NBA ALL-STAR PREDICTIONS USING DATA MINING TECHNIQUES

BY MICHAEL D'ORAZIO

BACKGROUND – THE NBA DRAFT

- Each year, the National Basketball Association holds an event referred to as the NBA Draft.
- The draft involves the selection of 60 prospective professional basketball players over the course of two rounds, primarily from college, international, and G-league basketball teams.
- For every organization, the aim of the draft is to select the best players possible.
- Ideally at least one whose abilities improve their team's overall success and develops into an NBA All-Star.

BACKGROUND – WHAT IS AN ALL-STAR?

- "All-Stars" are 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 a professional basketball player can aspire to achieve.
- In other words, All-Stars can be considered some of the best players in throughout the entire league.

BACKGROUND – FROM PROSPECT TO ALL-STAR?

- While the NBA Draft is a straightforward process, its results have historically been unpredictable.
- Draft "busts" and "booms" have occurred many times over, leading to players being selected either higher or lower than they likely should have been.
- Since some of the strengths of data mining techniques involve prediction, wouldn't it be great if we could use them to make things a little bit more predictable?

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.
- The problem is, that since a player's long-term potential is often difficult to predict, this is not always the case.
- Therefore, the problem I set out to solve with this research was that of draft unpredictability.
- My aim was to leverage data mining techniques to determine if it was possible to predict within a high level of accuracy that NBA prospects had the potential to develop into future All-Stars.

WHY PEOPLE SHOULD CARE ABOUT THE PROBLEM

- Draft unpredictability is an issue that affects both fans and organizations.
- For fans, analysis of draft prospects is beneficial as it allows them to assess players in an upcoming draft to see if they are a good pick for their team.
- For coaching staff and organizations, it is their job to assemble the best team possible and the presence of All-Star-level talent helps them to do so.
- In other words, people should care about this problem if they have a vested interest in basketball and want to see their favorite teams succeed.

