CAP5768 - Introduction to Data Science

NBA Data Analysis Report

Final Project Report

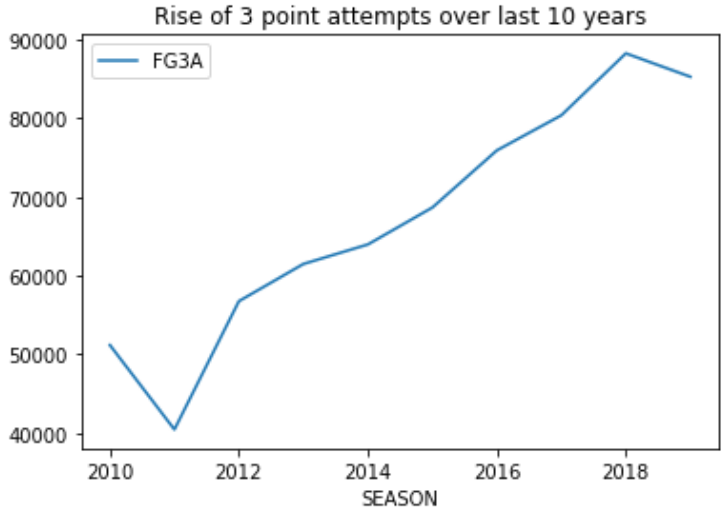
Andres Canaveral
Nestor Maysonet
Mohammed Aquib Khan
Charles Wardlow

Abstract

In this report, we will use data science to both examine coaching philosophies and predict injuries in the NBA. Our goal is to find possible common factors behind injury occurrences and create models that may help reduce these injuries in the future. The datasets used in the report contain about 600,000 rows of information about games, player statistics, and injury details. This project relied heavily on preprocessing the datasets so the machine learning algorithms could create an accurate classifier.

Project Definition

The NBA has embraced data analytics in recent years. Beginning with the three-point revolution, many teams are asking their players to shoot more and more three-point shots, even players in center and power forward positions are moving out to the three-point line during offensive possessions. This increase in 3-pointers can be seen in the graph below.



Coaches are being asked to rest their players more often to keep them healthy for the postseason, and teams are signing players with high efficiency numbers instead of counting stats such as points per game.

In this project we are hoping to explore possible uses of data science in the NBA, from how we can predict injuries with the data that is readily available in official NBA sites, as well as make other observations to explain how injuries happen in the NBA and how we can prevent them.

Data Sources Used

We have used three datasets namely Games, Games_Details and Injuries which we obtain from the following Kaggle links.

Games Dataset:

This dataset incorporates information about games such as the date the game took place, points scored by team, which team won, field goal percentage by team, three pointers by team, rebounds by team, assists by team, and free throws by team. This data set has a total count of 24678 records.

Below is the link to the dataset:

Lauga, N. (2021, November 18). *NBA games data*. Kaggle. Retrieved December 7, 2021, from https://www.kaggle.com/nathanlauga/nba-games?select=games.csv.

Games_Details:

This dataset contains more specific information regarding each game played broken down by player. The main features in this dataset include: points scored,, field goal percentage, three pointers attempted and made, rebounds, assists, turnovers, personal fouls, and free throws. This data set has a total count of about 600,000 records.

Below is the link to the dataset:

Lauga, N. (2021, November 18). *NBA games data*. Kaggle. Retrieved December 7, 2021, from https://www.kaggle.com/nathanlauga/nba-games?select=games_details.csv.

Injuries:

This dataset contains information on injury reports filed in the NBA from the 2010 season up to and including the 2019 season. Injury reports contain the player that was involved, the date the injury report was filed and a thorough description of the location of the injury and how long they are expected to be out.

Below is the link to the dataset:

Hopkins, R. (2020, October 13). *NBA injuries from 2010-2020*. Kaggle. Retrieved December 7, 2021, from https://www.kaggle.com/ghopkins/nba-injuries-2010-2018.

