Capstone Final Report:
Predicting All-NBA Team Selections: Insights from Basketball Player Statistic

Problem Statement

The National Basketball Association (NBA) is a professional basketball league in North America. Comprising 30 teams, the NBA showcases the world's best basketball talent and attracts a vast audience of fans. The league features highly skilled athletes competing at the highest level, with games characterized by fast-paced action, athleticism, and strategic gameplay.

The All-NBA teams honor the league's top players for each season. Being selected for an All-NBA team is a prestigious honor and primarily based on a player's performance and contribution to their team's success for the current season. By looking at a player's stats, we can help determine who will be chosen for these teams which is of great interest to team managers, coaches, players and fans. By understanding who is likely to make the All-NBA Teams, we can help inform strategic decisions such as player recruitment, contract negotiations, and marketing efforts.

By using predictive modeling techniques and analyzing player statistics, this project aims to uncover the key factors that influence All-NBA Team selections. These techniques can be used on future datasets of NBA player stats in the upcoming years.

Data Wrangling

The data set I used for this project was a NBA Player Stats dataframe from kaggle. It contained the 2022-2023 regular season NBA player statistics per game, and had a shape of 679 rows and 29 columns. Each row contained a player and their statistics for the season, including duplicate player names for each team a player had been on for that season. I dropped the team name column and grouped the multiple rows of each player, making sure I aggregated the rest of the columns by their average.

With 29 columns containing all different features, I chose to keep 10 columns that I thought would be the most important to help predict and model the data. The following columns are the columns that I believe are the most important based on their relevance to player performance and historical trends in All-NBA team selections:

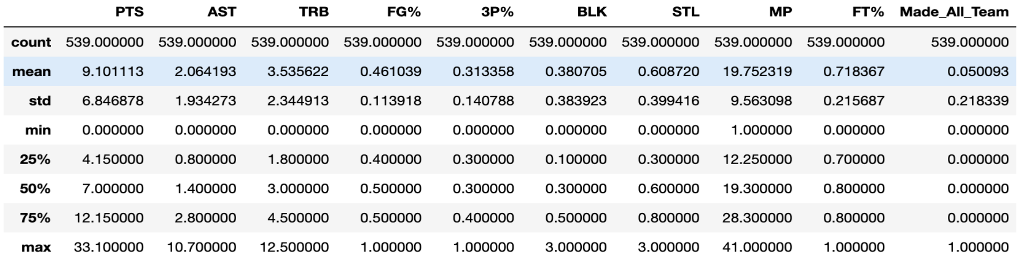
- Points per game (PTS): Players who score a high number of points per game are often considered for All-NBA selections, as scoring ability is highly valued.
- Assists per game (AST): Players who excel at facilitating their team's offense and creating scoring opportunities for teammates are highly regarded.

- Rebounds per game (TRB): Rebounding is a crucial aspect of the game, and players who contribute significantly in terms of total rebounds per game, including offensive and defensive rebounds, are often recognized.
- Field goal percentage (FG%): Efficiency in scoring, as reflected by field goal percentage, is important. Players who can efficiently convert their field goal attempts into points are highly valued.
- Three-point field goal percentage (3P%): In the modern NBA, the ability to shoot three-pointers efficiently is highly prized. Players with a high three-point percentage can stretch defenses and provide valuable floor spacing.
- Blocks per game (BLK): Defensive contributions, including shot-blocking, are significant factors in player evaluations. Players who excel at protecting the rim and altering opponents' shots are often considered for All-NBA defensive teams.
- Steals per game (STL): Defensive playmaking, such as stealing the ball, can also influence voters' decisions. Players who excel at generating steals contribute to their team's defensive success.
- Minutes played per game (MP): Players who log significant minutes demonstrate their importance to their teams and often have ample opportunities to make an impact on games.
- Free throw percentage (FT%): Although not as heavily weighted as other statistics, free throw shooting proficiency can still be a factor in player evaluations, especially in clutch situations.

After this, I conducted a thorough check to identify any missing values within the dataset. Fortunately, no null values were detected, indicating a complete dataset. Additionally, I checked for any duplicate rows, and yet again found none, affirming the dataset's cleanliness. The dataset was very clean and no cleaning procedures were necessary, proving to be a solid foundation for data analysis.

Exploratory Data Analysis

In order to help gain insights for the underlying patterns and relationships within the data, I conducted Exploratory Analysis on the dataset. I calculated descriptive statistics for all the numerical variables (all of the columns except the player names), and was able to retrieve measures of central tendency, standard deviation, range, and distribution.