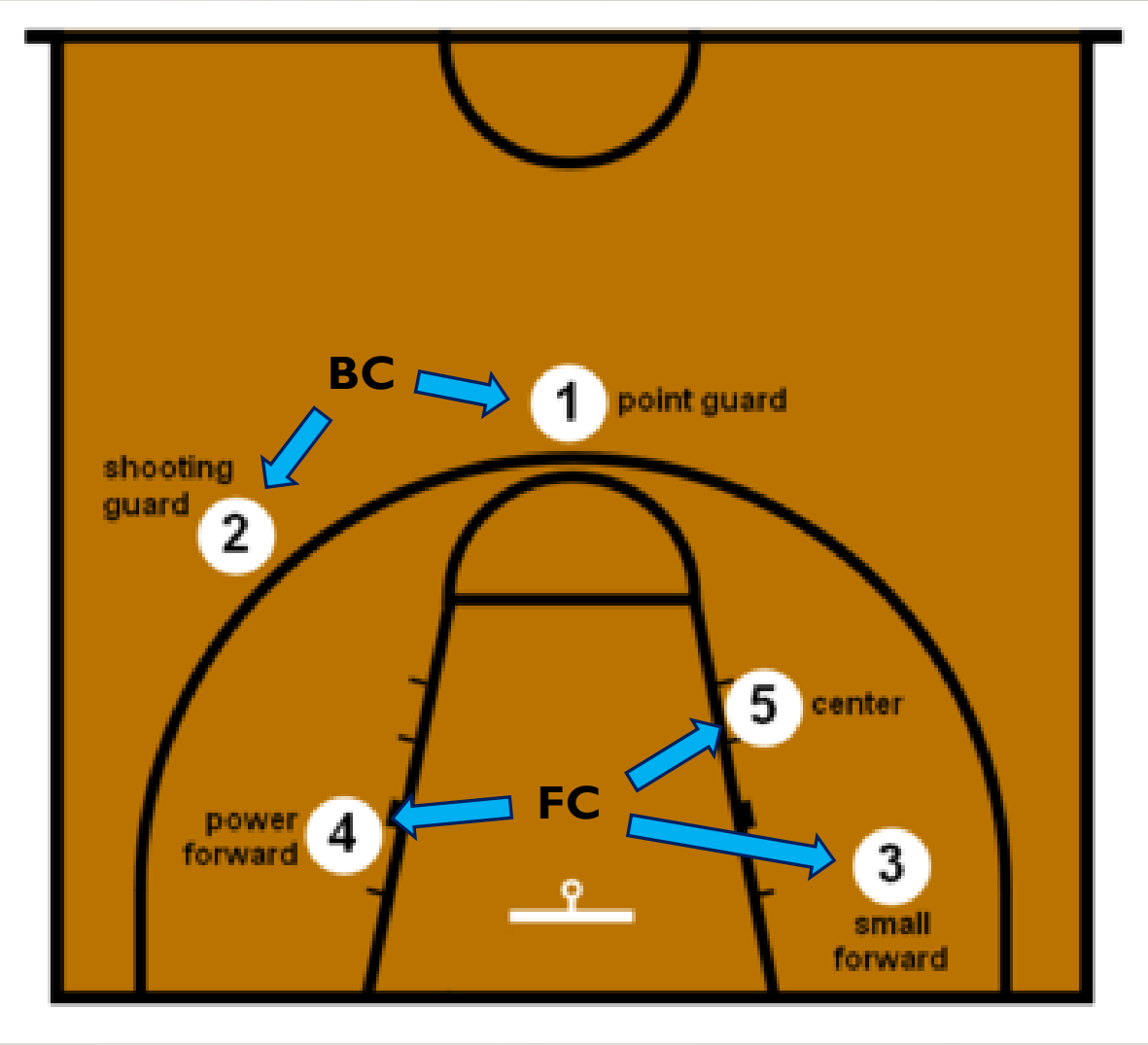
MY APPROACH TO SOLVING THE PROBLEM

- My approach to solving this problem involved leveraging data mining algorithms/techniques for the purpose of prediction.
- The algorithms I used in this project included Logistic Regression, SVM, Random Forest, and XGBoost.
- Before training the models, however, I made an important distinction between "Backcourt" and "Frontcourt" players that was crucial to my analysis.

MY APPROACH - BACKCOURT VS FRONTCOURT

- In the sport of basketball, there are a total of 5 positions: Point Guard, Shooting Guard, Small Forward, Power Forward, and Center.
- Backcourt players consist of Point Guards and Shooting Guards.
  - They are generally smaller and more skilled with the ability to play well beyond the arc.
- Frontcourt players consist of Small Forwards, Power Forwards, and Centers.
  - They are generally some of the tallest and most physically imposing players on the court.

MY APPROACH - BACKCOURT VS FRONTCOURT (CONTINUED)



Positions in Basketball [1]

MY APPROACH – THE DATA

- My dataset was created through the combination of multiple datasets harvested from data.world, NBA.com, basketballreference.com, and sports-reference.com.
- It is a balanced dataset that consists of 604 players, with a ratio of 302 All-Stars to 302 Non-All stars in equivalent proportions for every position.
- In its original form, it consists of columns representing a player's name, their All-Star classification, their position, distinction between whether or not they are a Backcourt player or Frontcourt player, 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.

MY APPROACH - TOTAL POINT VALUE (TPV)

							Per Game															Shooting		
Rk	Season	Lg	Age	Ht	Wt	G	MP	FG	FGA	3P	3PA	FT	FTA	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	FG%	3P%	FT%
1	2023-24	NBA	26.4	6-7	216	1230	241.4	42.2	88.9	12.8	35.1	17.0	21.7	10.6	33.0	43.5	26.7	7.5	5.1	13.6	18.7	114.2	.474	.366	.784
2	2022-23	NBA	26.1	6-6	216	1230	241.8	42.0	88.3	12.3	34.2	18.4	23.5	10.4	33.0	43.4	25.3	7.3	4.7	14.1	20.0	114.7	.475	.361	.782
3	2021-22	NBA	26.1	6-6	215	1230	241.4	40.6	88.1	12.4	35.2	16.9	21.9	10.3	34.1	44.5	24.6	7.6	4.7	13.8	19.6	110.6	.461	.354	.775
4	2020-21	NBA	26.1	6-6	217	1080	241.4	41.2	88.4	12.7	34.6	17.0	21.8	9.8	34.5	44.3	24.8	7.6	4.9	13.8	19.3	112.1	.466	.367	.778
5	2019-20	NBA	26.1	6-6	218	1059	241.8	40.9	88.8	12.2	34.1	17.9	23.1	10.1	34.8	44.8	24.4	7.6	4.9	14.5	20.8	111.8	.460	.358	.773
6	2018-19	NBA	26.3	6-6	219	1230	241.6	41.1	89.2	11.4	32.0	17.7	23.1	10.3	34.8	45.2	24.6	7.6	5.0	14.1	20.9	111.2	.461	.355	.766
7	2017-18	NBA	26.4	6-7	220	1230	241.4	39.6	86.1	10.5	29.0	16.6	21.7	9.7	33.8	43.5	23.2	7.7	4.8	14.3	19.9	106.3	.460	.362	.767
8	2016-17	NBA	26.6	6-7	221	1230	241.6	39.0	85.4	9.7	27.0	17.8	23.1	10.1	33.4	43.5	22.6	7.7	4.7	14.0	19.9	105.6	.457	.358	.772
9	2015-16	NBA	26.7	6-7	222	1230	241.8	38.2	84.6	8.5	24.1	17.7	23.4	10.4	33.3	43.8	22.3	7.8	5.0	14.4	20.3	102.7	.452	.354	.757
10	2014-15	NBA	26.7	6-7	223	1230	242.0	37.5	83.6	7.8	22.4	17.1	22.8	10.9	32.4	43.3	22.0	7.7	4.8	14.4	20.2	100.0	.449	.350	.750
11	2013-14	NBA	26.5	6-7	223	1230	242.0	37.7	83.0	7.7	21.5	17.8	23.6	10.9	31.8	42.7	22.0	7.7	4.7	14.6	20.7	101.0	.454	.360	.756
12	2012-13	NBA	26.7	6-7	223	1229	241.9	37.1	82.0	7.2	20.0	16.7	22.2	11.2	31.0	42.1	22.1	7.8	5.1	14.6	19.8	98.1	.453	.359	.753
13	2011-12	NBA	26.6	6-7	223	990	241.9	36.5	81.4	6.4	18.4	16.9	22.5	11.4	30.8	42.2	21.0	7.7	5.1	14.6	19.6	96.3	.448	.349	.752
14	2010-11	NBA	26.6	6-7	223	1230	241.9	37.2	81.2	6.5	18.0	18.6	24.4	10.9	30.5	41.4	21.5	7.3	4.9	14.3	20.7	99.6	.459	.358	.763
15	2009-10	NBA	26.6	6-7	222	1230	241.7	37.7	81.7	6.4	18.1	18.6	24.5	11.0	30.8	41.7	21.2	7.2	4.9	14.2	20.9	100.4	.461	.355	.759