Methodology

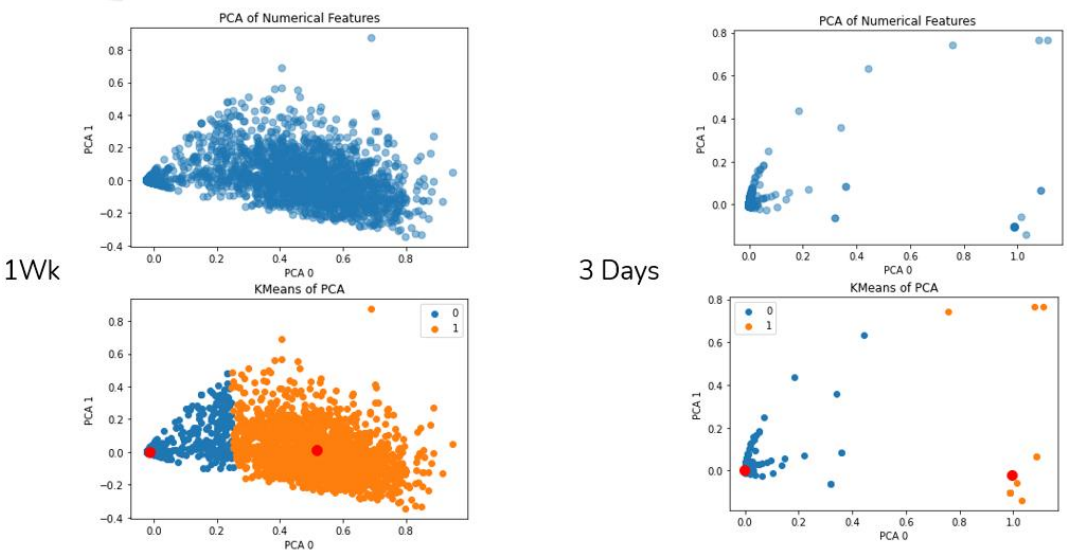
For our project we used tools such as Jupyter Notebook and Python programming language, leveraging various libraries such as NumPy and Panda to analyze our datasets, Matplotlib and Seaborn for visualization, and Scikit-Learn for machine learning.

Pre-processing

With the tools and libraries listed above we started off by merging the game_details.csv and games.csv into a single DataFrame and then merged this dataframe with the injuries dataset.

At first we tried to create a severity value for injuries which would take into account how long players would be out for. We did this by reading the string of the injury report for key words such as "out for season", "out indefinitely" or "DNP". This proved pointless however because the machine learning algorithms worked better with a simple binary class label, injured or not injured. Having said that, the ability to label the severity is there if the opportunity makes sense for the model. It is possible that with help from medical experts and a better labeling scheme, injuries types could be identified and related to the available data models; however, without a successful binary classifier for injury occurrences it seems unlikely to be able to develop one for injury severity.

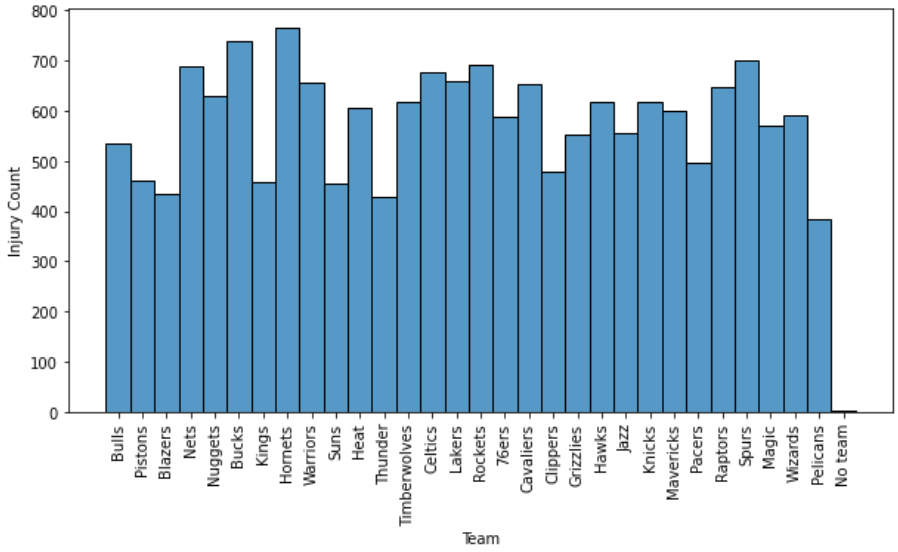
Merging between the DataFrame of games and games_details was easy because we could merge on the basis of game_id which was unique for every game from 2003-2020. However, merging this dataset with the injury dataset presented many issues. Since the date an injury is reported could be any time between the time it occurred and the next game that was played, we needed a way to attach these injuries to a game. Our solution was to use a merge_asof on the basis of date. This means injuries were attached to the appropriate player and game if the game took place within a certain time-frame of the injury report. After considering several time-frames, we ended up settling on 3-days due to there being some separation in the data when performing PCA as seen below.



