**BAIS 3250 Final Project Report**

**NBA Player Performance Vs. Travel Schedule**

**1. Introduction**

In the National Basketball Association (NBA), teams often face grueling travel schedules and condensed game calendars that put their bodies and minds under a lot of pressure. The conversation around players intentionally sitting out of games for rest has been an increasingly popular point of discussion among fans and analysts. However, quantitative studies that explore the effects of travel fatigue remain limited. This project investigates the extent to which travel-related factors—such as distance traveled, back-to-back games, rest days, and game location—affect individual player performance across a season. Furthermore, we explore whether factors like age and player position influence one’s susceptibility to travel fatigue. The insights gained from this analysis could benefit teams, coaches, and analysts by informing load management strategies and optimizing performance. Listed below are the original research questions that drove my project:

1. Do travel related fatigue factors negatively impact player performance?
2. Are specific player positions (e.g., guards vs. centers) more affected by travel fatigue? Are older players more affected by travel fatigue than younger players?
3. Can we accurately predict game-to-game performance changes based on travel fatigue factors?

**2. Data**

This project uses two primary sources of data as listed below.

**Sources:**

* **NBA Player Game Logs (Season: 2021-2024):** Imported from NBA stats API
* **NBA Team Travel Information:** Scraped from [ESPN.com](https://www.espn.com/nba/schedule)

*2.1 NBA Player Game Logs*

Using the NBA stats API from NBA.com, I imported the stat lines of each player that played in a game across four seasons. This data contained key performance metrics from each game like points, assists, turnovers, plus-minus, etc. It also included each player’s name and player ID, which I used to sift through the NBA stats API and pull each player’s age and position for the given season. I then added these two new columns onto the existing DataFrame.  
This DataFrame didn’t require much cleaning aside from renaming the columns, remapping the position names, and dropping an abundance of irrelevant and unnecessary columns. After that, it was ready to be merged with my second DataFrame.

*2.2 NBA Team Travel Information*

I web scraped ESPN.com to obtain the travel schedule for each NBA team. This required scraping 30 pages four times (one per team per season). From each page, I obtained the team’s name, their opponent, game date, and the year of the season.  
This DataFrame required a lot of data cleaning. I did some feature engineering to add a Home/Away column based on the symbol (@ or vs) displayed in front of the opposing team’s name. I remapped all of the team names, dropped the extra symbols in the opponent column, and converted the game date to a proper format.

*2.3 Dataframe Merging*

I performed a horizontal integration of these two DataFrames, specifically using a left merge on *Team* and *Game Date*. Following the merge, I did a little more feature engineering to set myself up for a more insightful analysis. I created a *Rest Days* column by calculating the difference between game dates and made sure to account for special cases (first game of the season, first game back from injury, etc.). I then created a *Back\_to\_Back* column by checking if the *Rest Days* column was equal to 0. I used the latitude and longitude coordinates of each team’s stadium to determine the distance they must travel from one game to the next and labeled this column as *Travel Distance*.

Lastly, I created season average columns for various performance metrics to compare a given game's performance to during our classification machine learning models. Each of these columns had an accompanying “is higher” binary column to label whether a player performed better or worse than his season average in each game. To keep the analysis semi-simple, I only used *Points*, *Plus-Minus*, *Field Goal Percentage*, and *3PT Percentage* as the target variables in my later analysis, which is why only those stats have average binary columns. I decided to leave the other statistics in the DataFrame so that they are available for future work.

*Table 1 Data Dictionary*

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Description |
| Season | str | NBA season in which the game was played |
| Player | str | Full name of the player |
| Team | str | Full name of the player's team |
| Game Date | datetime | Date when the game was played |
| Minutes Played | float | Total minutes the player was on the court |
| Field Goals Made | int | Total number of field goals made |
| Field Goals Attempted | int | Total number of field goal attempts |
| Field Goal Percentage | float | Shooting accuracy from the field |
| Three Pointers Made | int | Number of 3-point field goals made |
| Three Pointers Attempted | int | Number of 3-point field goals attempted |
| Three Point Percentage | float | 3-point shooting percentage |
| Free Throws Made | int | Number of free throws made |
| Free Throws Attempted | int | Number of free throws attempted |
| Free Throw Percentage | float | Free throw shooting percentage |
| Offensive Rebounds | int | Number of rebounds collected on offense |
| Defensive Rebounds | int | Number of rebounds collected on defense |
| Total Rebounds | int | Total number of rebounds |
| Assists | int | Number of assists recorded |
| Turnovers | int | Number of turnovers committed |
| Steals | int | Number of steals recorded |
| Blocks | int | Number of blocked shots recorded |
| Personal Fouls | int | Number of fouls committed |
| Points | int | Total points scored in the game |
| Plus Minus | int | Point differential while the player was on the floor |
| Age | int | Age of the player at the time of the game |
| Position | str | Player’s primary listed position |
| Opponent | str | Full name of the opposing team |
| Location | str | Indicates if the game was played at Home or Away |
| Rest Days | float | Number of days since the player’s previous game. |
| Back-to-Back | int (binary) | Whether the game second night of a back-to-back |
| Travel Distance | float | Distance (in miles) traveled by the player since the previous game. |
| Season Avg Points | float | Player’s season average points per game |
| Season Avg Field Goal Percentage | float | Player’s season average field goal percentage |
| Season Avg Three Point Percentage | float | Player’s season average 3-point shooting percentage |
| Season Avg Plus Minus | float | Player’s season average plus-minus value |
| Points Higher | int (binary) | Whether the player scored more points than their season average in the game |
| FG% Higher | int (binary) | Whether the player’s field goal percentage in the game exceeded their season average |
| 3P% Higher | int (binary) | Whether the player’s 3-point percentage in the game exceeded their season average. |
| Plus Minus Higher | int (binary) | Whether the player’s plus-minus in the game was greater than their season average |