Average NBA Stats Per Game for the Past 15 Seasons [2]

- Avg. FG% = 0.459
- Avg. 3P% = 0.358
- Avg. FGA = 85.4
- Avg. 3PA = 26.9
- Avg. 2PA = 85.4 – 26.9 = 58.5
- 3PA% = 26.9 / 85.4 = 0.315
- 2PA% = 58.5 / 85.4 = 0.685
- REB_Value = 2(0.459 x 0.685) + 3(0.358 x 0.315) = 0.967
- AST_Value = 2(0.685) + 3(0.315) = 2.32
- TPV = PTS + (REB_Value x REB) + (AST_Value x AST)
- TPV = PTS + 0.967REB + 2.32AST

WHY I CHOSE MY DESIGN

- I chose Logistic Regression, SVM, Random Forest, and XGBoost for two main reasons.

1) These four models are well-known for predictive purposes.

2) I wanted to use an even split of models that considered a limited number of features(Logistic Regression and SVM) and a larger number of features (Random Forest and XGBoost).

- Given these details, I figured that trying a variety of models with differing levels of inputs could help broaden my analysis.
- I also realized that since I had to essentially run the models twice (once each for Backcourt and Frontcourt players), the results would be even more insightful.

DETAILS OF THE PROCESS

1) To train the Linear Regression, SVM, Random Forest, and XGBoost models, separate dataframes representing Backcourt and Frontcourt players were created.

2) For Linear Regression and SVM where a limited number of features were required, a pairplot was made in Python to determine what the most separable features were to use for these models.

3) For Random Forest and XGBoost, however, all features were used.

4) A small dataset of four prospective players was generated to use for predictions.

5) Accuracy, Precision, Recall, and F1 Scores were obtained for each model.

6) Afterwards, the player data was used, and predictions were assessed.

DETAILS – LR & SVM SCATTERPLOTS (BC)

- **Fig. 1** displays a scatterplot of All-Star and Non-All-Star Backcourt players by PTS and Vertical_Max_Reach.
- It was obtained from the aforementioned pairplot due to PTS and Vertical_Max_Reach being the two variables that made the class data the most separable.
- It also yielded the highest accuracy for Logistic Regression & SVM models out of all feature pairs, but more on that later.

