**ST494 Final Report: *A Report on Scheduled Losses***

Wilfrid Laurier University - ST494: Final Project

Owen Bell - 200449530, Marcus Rilling - 200656520

Dr. Devan Becker

April 12th, 2023

### Abstract

This report looks at the idea of using fatigue as a factor to predict the outcome of an NBA game. Using previous ideas used by ESPN writer Baxter Holmes, we expand on them using other features and methods learned from this class. Using data from the 2018-2019 NBA season, we were able to create logistic regressions that predicted a subsection of games at above 80%. In addition, using these predictions to look at potential profits if money was placed on oddsmaker's lines.

We decided through our process to see if we could use whether or not a game is classified as Baxter as a binary predicting feature. Using linear regression and three types of decision trees, we were able to get a top accuracy of 71.1% from the logistic model. Grouping this all together, we found through our process that fatigue is in fact a viable tool for helping predict NBA games.

### Introduction

Fatigue is an issue for many of us, although some are able to work while tired, many cannot. This is also true for many athletes, especially in sports with a high volume of games in a short amount of time. Each NBA team plays eighty-two games in one hundred and seventy-three days while travelling across the United States and Canada. Although these athletes are at the highest level, this does not exempt them from the fatigue felt throughout the season. In fact, the NBA has had recent struggles with star players not playing in regular season games due to injury prevention and rest. The NBA has taken steps to reduce this by reducing difficult travel processes, providing more rest for the players, and even starting the season two weeks earlier to lessen the burden. With all this being said, there are still very difficult travelling stretches for teams: back-to-back games, four games in five days, five in seven, and even extreme travelling distances, for example, playing in New York than the next day in San Francisco.

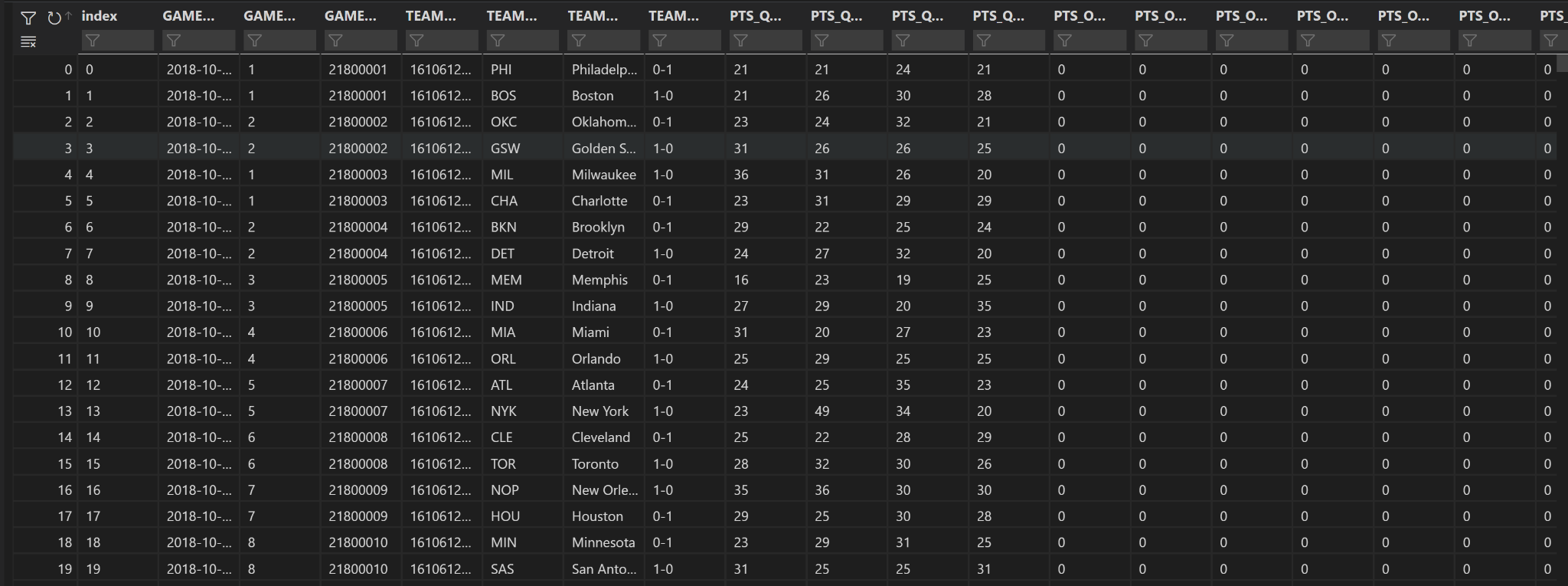
Baxter Holmes, a writer for ESPN, wrote about this idea of losing games just due to the scheduling done by the NBA. Baxter, with the aid of sleep consultant Cheri Mah, constructed a formula in order to identify potential *"schedule losses".* The statistic called "MAH SCORE" helped to pick out games based on eight fatigue factors, creating a scale from 0 to 12, with a higher score indicating a greater disadvantage. With this model, they were able to identify games and predict wins or losses at a rate above 76.5% for their red alert games (MAH score above 7.5), although they used only 17 games during the 2016–2017 NBA season. A note to be made is that the MAH score did not take prior team performance into account and selected a low number of games (fifty-nine games in 2016–2017 and about 63 in 2017–2018).

