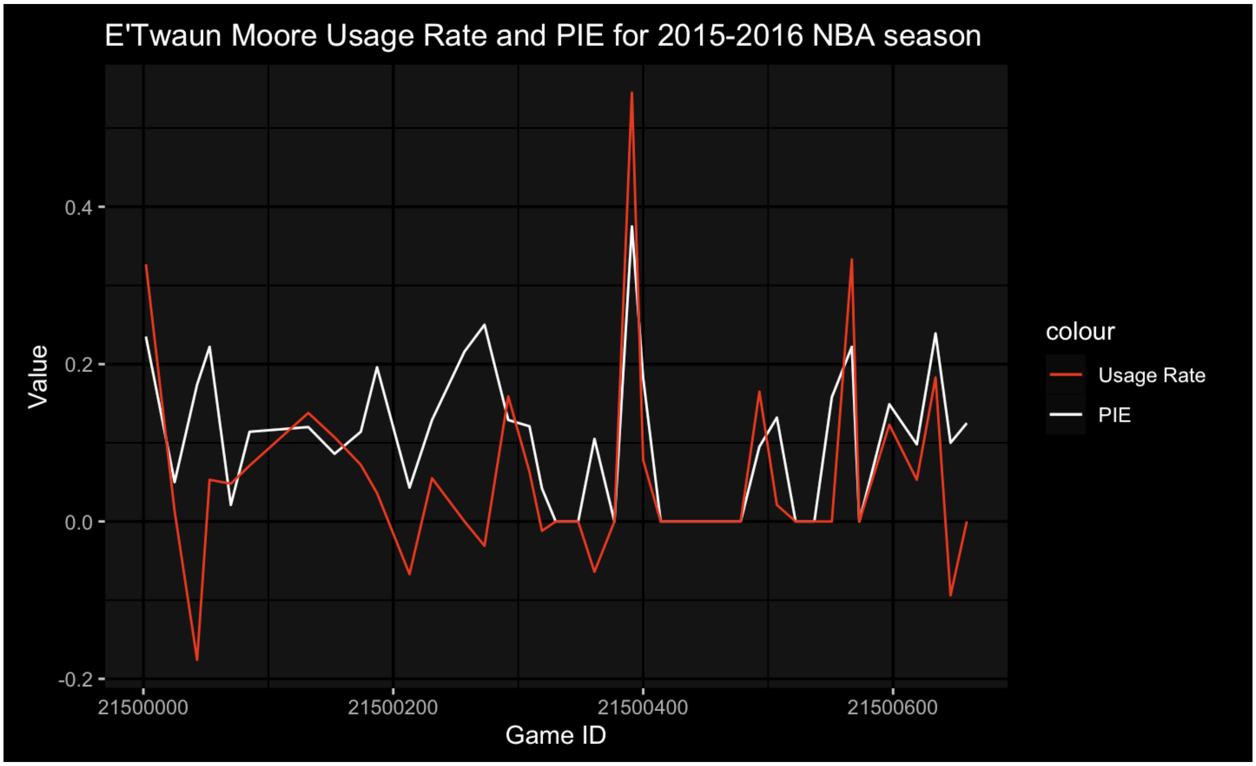
Derrick Rose's usage rate was relatively high throughout the season, as one would expect because he was the main cog in the Bulls offense this season. However, just like with Joakim Noah's injury, Rose got hurt shortly after his highest peak in usage rate as well. Two of the team's biggest stars and important players suffered injuries shortly after their moment of highest usage on offense during the season.



Derrick Rose's usage rate and PIE were closely correlated throughout most of this sample of games. This suggests that his main role was offensive, as most of his production overall peaked at the same time as his offensive metrics. This further illustrates the concept that



E’Twaun Moore’s usage rate was in the range between about 0.03 and 0.25 for most of the first half of the data points, and then sees a huge surge that goes to almost 0.40. Shortly after this outlier of a peak, Moore suffered an injury. The same pattern can be seen in a lot of these graphs and usage rate seems to be a more and more telling predictor of injury.