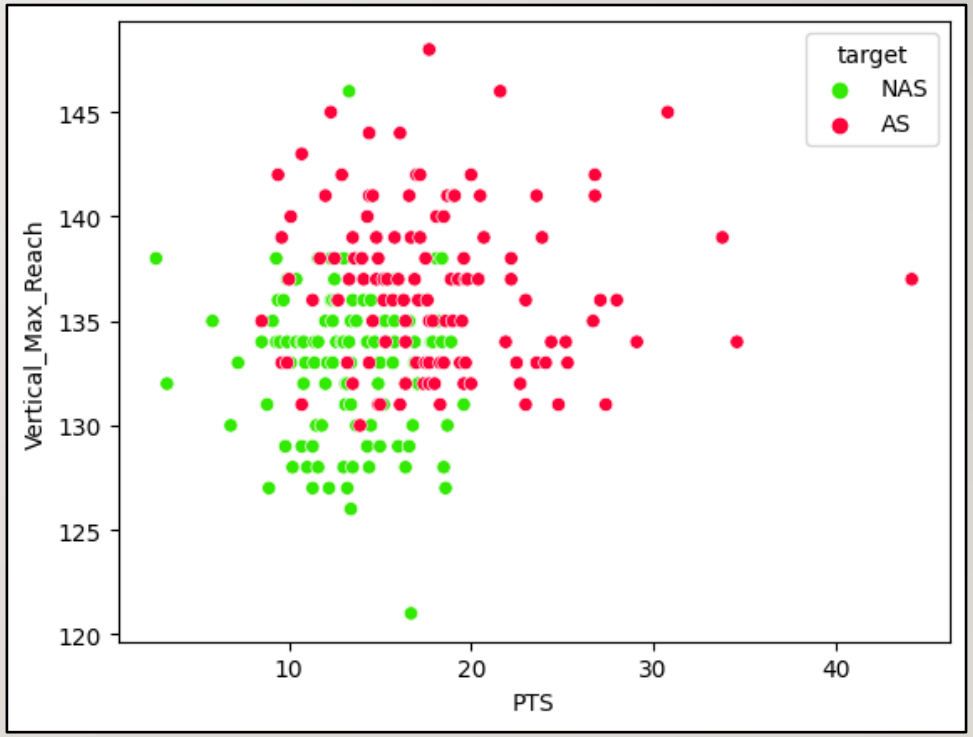


Fig. 1: Scatterplot of Backcourt Players by PTS & Vertical_Max_Reach

DETAILS – LR & SVM DECISION BOUNDARIES (BC)

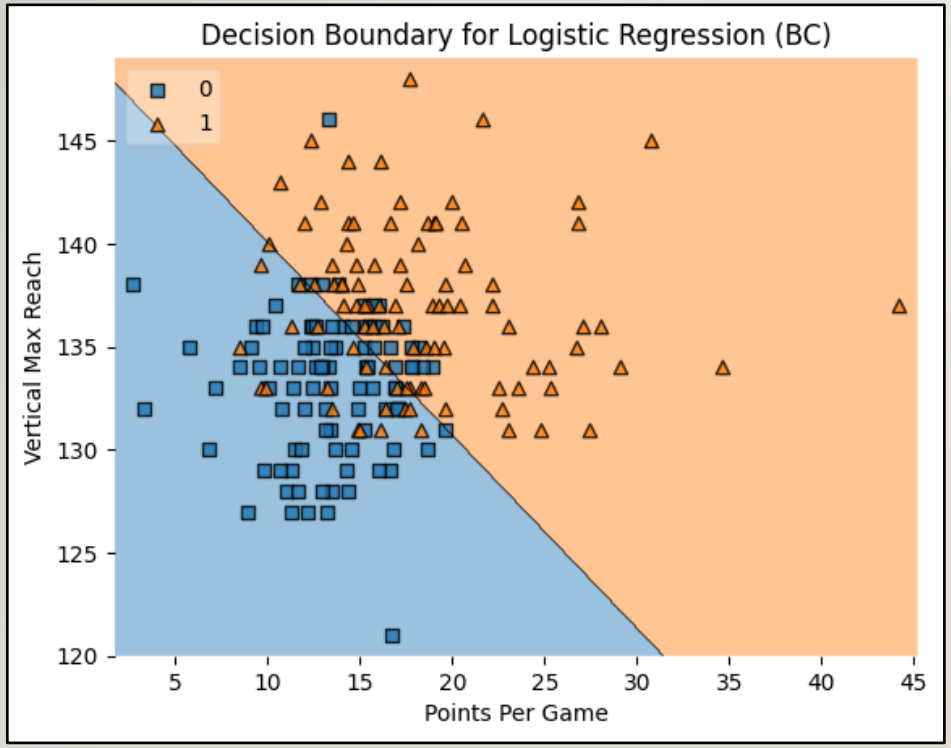


Fig. 2: Decision Boundary for Logistic Regression (BC)