We thought Baxter Holmes had a great idea of "schedule losses" and could try to recreate finding these games through some of the concepts we learned in this course. However, there were gaps in his model, as it did not take into account the prior performance of the teams playing as well as a smaller sample size. With no background expertise in sleep science, it would be very difficult to completely recreate the MAH statistic, but there were ways to look at other fatigue factors such as tracking games in a certain amount of days and team travel. These games which we found to be “scheduled losses” were labeled as Baxter games in the datasets. We also wanted to see if there was any way we could apply the models we made for this project to betting odds to see if Vegas does take fatigue into account when creating odds for each team or if there was an opportunity. This report details our process and works in bringing Baxter Holmes’ idea to life to see how much an NBA team's performance can be affected by scheduling and fatigue.

### Data

#### Collecting Data

When finding the datasets we needed to conduct analysis and modeling, we were very fortunate to only require a couple of datasets with everything we needed. The first dataset used was one from a collection of datasets from Kaggle that contained a multitude of information surrounding the NBA’s 2018-2019 season, which we referred to as the box score dataset. Although the Kaggle site also included game-by-game information for the 1940s, we felt this would be difficult to model due to the large volume of data and large changes within the data. Not only has the game of basketball evolved greatly, but the outside factors have also greatly changed. With teams looking deeper into improving player performance, there has been a significant improvement in team travel and player rest. We thought the data set we found was enough to test our hypothesis. We also used the most recent season available as we wanted to have our data most closely resemble what the NBA is like currently.



Pictured here is what the raw csv loaded into a data frame prior to data cleaning and transformation

Secondly, through the same Kaggle dataset, we were able to find data including a wide variety of betting information from every game in the 2018–19 season. This was incredibly useful in our analysis of our models compared to the odds of NBA games at the time and allowed us to do our calculations to see how much we could have profited from a model like this.

Finally, we found location data for every NBA stadium in the form of longitude and latitude coordinates. We wanted to use this data in order to get distances from each NBA stadium to use in our calculation of potential Baxter games as well. This dataset included almost all the information we needed except for the Toronto Raptors stadium, and the Scotiabank Arena, which was easy to find and add to the dataset.

#### Data Cleaning/Transformation

While the data we found included much of the information we needed, there were still some issues with cleaning and transforming the data in a way that made it easy to use in the model. As we mentioned beforehand, the data we found for the longitude and latitude of each NBA stadium was missing the Toronto Raptors information and used some different abbreviations for certain teams compared to our other data.

The main bulk of the work came with cleaning the box score dataset as each game took up two rows, with the away team coming first with all their stats and the home team second. This needed to be fixed to keep each game we had on one row for better game-by-game comparison. This dataset also had more traditional basketball stats for each game, such as FG\_Perc (Field Goal Percentage) and Points\_Scored. We did some more transformations in order to make these stats into running averages over the entire season so you can compare and see how a team's performance in many metrics compares over the course of a season.

Finally, the data frame with Vegas odds required some cleaning and transformation as well. Like the box score, it also splits each game into two lines, storing betting information for the two teams in each row. We also had to combine these stats into a one-game, one-row idea to better match what we had already done for the box score data. Luckily, the unique identifier, Game\_ID, was the same in both the box score and betting data, making joining the two datasets together or comparing them with one another incredibly easy. Pictured below is a preview of the cleaned and transformed box score data.

