COMP9417: Machine Learning Project

z5113817

University of New South Wales — August 1, 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 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* in any given season. Suffering a major injury will be defined 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-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**. The code used to execute this is stored at **scripts/-clean_data.py**.

Processing Data

The Injury List

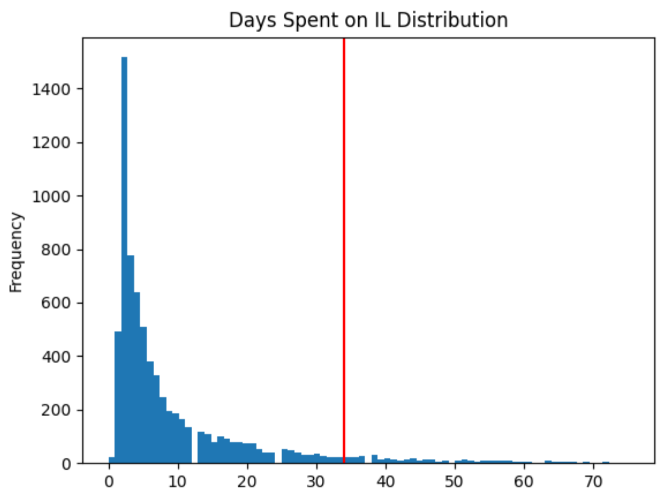
This section explains the reasoning and steps take to produce the **data/processed/physical_injuries_2010_2021.csv** file and what each of the columns mean. The code used to execute this is stored at **scipts/process_injury_list.py**.

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 list [1]. I selected this to be the boundary between a minor and major. Even though this was done somewhat arbitarily (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.

The dataset **data/cleaned/player_stats_2010_2021.csv** contains a general overview of player statistics for a given season. We can perform a *left join* on this dataset with the processed physical injuries dataset so that each injury is now linked to the player's performance during the season that they were injuried.

[1]Some 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

Player	Year	Season	Position	Age	Injury Date	Duration	Notes
Malik Allen	2010	regular	PF	32	2010-10-28	9	placed on IL
Malik Allen	2010	regular	PF	32	2010-11-10	5	placed on IL
Malik Allen	2010	regular	PF	32	2010-11-20	4	placed on IL
Malik Allen	2010	regular	PF	32	2010-11-30	1	placed on IL
Malik Allen	2010	regular	PF	32	2010-12-23	43	placed on IL with sprained left ankle

Table 1: Dataset showing Malik Allen suffering multiple injuries in the 2010 regular season

But what about players who were injured multiple times during the same season? In table 1, we can see that Allen suffered a few injuries in the 2010 season. Each of these rows also contain data about Allen's performance for the 2010 season, which are all identical in their values, since it's all for the 2010 season. Recall that the aim of this model is to predict the likelihood that a player will suffer a major injury in a given season. Feeding Allen's data into a model would confuse the model since it has 4 examples where a player didn't suffer a major injury and 1 example where they did, all with the *exact same* stats for the season.

We want to group these examples together so that the model doesn't have contradicting data, but we also want to keep the information that Allen suffered *minor* injuries prior to his major injury.

Let's create a new field **Recent Minor Injury Count** which for each season for a given player, counts the number of minor injuries they received in that season. A *minor injury* will be defined as an injury that wasn't a major injury.

Let's also create another field **Previous Major Injury Count** which counts the number of *major injuries* that a player has suffered strictly prior to that season. The new dataset is created at **data/processed/aggregated.csv**.

Tokenising

Fields such as the player's position and whether the season is post season or regular are string values. We want to encode these to a numerical value and store their encoding somewhere.

The script **tokenize_data.py** performs these operations and stores the new dataset at **data/processed/-tokenized.csv**.

Normalising and Standardising

This section explains the steps in **normalise.py** which is used to create the dataset at **date/processed/nomarlised.csv**.

Some players play more minutes per game than others. Because of this, some of fields such as *FGA* (Field Goal Attempts per game) will be greatly skewed by the average amount of minutes a player spends in a game. As such, the first step I have taken to normalise is make all of "per game" stats a ratio with the average minutes played [2].

Since the data has a lot of continuous values, it is a good idea to nomralise these between 0 and 1. I don't want fields such as the age of a player to be weighted significantly higher than a player's average offensive rebounds per game, just because the value of a player's age is always much larger.

Also, I will assess various models and I do not know the distribution of the dataset across all of its values. As such, standardising is a good idea.

[2]Note that all normalising and standardising after this point is done *after* training and testing data has been established

Data Selection

I now have a dataset with a lot of meaningful values - almost too many. In the script, **select_data.py** I clean out datapoints that I think will mislead the model (for example, the player's name) or are correlated and such, provide no further information (for example, Field Goals per game with Field Goal Attempts per game). This results in a final processed dataset at **data/processed/final.csv** which contains the data that I will be using to train various models on.