	PTS	AST	TRB	FG%	3P%	BLK	STL	MP	FT%	Made_All_Team
count	539.000000	539.000000	539.000000	539.000000	539.000000	539.000000	539.000000	539.000000	539.000000	539.000000
mean	9.101113	2.064193	3.535622	0.461039	0.313358	0.380705	0.608720	19.752319	0.718367	0.050093
std	6.846878	1.934273	2.344913	0.113918	0.140788	0.383923	0.399416	9.563098	0.215687	0.218339
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000
25%	4.150000	0.800000	1.800000	0.400000	0.300000	0.100000	0.300000	12.250000	0.700000	0.000000
50%	7.000000	1.400000	3.000000	0.500000	0.300000	0.300000	0.600000	19.300000	0.800000	0.000000
75%	12.150000	2.800000	4.500000	0.500000	0.400000	0.500000	0.800000	28.300000	0.800000	0.000000
max	33.100000	10.700000	12.500000	1.000000	1.000000	3.000000	3.000000	41.000000	1.000000	1.000000

Figure 1: Column Statistics

From graphing boxplots with the data, I was able to explore the distribution, variability and outliers within each feature of the dataset. The most valuable information I was able to retrieve was through detecting outliers. Since outliers are data points that deviate significantly from the rest of the dataset, these outliers may help distinguish exceptional performances from players.

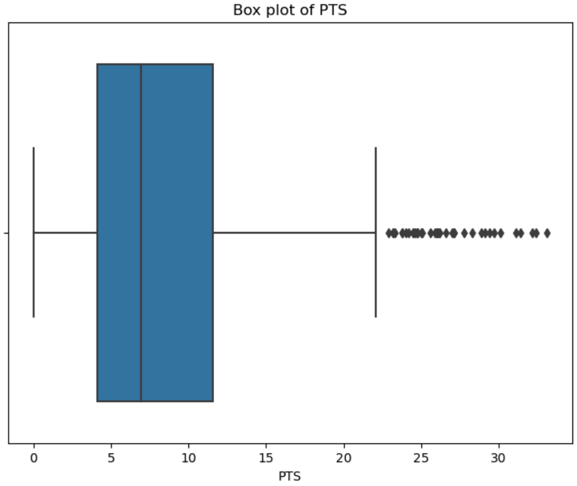


Figure 2: Boxplot of Points Per Game

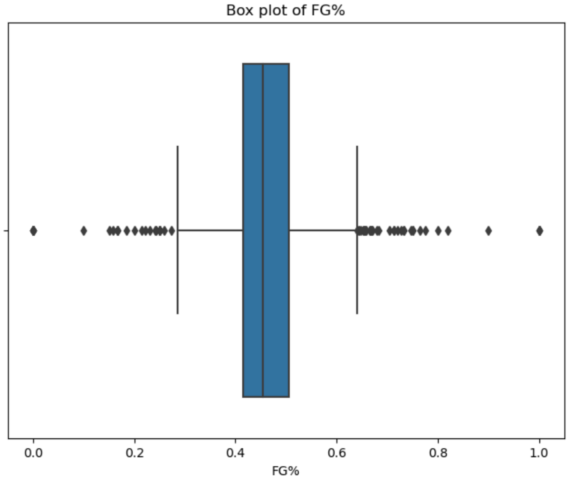
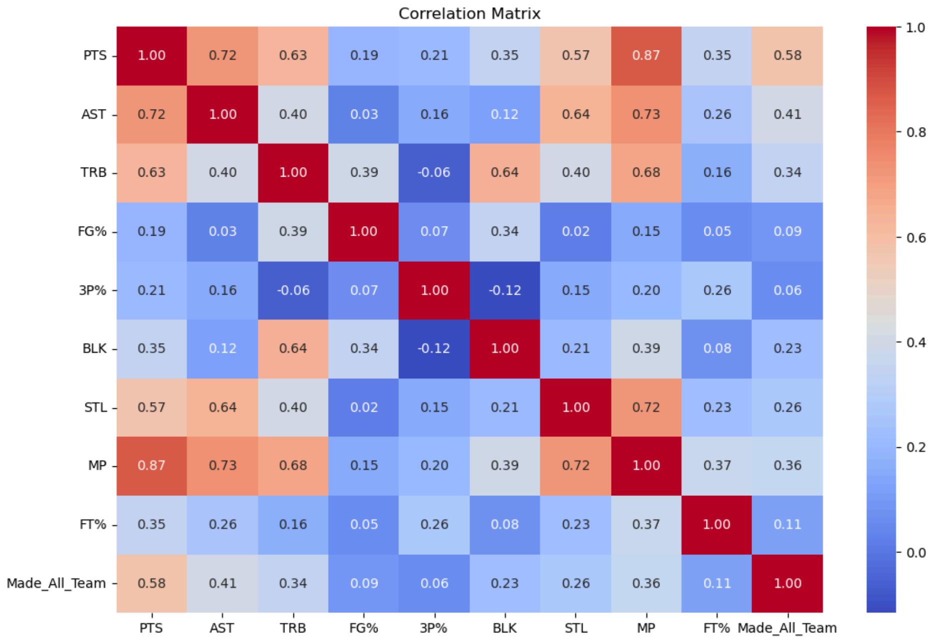


