**EE 534: BIG DATA**

**CHELSEA GORIUS**

**GAVIN HEMMERLEIN**

**CLASS 11530**

**FALL 2020**

**NBA PERFORMANCE AND INJURY**

1. Team

Our team will consist of Chelsea Gorius (cgorius) and Gavin Hemmerlein (ghemmer). Both members are students in the ENGR E534 course. Chelsea and Gavin are also IU Masters students pursuing a degree in Data Science.

1. Topic

The topic to be investigated is basketball player performance as it relates to injury. The topic of injury and recovery is a multi-billion dollar industry. The Sports Medicine field is expected to reach $7.2 billion dollars by 2025 [1]. The scope of this effort is to explore National Basketball Association(NBA) teams, but the additional uses of a topic such as this could expand into other realms such as the National Football League, Major League Baseball, the Olympic Committees, and many other avenues. For leagues with salaries, projecting an expected return on the investment can assist in contract negotiations and cater expectations.

1. Dataset

To compare performance and injury, a minimum of two datasets will be needed. The first is a dataset of injuries for players [2]. This dataset will create the samples necessary for review.

Once the controls for injuries are established, the next requirement will be to establish pre-injury performance parameters and post-injury parameters. These areas will be where the feature engineering will take place. The datasets needed must dive into appropriate basketball performance stats to establish a metric to encompass a player’s performance. One example that ESPN has tried in the past is the Player Efficiency Rating (PER). To accomplish this, it will be important to review player performance within games such as in the “NBA games data” [3] dataset. There is a potential to pull more data from other datasets such as the “NBA Enhanced Box Score and Standings (2012 - 2018)” [4]. It is important to use the in depth data from the “NBA games data” [2] dataset because of how it will allow us to see how the player was performing throughout the season, and not just their average stats across the year. With in depth information about each game of the season, and not just the teams and players aggregated stats, added to the data provided from the injury dataset [2] we will be able to compose new metrics to understand how these injuries are actually affecting the players performance.

Along the way we look forward to discovering if there is also a causal relationship to the severity of some of the injuries, based on how the player was performing just before the injury. The term “load management” has become popular in recent years to describe players taking rest periodically throughout the season in order to prevent injury from overplaying. This new practice has received both support for the player safety it provides and also criticism around players taking too much time off. Of course not all injuries are entirely based on the recent strain under the players body, but a better understanding about how that affects the injury as a whole could give better insight into avoiding more injuries. It is important to remember though that any pattern identification would not lead to an elimination of all injuries, any contact sport will continue to have injuries, especially one as high impact as the NBA. There is value to learn from why some players are able to return from certain injuries more quickly and why some return to almost equivalent or better playing performance than before the injury. This comparison of performance will be made by deriving metrics based on varying ranges of games immediately leading up to injury and then immediately after returning from injury. In addition to that we will perform comparisons to the players known peak performance to better understand how the injury affected them. Another factor it will be important to include is the length of time recovering from the injury. Different players take differing amounts of time off, sometimes even with similar injuries. Something will be said about the player’s dedication to recovery and determination to remain at peak performance, even through injury, when looking at how severe their injury was, how much time was taken for recovery, and how they performed upon returning.

These datasets were chosen because they allow for a review of individual game performance, for each team, throughout each season in the recent decade. Aggregate statistics such as points per game (ppg) can be deceptive because duration of the metric is such a large period of time. The large sample of 82 games can lead to a perception issue when reviewing the data. These datasets include more variables to help us determine effects to player injury, such as minutes per game (mpg) to understand how strenuous the pre-injury performance or how fatigue may have played a factor in the injury. Understanding more of the variables such as fouls given or drawn can help determine if the player or other team seemed to be the primary aggressor before any injury.

1. Objective

The objective of this project is to develop performance indicators for injured players returning to basketball in the NBA. It is unreasonable to expect a player to return to the same level of play post injury immediately upon starting back up after recovery. It often takes a player months if not years to return to the same level of play as pre-injury, especially considering the severity of the injuries. In order to successfully analyse this information from the datasets, a predictive model will need to be created using a large set of the data to train.

From this point, a test run will be used to gauge the validity and accuracy of the model compared to some of the data set aside. The model created will be able to provide feature importance to give a better understanding of which specific features are the most crucial when it comes to determining how bad the effects of an injury may or may not be on player performance. Feature engineering will be performed prior to training the model in order to improve the chances of higher accuracy from the predictions. This model could be used to keep an eye out for how a player's performance intensity and the engineered features could affect how long a player takes to recover from injury, if there are any warning signs prior to an injury, and even how well they perform when returning.

Sources

1. A. Mehra, *Sports Medicine Market worth $7.2 billion by 2025,* Markets and Markets.

https://www.marketsandmarkets.com/PressReleases/sports-medicine-devices.asp

2. R. Hopkins, *NBA Injuries from 2010-2020,* Kaggle. https://www.kaggle.com/ghopkins/nba-injuries-2010-2018

3. N. Lauga, *NBA games data,* Kaggle. https://www.kaggle.com/nathanlauga/nba-games?select=games\_details.csv

4. P. Rossotti, NBA Enhanced Box Score and Standings (2012 - 2018), Kaggle. https://www.kaggle.com/pablote/nba-enhanced-stats