Jonathan Greenberg

Milestones #1 and #2 Report

3/15/2020

1. Gain complete understanding of the business challenge. Iteratively refine project plan.

I originally was considering further researching the “hot hand” in basketball, and how physical attributes play a factor in an NBA players’ success. Two topics that I would love to return to at some point in my life. However, I have decided to research a topic that has recently gained much media attention and has not been highly researched to date:

Topic I have chosen:

NBA Injuries: An analysis on resting players, and the injuries we save them from.

Questions I am looking to answer:

Does load management/resting players work to keep them from getting injured?

What are the different injury types in the NBA? How can I best visualize this?

What causes what types of injuries?

When is it best to rest a player in the NBA?

Can I predict when injuries will occur to certain players?

These different questions will lend themselves to different models. I can have a classification model for the prediction of an injury. I can have a regression model for the injury risk on any given night. I can have a decision tree for the types of injuries, and I can have various multidimensional visualizations for the details of a particular injury.

In order to answer these questions I will be creating more columns based on the ‘notes’ column which contains details of each injury. Examples are the type of injury (soreness, broken, fractured), the body part that was injured (hamstring, finger, foot), and the severity of the injury (DTD, Indefinitely). How to split this column up I am still working on. There will be a lot of diving into the data myself and determining the semantics on my own of what goes into what column. My own domain knowledge of the NBA will have to come into play here.

In order to answer a few of these prediction questions I will really need to acquire more data that costs $260 and I would like to send out an email as soon as possible to whomever can help facilitate this acquisition.

2. Data acquisition & understanding

I am still looking to acquire more data, and I need to email the provost about acquiring this data.

There are two datasets that I would like to purchase for my capstone project:

1. Historical team schedule: this dataset will provide for me the amount of rest days each team has had in the last 10 years. This will give me extra features such as, number of back to back games prior to injury, or number of rest days prior to injury, and will help in predicting future injuries.
2. Historical player data: This dataset will provide for me every single game statistic for each player, so I will be able to extract all the statistics from a player’s game leading up to the date of their injury. The amount of minutes played, how well they shot the ball, and other factors may contribute to predicting their injury.

The data that I do have is from kaggle, and here is the link:

<https://www.kaggle.com/ghopkins/nba-injuries-2010-2018>

It has all of the injury logs of NBA players from 2010 to 2020

There were originally 5 columns, and here is the [dataframe.info](http://dataframe.info)() of the dataframe after a little bit of preprocessing.

Int64Index: 9186 entries, 0 to 11249

Data columns (total 6 columns):

Date 9186 non-null object

Team 9184 non-null object

Notes 9186 non-null object

Player 9186 non-null object

dateTime 9186 non-null datetime64[ns]

Year 9186 non-null int64

dtypes: datetime64[ns](1), int64(1), object(4)

This shows there are 9,186 rows, and only 2 rows have a null value in the Team column.

Here is the link to the Colab notebook that I have to date:

https://colab.research.google.com/drive/1InZ1b1YDH97RYYncTBs5TKvWcNoX9PPr