Evan Chang
Prof. Chen
CSCI 184
7 Dec 2024

Fantasy Basketball Trade Analyzer/Performance Analyzer

Abstract: *In this paper, we will investigate using machine learning algorithms to predict the ranking of different fantasy basketball players and trade value between fantasy basketball managers. Basketball has a large amount of raw data, including points, assists, rebounds, turnovers, and many different statistics being considered for each game. This combination of available data online and a large number of games make basketball a great data to test. We use statistics from 2022 individual statistics to predict the ranking of potential 2024-25 ranking for each player and value them to create a functional trading analysis machine.*

Background:

Fantasy basketball is a strategic statistical game where managers predict and draft 13 NBA players on a team that you believe can outperform others in a season-long performance. There are numerous features in this statistical game during the season where you can trade your current players with other managers' players however you see fit. The initiative involves the necessity to have advanced prediction and analytics skills to have a good fantasy basketball season. However, by putting numerous online datasets from authoritative basketball sources, including the official NBA sites, this project aims to compile a structured dataset and models to display crucial information for predicting player performance.

Since the machine learning algorithm will have a dataset that includes the past season ranking and all individual statistics, this rich and detailed dataset will form the basis for our predictive model. Having diverse features ranging from player averages and performance will allow the algorithm to accurately project an individual's short-term feature and where they stand in the ranking of players. Since the 2024-25 season is live right now, we can gain access to the live ranking and immediately put this project to use. Since I am a long-time fan of fantasy basketball and NBA, this project will be able to offer me valuable insights into the dynamics of basketball players. It can help us gain a deeper understanding of the contributing factors influencing player performance and allow a more strategic approach to our future fantasy team selection.

In this project, I plan to use a decision tree, KNN, and linear regression model to test the accuracy of machine learning when comparing the raw data of players' statistics. Then based on the result of the algorithm, I will choose an algorithm to perform my fantasy trade analysis simulation.

Introducing Fantasy Basketball

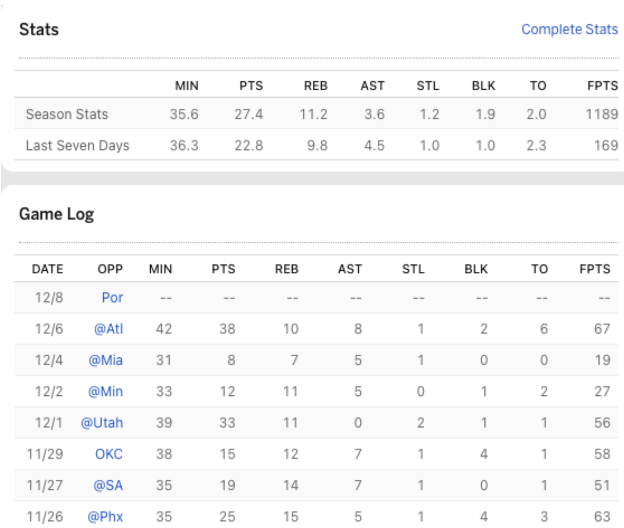
This section is a short overview of fantasy basketball and how it relates to our final project.

Fantasy basketball is an online platform statistical game that allows NBA enthusiasts to participate in a competition league. In fantasy basketball, all participants act as team managers and draft their own virtual teams composed of current NBA players during the NBA season.

Currently, we are in the 2024-25 NBA season. The drafted teams would then compete against each other based on the live statistical performance of the players in actual NBA games.

The performance of these drafted players in real NBA games directly impacts the fantasy team's success. Points are awarded based on various statistical categories such as points scored, rebounds, assists, steals, blocks, and shooting percentages. Each victory is based on the final score of the week and the player with more points at the end of the week wins. To receive points, the drafted NBA players will have to play in the live NBA game and receive points through their in-game performance.

Throughout the season, managers can make roster changes, trades, and other moves to optimize their team's performance to potentially lead in ranking and win the season-long tournament.



Stats

Complete Stats

	MIN	PTS	REB	AST	STL	BLK	TO	FPTS
Season Stats	35.6	27.4	11.2	3.6	1.2	1.9	2.0	1189
Last Seven Days	36.3	22.8	9.8	4.5	1.0	1.0	2.3	169

Game Log

DATE	OPP	MIN	PTS	REB	AST	STL	BLK	TO	FPTS
12/8	Por	--	--	--	--	--	--	--	--
12/6	@Atl	42	38	10	8	1	2	6	67
12/4	@Mia	31	8	7	5	1	0	0	19
12/2	@Min	33	12	11	5	0	1	2	27
12/1	@Utah	39	33	11	0	2	1	1	56
11/29	OKC	38	15	12	7	1	4	1	58
11/27	@SA	35	19	14	7	1	0	1	51
11/26	@Phx	35	25	15	5	1	4	3	63

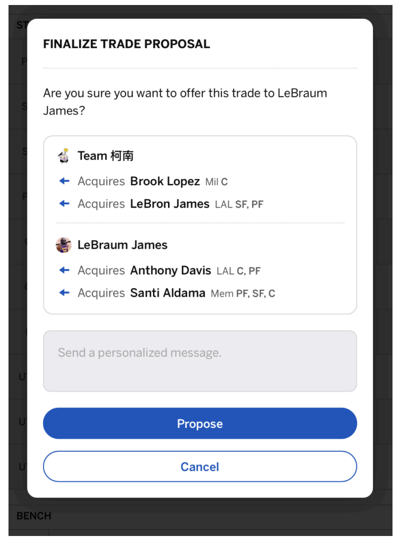


Figure 1: This is an example of Anthony Davis' performance, one of the top fantasy players, transferring his performance to fantasy points(screenshot from ESPN fantasy)
Figure 2: This is an example of requesting a trade between two fantasy managers.