E’Twaun Moore’s usage rate and PIE, for the most part, were in sync and peaked together at the same time. This concurs with all of the other players’ usage rate vs PIE graphs, besides Joakim Noah, which suggests that usage rate and PIE are heavily positively correlated in most cases. Exceptions like Joakim Noah should be taken with a grain of salt and have their play style factor into the correlation between the two variables.

Several recommendations can be made from looking at these basic line graphs. The biggest thing that stands out in all of the usage rate graphs is that injuries typically follow abnormally high peaks in usage rate. This information can be super helpful in implementing effective load management amongst NBA players. More balanced offensive attacks, especially over the course of the regular season, can not only help overall team performance and development, but it can also help cut down on player’s usage rate. This will, in turn, also lower the player's individual cumulative loads. On the flip side, if a player is asked to do more in a close game or against a higher quality opponent, maybe coaches should give them less minutes in the next game, especially if it is the second game of a back to back. While PIE seems to be a good overall metric to use, it is not as good as usage rate, as PIE factors in defensive performance as well, so primarily defensive players like Joakim Noah can throw this metric off.

Integration of Model Insights into Player Management Systems:

Variables like “max total distance” and “time on offense” can be pinpointed as effective variables to look at when trying to come up with a gameplan. Having players run more in practice and in a controlled environment can help them to prevent shock load fatigue. Gameplans can be adjusted to possess the ball for more of the shot clock rather than take the first shot available to better prevent fastbreak opportunities and injuries. On the other hand, if players have a tendency to struggle with higher loads or with more usage on offense, they should be dealt with accordingly and on a player-by-player basis. Practices can be designed to limit the amount of explosive or high change in acceleration plays that happen to prevent the amount of injuries that occur in practice. Better warmups and a gradual ease into higher levels of acceleration can ensure proper game-like practice, while also not jeopardizing the health of the players.

Real-Time Monitoring and Adjustments:

Incorporating some sort of monitoring tool or device during practices and games can help cut down on injuries. Ideally, these devices would alert the coaching and training staff when high-risk thresholds are being neared or surpassed. Wearable technology in the form of a watch that could track all of these metrics would be a great way to optimize player surveillance and wellness. Coaches can use the metrics provided from these devices in practice and in previous games to devise a game plan for an upcoming game that would seek to minimize injury concerns for at-risk players.

Collaborative Workshops and Training:

All these analytical models are great and the insights we found are valuable, but unless the coaching staff and the players buy into it, the analytics are basically useless. Having workshops that explain our predictive models and how prescriptive analytics can be used based on these insights to come up with plans that help mitigate injury risk can be a great way to bridge the gap in analytical knowledge and get the team on board. The workshops can help explain the complicated models and SHAP plots included in our project that may go over the head of an analytics layman. This would ultimately help everyone to realize what the problem areas are and why they need to be targeted for their own good.

Conclusion

The findings of this project were that max distance was a strong deterrent of injury. This can be explained by the fact that fitter players likely play more minutes, and they are more accustomed to the load of playing a lot. This goes hand in hand with distance covered, which makes sense with the results of our xG boost model. Time on offense, in a single game, seems to be linked with lower injury rates. However, usage rate was highly positively correlated with injury in the box-score data, which suggests being on offense is linked to a higher injury rate. One rationalization of these two juxtapositions is that in an individual game, you are more likely to get hurt on defense, but over the course of a season, carrying a team's offensive load to a level that a player is not generally accustomed to can be a big predictor in chronic injuries. According to the SHAP plot, mean acceleration was also linked to lower injury rates. This can be attributed to a player being used to a certain pattern of acceleration being less prone to injury because they are used to that level of acceleration. The problem with acceleration is high amounts of changes in acceleration that puts too much wear and tear on lower extremity joints like knees and ankles. Within a game, coaches should try to avoid putting players in situations where they have to carry an offense if they are not used to it. Coaches should gradually increase the cumulative load of a specific player to best avoid the risk of injuries. From game to game, coaches should monitor a player's usage rate in previous games and see if there have been any anomalies in recent games. If that is the case, a coach should consider resting a player on the second leg of a back to back game or at least restricting the number of minutes that player plays in the second game. If a player has seen a relatively consistent usage rate in recent games, they are likely to proceed as normal and not have issues. While our group sought out to seek an alternative to load management altogether, it seems like preventative measures need to be taken to avoid players completely sitting out games. There is no perfect solution for players, coaches, and the fans, but using the metrics we found significantly important in our project can be the future in making "load management" better and more effective for everyone involved.