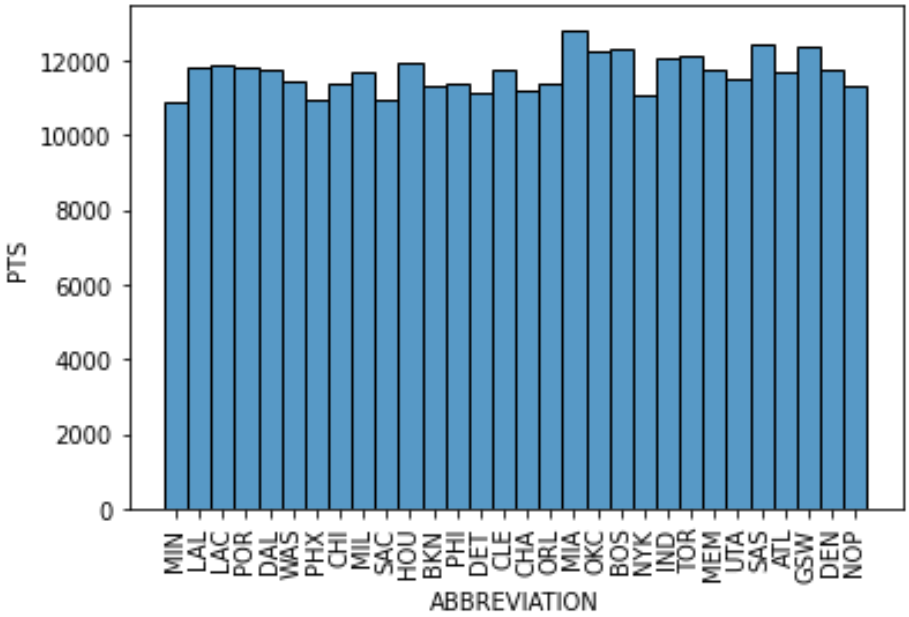
We performed and checked whether all the data points are normally distributed, however because many of the players in our dataset may play 0 minutes and put up 0 stats this didn't tell us much.

Analysis

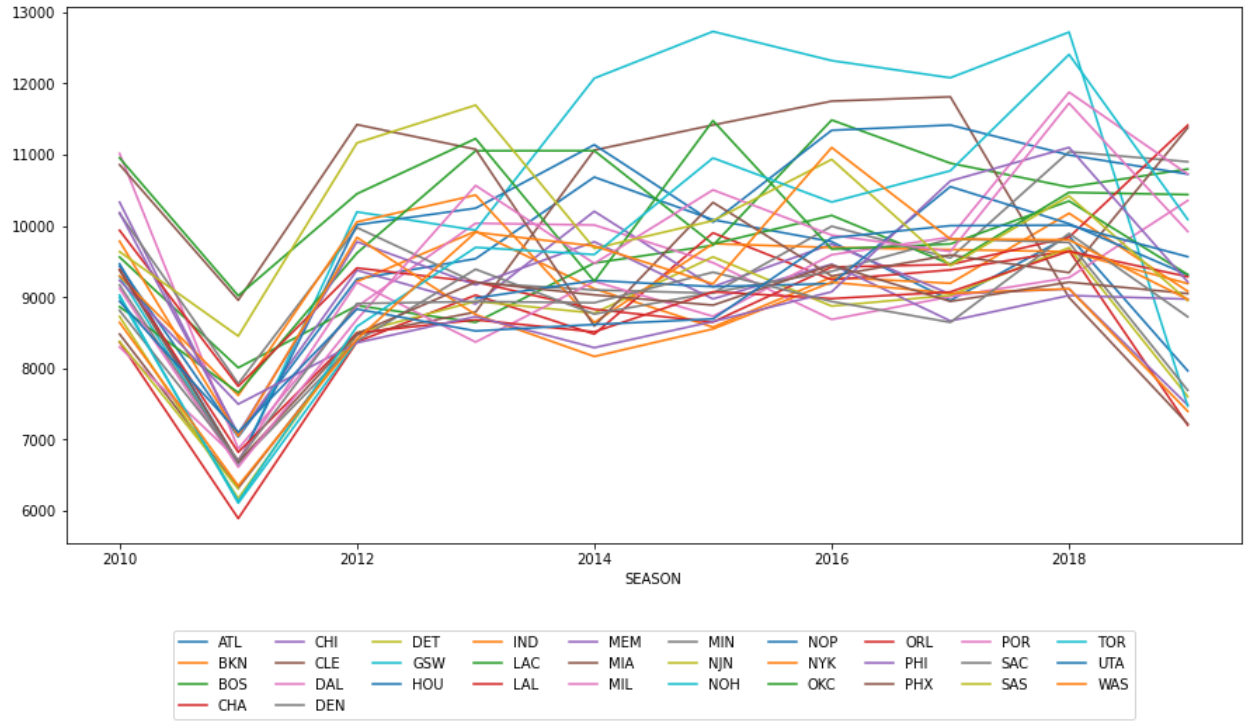
After creating our dataset, we moved onto visualizing and analyzing the data. The first part of this involved making charts and seeing if any trends were present or obvious.



First, we look at injuries by team to see if there are any obvious outliers. In the bar plot, it is apparent that all teams seem to have a similar amount of injuries. The most injury-prone teams seem to be the Hornets, Bucks, and Nets. The least injury-prone teams seem to be the Pelicans, Thunder, and Blazers. These teams seem to have varying levels of success so there is no obvious correlation between injuries and wins.



Next, we look at the total points scored in the NBA in the last 10 years by team. The highest-scoring teams are the Heat, Spurs, and Nuggets. The Spurs and Heat had a very successful decade but the Nuggets haven't. The lowest-scoring teams are the Timberwolves, Suns, and Kings. These teams have all been very bad in the 2010's so there is something to be said about low-scoring teams being unsuccessful. Again it is difficult to draw any meaningful conclusions from this chart alone, but it does tell us a little more than the injury counts.