An explanation on all of the columns contained in this dataset and what they each mean can be found in appendix .

Exploratory Analysis

In previous sections of this report I have explored the raw data and used my findings to derive new fields and select the final data. This section will focus on exploring the final dataset to reason about which models can be used. The script **exploratory_analysis.py** was used to produce this information.

There are **7,820** examples in the dataset with **22** features. Of which, **1,094** are examples with major injuries. Looking at the correlations between features of the final dataset 1, we can see that there are no major correlations between our dependent variable (Major Injury) and the rest of the dataset.

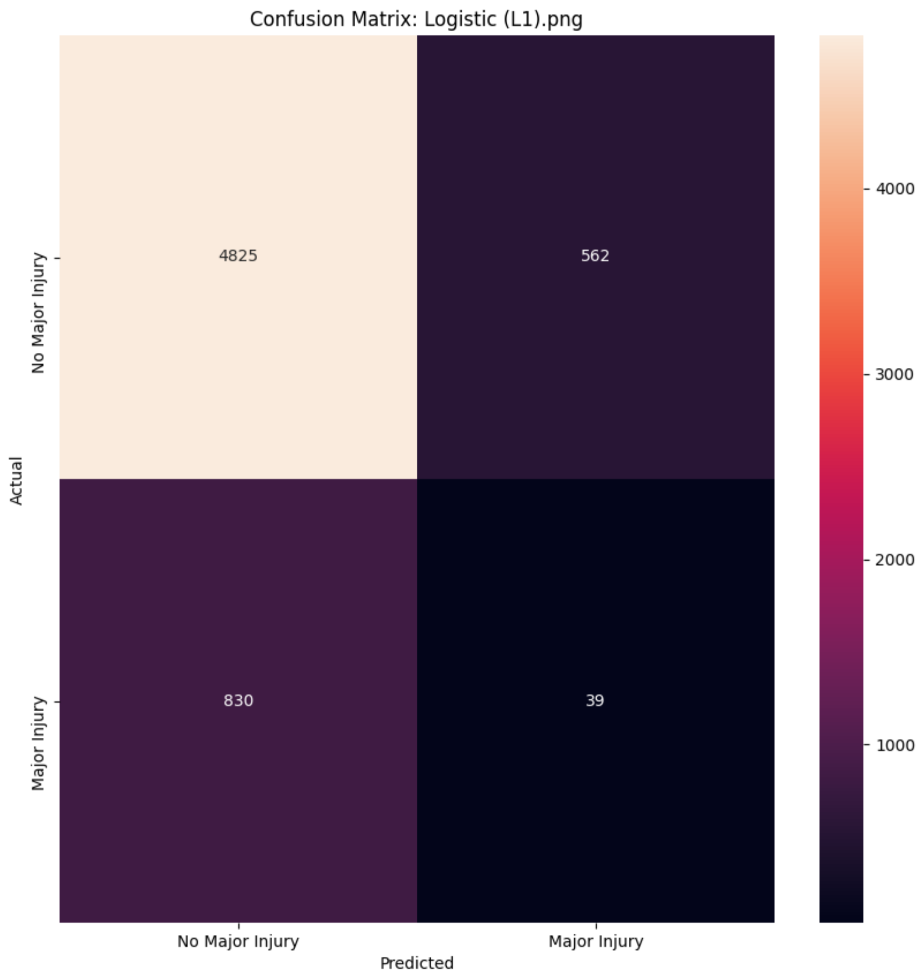
Model Selection

The **model_selection.py** script contains the code that was used to overview a selection of models. The models were overviewed by training against 80% of the dataset and then tested against the remaining 20%. An *accuracy score* and *logloss* was calculated for each of the models to yield the following results:

