Michael D'Orazio

Data Mining 2

April 29, 2024

Rowan University

NBA All-Star Predictions Using Data Mining Techniques

1. Introduction

Each year, the National Basketball Association holds an event referred to as the NBA Draft. The NBA Draft involves the selection of prospective professional basketball players, primarily from college, international, and G-league basketball teams in more recent years. For every organization, the aim of the draft is to select the best players from the available pool of prospects, and ideally those whose strengths and abilities improve their team's overall success and stand as top performers. With a two rounds making up the selection process, a total of 60 players are picked by organizations on draft night, with players selected in the first round ideally being considered the cream of the crop. Historically, however, this does not always pan out as intended, and draft "busts" occur. A bust can be defined as a player that provides minimal impact to the team they are drafted for relative to the pick that was used to select them originally. Conversely, players selected with relatively low draft picks have historically developed into impressive athletes when no one ever expected them to by performing significantly above their projections or draft positions. In contrast to draft busts, these types of players are more commonly referred to as draft "booms".

Ideally, regardless of their draft position, teams wish to draft a boom over a bust in virtually every situation. A great way to measure whether or not an individual player has achieved a high level of performance-based distinction is via All-Star selection later on in their career. "All-Stars" are considered players that are representative of the league's top performers for a given year that they are awarded this title. Apart from the league MVP or Finals MVP award, an All-Star selection is among one of the highest individual distinctions that an NBA player can aspire to achieve.

1A. Statement of the Problem

As one would imagine, it is within the hopes of management, coaching staff, and fans, alike, that players selected with high draft picks develop and evolve over the course of their careers to an All-Star level of excellence, though this is not always the case, as a player's long-term potential is often difficult to predict. Therefore, the problem that was set out to be solved with this research was that of draft unpredictability. The aim was to leverage data mining techniques to determine whether or not it was possible to predict with a relatively high level of accuracy that arbitrary prospects had the potential to develop into NBA All-Stars based on shared pre-draft metrics and statistics that previous All-Stars had recorded before they entered the league.

1B. Why People Should Care About the Problem

Draft unpredictability is a problem that affects both fans and organizations alike. For fans, analysis of draft prospects is extremely beneficial, as it allows them to assess players in an upcoming draft to see if they are a good pick for their team or analyze their team's official pick after their draft selection is finalized.

For organizations, on the other hand, it is their job to assemble the best team possible in order to drive ticket and merchandise sales up in the presence of All-Star-level talent. For this reason, obtaining franchise-level talent is most important for organizations, as individuals' jobs and sometimes even their reputations depend upon their ability to make strong draft picks that will develop into successful players over the course of their careers.

Overall, however, this problem is one that primarily invokes the care of numerous parties, including NBA fans, coaching staff, and upper management, due to the time and potential career investment that rides on the success of their team. As the desire to acquire top talent is one that spans across numerous parties, it gives a degree of added weight to those who have a vested interest in a particular basketball team.

2. Methodology

In this section, both the approach to solving the problem outlined in the problem statement, reasons for why the design was chosen, and details of the process taken to carry out this research are expanded upon. To run these algorithms, I first had to obtain my dataset. I did this by combining multiple datasets harvested from data.world, NBA.com, basketballreference.com, and sports-reference.com. As a result, the dataset consists primarily of pre-draft combine metrics for players who became All-Stars and those who have not. It is a balanced dataset that consists of 604 players, with a ratio of 302 All-Stars to 302 non-All-Stars. In summary, the dataset I created for this research consists of columns representing a player's name, their classification as an All-Star or not, their position, distinction between whether or not they are a Backcourt player or Frontcourt player (more on this in **Section 2B**), height without shoes, height with shoes, wingspan, standing reach, max vertical, max vertical reach, no step vertical, no step vertical reach, weight, reach-to-height ratio, wingspan-to-height ratio, BMI, pre-draft points per game, pre-draft rebounds per game, pre-draft assists per game, and pre-draft total point value to quantify their impact on the game.

2A. Approach to Solving the Problem & Why the Design was Chosen

To solve the problem of NBA draft unpredictability, Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost algorithms were chosen. This approach was deemed appropriate for solving the problem of draft unpredictability as these types of data mining techniques are capable of making predictions on both a limited number of features and a larger number of features, and the desire to assess how predictions varied across models was something that was thought of early on. In other words, the design for this research was chosen based on the use of a variety of models. This was done in order to assess the overall performance

of a multitude of models in order to determine which worked best for solving the problem of draft unpredictability and assessing All-Star potential in prospective NBA players.

Also worth mentioning, a new metric was created to represent a player's impact on the game called "Total Point Value" which was represented as "TPV" in the dataset and analysis. TPV was calculated using the following equation:

$$TPV = PTS + 0.967REB + 2.32AST \quad \textbf{(1)}$$

As shown in **Equation (1)**, Total Point Value can be calculated by taking the sum of a player's Points Per Game, Rebounds Per Game multiplied by 0.967, and Assists Per Game multiplied by 2.32. For further context, the coefficients of 0.967 and 2.32 were not chosen at random and were instead derived from statistics stretching back to the past 15 NBA seasons. In order to derive these coefficients, quantities for Avg. FG% (field goal percentage), Avg. 3P% (3-point percentage), Avg. FGA (field goals attempted), Avg. 3PA (3-pointers attempted), Avg. 2PA (2-pointers attempted), 3PA% (percentage of 3-pointers attempted), and 2PA% (percentage of 2-pointers attempted) were calculated in order to determine the true value of a rebound and an assist. Since a rebound is the equivalent of obtaining a possession, its value was calculated via the following equation:

$$REB\_Value = 2(FG\% \; x \; 2PA\%) + 3(3P\% \; x \; 3PA\%) \quad \textbf{(2)}$$