The chart above breaks down scoring by season and team. No team stands out except the Golden State Warriors who had some of the highest-scoring seasons in NBA history during the latter half of the decade. They also won 3 championships and appeared in 5 during this timespan. This confirms the obvious suspicion that more points lead to more wins, however again we can't tell anything about injuries from this information.

We leveraged various Machine Learning algorithms such as Logistic Regression, DecisionTreeClassifier, LinearSVC, SGDClassifier, RandomForestClassifier, Linear Discriminant Analysis for forecasting the possible outcomes.

Before diving into the implementation of models, we normalized the data using StandardScaler and Normalizer functions and have made separate data matrices for assignment of train values and test values.

Through our analysis we found out that we have to reduce the dimensionality of the final dataset used for Machine Learning models with the sole intention of improving the functionality of the model as much as possible. For that we performed Principal Component Analysis on the Train dataset/data matrices. Then we performed KMeans Clustering on the reduced dimension of X_Train.

Using Kfold and executing all the Machine Learning models at once, we checked for the scores of all the models.

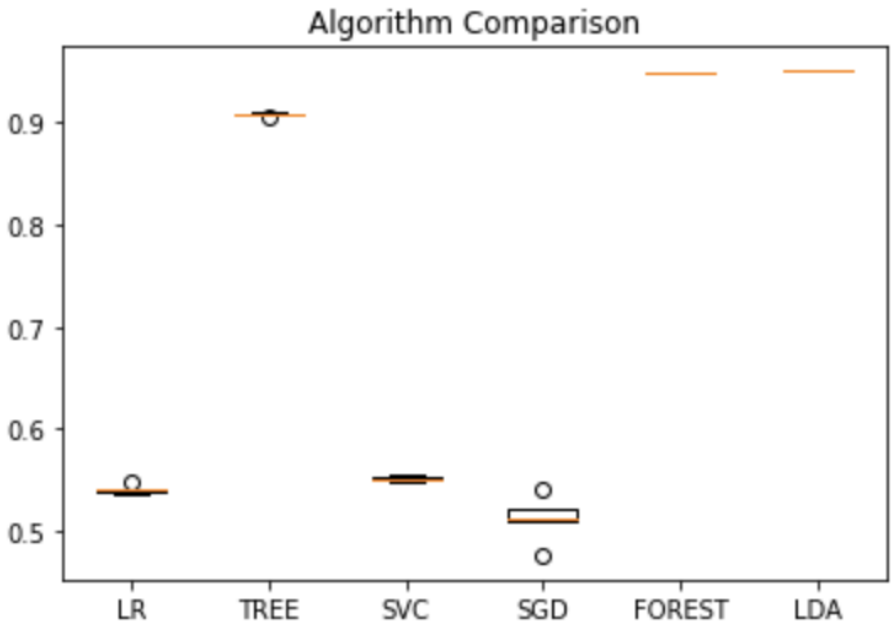
In [21]:
```
#KFold
results = []
names = []
for name, model in models:
    kfold = StratifiedKFold(n_splits=5, random_state=24, shuffle=True)
    cv_results = cross_val_score(model, X_train, y_train, cv=kfold, scoring='accuracy')
    results.append(cv_results)
    names.append(name)
    print('%s: %f (%f)' % (name, cv_results.mean(), cv_results.std()))
```

```
LR: 0.541680 (0.003730)
TREE: 0.909682 (0.000988)
SVC: 0.551599 (0.002130)
SGD: 0.512511 (0.021119)
FOREST: 0.949355 (0.000233)
LDA: 0.950622 (0.000064)
```

Here we found out that Random Forest Classifier provides the most accurate results with the maximum accuracy score.

Finally, we extensively checked for the output of each model used by using Confusion Matrix, Classification Report and a accuracy score

Going further, we visually compared the algorithm by plotting the results of every model into a boxplot.