#### 

#### Feature Selection

The majority of the features we used in the models were created in the data cleaning and transformation process using our background knowledge to help us. However, one of the main ideas we wanted to tap into was the distance travelled in a certain number of days, as well as the number of games played in a certain number of days. Using the longitude and latitude data, we were able to sort through every team’s games in a season and compare the distances of a team's previous and current game using a formula to calculate distances using just the coordinates. From this, we were able to map the distance every team travelled for each game in the 2018–19 NBA season. From these game-to-game distance calculations, we created another function that enabled us to calculate the distance travelled over a certain number of days. However, these statistics assume no unnecessary travel by teams in a season, such as returning back to their home city in between a group of two away games, since that is nearly impossible to track. Finally, using a similar algorithm, we calculated the number of games a team has played in the last 2, 5, and 7 days.

Using these features that we created, we were finally able to define what a Baxter game was. What we did was find games that had one or more of these qualifying features. For distance travelled in the latest 2, 5, and 7 days, we used distances that were in the 97.5% quantile. We also included the criteria of back-to-back games, four games in the last five, and five in seven. If a game had a team that qualified for any of these criteria, we designated the game as a Baxter game. We were able to choose these numbers through a mix of trying to capture around 10–15 percent of a total NBA season's games as a Baxter Game and previous knowledge of what a tough schedule looks like. For our purposes, in the 2018-19 season, we captured 147 games with this criteria or just under 12% of a total NBA season.

Due to teams coming into games not having played the same number of games, a win percentage difference was used to compare how ‘good’ a team was compared to one another. The win percentage difference feature was created by subtracting the home team's winning percentage from the away team’s.

For the features surrounding the Vegas dataset, we decided to look at the money line for both the home and away teams. The money line is essential because it tells which team the oddsmaker favours. If the money line is positive, that means the team is not favoured, and the positive value associated is how much money would be made if a $100 wager was placed. The opposite is true of a negative line: the oddsmaker favours the team, while the value is associated with how much money must be wagered in order to make $100. The money line is a fitting feature due to the fact that it only looks at the binary outcome of the game: does a win or a loss occur? Due to the nature of the dataset having information for several oddsmakers, an average money line for both the home and away teams was used. This was a feature provided by the data set, which saved us a lot of effort and time. The average line spread for both home and away teams was planned to be used but was not due to computational complications.

### Methods

We decided to use logistic regression as our primary method. This was appropriate due to the fact that we wanted to look at the binary outcome of whether the home team had won or lost a game. We knew from previous knowledge obtained in this class that linear regression models wouldn't be as appropriate in terms of fitting a binary outcome as logistic ones are. We used ‘Home\_Win’ as our response variable and used every binary and numerical predictor as a feature as a baseline for training the model. The model was then trained and tested on the 147 Baxter games found. Feature selection was done in hopes of improving the original model; using Python's built-in function SelectKbest and Chi squared as a score function, the ten most important features were found. Seven features were chosen mainly based on scoring above 300 on the method mentioned above.

Subsequently, using the predicted and actual outcomes, average money lines for both home and away teams were used to create a data frame that would enable us to look at the potential profit. We were able to find profit by running a series of if statements based on whether a true prediction was made or not, and then determining which money line to use based on whether we predicted the home to win or lose (home money line was used in the case of true, and away money line was used in the case of false).