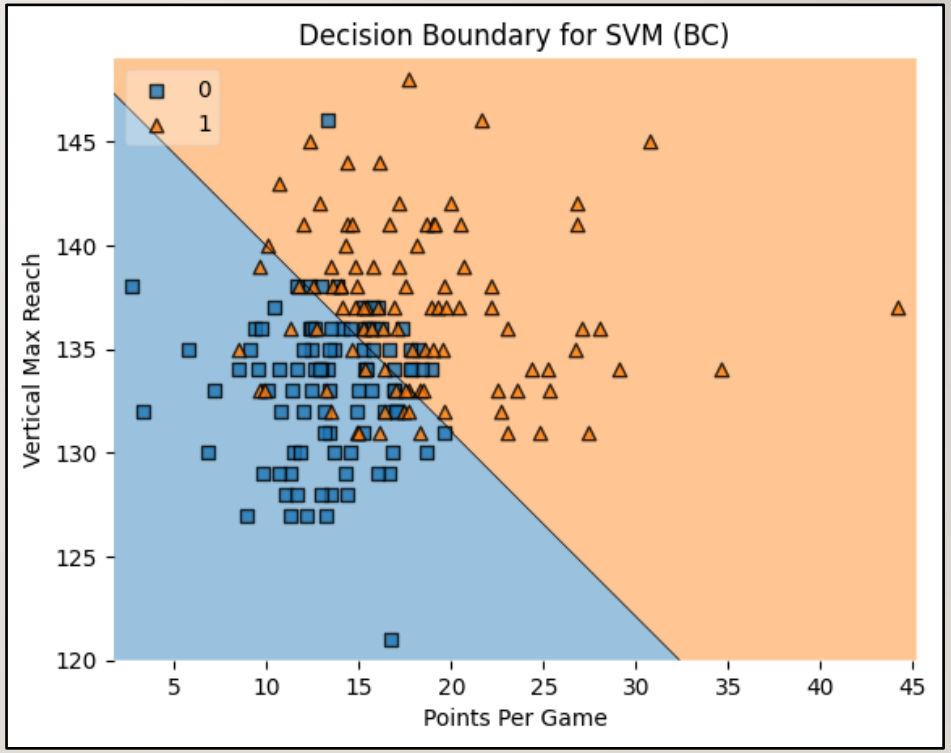


Fig. 3: Decision Boundary for SVM (BC)

DETAILS – LR & SVM SCATTERPLOTS (FC)

- **Fig. 4** displays a scatterplot of All-Star and Non-All-Star Frontcourt players by TPV and Vertical_Max_Reach.
- It was also obtained from the aforementioned pairplot due to TPV and Vertical_Max_Reach being the two variables that made the class data the most separable, similarly to **Fig. 1**.
- This makes sense as the role of Frontcourt players is more all-around, so a combined statistic is more representative of their expected skillset.

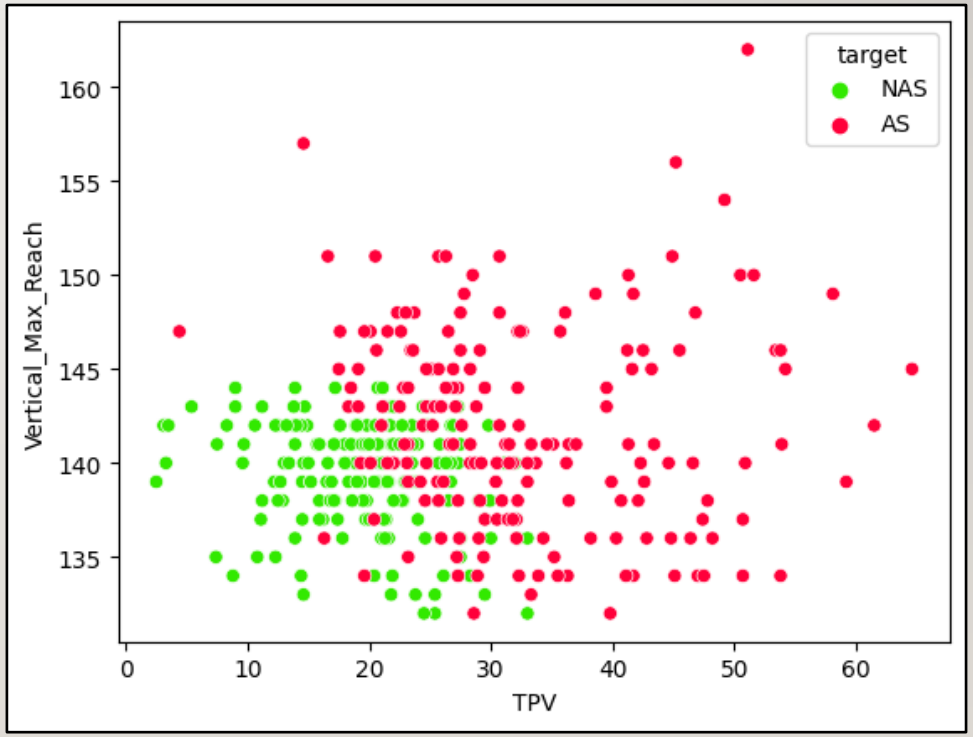


Fig. 4: Scatterplot of Frontcourt Players by TPV & Vertical_Max_Reach

DETAILS – LR & SVM DECISION BOUNDARIES (FC)