In **Equation (2)**, the value of a rebound is based on the probability that a possession converts into points. As free throws are not considered to be a part of a possession (the period of time when a player is holding, dribbling, or passing the ball and the shot clock is active), the probabilities of a player or their teammate scoring only either a 2 or 3-pointer were considered to quantify this value. Since over the past 15 years, Avg. FG% was determined to be 0.459, Avg. 2PA% was determined to be 0.685, Avg. 3P% was determined to be 0.358, and Avg. 3PA% was determined to be 0.315, the value of a rebound was calculated to be 0.967 for the purpose of this research. Apart from determining the value of a rebound, the value of an assist was also calculated in order to fully flesh out the TPV equation. The value of an assist was calculated via the following equation as well:

$$AST\_Value = 2(2PA\%) + 3(3PA\%) \quad \textbf{(3)}$$

As assists represent an instance where a player passes the ball to a teammate that directly leads to them scoring, assists are treated as a guaranteed point conversion. Since free throws do not count as assists in the NBA, only 2 and 3-pointers were considered in this equation as only one of the two can be scored in a possession. With these considerations in mind, **Equation (3)** was developed accordingly. As a result, the value of an assist was calculated by taking the sum of the available scoring options, respectively multiplied by the odds that their teammate will attempt a shot which yields that particular value. The resultant assist value when following this methodology came out to be 2.32.

In addition to total point value (TVP), other variables were generated mathematically as well such as reach-to-height and wingspan-to-height ratio, though these are more general metrics

that have been used in sports statistics in the past and therefore did not require any derivation or advanced equations to implement ahead of time.

2B. Details of the Process

A key distinction that was made early on in the process was to acknowledge the stark difference between Backcourt and Frontcourt players. In basketball, Backcourt players are those who occupy the Point Guard and Shooting Guard positions. Historically, these players are generally shorter, and typically play beyond the arc to facilitate an outside presence on both offense and defense. As they are usually smaller than other players on the basketball court, they rely largely on their skill and athleticism to overcome size deficiencies.

Apart from Backcourt players, those representative of positions on the opposite end of the coin are referred to as Frontcourt players. Frontcourt players are those who occupy the Small Forward, Power Forward, and Center positions. These players are almost always some of the tallest players on the court and play well in the half-court where they are expected to both an offensive and defensive threat due to their imposing stature and strength.

Due to this distinction and the fact that different qualities make players good Frontcourt players that may not necessarily make them good Backcourt players, the Pandas library was used to create two separate dataframes for Backcourt and Frontcourt players, respectively. Using these dataframes, each model was trained separately for Backcourt and Frontcourt players to allow for the most effective predictions possible. After the models were fully trained, an assessment was done to determine the accuracy, precision, recall, and F1 scores of each model in order to rank them from best to worst.

Afterwards, data was created to represent prospective players and supplied to the models in order to make predictions and to see which were consistent across models. As mentioned in **Section 2A**, Logistic Regression and SVM were trained on fewer features when compared to Random Forest and XGBoost which were trained on all features in the Backcourt and Frontcourt dataframes. This distinction is worth mentioning because during the prediction phase, Logistic Regression and SVM only required a limited number of features to be supplied into them for prospective players when compared to Random Forest and XGBoost which required more features to be accounted for. **Sections 3** and **4** of this report further expand upon the results and conclusions made from following this process through graphs, tables, and an explanation of the analysis.

3. Results and Discussion

In this section, the research's results are fully explained. These results are supported by tables and graphs that serve to visually present findings in a way that fosters ease of understanding. This section in particular presents the majority of tables and figures along with brief explanations discussing what is shown and how it relates to the overall research itself.

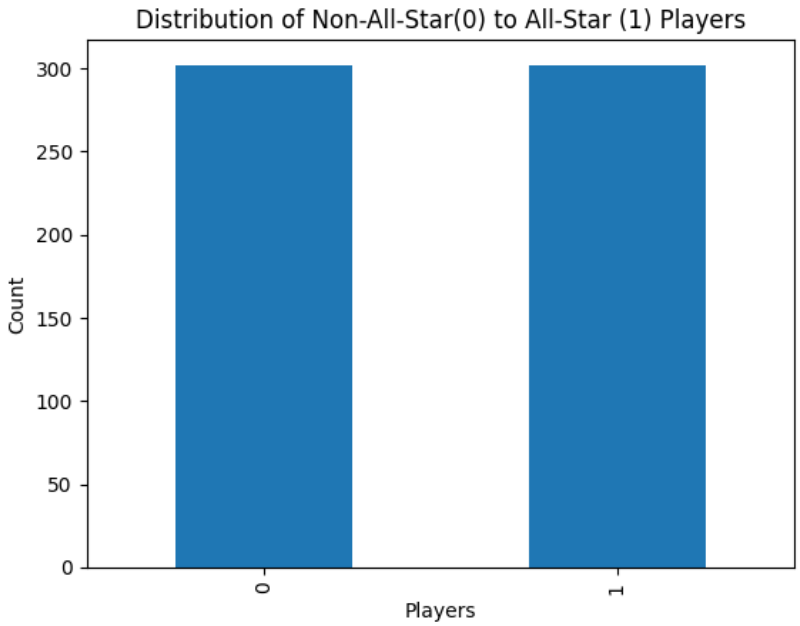

Fig. 1: Distribution of Non-All-Star to All-Star Players

Beginning with the first visualization in this section, **Fig. 1** was created to show the initial dataset's overall balance. It displays a bar chart used to illustrate the count of players in the Non-All-Star (0) and All-Star (1) classes. In total, there are 302 Non-All-Stars to 302 All-Stars in the initial dataset used for this research, thus illustrating its overall balance. For additional context, there exists a 1:1 ratio between individual player positions as well, meaning that there are 54 Non-All-Star Point Guards to 54 All-Star Point Guards, 65 Non-All-Star Shooting Guards to 65 All-Star Shooting Guards, 57 Non-All-Star Small Forwards to 57 All-Star Small Forwards, 74 Non-All-Star Power Forwards to 74 Non-All-Star Power Forwards, and 52 Non-All-Star Centers to 52 All-Star centers. Once the dataset was split based on Backcourt and Frontcourt distinction, the 1:1 ratio between the two classes was preserved with there being 119 Non-All-Star Backcourt players to 119 All-Star Backcourt players and 183 Non-All-Star Frontcourt players to 183 All-Star Frontcourt players. Due to the data being balanced, a technique like SMOTE did not need to be applied to rectify a class imbalance.