Figure 3: Boxplot of Field Goal Percentage

In both of these graphs above, they have many outliers that consist of higher values. An NBA player who consistently scores significantly higher points per game (PTS) or an exceptionally high field goal percentage (FG%) compared to other players may be considered an outlier. We also tend to see this with many of our other features in the dataset. With such high valued outliers, those features could be a key factor in All-NBA Team selection.

Lastly, I did a correlation heatmap matrix to look into the relationships and interactions between player performance metrics and All-NBA Team selections. Here are some of the stronger correlations I found:

- Minutes Played (MP) and Points per Game (PTS): This strong correlation of 0.87 suggests that players who log more minutes tend to score more points per game. It indicated that playing time is a significant factor in player scoring output.
- Minutes Played (MP) and Assists per Game (AST): Similarly, this correlation of 0.73 suggests that players who play more minutes tend to have more assists per game. Indicating that increased playing time provides players with more opportunities to help facilitate scoring opportunities for their teammates.
- Made All Team (‘Made_All_Team) and Points per Game (PTS): This correlation of 0.58 suggest that players selected for All-NBA Teams tend to score more points per game, implying that scoring proficiency yis a significant factored consisted by voters when selecting players for All-NBA honors.



	PTS	AST	TRB	FG%	3P%	BLK	STL	MP	FT%	Made_All_Team
PTS	1.00	0.72	0.63	0.19	0.21	0.35	0.57	0.87	0.35	0.58
AST	0.72	1.00	0.40	0.03	0.16	0.12	0.64	0.73	0.26	0.41
TRB	0.63	0.40	1.00	0.39	-0.06	0.64	0.40	0.68	0.16	0.34
FG%	0.19	0.03	0.39	1.00	0.07	0.34	0.02	0.15	0.05	0.09
3P%	0.21	0.16	-0.06	0.07	1.00	-0.12	0.15	0.20	0.26	0.06
BLK	0.35	0.12	0.64	0.34	-0.12	1.00	0.21	0.39	0.08	0.23
STL	0.57	0.64	0.40	0.02	0.15	0.21	1.00	0.72	0.23	0.26
MP	0.87	0.73	0.68	0.15	0.20	0.39	0.72	1.00	0.37	0.36
FT%	0.35	0.26	0.16	0.05	0.26	0.08	0.23	0.37	1.00	0.11
Made_All_Team	0.58	0.41	0.34	0.09	0.06	0.23	0.26	0.36	0.11	1.00

Figure 4: Correlation Matrix for 2022-203 NBA Player Stats

Model Selection

Since the primary objective of the project is to predict which basketball players will be chosen for the All-NBA Teams, modeling will help us develop predictive algorithms that can analyze player statistics and accurately forecast which players are most likely to receive this prestigious honor. The three models I chose to use to predict NBA All-Team selection were Logistic Regression, Random Forest, and Gradient Boosting Machine. Logistic Regression suits the binary classification of predicting whether a player will make the NBA All-Team or not, and provides a simple and interpretable model. Random Forest is a great ensemble learning method that works well with structured data like player statistics and works well without much hyper parameter tuning. Gradient Boosting Machines can capture complex relationships in the data and typically provide high predictive accuracy.