**3. Analysis**

*3.1 Descriptive Statistics*

I wanted to visualize different parts of the data to see how the travel-related factors were distributed. Below are univariate visualizations of *Rest Days* and *Travel Distance*. I accounted for outliers in both visualizations to make the plots more interpretable. The data shows that the average number of rest days a player gets between games throughout a season is 1. Additionally, the average distance that a player must travel for away games is between 250–1000 miles. These visualizations create a comprehensive understanding of how much players travel and how little time they have between games.

A graph of a distribution of travel distance

AI-generated content may be incorrect.

A graph of a number of bars

AI-generated content may be incorrect.

After visualizing some of the data, I wanted to see how travel distance impacted different types of players. I compared the performance of different player age groups and different player positions with travel distance using bivariate visualizations. For these, I used *Plus-Minus* as the performance metric, as it gives a comprehensive overview of a player’s impact on the game.

A graph of a graph showing the average plus minus by travel distance

AI-generated content may be incorrect.

A graph of a graph with lines and dots

AI-generated content may be incorrect.

*3.2 Hypothesis Tests*

As you can see, younger, smaller players tend to be more negatively impacted by travel. However, it doesn’t seem to be by a large amount, especially considering that they are the lower-performing group on average anyway. To further determine the significance of travel distance, I ran ANOVA hypothesis tests to assess whether travel distance is truly an impactful variable.

A black text on a white background

AI-generated content may be incorrect.

**A black text on a white background

AI-generated content may be incorrect.**

With wildly low p-values, it is clear that travel distance does impact players of different ages and positions. We can be confident that performance changes do occur as travel distance varies, but I wanted to take this analysis one step further. I wanted to determine if we could predict player performance based on travel-related factors.

*3.3 Machine Learning Models*

Originally, I planned to perform regression models to predict exact performance metrics of players. However, these yielded very insignificant, ambiguous, and confusing results. Therefore, I took the advice of my professor and decided to perform classification analysis to determine whether a player would perform better or worse than their season average. I created and tested three different models: Logistic Regression, Decision Tree, and XGBoost. I made sure to only use travel-related features for the X sets and set the target variable to one of the following binary performance metrics: Points Higher, FG% Higher, 3P% Higher, or Plus Minus Higher. I also made sure to preprocess the models accurately using standard scaling and a proper train/validation/test split. Lastly, I used various model performance metrics, like F1-score and ROC AUC, to determine model performance since this is a classification problem. Below are the results for the Points Higher target variable:

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A graph of a number of different colored lines

AI-generated content may be incorrect.

The three models all achieved similar results for all four target variables, with an average ROC AUC score of around 0.53. This score indicates that the models performed just slightly better than random guessing when determining whether a player would perform better or worse than their season average. These results indicate that travel metrics alone are insufficient for reliable predictions.

*3.4 Time Series Forecasting*

Lastly, I performed a time series analysis forecast to see if I could predict a given player's points over time. I used two different players for this analysis: LeBron James (a superstar) and Josh Hart (a more consistent role player) to introduce variation amongst the models. Both achieved low RMSE scores of 6.32 and 4.53, respectively. This reinforces the findings and interpretations from our previous ML models.

**A graph with blue and orange lines

AI-generated content may be incorrect.**

**A graph with blue and orange lines

AI-generated content may be incorrect.**

**4. Conclusion**

In this project, I explored the impact of travel fatigue factors such as rest days, travel distance, back-to-back games, and home vs. away status on NBA player performance. While general logic would suggest these variables should significantly alter player performance, our modeling results tell a different story. Across various hypothesis tests and multiple classification models, we evaluated the significance of travel-related variables and their predictive power.

The answers to our research questions are as follows:

1. **Do travel-related fatigue factors negatively impact player performance?**  
   Yes, there is a slight dip in player performance across various metrics, but the change is small. However, the inconsistency and small impact of these effects suggest that travel fatigue is generally well-managed by professional athletes and teams.
2. **Are specific player positions (e.g., guards vs. centers) more affected by travel fatigue? Are older players more affected by travel fatigue than younger players?**  
   Younger players and smaller players are more affected by travel fatigue. This is likely due to their inexperience with consistent travel and the larger movement requirements of smaller positions like guards, compared to centers who don’t move as much in-game.
3. **Can we accurately predict game-to-game performance changes based on travel fatigue factors?**  
   No, we cannot predict game-to-game performance changes consistently by solely relying on travel fatigue factors. Additional factors, model changes, and fine-tuning may produce different results in future analyses.

Ultimately, travel-related factors do make a difference in a player's performance, but not on a large enough scale to create strongly predictive models. We can conclude that a player’s game performance appears to be more strongly influenced by intrinsic factors that happen on the court rather than external travel conditions. While fatigue may play a role in extreme situations, it is not a dominant factor in determining player performance outcomes across a season.

These findings reinforce the resilience and adaptability of these elite athletes. Playing basketball is their livelihood, job, and for many, the most important part of their lives. At the end of the day, they will play and perform to the best of their abilities, regardless of where they are or when they played last.