Beginning with Logistic Regression and SVM, the first step that was taken was to determine which two features presented the most divisible between All-Stars and Non-All-Stars so that a decision boundary could be formed between them. In order to do this, a pair plot was created for all features, and those that were the most separable were used to train these two

models. For Backcourt players, these features were PTS and Vertical_Max_Reach, whereas for Frontcourt players, these features were TPV and Vertical_Max_Reach.

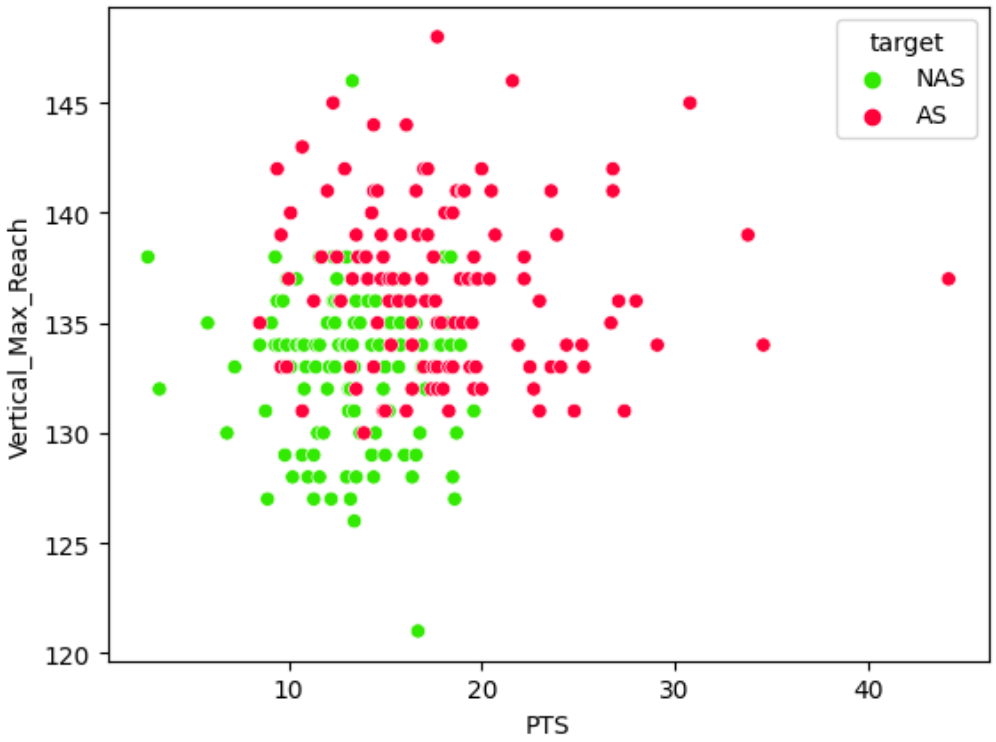

Fig. 2: Scatterplot of Backcourt Players by PTS & Vertical_Max_Reach

Fig. 2 displays a scatterplot of All-Star and Non-All-Star Backcourt players by PTS and Vertical_Max_Reach. PTS is a variable that represents how many points per game a player averaged before being drafted into the NBA, and Vertical_Max_Reach is a player's standing reach plus their max vertical jump. The reason this pair of features was chosen for Backcourt players is because they allowed for the most separation between variables in anticipation for a decision boundary. A viewer can visibly decipher that past a certain point, Non-All-Star points become All-Star points, making it possible to model these features further through Logistic Regression and SVM. As the separability of variables is a crucial factor for both Logistic Regression and SVM for effective classification, the variables that produced the most separable pair plot (PTS and Vertical_Max_Reach) were the ones that were selected. To help illustrate further why the attempt to plot the classifiers in the most separable way possible, decision boundaries were graphed based on the results of the Logistic Regression and SVM models in **Figs. 3** and **4**.