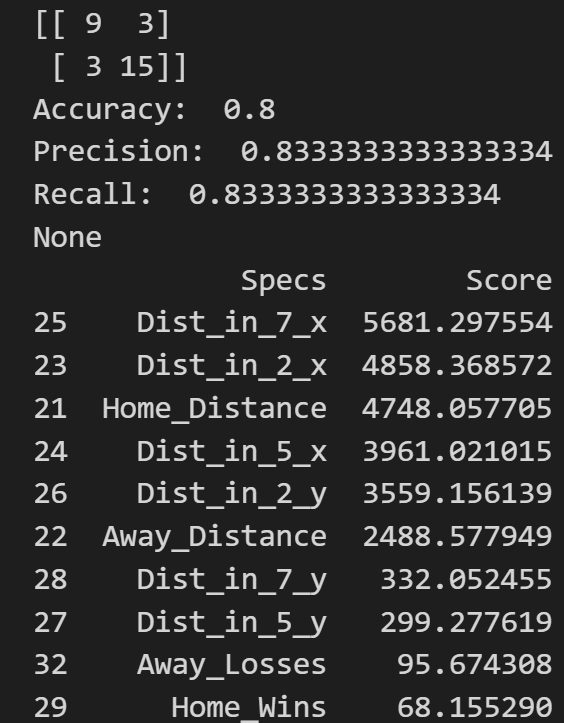
We next decided to look at using this Baxter identifier for games over a season as a feature in logistic regression as well as three types of classification trees to predict games over the entire course of a season. We decided on logistic regression and the three types of classification trees since the response variable is binary in nature (does the home team win, true or false). We also wanted to keep with the trend of a logistic model since that has been the model we have previously used to predict only Baxter games; this would also allow us to compare the accuracy of each model. The three classification trees we used were a general decision tree, a pruned version of a general decision tree, and finally a bagged tree. We chose these models as they can offer us a bigger breakdown of how a model can use its features to predict which features are valuable for predicting whether a team wins or not.

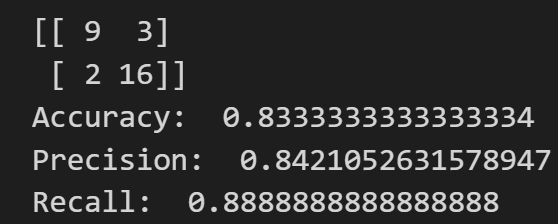
For the logistic regression model on the entire season, we decided to use four features in order to predict the resulting outcome of games: the win percentage difference, the field goal percentage of both teams and finally the Baxter variable we created. We decided on these features since we believed that these four could have the biggest impact on predicting game outcomes. This model was again predicting the "*Home\_Win"* variable of our data to keep everything consistent across all our different models.

For the three trees, we opted to use every numerical statistic created in the data frame as predicting features for the classification tree. This enabled us to take a look at all the statistics we found and created to see which were the biggest predictors of a team's winning output. From these, we can see how impactful the Baxter stat truly was in comparison to the other already established statistics in the game of basketball.

### Results

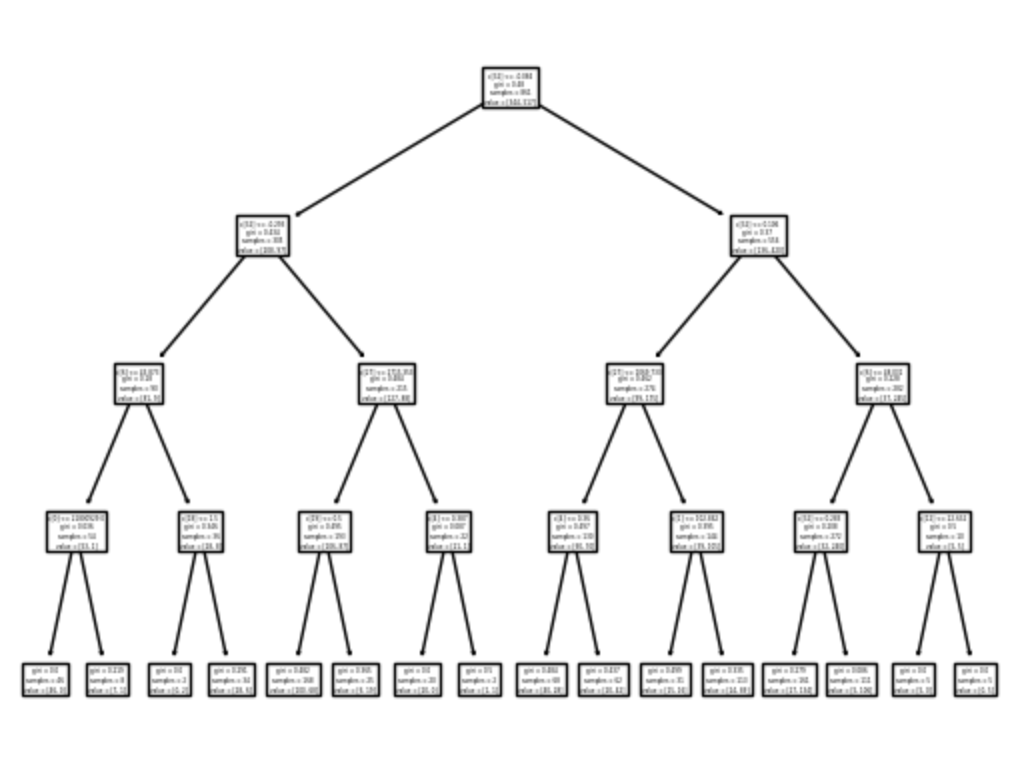
Measuring the performance of the logistic models was fairly easy using Python's built-in methods from Scikit-Learn. We decided to use the three measurements of accuracy, precision, and recall due to the ability of outside readers with some statistical background to make conclusions about our model. The first model trained performed really well, scoring above 80% on accuracy, precision, and recall. Although this was a great outcome, it was improved once the feature selection was done. Using 38 fewer features, each of the three measuring features improved to score about 83%, even getting an 89% on recall.