Data and Choice of Feature:
Through online research, I found a few large databases on an official NBA yearly update website (https://www.nbastuffer.com/nba-stats/player/). I ended up using a ranked database of 2022 nba individual performance and a weekly updated current season database. With both of these raw data, I decided to allow my algorithm to calculate some features that are used in the official NBA fantasy game.
Offensive points

- MPG(minutes per game), FTA(Free Throw Attempt), 2PA(2 point attempt), 3PA(3 point attempt), PPG(points per game), RPG(Rebound per game), APG(assist per game),SPG(steal per game), BPG(block per game), TPG(turn over per game)

Individual Percentage

- TO%(turn over %), FT%(Free Throw Percentage), 2P%(2 point percentage), 3P%(3 point percentage)

	RANK	Rank	NAME	TEAM	POS	AGE	GP	MPG	USG%	TO%	...	APG	SPG	BPG	TPG	P+R	P+A	P+R+A	VI	ORtg	DRtg
0	NaN	1	Joel Embiid	Phi	C-F	29.1	66	34.6	37.0	14.5	...	4.2	1.0	1.7	3.4	43.2	37.2	47.4	13.0	124.4	104.1
1	NaN	2	Luka Doncic	Dal	F-G	24.1	66	36.2	37.7	14.0	...	8.0	1.4	0.5	3.6	41.0	40.4	49.0	14.4	120.0	109.2
2	NaN	3	Damian Lillard	Por	G	32.7	58	36.3	33.8	13.7	...	7.3	0.9	0.3	3.3	36.9	39.5	44.3	11.5	126.4	117.6
3	NaN	4	Shai Gilgeous-Alexander	Okc	G-F	24.7	68	35.5	32.8	12.2	...	5.5	1.6	1.0	2.8	36.2	36.9	41.7	10.6	124.9	109.8
4	NaN	5	Giannis Antetokounmpo	Mil	F	28.3	63	32.1	38.8	16.2	...	5.7	0.8	0.8	3.9	42.9	36.8	48.6	15.9	116.8	100.6

Figure 3: nba2022.csv - List of ranked players from the 2022 season.

	RANK	NAME	TEAM	POS	AGE	GP	MPG	USG%	TO%	FTA	...	APG	SPG	BPG	TPG	P+R	P+A	P+R+A	VI	ORtg	DRtg
0	NaN	Joel Embiid	Phi	C	30.0	34	34.0	39.2	14.3	403	...	5.7	1.1	1.8	3.7	46.6	41.0	52.3	15.5	125.2	106.0
1	NaN	Luka Doncic	Dal	G	25.0	57	37.5	36.0	14.2	510	...	9.9	1.5	0.6	3.9	43.6	44.4	53.5	15.6	123.6	111.9
2	NaN	Shai Gilgeous-Alexander	Okc	G	25.7	64	34.4	33.1	9.7	564	...	6.4	2.1	1.0	2.2	36.7	37.5	43.1	12.1	130.7	107.8
3	NaN	Giannis Antetokounmpo	Mil	F	29.3	63	35.0	33.0	15.3	696	...	6.4	1.2	1.0	3.4	42.0	37.2	48.4	14.9	126.5	106.9
4	NaN	Kevin Durant	Pho	F	35.5	58	37.3	29.8	14.6	376	...	5.5	0.9	1.3	3.3	35.2	33.9	40.7	10.9	120.1	113.0

Figure 4: nba2024.csv - List of unranked players from the current 2024-25 season.

The nba2022.csv will act as the training dataset while the list of unranked 2023-25 seasons will be the testing dataset and the final ranking the algorithm will calculate. I believe that the most recent performances of players are essential, therefore, I found the most recent individual dataset I have access to as training data to have a more accurate prediction of their performance.

Approach:
After researching machine learning and gaining an understanding of the analytics behind NBA and sports performance in general, it became evident that raw statistical data often contains a considerable amount of noise and irrelevant information. To address this issue, we used the Seaborn library along with Matplitlib in Python to generate a heatmap. This heatmap aids in visualizing the correlations between selected columns within our training dataset.

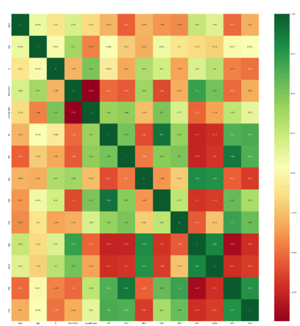


Figure 5: Generate Heatmap(clearer visual on source code)

I tested 3 different machine learning algorithms: linear regression, K nearest neighbour, and decision tree. The decision tree was chosen because it would display the clearest visual between the features and the correlation between player ranking. I added KNN and linear regression to compare our results and also see the differences between the prediction ranking.