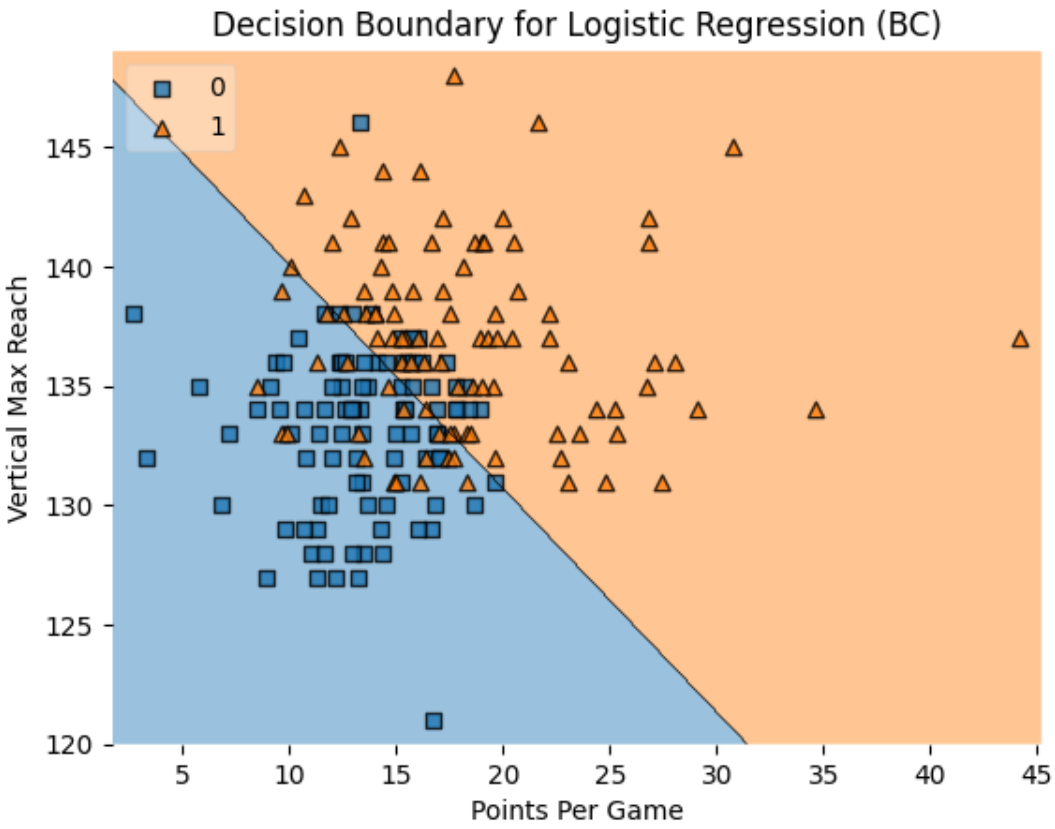

Fig. 3: Decision Boundary for Logistic Regression (BC)

Fig. 3 illustrates the decision boundary for Logistic Regression for Backcourt players. Similarly to **Fig. 2**, this chart displays Points Per Game on the X-axis and Vertical Max Reach on the Y-axis. Using both color and shape encoding to differentiate data points, orange triangles represent All-Star players (1), while blue squares represent players who had never been named an All-Star in their career (0). The decision boundary plots a conceptual line that was learned during the training process of the Logistic Regression model and in the case of this graph, it takes the shape of a line in 2-dimensional space. This boundary is important as it establishes where new data points will be classified depending on which side of the boundary line they fall. If they fall in the blue region, they will be classified as 0 (Non-All-Star), and if they fall in the orange region, they will be classified as 1 (All-Star). In analyzing the chart, the key takeaways are that if a prospective player averages a higher the number of points per game before entering the draft, the model will classify them as an All-Star even if their maximum vertical reach is relatively low to the rest of the competition. Conversely, if a player has a very high maximum vertical reach and a low scoring average, they will also be classified as an All-Star when the Logistic Regression model makes its prediction. In other words, the decision boundary helps viewers interpret how a model (in this case Logistic Regression) makes decisions and a provides a good generalization of its performance and behavior.

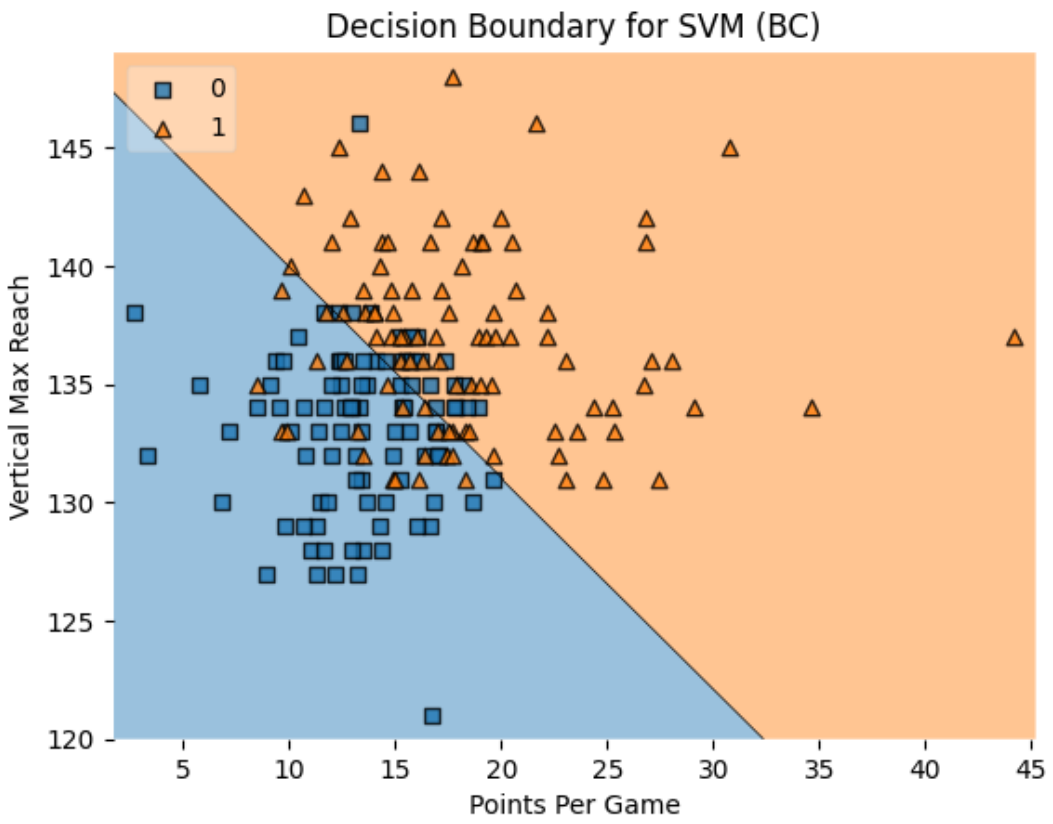

Fig. 4: Decision Boundary for SVM (BC)

Similarly to **Fig. 3**, **Fig. 4** illustrates the decision boundary for SVM for Backcourt players. Similarly to **Figs. 2** and **3**, this chart displays Points Per Game on the X-axis and Vertical Max Reach on the Y-axis. Though it appears almost identical to the decision boundary plotted in **Fig. 3**, a closer look shows that it is distinct in its own right. Though the difference is small, the decision boundary generated to represent SVM requires a slightly higher average points per game and a slightly lower maximum vertical reach to classify new data into either the All-Star (1) or Non-All-Star (0) regions.

