COMP9417: Machine Learning Project

z5113817

University of New South Wales — July 25, 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 lossed 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 recommened recovery time, the rest of Bowie's career was undermined by the recurring injury. In 10 seasons with the NBA, Bowie only appears in 511 games.

The question is, even though injuries in sport and in the NBA can seem random, can a machine learning model be used to eliminate some of the unpredictability and determine 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 can be defined as a physical injury that leaves the player missing 15 or more games.

The Data

Datasets

The datasets were scraped from various sources such as prosporttransactions and basketball-refrence. 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**.

Processing Data

The Injury List

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**: TODO: obtain number of entries that satisfy this criteria.
- **COVID-19**
- **personal reasons**: e.g birth of child.

TODO: Make a new dataset from this (injury_list_cleaned)

For this project, I have defined a "major injury" as a physical injury that results in a player missing 15 or more games.

Additionally, the Injury List contains when a player was acquired on the injury list as well as relinquished. TODO:

New Fields

So now we have X entries where players were acquired on the injury list and missed 15 or more games. We want to tell a bigger story with these entries by adding some more data around this.

We want to answer questions such as: How was the player performing before this injury? How intense was their team's schedule? Had they suffered an injury prior to this?

In his sophomore year, Bowie had suffered a stress fracture in his left tibia. Even though he gave this

Exploratory Analysis

Model Selection

Model Evaluation

Conclusion