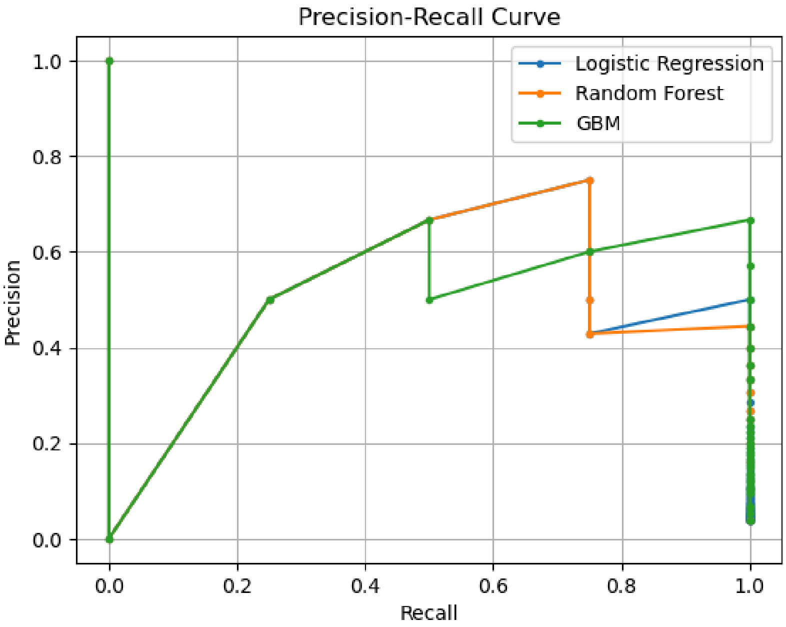


Figure 5: Precision Recall Curve Comparison

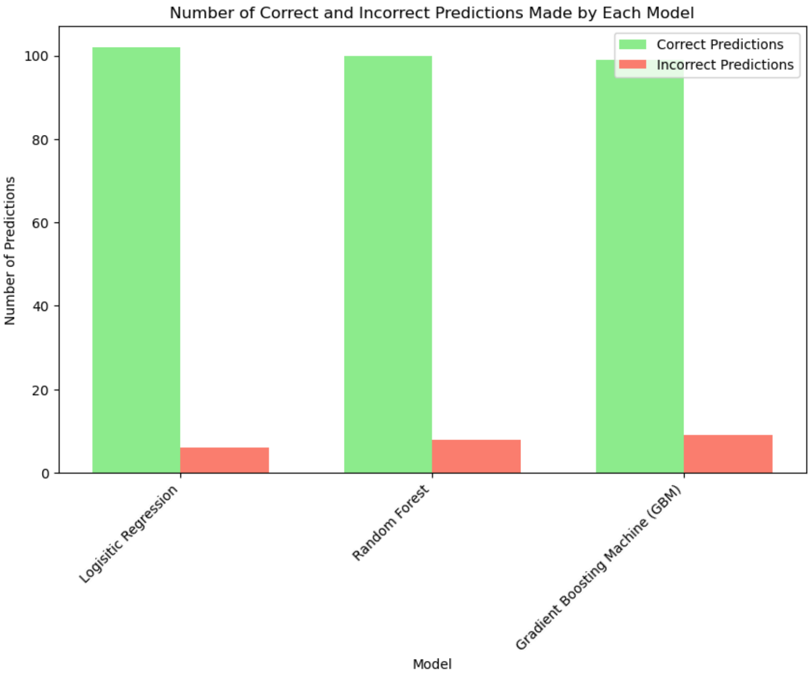


Figure 6: Correct vs Incorrect Selection per Model

When it came to the models, the Random Forest Classifier performed the best. The Random Forest model achieves the highest accuracy among the three models, indicating that it correctly predicts the All-NBA Team selections for a large proportion of instances. The precision and recall for the players selected for the All-NBA team are both high (0.75), suggesting that the model can effectively identify true positives while minimizing false positives. Random Forest models are also known for their robustness to overfitting, so they are less sensitive to outliers and noisy data compared to other models.

Conclusion

This study aimed at predicting All-NBA Team selections in basketball, where we evaluated various machine learning models to learn their effectiveness. Among the three models tested, the Random Forest Algorithm was the top performer, demonstrating the highest accuracy in identifying players chosen for the prestigious All-NBA Teams. Through our analysis, we uncovered key factors that influence team selections, highlighting player statistics such as points per game and field goal percentage as pivotal in predicting All-NBA Team placements.

While our models provide valuable insights, it's essential to acknowledge their limitations, including the need for additional data and the dynamic nature of player performance. Moving forward, we can further explore additional factors that impact All-NBA Team selections.These factors can enhance our ability to predict team compositions more accurately and create more advancements in basketball analytics. Overall, our data-driven approach contributes to the sophistication of basketball analysis, providing teams with actionable insights to perform strategic decision-making processes.