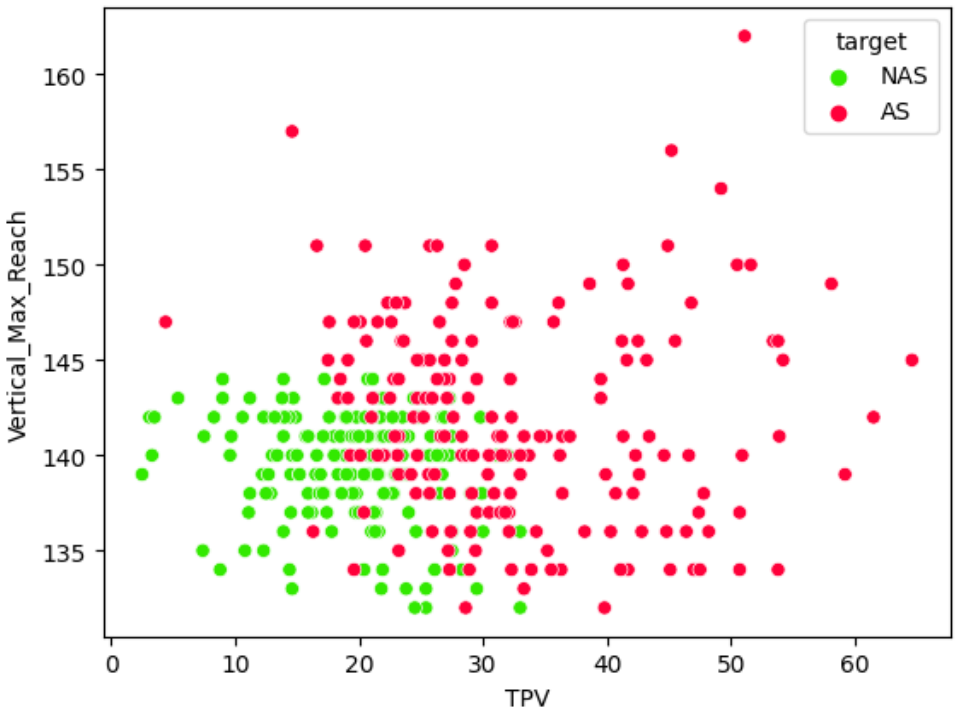

Fig. 5: Scatterplot of Frontcourt Players by TPV & Vertical_Max_Reach

Fig. 5 displays a scatterplot of All-Star and Non-All-Star Frontcourt players by TPV and Vertical_Max_Reach. TPV is a variable that represents the total point value that a player produced per game before being drafted into the NBA and, as previously established, Vertical_Max_Reach is a player's standing reach plus their max vertical jump. Similarly to **Fig. 2**, this scatterplot is shown so that a viewer can visibly imagine that a decision boundary could be formed which divides an All-Star and Non-All-Star region. The key difference between **Fig. 5** and **Fig. 2**, however, can be seen on the axis representing a player's maximum vertical reach. A quick look between the two shows that the higher maximum vertical reach measurements for Frontcourt players are significantly higher than those shown for Backcourt players. This is because as previously mentioned, Frontcourt players are larger individuals in turn, have a higher standing reach to begin with. Coupled with comparable vertical jumping ability, their overall maximum vertical reach is bound to eclipse values typically associated with Backcourt players to a noticeable degree. Even though this is only one visualization between two features, it helps to justify the decision to separate Backcourt and Frontcourt players into their own dataframes for model training purposes, as it hammers home the idea that certain features that make great Backcourt players may not necessarily be enough for them to excel in positions that Frontcourt players encompass. Additionally, it was interesting to see that total point value was among one of the variables that helped to best separate the classifiers. This may be due in part to the real-world function of Frontcourt players and their combined value to score, rebound, and occasionally assist by passing the ball out to an open perimeter player for a shot. In other words, as TPV is a combined statistic, it fits better with the overall function of Frontcourt players than individual points, rebounds, or assists per game statistics typically do.