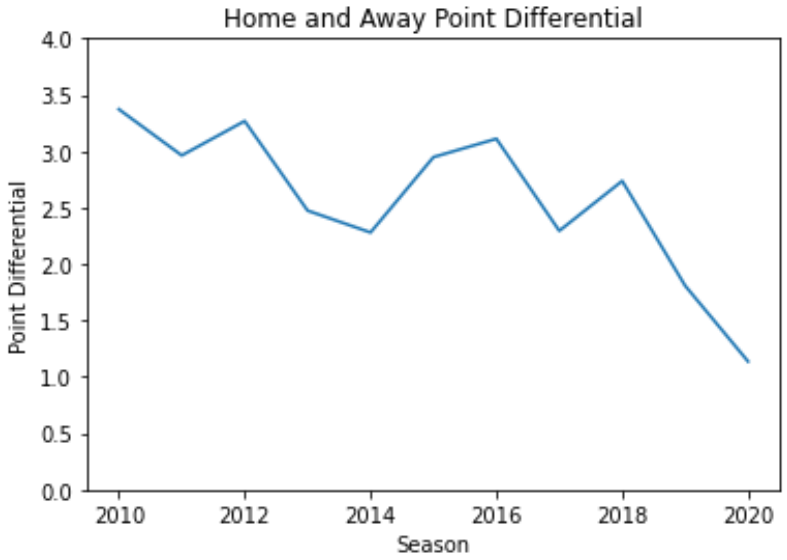


Results

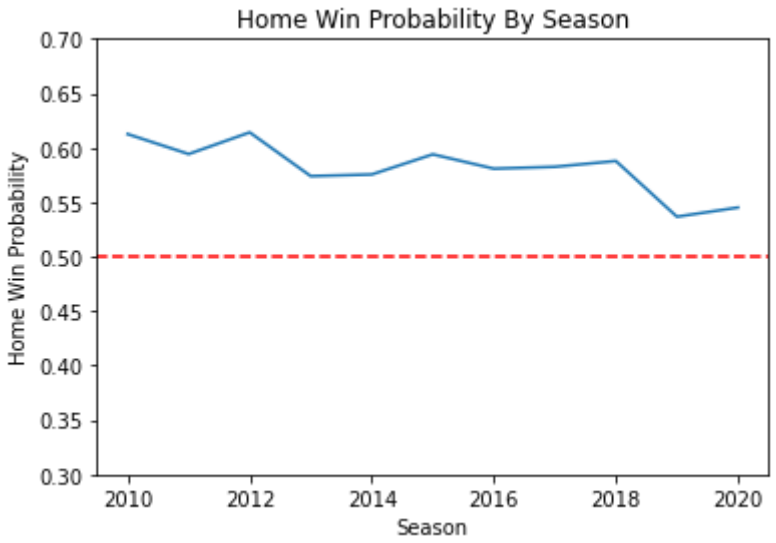
Winning in the Regular Season

After looking at this data, we wanted to examine one of the bigger debates in the NBA community regarding player health, "How important is winning in the regular season?". Currently in the NBA there appears to be two schools of thought. On one end of the spectrum, many argue that winning in the NBA regular season is very important because a better regular season record gives you a higher playoff seed. This means more home games and matchups against potentially weaker competition when the postseason comes around. The other school of thought is that making the playoffs is important but the benefit of the higher seed is negligible. This is especially true if it means that players can rest more often throughout the regular season preventing injury and fatigue come playoffs. Making the playoffs in the NBA is also fairly easy since 16 out of 30 teams earn a chance to compete for the championship. Coasting through the regular season is possible because the majority of teams end up in the playoffs. To break down both sides of this argument, first we'll take a look at home and away splits in the NBA throughout the past few years.

Is Home-Field Advantage Real?



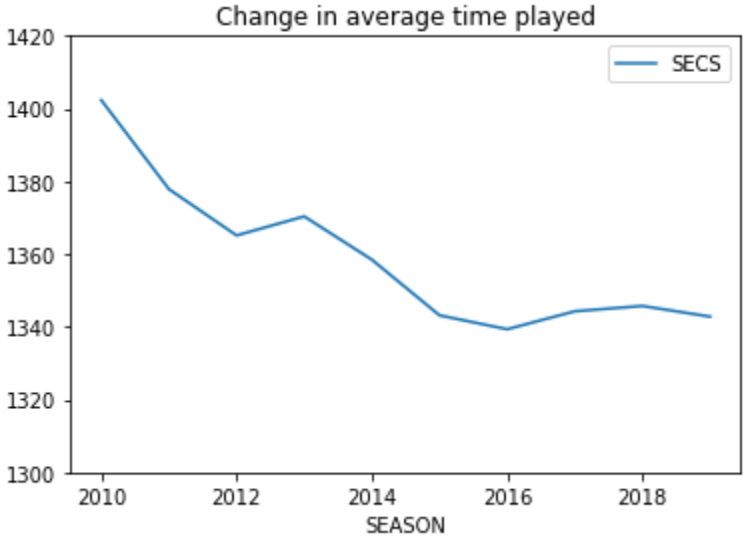
Looking at the chart above we can see that home teams on average scored more points than away teams every season. This decreased in 2019 and 2020 because of Covid. Teams were playing at home without fans due to local and state laws or in the NBA "Bubble", the term given to the games played at the end of the 2019-2020 season in Disney World. In a typical season, home teams seem to have a huge advantage when it comes to scoring points at least in part due to the crowd.



The graph above shows that home-field advantage also translated to wins. Again it is apparent that the numbers dropped off in the 2019 and 2020 seasons due to the pandemic. On average, even including the pandemic seasons, the home team wins 58% of their games. Given these numbers, home-field advantage seems to be a real factor in deciding the outcome of games. This is huge in sports where every possible advantage within the rules should be taken into account.