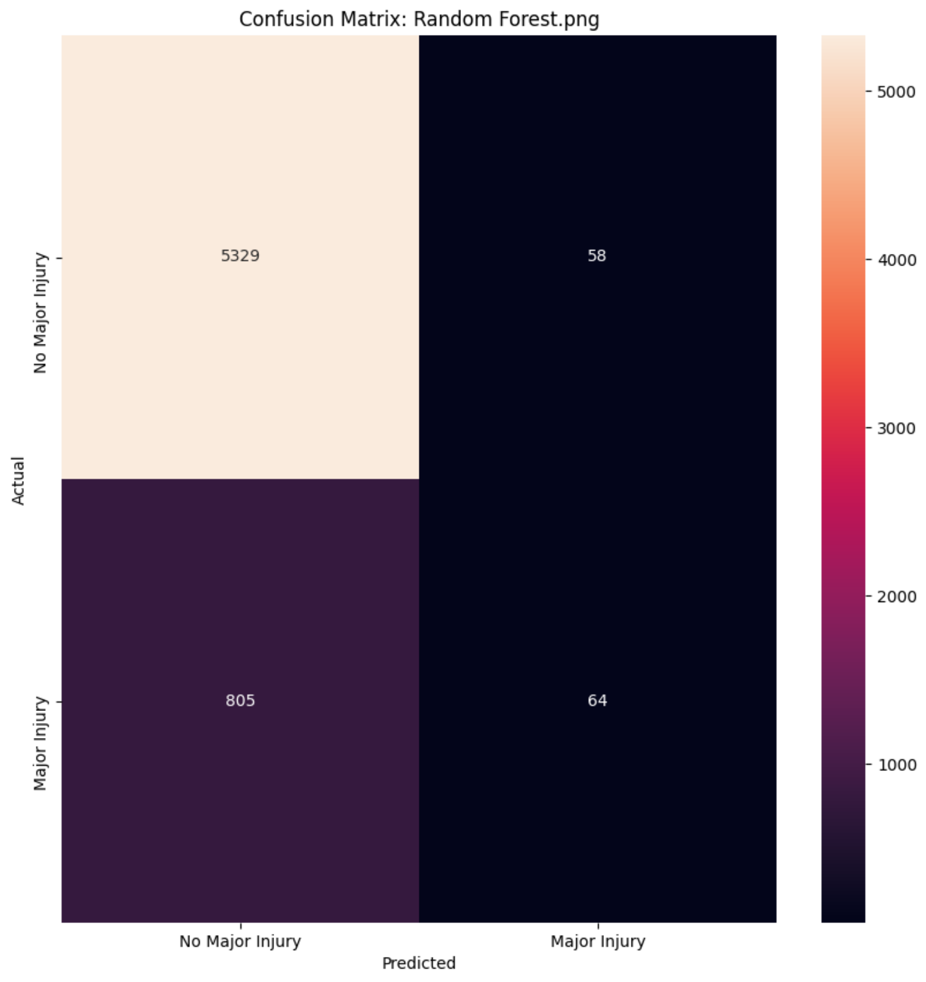
- Logistic (L1): 0.8566176470588235 accuracy and 7.685170414877148 log-loss
- Logistic (L2 Saga Solver): 0.8562979539641944 accuracy and 7.7403808623708885 log-loss
- Linear SVC (linear): 0.8610933503836317 accuracy and 4.79766571086595 log-loss
- Linear SVC (gamma): 0.8607736572890026 accuracy and 4.79766571086595 log-loss
- Naive Bayes (Multinomial): 0.8610933503836317 accuracy and 4.78662697006488 log-loss
- Random Forest: 0.8639705882352942 accuracy and 4.698297616097271 log-loss

Immediately, the accuracies (ratio of correct predictions to incorrect) are all around 85% which seems promising. Considering, however, that that approximately 86% of the dataset is filled with examples of players who weren't majorly injured during the season, you could expect a dumb model that guesses "no major injury" for every example to have similar accuracy. A confusion matrix tells a better story about each model.

In these confusion matricies, the top left and bottom right squares are the number of correct predictions broken down by classification. The top right is the number of false positives and the bottom left is the number of false negatives.



Looking at the confusion matrix for the Logistic (L1) model, we can see that the model missed 830 major injury predictions and of the 601 major injury predictions that it made, only 39 were correct.



Similarly with the Random Forest, the model missed 805 major injuries however it was much more accurate when it did make a prediction.

The SVM models didn't make any predictions for a major injury and the Bayesian models had similar results to the Random Forest, except it was more reserved with its positive predictions.

From this overview, the Random Forest model performs the best, not only in having less incorrect prediction (albeit marginally) to the other models, but there is also a higher level of accuracy when it does predict a major injury.

Model Evaluation

Conclusion

Nice to haves

Would've been nice to incorperate the intensity of the schedule some more. Do some NLP analysis on the Notes of the injury.