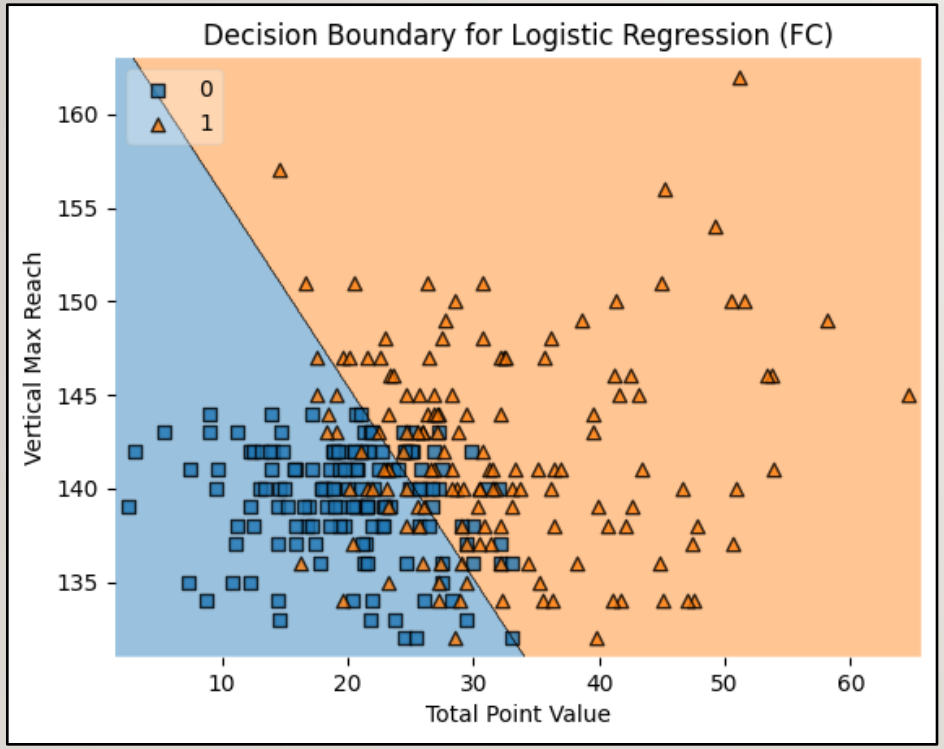


Fig. 5: Decision Boundary for Logistic Regression (FC)