The first photo shows the three metric scores of our first model and the feature scores. The second photo is the scores for the model with feature selection

As mentioned above, we found a method in order to discover how much money we can make based on if we had put money into the 30 test set games. After running the code, we discovered that we would have profited $1478.40 if a $100 wager was placed on each game. Out of curiosity, we repeated this process for the 147 Baxter games (including the training set of data), and a profit level of $8305.1 was found.

In general, for the logistic regression model on the entirety of an NBA season, we used accuracy to measure how well our model fit the data. After splitting the data and fitting/predicting the model, we received an accuracy of 71.1%. All things considered, we were really happy with the number output by our model. Though it is not an outstanding accuracy in terms of the 80%–90% range, for predicting games across an entire NBA season and with the amount of randomness seemingly sports have, this is quite impressive. One thing we noticed about this model was the quite high false-positive rate. When we broke it down into its confusion matrix, we had a false-positive rate of nearly 47%. This can again be explained by the randomness of predicting sporting events and by underdogs seemingly upsetting teams at random.

For our decision tree model, we decided to use a total of three trees to see which one would give the highest accuracy. When comparing the general decision tree, a pruned decision tree, and a bagged decision, we were given the approximate accuracy of 61%, 69%, and 67%, respectively, with different passes of the models giving us values in the range of these percentages. We decided to visualise the structure of the pruned tree as it is depicted on the left; we pruned it to provide a maximum depth of only four and decision criteria based on the Gini index. We found that using these parameters for the model gave, on average, the best accuracy when predicting games. Breaking down this tree, we found that the win percentage difference was the main predicting factor in the outcome of games. This doesn’t surprise us, as the win percentage difference has been an important feature in almost all of our models. Though these trees did not outperform the logistic regression, they gave us great insight into what made our models predict as well as the breakdown of how they did.

Conclusion

The main takeaway from this work is that fatigue is not only a factor for us as students but also for athletes at the highest levels of their sports. Through our methods, we were able to predict with 80% accuracy the outcome of an NBA game through a season based on extreme scheduling. From these games, we were also able to use this prediction accuracy to make money by looking at the oddsmaker's lines and betting on teams that our models had predicted. Finally, using a classifier for a Baxter game, we were able to predict the results of an entire NBA season with 71.1% accuracy.

### 

### 

### 

### 

### 

### 

### 

### 

### References

* Link to NBA Dataset from Kaggel:

https://www.kaggle.com/datasets/erichqiu/nba-odds-and-scores

* Baxter Holmes Article References

Holmes, B. (2017, November 1). Schedule alert! every game tired teams should lose this month. ESPN. Retrieved April 11, 2023, from https://www.espn.com/nba/story/\_/id/21236405/nba-schedule-alert-20-games-tired-teams-lose-november

### Delineation of Work

Owen:

* Calculated and worked on distance function as well as win percentage difference
* Part of cleaning longitude/latitude stadium data
* Feature Selection for models over entire NBA season
* Logistic Regression Models over entire NBA season
* Decision Tree Models over entire NBA season
* Betting algorithm to calculate net profit/loss

Marcus:

* Box score information,
* Cleaning of Box Scores and Vegas data frame
* All work done for the logistic regression model on the Baxter games
* Creating Baxter Dataframe as well as Baxter statistics
* Calculating profits