

COMP9417: Machine Learning Project

z5113817

University of New South Wales — July 29, 2021

Motivation

In the 1984 NBA draft, Sam Bowie was drafted as the number 2 pick to the Trail Blazers. It might be shocking to hear that Bowie was drafted one place *above* the hall of fame superstar Michael Jordan. This was because the Trail Blazers needed a new superstar "big man" to replace the Center they lost the season before. Bowie had an impressive 76 games with the Trail Blazers until a fracture in his left tibia put him out for the season. Even though Bowie followed the recommended recovery time, the rest of Bowie's career was undermined by the recurring injury. In 10 seasons with the NBA, Bowie only appeared in 511 games.

The question is, even though injuries in sport are seen as an unforeseeable tragedy, can a machine learning model be used to eliminate some of the unpredictability and quantify the likelihood that a player will suffer a major injury in the current season?

The Goal

Create a model that assigns a likelihood that a player will *suffer a major injury* given their current performance. Suffering a major injury will be defined as a physical injury that leaves a player on the injury list for more than 34 days.

The Data

Datasets

The datasets were scraped from various sources such as [prosporttransactions](#) and [basketball-reference](#). The scrapers were sourced from [elap733's repository](#) found [here](#).

The following datasets can be found in the **data/raw** directory:

- **player_stats**: Contains every NBA player's basic statistics for a given season.
- **injury_list**: Contains information on when players were acquired on and relinquished from NBA injury list.
- **missed_games**: Contains information on games which players missed (not necessarily due to injury).
- **all_games_schedule**: The schedule for every NBA team.

The data ranges from 2010 to today.

Cleaning

For the most part, the data contains pretty clean data with no missing entries or gibberish values. The datasets were last scraped in 2019 however, so I had to do some additional scraping to obtain the latest data. Luckily, the scrapers still worked just fine.

For cleaning, all that I had to do was append the latest scraped data with the existing datasets. The cleaned datasets can be viewed at **data/cleaned**. The code used to execute this is stored at **scripts/clean_data.py**.

Processing Data

The Injury List

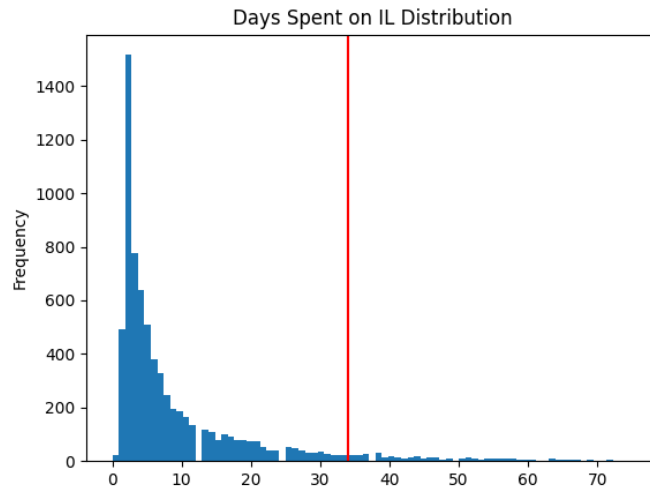
This section explains the reasoning and steps take to produce the **data/processed/physical_injuries_2010_2021.csv** file and what each of the columns mean. The code used to execute this is stored at **scripts/process_injury_list.py**.

The injury list in the NBA isn't quite as the name suggests. Rather than being a list of players who are currently suffering from physical injury, players who miss games for other reasons can be placed on this list. Other reasons include:

- **illness**: 413
- **surgery**: 253
- **COVID-19**: 11
- **personal reasons**: 11

In addition, the list includes players who are *relinquished* from their team (put on the injury list) and players who are *acquired* to their team (removed from the injury list). This means there are essentially 2 entries for each individual injury. We want to move the date of this second entry to its own column so that for each injury we know the dates that player was put on and removed from the injury list.

Earlier I defined a "major injury" as a physical injury that leaves a player on the injury list for more than 34 days. I defined it in this way by looking at the distribution of days spent on the injury list in our dataset.



The red line shows the 80th percentile of this distribution which is at **34** days spent on the injury list¹. I selected this to be the boundary between a minor and major. Even though this was done somewhat arbitrarily (I manually decided to use the 80th percentile), I believe this is a reasonable number since it represents the injuries that left a player on the injury list for just over a month and represents only 20% of all NBA injuries since 2010.

The processed injury list dataset is stored at **data/processed/physical_injuries_2010_2021.csv**. There are **8,552** entries in this dataset and **1,711** are major injuries.

Improving the story

Prior to his career in the NBA, Bowie had developed a stress fracture in his left tibia. Even though he rested and took the recommended amount of time off for this minor injury, this would be the same place that Bowie suffered his first major injury when he joined the NBA.

Now that we have 1,711 examples where a player were acquired on the injury list for longer than 34 days, we want to tell a bigger story with these entries by attaching more data to these injuries. The case of Bowie can inspire us to ask questions such as:

- How was the player performing before this injury?
- What was their average gametime?
- How intense did the player play?
- How intense was their team's schedule?
- Have they suffered an injury prior?

The script **scripts/derive_new_fields.py** attempts to answer some of these questions by joining the existing dataset at **data/cleaned/player_stats_2010_2021.csv** with the processed injury list at **data/processed/physical_injuries_2010_2021.csv**.

¹Some injuries were marked as "player out for season" which means they don't have another entry in the dataset showing their return. For this case, the player was marked down as being on the IL for 100 days. These players are not included in the histogram however do play a role in this percentile calculation

Exploratory Analysis

Correlated fields

Model Selection

Model Evaluation

Conclusion

Nice to haves

Would've been nice to incorporate the intensity of the schedule some more.