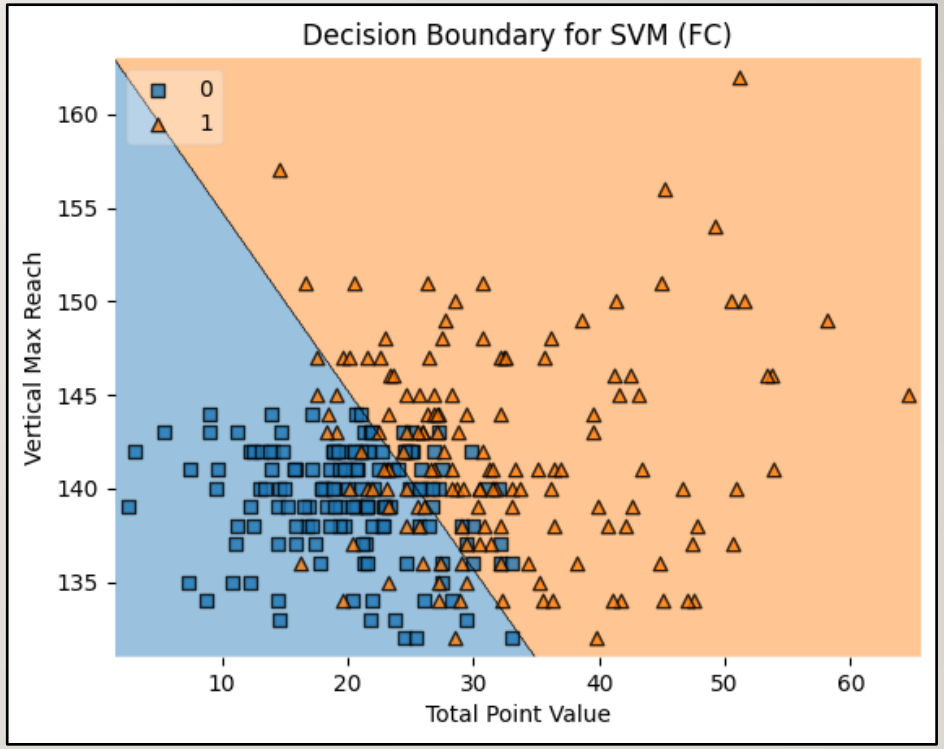


Fig. 6: Decision Boundary for SVM (FC)

DETAILS – RANDOM FOREST

- For full disclosure into the specifics of the Random Forest model, the details for it are shown in **Fig. 7**.
- It possesses a max depth of 10, has 200 decision trees, and has its random state set to 42.

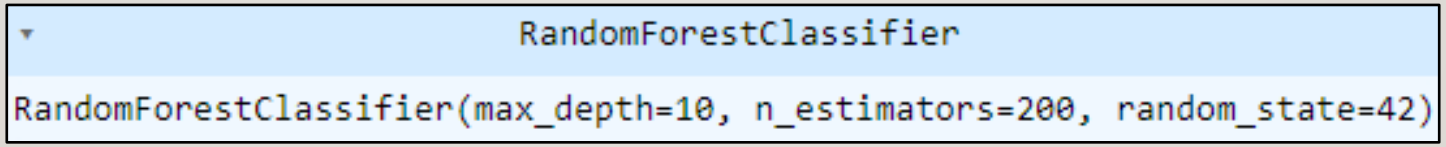


Fig. 7: Details for the Random Forest Classifier

DETAILS – XGBOOST

- Additionally, for full disclosure into the specifics of the XGBoost model, the details for it are shown in **Fig. 8**.
- 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, a gamma of 0.5, and a seed value of 42 (though not shown in the figure).

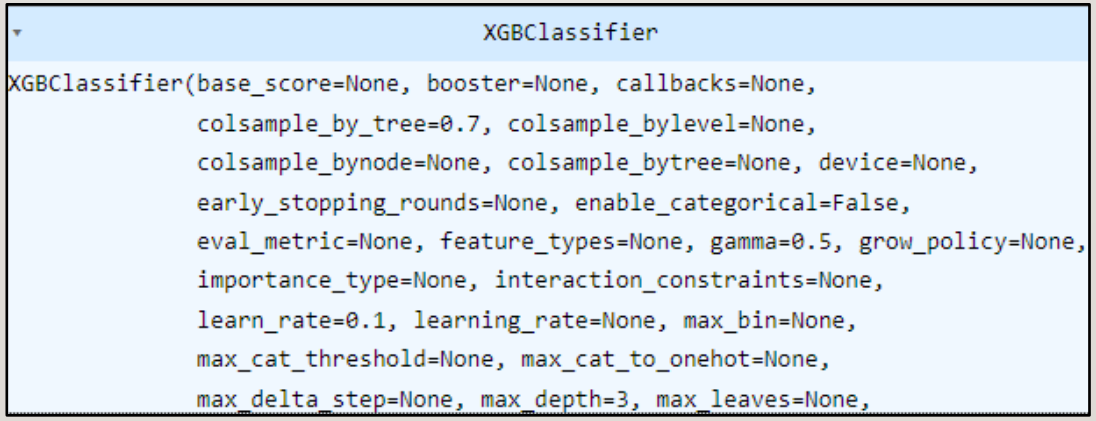


Fig. 8: Details for the XGBoost Classifier

DETAILS – TABLE 1: PROSPECTIVE PLAYERS (BC)

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: A table showcasing the custom data for prospective Backcourt players to be used for prediction.

Player 1: A 6'1, 189-pound player with a 33-inch maximum vertical jump that averaged 8 points, 2 rebounds, and 3 assists, with a total point value of 16.9.

Player 2: A 6'5, 205-pound player with a 38-inch maximum vertical jump that averaged 22 points, 4 rebounds, and 8 assists, with a total point value of 44.4.

DETAILS – TABLE 2: PROSPECTIVE PLAYERS (FC)

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: A table showcasing the custom data for prospective Frontcourt players to be used for prediction.

Player 3: A 6'5, 220-pound player with a 33-inch maximum vertical jump that averaged 12 points, 8 rebounds, and 2 assists, with a total point value of 24.4.

Player 4: A 6'10, 210-pound player with a 37-inch maximum vertical jump that averaged 18 points, 7 rebounds, and 4 assists, with a total point value of 34.