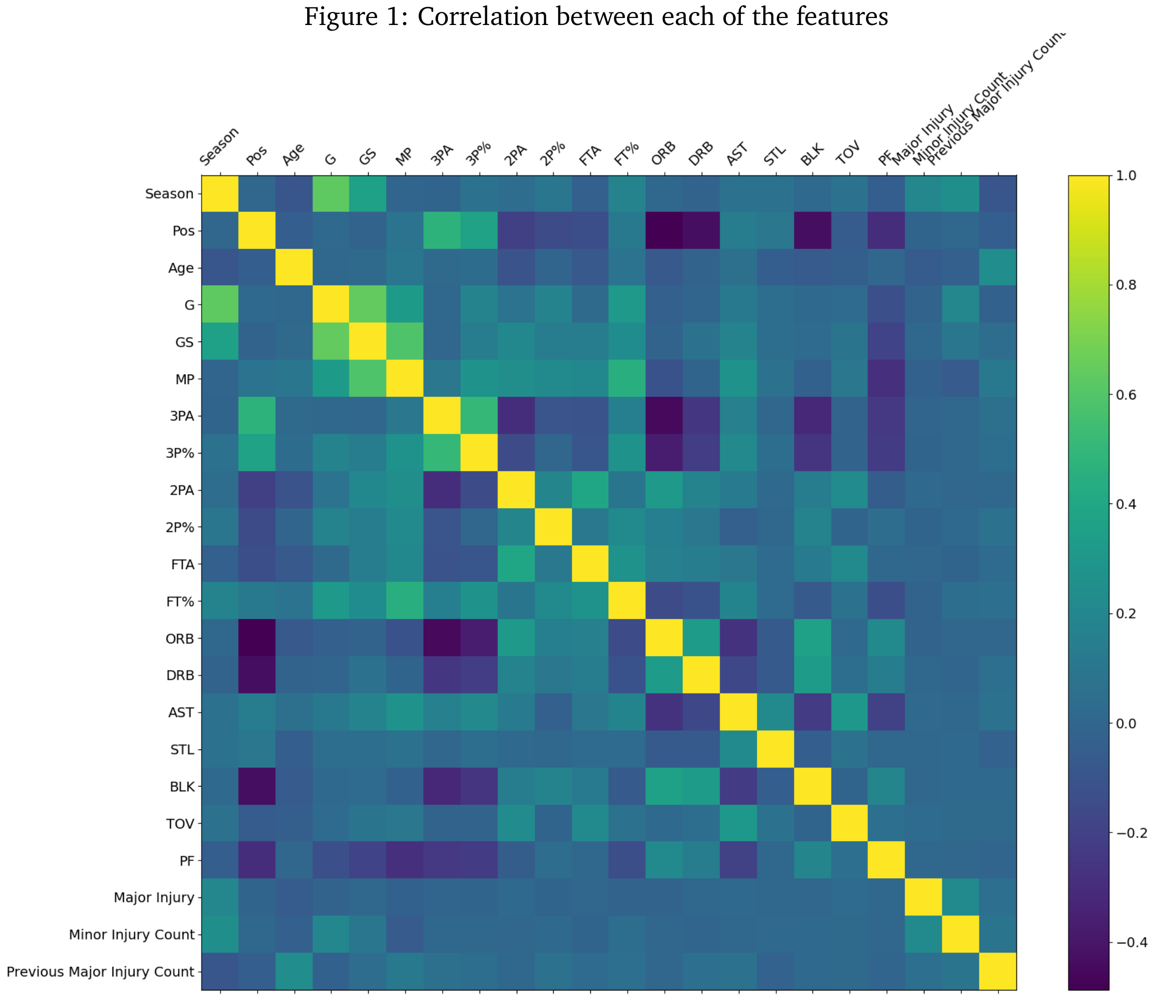
Appendix

Final Dataset

Each row represents a player's statistics during a given season. Each row has the following columns:

- **Season**: whether the statistics are from regular (1) or post (0) season
- **Pos**: the position of the player
- **Age**: the age of the player
- **G**: the number of games played
- **GS**: the number of games started
- **MP**: the average number of minutes played per game
- **3PA**: the average number of 3-point attempts per game per minute
- **3P%**: the percentage of 3-point attemps that were successful
- **2PA**: the number of 2-point attempts per game per minute
- **2P%**: the percentage of 2-point attemps that were successful
- **FTA**: the number of Free Throw attempts per game per minute
- **FT%**: the percentage of Free Throw attempts that were successful
- **ORB**: the number of Offensive Rebounds per game per minute
- **DRB**: the number of Defensive Rebounds per game per minute
- **AST**: the number of assists per game per minute
- **STL**: the number of steals per game per minute
- **BLK**: the number of blocks per game per minute
- **TOV**: the number of turn-overs conceeded per game per minute
- **PF**: the number of personal fouls against the player per game per minute
- **Minor Injury Count**: the number of minor injuries suffered by the player during that season
- **Previous Major Injury Count**: the number of major injuries suffered by the player prior to that season
- **Major Injury**: whether the player suffered a major injury (1) during that season or not (2)

Figure 1: Correlation between each of the features


Season
Pos
Age
G
GS
MP
3PA
3P%
2PA
2P%
FTA
FT%
ORB
DRB
AST
STL
BLK
TOV
PF
Major Injury
Minor Injury Count
Previous Major Injury Count