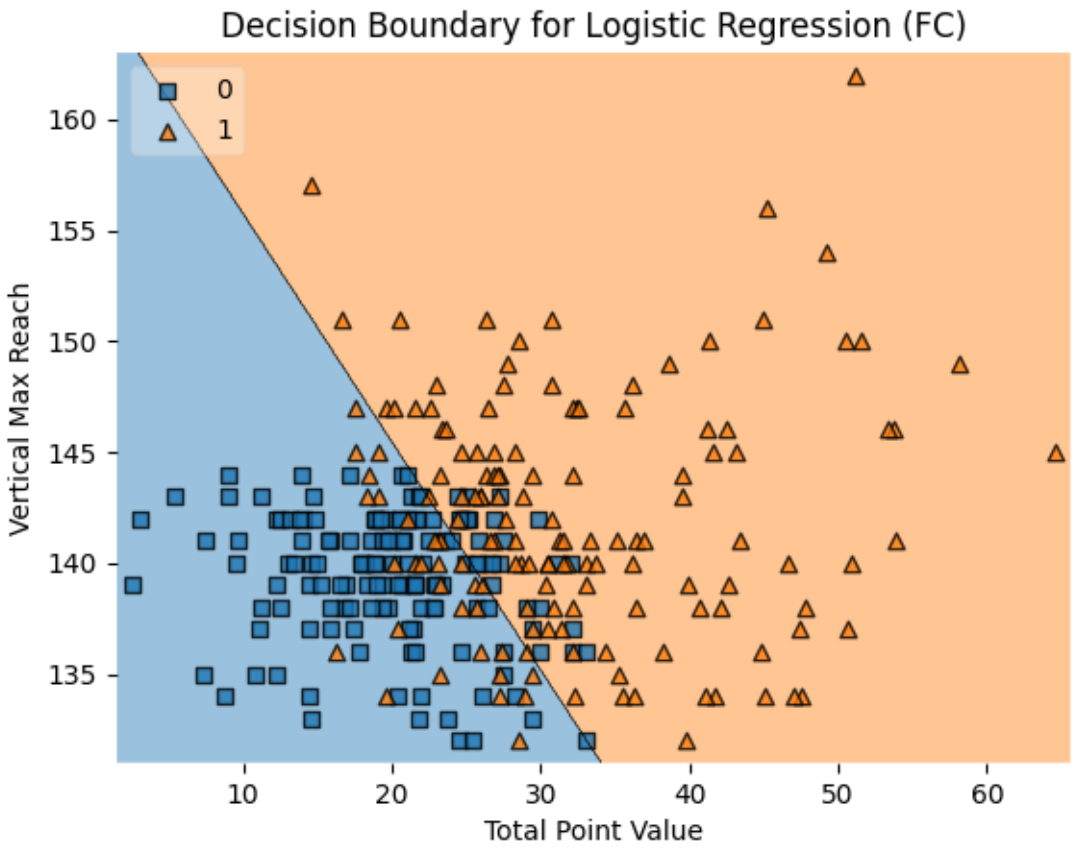

Fig. 6: Decision Boundary for Logistic Regression (FC)

Fig. 6 illustrates the decision boundary for Logistic Regression for Frontcourt players. As previously mentioned, Frontcourt players are generally taller, meaning that greater height is a variable that most sports analysts believe plays greatly into the overall effectiveness of a Small Forward, Power Forward, or Center in the NBA. This greater height leads to a proportionally greater maximum vertical reach since their standing reach is usually elevated to match their increased height. In viewing the chart, one can see that the decision boundary for Logistic Regression slightly favors maximum vertical reach over total point value. Nevertheless, this chart makes it easy to distinguish that, in general, players who have greater standing reaches, possess a high maximum vertical jump, and have a strong positive impact on the game are more likely to develop into All-Stars than those who do not possess such qualities.

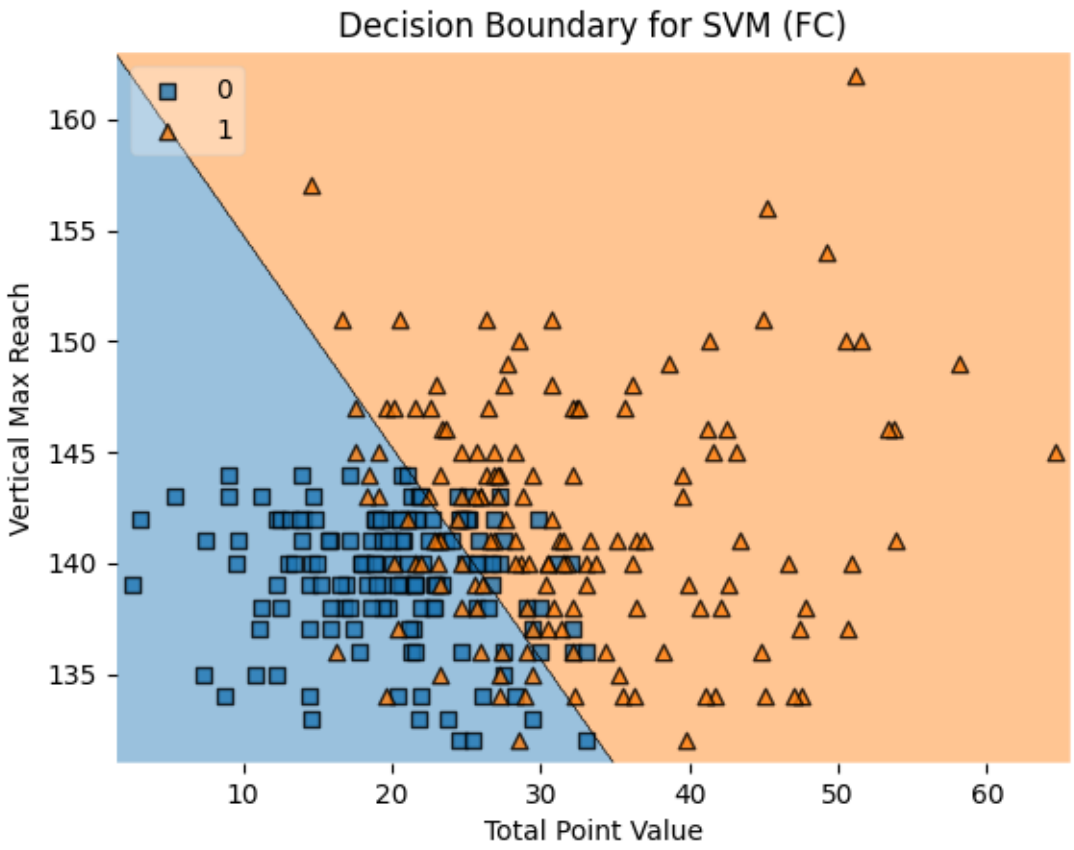

Fig. 7: Decision Boundary for SVM (FC)

Fig. 7 illustrates the decision boundary for SVM for Frontcourt players. As previously mentioned, Frontcourt players are generally taller, meaning that greater height is a variable that most sports analysts believe plays greatly into the overall effectiveness of a Small Forward, Power Forward, or Center in the NBA. This greater height leads to a proportionally greater maximum vertical reach since their standing reach is usually elevated to match their increased height. In viewing the chart, one can see that the decision boundary SVM slightly favors total point value over maximum vertical reach. This is in contrast to **Fig. 6** which slightly favors maximum vertical reach as a distinguishing variable to base the boundary off of. Nevertheless, this chart makes it easy to distinguish that, in general, players who have greater standing reaches, possess a high maximum vertical jump, and have a strong positive impact on the game are more likely to develop into All-Stars than those who do not possess such qualities.