DETAILS – TABLE 3: MODEL SCORES (BC)

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: A table showcasing each model's Accuracy, Precision, Recall, and F1 Score for Backcourt Players.

DETAILS – TABLE 4: MODEL SCORES (FC)

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: A table showcasing each model's Accuracy, Precision, Recall, and F1 Score for Frontcourt Players.

DETAILS – TABLE 5: MODEL PREDICTIONS

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

Table 5: A table showcasing each model's prediction based on the prospective player data.

1 = All-Star
0 = Non-All-Star

WHAT IS ORIGINAL ABOUT THIS RESEARCH?

- This research is original in a handful of ways.
- For starters, attempting to solve the problem of draft predictability appeared to be a rather novel application of data mining techniques in and of itself, as datasets made for this specific purpose are virtually nonexistent, at least on the public level.
- As a result, the dataset created for this research was wholly original and based on authentic data consisting of pre-draft metrics and statistics.
- Additionally, the implementation of the **total point value (TPV)** stat 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 for running basketball-related classification models.

CONCLUSIONS

- Out of the four models, XGBoost and Random Forest were the top performers followed by SVM and Logistic Regression.
- For both Backcourt and Frontcourt players, XGBoost and Random Forest excelled.
- For Backcourt players, XGBoost obtained a 91.67% Accuracy Score.
- For Frontcourt players, both XGBoost and Random Forest obtained a 90.54% Accuracy Score.
- Both formed consistent predictions indicating that it is better to use models that take a multitude of features into account when predicting NBA All-Stars.

LESSONS LEARNED

- For All-Star predictions it is best to take multiple features into account rather than relying on only two.
- Additional metrics 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 sense there may also exist many unquantifiable variables in sports that no model can predict, so it is important to assess a model through the data from which it is trained.

➢ These include details like whether or not a player is will suffer an injury early on in their career or understanding their level of their overall work-ethic.

➢ As it is difficult to quantify these variables, they may stifle or elevate certain players in ways a model can't predict.

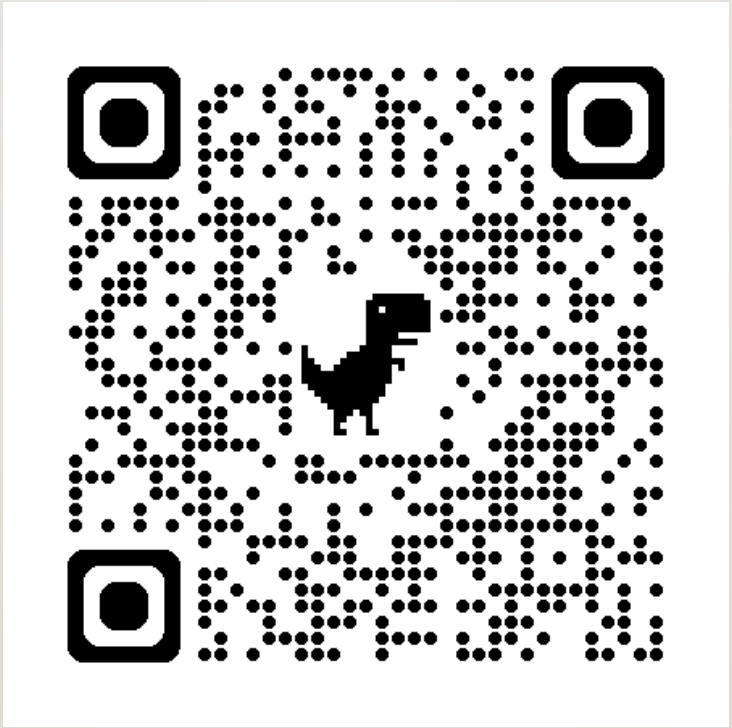
IMPLICATIONS

- Nevertheless, if models based on physical attributes and historical statistics like the ones in this research were to carry over into the real-world to a similar level of accuracy, it could revolutionize how teams throughout the NBA draft players going forward.
- As a result, the value of obtaining high draft picks could directly go up, as taking a chance on a pick would be less of a gamble than it has been in the past.

TOPICS I APPLIED FROM DATA MINING 2

- Support Vector Machine (SVM) – Weeks 2 & 3
- Random Forest – Week 3
- XGBoost Algorithm – Week 7

QR CODE TO DATASET ON GITHUB



QUESTIONS?

REFERENCES

- [1] https://en.wikipedia.org/wiki/Basketball_positions
- [2] https://www.basketball-reference.com/leagues/NBA_stats_per_game.html

THANKS FOR WATCHING!