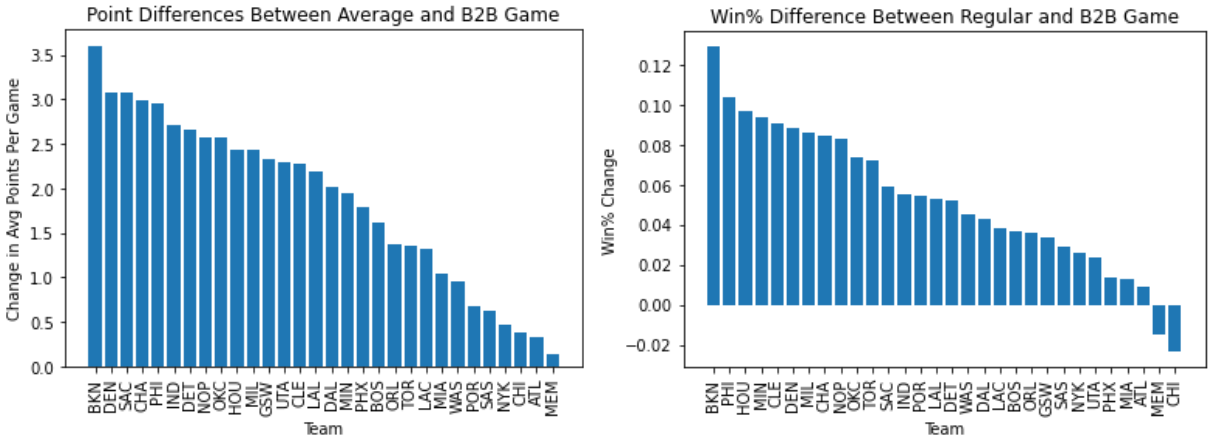
When do players rest?

Next, we will look at how minutes and games played are changing in the NBA as some teams are de-emphasizing the importance of the regular season.

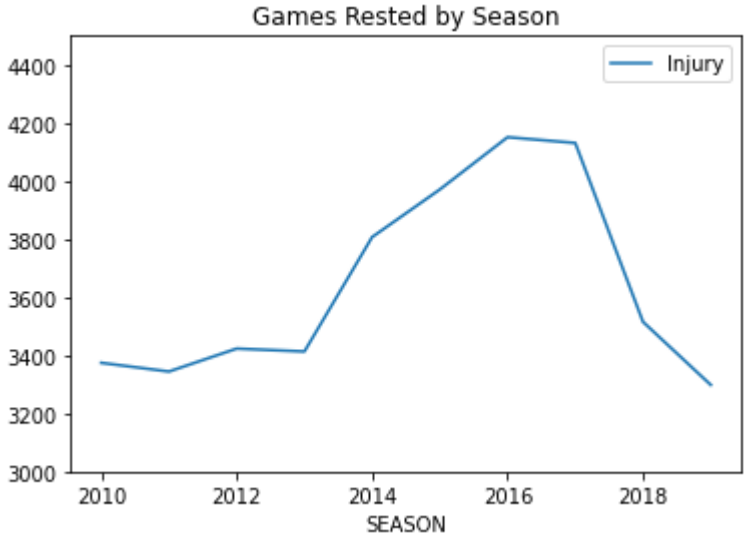


The graph above shows the average seconds played per player from the 2010-2019 season. As you can see, the amount of time spent playing per game is decreasing. Coaches are relying more on their bench, sometimes playing their 9th, 10th, and 11th players in the lineup double digit minutes which allows their stars to sit and rest more often. The idea is that this will translate to larger contributions from star players in important matchups such as rivalry games or the postseason. Players not only

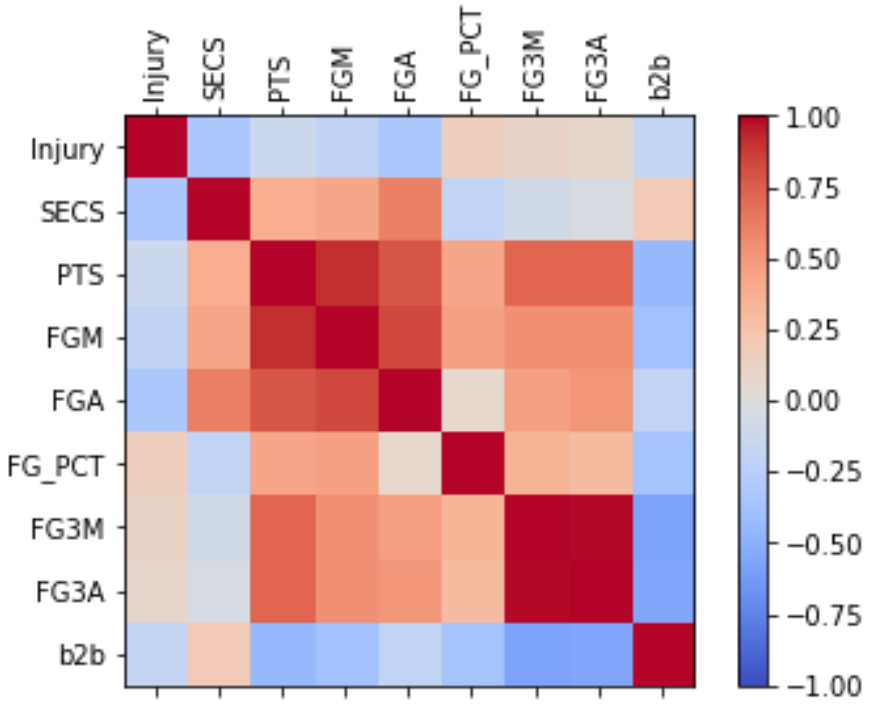
rest during the games, many have also begun taking off the second game of back to back matches. Back to back matches are games that take place within 1 day of each other.