Future Work

In the future, we would like to look at several different variables that can affect injury, the first being integrating play types into our model. A player who takes more catch and shoot opportunities rather than driving to the basketball receives a lot less contact that can lead to injuries. The same can be said for a point guard versus a center. Center's have a higher amount of paint touches where there are more collisions, start and stop opportunities, and jumps. All of these factors can increase the chance of injury. We would also like to look at the body orientation of players. More often than not, offensive players will be facing the basket, while defenders face the ball handler with their back to the basketball. There could be more injuries when defenders are going backwards and trying to cut different ways to guard the ball. This would be very interesting to explore, and is an area within basketball analytics that does not have much prior research. Next, future studies could evaluate the impact of the latest advancements in wearable technology on tracking player workload and predicting injury. This could include devices that measure more detailed physiological responses such as heart rate variability and sleep patterns to optimize recovery and performance. Another area further research would be valuable is integrating biomechanical data such as joint angles, muscle activation patterns, and force exertion to gain deeper insights into the specific movements that lead to injuries. This type of data could help in developing highly personalized training and recovery programs that address the biomechanical weaknesses of individual players. Additionally, we might be able to use heatmaps describing player location to create insights from player location tendencies. For instance, if a player shows less movement but stays in high contact areas of the court, such as under the key, we might be able to determine the likelihood of contact injuries from hard fouls or driving to the hoop. This could also be used for positions with less contact, like guards, who may have more non-contact injuries due to a higher rate of dynamic movement. Finally, investigating the effectiveness of various recovery modalities (e.g., cryotherapy, compression garments, active recovery) in conjunction with load management could offer valuable insights. This research could determine which modalities best reduce injury risk and enhance recovery when combined with strategic workload management. Additionally, our next steps involve applying more of a weight to a player's athleticism by converting their physical metrics into percentiles of their own movements because different players have more athletic bodies and have different thresholds for affecting workloads. For example, a player that has a higher velocity average doesn't always have to increase their injury risk.

Contributions

George Cole:

Contributed to problem identification and project planning.

Assisted in the initial investigation of the dataset.

Played a role in data preprocessing and transformation.

Participated in model selection.

Contributed to the analysis of the results.

Jack Crilly:

Contributed to problem identification and project planning.

Conducted a review of previous work in the project area.

Played a significant role in data collection and dataset selection.

Assisted in data preprocessing and cleaning.

Participated in model selection and evaluation.

Reuben Dayal:

Actively participated in problem framing and creating the project plan.

Assisted in creating visualizations to support the project's conclusions.

Cleaned the box-score dataset and prepared it for use with the other datasets

Played a role in the final report preparation.

Brought knowledge of basketball to help make sense of data

Brian Papiernik:

Contributed to problem identification and project planning.

Helped in data collection and dataset selection.

Contributed to model selection and hyperparameter tuning.

Played a role in the analysis of the results.

Assisted in the creation of visualizations for the project.

Contributed to the project presentation.

David Sobek:

Participated in problem framing and creating the project plan.

Took a lead role in data preprocessing and transformation.

Played a key role in model selection and evaluation.

Helped in the preparation of the final report.

Helped preprocessing and transformation.

● Bibliography (Does not count towards limit) - List of sources used for the project.

● R-code (Does not count towards limit) - This should be provided separately as an r-script and should allow the findings and graphs in your project report to be recreated.