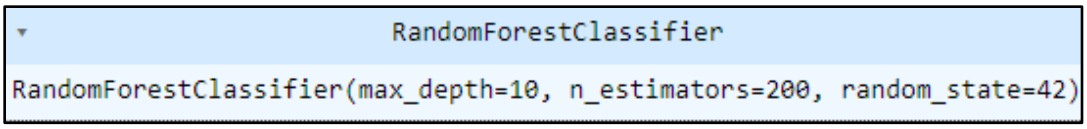

Fig. 8: Details of Random Forest Classifier

For full disclosure into the specifics of the Random Forest model, **Fig. 8** is shown. These are the specific details that the Random Forest model was tuned to. It possesses a max depth of 10, 200 decision trees, and a random state of 42.

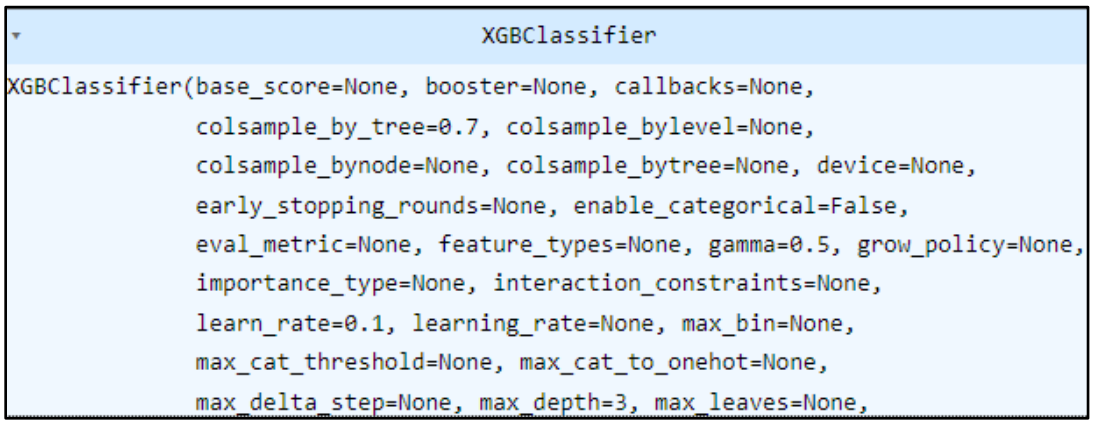

Fig. 9: Details of Random Forest Classifier

Additionally, for full disclosure of the specifics of the XGBoost model, **Fig. 9** is shown. These are the specific details that the XGBoost model was tuned to. A Grid Search was run to determine the best parameters. It possesses a learn rate of 0.1, a max depth of 3, a column subsampling by tree of 0.7, and a gamma of 0.5. Though not shown in **Fig. 9,** the seed value for the XGBoost model is 42.

Table 1: Backcourt Player Prospects

Player	HNS	HS	WS	SR	VM	VMR	VNS	VNSR	W	R:HS	WS:HS	BMI	PTS	REB	AST	TPV
1	72	73	75	97	33	130	27	124	189	1.33	1.03	24.9	8	2	3	16.9
2	76	77	83	102	38	140	32	134	205	1.4	1.08	24.3	22	4	8	44.4

Table 1 represents the prospective players that would be used to test the models representing Backcourt players. **Player 1** represents an individual that possesses a height of 72-inches with no shoes, 73-inches with shoes (6'1), a 75-inch wingspan, a 97-inch standing reach, a 33-inch maximum vertical jump, a 130-inch maximum vertical reach, a 27-inch no step vertical jump, a 124-inch no step vertical reach, a 189-pound weight, a 1.33 reach-to-height ratio, a 1.03 wingspan-to-height ratio, a BMI of 24.9, and who averaged 8 points, 2 rebounds, and 3 assists per game, giving them a total point value of 16.9. On the other hand, **Player 2** represents an individual that possesses a height of 76-inches with no shoes, 77-inches with shoes (6'5), an 83-inch wingspan, a 102-inch standing reach, a 38-inch maximum vertical jump, a 140-inch maximum vertical reach, a 32-inch no step vertical jump, a 134-inch no step vertical reach, a 205-pound weight, a 1.4 reach-to-height ratio, a 1.08 wingspan-to-height ratio, a BMI of 24.3, and who averaged 22 points, 4 rebounds, and 8 assists per game, giving them a total point value of 44.4. **Table 1** was displayed in this report to provide context into how predictions for prospective players were made and which variables the models consider in order to determine whether or not a particular player may develop into an All-Star or not.

Table 2: Frontcourt Player Prospects

Player	HNS	HS	WS	SR	VM	VMR	VNS	VNSR	W	R:HS	WS:HS	BMI	PTS	REB	AST	TPV
3	76	77	80	105	33	141	28	136	220	1.4	1.08	26.1	12	8	2	24.4
4	81	82	86	110	37	147	33	143	210	1.34	1.05	22	18	7	4	34

Table 2 represents the prospective players that would be used to test the models representing Frontcourt players. **Player 3** represents an individual that possesses a height of 76-inches with no shoes, 77-inches with shoes (6'5), an 80-inch wingspan, a 105-inch standing reach, a 33-inch maximum vertical jump, a 141-inch maximum vertical reach, a 28-inch no step vertical jump, a 136-inch no step vertical reach, a 220-pound weight, a 1.4 reach-to-height ratio, a 1.08 wingspan-to-height ratio, a BMI of 26.1, and who averaged 12 points, 8 rebounds, and 2 assists per game, giving them a total point value of 24.4. On the other hand, **Player 4** represents an individual that possesses a height of 81-inches with no shoes, 82-inches with shoes (6'10), an 86-inch wingspan, a 110-inch standing reach, a 37-inch maximum vertical jump, a 147-inch maximum vertical reach, a 33-inch no step vertical jump, a 143-inch no step vertical reach, a 210-pound weight, a 1.34 reach-to-height ratio, a 1.05 wingspan-to-height ratio, a BMI of 22, and who averaged 18 points, 7 rebounds, and 4 assists per game, giving them a total point value of 34. Similarly to **Table 1**, **Table 2** was also displayed in this report to provide context into the player data that was supplied to the trained models. These metrics were kept consistent throughout, although for Logistic Regression and SVM, only a limited number of features were needed, while for Random Forest and XGBoost, all features had an impact on predictions.