The graphs above show that all teams score less points on the second half of a back to back and all except the Grizzlies and Bulls lose more often as well. This means that teams are either resting their best players and thus essentially punting these games or performing worse without time between games. To check if teams are resting players more often, let's look at the number of games where players are out due to 'rest' over the past couple of seasons.



In 2017, the NBA introduced a new policy that may fine teams that rested players without health concerns. Teams may have circumvented these rules by simply changing the injury designation and not including rest as a reason which could explain the drop off in 2018. At the end of the day, whether it is due to fatigue or resting their players, back to back games are essentially lost before they begin and this hurts overall regular season standings. However, the benefit to resting players is supposedly a prevention of injuries. To check this, let's look at the heatmap below of the correlation between injury and other statistics.



Heat Map of Correlations

There doesn't appear to be any interesting correlations in the chart. In fact, there appears to be a slightly negative relationship between back to back games and injury.

Heading into this analysis, we were fully expecting to side with the modern coaching philosophy of resting players in order to prevent injury and save their energy for the playoffs. However, after looking at the numbers, home-field advantage seems to be a real factor while the benefits of resting players seem to be almost non-existent. Of course, this is purely from a numbers standpoint. Star players such as Lebron James have noted and complained about a large number of back to back games in their teams' schedule in the past.

Machine Learning

Inexact matching schemes like those used in merge_asof have the potential to obfuscate correlations in the data. Since the algorithms applied to the dataset had to not only weigh a minority class higher than a majority class and find a way to learn based on likely obfuscated correlations, it is probable that our training results are skewed. Take note that in our Decision Tree classifier we have a fairly high weighted average, but a low macro average and an abysmal predicting accuracy of the minority class. Even with tuning the weights of these algorithms, the classifiers could not classify the minority class effectively.

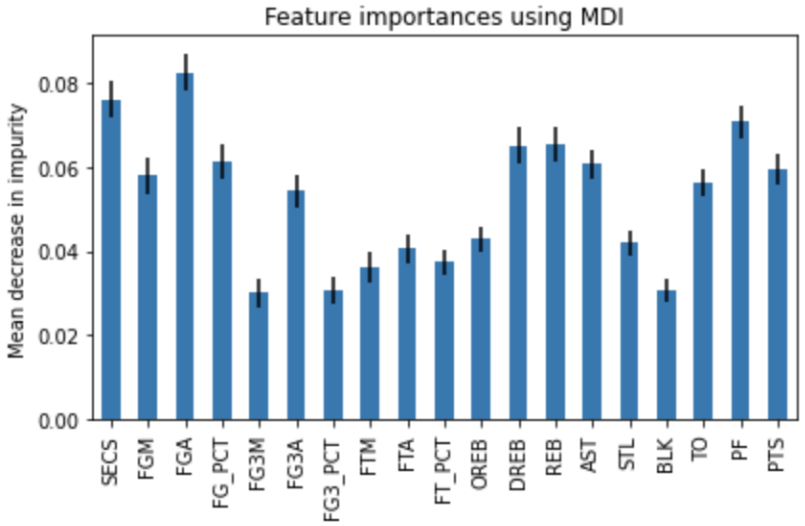
```
Accuracy for TREE is: 0.9109278198186789
Confusion Matrix for TREE is:
 [[172139   8245]
 [  8634    480]]
Classification Report for TREE:
              precision    recall  f1-score   support

           0       0.95      0.95      0.95    180384
           1       0.06      0.05      0.05      9114

    accuracy                           0.91    189498
   macro avg       0.50      0.50      0.50    189498
weighted avg       0.91      0.91      0.91    189498
```

Decision Tree Report

Even with the less than ideal results of the training, we can look at the feature importances derived from our Decision Tree classifier and observe that time played in seconds (SECS), field goals attempted (FGA), had some of the highest importances. It is possible that by culling other features that may be confusing the learning of these algorithms that these results may improve.



Discussion

Finding the cause of injury in sports is tricky from a data science perspective. One of the limiting factors of our experiments is the simple fact that injury reports are not honest. Coaches may announce injury reports at timings convenient to their chances of winning, resulting in time lag from when an injury occurred to when it is recorded for the public. Furthermore, there are multitudes of external factors that may have caused reported injuries that are completely unrelated to any of the games played. Sometimes injuries likely not probably caused by game related stress can be identified easily, such as an injury report that details a COVID-19 infection. Other times, injury reports can be frustratingly unspecific and only detail a surgery in passing. In the original injury severity problem that we tried to cover, this ended up causing us to drop the topic altogether, as the variety of unspecific, unclear

injuries and the fact that we could not exactly know when these injuries truly occurred meant that such a multiclass problem would be infeasible.

One of the goals at the outset of this project was to recommend policy changes based on the results. As of now, it is unclear if we can make any policy changes backed up by our data. Observations we can make, as mentioned in our results, include that head coaches likely will continue to rest their best players on home games and field them aggressively on away games, as they already have a significant home field advantage. This home field advantage according to what we have been able to glean currently outweighs any disadvantage back to back games could have. Therefore, in a situation where a team needs to play multiple away games back to back, coaches will push their players to play regardless.

During the course of our research into this problem and reviewing the work of others, we came across interesting conclusions that, "No correlations were found between injury rate and player demographics, including age, height, weight, and NBA experience." (Drakos et al., 2010, 284) This is counterintuitive, most people would assume the larger the frame or the heavier the player, the more risk of injury. According to another study, the odds of injury increase by 2.87% for each 96 minutes played and decrease by 15.96% for each day of rest. (Lewis, 2018, 503). Our feature importance results seem to also conclude that time spent on the playing field is a large factor in classifying the data. However, seeing as the other statistics such as the player stats that determine their overall performance, statistics commonly used by fantasy basketball players and sports gambling institutions alike, are irrelevant in determining injury, when planning for injuries whether in situations real or imagined it seems that the factors that affect injury are not unique to professional basketball.

We believe that the primary reason that our dataset did not show these correlations was due to the unfortunate obfuscation due to our merging methods which were necessary to compose a functional dataset.

Conclusion

According to our findings, we were not able to establish a clear correlation between the features we selected but we have identified a number of limiting factors, such as our merging method using pandas.merge_asof and the quality of the data, particularly in the Injuries dataset we obtained. Injury reports are not necessarily honest nor are reports released at the exact time of injury. Via estimation of injury occurrences we were able to chart out visualizations to help us understand the characteristics of the dataset, but when engaging in deeper analysis by picking apart at correlations, it seems that the relations are obfuscated by time lags. The dataset also suffers from a minority class problem that was not alleviated sufficiently by adjusting weights in our classifiers. We believe that a more sophisticated merging and aggregation scheme could preserve the correlations of the data. There are also techniques such as undersampling which could be tried to solve the minority class problem. Applying a better understanding of the individual classifiers or simply trial and error application of other weights could also improve classifier performance. An attempt at

restructuring the problem as a regression estimating injury rates as injury occurrences divided by games played with that statistic normalized could also be a possible avenue for further study.

References

Boone, K. (2017, September 20). *Charles Barkley calls out 'poor babies' in the NBA who complain about back-to-backs*. CBS Sports. Retrieved December 7, 2021, from https://www.cbssports.com/nba/news/charles-barkley-calls-out-poor-babies-in-the-nba-who-complain-about-back-to-backs/

Drakos, M. C., Domb, B., Starkey, C., Callahan, L., & Allen, A. A. (2010, July). Injury in the National Basketball Association. *Sports Health*, *2*(4), 284-290. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3445097/

Helin, K. (2020, December 7). *NBA memo warns of large fines for resting players for nationally televised games*. NBC Sports. Retrieved November 21, 2021, from https://nba.nbcsports.com/2020/12/07/nba-memo-warns-of-large-fines-for-resting-players-for-nationally-televised-games/

Lauga, N. (2020, October 13). *Games.csv*. www.kaggle.com. Retrieved November 18, 2021, from https://www.kaggle.com/nathanlauga/nba-games?select=games.csv

Lewis, M. (2018, May). It's a Hard-Knock Life: Game Load, Fatigue, and Injury Risk in the National Basketball Association. *Journal of Athletic Training*, *53*(5), 503-509. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6107769/

McCarthy, M. (2017, March 20). *ESPN rips 'one-way' relationship with NBA stars who sit out showcase games*. Sporting News. Retrieved December 7, 2021, from https://www.sportingnews.com/us/nba/news/espn-nba-abc-resting-players-turner-sports-

tnt-lebron-james-kyre-irving-kevin-love-cavaliers-clippers-jeff-van-gundy/1ablku36z1m5c1it6tqafyp5t1

NBA games data. (2021, November 18). Kaggle. Retrieved November 18, 2021, from https://www.kaggle.com/nathanlauga/nba-games?select=games_details.csv

NBA Injuries from 2010-2020. (2020, October 13). Kaggle. Retrieved November 18, 2021, from https://www.kaggle.com/ghopkins/nba-injuries-2010-2018

NBA Rules Archives - NBA.com. (n.d.). NBA Communications. Retrieved November 18, 2021, from https://pr.nba.com/category/nba-rules/