Table 3: Model Scores for Backcourt Players

Model	Accuracy	Precision	Recall	F1 Score
LR	83.33%	85.71%	78.26%	81.82%
SVM	81.25%	85.0%	73.91%	79.07%
Random Forest	89.58%	86.36%	90.48%	88.37%
XGBoost	91.67%	96.0%	88.89%	92.31%

Table 3 shows the results for accuracy, precision, recall, and F1 score for each model trained on Backcourt player data. In terms of overall ranking of accuracy from best to worst, XGBoost leads the pack, followed by Random Forest, Logistic Regression, and SVM. From these results, one can see that XGBoost is likely the best model to use to predict whether or not Backcourt players may develop into NBA All-Stars when compared to the others. However, Random Forest is still an exceptionally strong candidate model to use as well.

Table 4: Model Scores for Frontcourt Players

Model	Accuracy	Precision	Recall	F1 Score
LR	87.84%	88.98%	86.49%	87.67%
SVM	86.49%	88.57%	83.78%	86.11%
Random Forest	90.54%	94.44%	87.18%	90.67%
XGBoost	90.54%	89.74%	92.11%	90.91%

Table 4 shows the results for accuracy, precision, recall, and F1 score for each model trained on Frontcourt player data. In terms of overall ranking of accuracy from best to worst, XGBoost and Random Forest are roughly tied for first place, followed by Logistic Regression and SVM. In the case of Frontcourt players, it is interesting to see that Random Forest and XGBoost are so relative to one another, with the only real difference being a higher precision for Random Forest and a higher Recall for XGBoost.

Following the model scoring process, predictions were made using the prospective player data in order to determine which predictions were more consistent or variable. **Table 5**, below, shows the results of all model predictions for players **1** through **4**.

Table 5: Prediction Results for Players 1-4

Player	LR	SVM	Random Forest	XGBoost
1	0	0	0	0
2	1	1	1	1
3	1	0	0	0
4	1	1	1	1

As mentioned, **Table 5** shows the representative results for players **1** through **4** showing how the Logistic Regression, SVM, Random Forest, and XGBoost models predicted whether or not they would develop into an All-Star. In this table, 0's indicate that a model predicts that a player will not become an All-Star, while 1's indicate that a model predicts that a player will become an All-Star. The majority of the models predicted the same results, however, Logistic Regression had a different opinion on **Player 3** than the others did, as it predicted that they could develop into an All-Star where the others did not. It is worth noting, however, that both the Logistic Regression and SVM models were trained using limited features, therefore, their predictions may not be as accurate as others that take more features into account, such as Random Forest and XGBoost. As both Logistic Regression and SVM's exhibited comparatively lower accuracy, precision, recall, and F1 scores when compared to Random Forest and XGBoost, a key takeaway may be that more features beyond only two may lead to better model performance for the particular problem of unpredictability and assessing All-Star potential in prospective NBA players.

4. Conclusions

The conclusions of this research are that the XGBoost and Random Forest models performed the best when it came to accuracy, precision, recall, and F1 scores. As a result, these models would be the best to use for the purpose of All-Star prediction of prospective basketball players. While Logistic Regression and SVM were also evaluated at moderately high levels given the limited features that went into their creation, they are not nearly as effective as the XGBoost and Random Forest models were in the end. For Backcourt players, XGBoost exceeded expectations, obtaining a 91.67% accuracy score. For Frontcourt players, both XGBoost and Random Forest obtained a 90.54% accuracy score. Both formed consistent predictions that indicated that the models were of similar quality to one another. As a result, it was concluded that the best models to use for the prediction of future NBA All-Stars were XGBoost and Random Forest.

4A. What is Original About this Research

This research's originality owes itself to a multitude of areas. For starters, attempting to solve the problem of draft predictability appeared to be a rather novel application of data mining techniques and machine learning algorithms in and of themselves, as datasets tailored to this specific purpose are virtually nonexistent, at least on the public level. However, assuming that NBA organizations have data science teams working on similar projects related to draft predictability, it is likely that similar research has been performed in the past and continues to be done year after year to optimize draft picks. Nevertheless, the dataset created for this research was wholly original and based on authentic data consisting of pre-draft metrics and statistics, as previously mentioned in **Section 2** of this report.

Additionally, the implementation of the total point value (TPV) statistic is something that was created novelly for the purpose of this research, therefore, it is the first of its kind as it implements a new statistic when running classification models.

4B. Further Conclusions, Lessons Learned, and Overall Implications

Further conclusions made from this research are that while more quantifiable features aid in the creation of models used to predict NBA All-Stars, it is important to note which features are most important developmentally. While the dataframes used in this research contained a total of 16 features, it is important to consider the possibility that other features may exist, which could make the models even more accurate. Statistics relating to speed, agility, and endurance, for example, may have altered results if such statistics could be obtained or were publicly available for every player in the dataset. In a similar vein, there also exist many unquantifiable variables in sports that no model can predict. These include simple instances like whether or not a player will experience injuries early on in their career or even the level of their overall work-ethic and drive to be great. Therefore, a lesson learned is that as one continues to research a topic, they can continuously think of ways to make it better as they go.

Nevertheless, as the models trained in this research were based primarily on physical attributes and historical statistics, they were able to make predictions based on those features to a high-level of accuracy. The overall implications of research like this could revolutionize how NBA teams draft players going forward. It could potentially save a lot of time, generate greater revenue for organizations, and enhance the value of higher ranked draft picks. If research like this grew to a point where the real-world results matched up with the level of accuracy that these models display, then they could be highly used throughout basketball in the future by fans, coaches, and management, alike.