The accuracy, mean square error, mean absolute error, and r^2 are calculated through another dataset nba_statistic_2024_rank.csv. (found on: Kaggle.com) This is another weekly updated dataset of the current 2024-25 season with similar individual stats but is ranked through the constant changes and improvements of individual performance.

Testing Algorithm:
K Nearest Neighbor:
We started by making a KNN algorithm model to predict the ranking by training model using the ranked dataset as the training dataset with the feature we chose. The model will learn the patterns and similarities between feature combinations and ranking. For KNN, the algorithm will compare the player's features to the unranked dataset to identify the k-nearest neighbours. The K is set to 5 and the rankings of the nearest neighbors will predict the ranking of the unranked players in nba2024.csv. This algorithm will give us an estimated ranking for the unranked dataset based on the patterns observed in the ranked dataset.

```
Rank: -168.46774549438783, NAME: Joel Embiid
Rank: -133.27241790200276, NAME: Luka Doncic
Rank: -106.98292013890773, NAME: Kevin Durant
Rank: -84.403703889437, NAME: Donovan Mitchell
Rank: -73.0265371694428, NAME: Stephen Curry
Rank: -72.14841539921849, NAME: Kyrie Irving
Rank: -68.96553557417747, NAME: Miles Bridges
Rank: -65.32495442579204, NAME: Desmond Bane
Rank: -64.68146575628566, NAME: Jayson Tatum
Rank: -56.81353830973376, NAME: Ja Morant
Rank: -56.494882197338484, NAME: De'Aaron Fox
Rank: -52.45007129051271, NAME: Devin Booker
Rank: -51.89713408021498, NAME: Julius Randle
Rank: -45.654814441157896, NAME: Tyrese Maxey
Rank: -45.48346152134559, NAME: Anthony Edwards
Rank: -43.047659997017035, NAME: Kawhi Leonard
```

RANK	NAME	TEAM	POS
1	Giannis Antetokounmpo	Mil	F
2	LaMelo Ball	Cha	G
3	Nikola Jokic	Den	C
4	Shai Gilgeous-Alexander	Okc	G
5	Paolo Banchero	Orl	F
6	Luka Doncic	Dal	F-G
7	Jayson Tatum	Bos	F-G
8	Anthony Davis	Lal	F-C
9	De'Aaron Fox	Sac	G
10	Anthony Edwards	Min	G

Figure 6: Final result of K-Nearest Neighbors Analysis(left) compared to actual ranking(right)

```
Relaxed Accuracy (±10): 31.84%
Mean Squared Error (MSE): 10914.2561745828
Mean Absolute Error (MAE): 75.94531450577664
R-squared (R²): 0.6415029453774208
```

Figure 7: Accuracy, MSE, MAE, and R^2 of KNN algorithm result.

The outcome achieved from the KNN algorithm is promising from a fantasy player's perspective. When compared to an official NBA player ranking(shown in Figure 8), it similarly predicted the current top 10 players. But when compared to the accuracy of the entire ranking, The accuracy was only roughly 40%. Since the accuracy of an entire ranking is difficult and will result in a low accuracy because of the high precision it requires in a ranking, I used a relaxed accuracy method that gives every player a plus-minus of 10 ranking positions to determine the accuracy.

Linear Regression:
Linear Regression, a supervised learning algorithm, was also tested to predict ranking by using a linear regression model using the ranked dataset as training data. The model uses the trained machine model and creates a data frame that contains the predicted ranking as a single column named "predicted_ranking". Then it creates a new datagram "predicted_ranking_df" that combines the player name with their ranking and then sorts them into ascending order to display the actual ranking and their ranks. Similar to KNN, we received similar results.

```
Rank: -175.76, NAME: Joel Embiid
Rank: -142.20, NAME: Luka Doncic
Rank: -115.68, NAME: Kevin Durant
Rank: -75.66, NAME: Donovan Mitchell
Rank: -73.73, NAME: Devin Booker
Rank: -68.20, NAME: Kyrie Irving
Rank: -66.78, NAME: Shai Gilgeous-Alexander
Rank: -64.22, NAME: Jayson Tatum
Rank: -62.38, NAME: Tyrese Maxey
Rank: -61.70, NAME: Ja Morant
Rank: -60.09, NAME: Giannis Antetokounmpo
```

Figure 8: Result of Linear Regression testing algorithm

```
Mean Squared Error (MSE): 36.17
Mean Absolute Error (MAE): 5.01
R² Score: 1.00
Relaxed Accuracy (±10): 92.16%
```

Figure 9: Accuracy, MSE, MAE, and R^2 of Linear Regression

For the result of linear regression, while the ranking and name don't seem to have an issue, there are unexpectedly high metrics of accuracy, mse, mae, and r^2. This could be due to data leakage where the test set might have inadvertently influenced the training process. Another reason could possibly be the overfitting of the testing/training dataset. The linear regression model memorized the pattern in training due to excessive complexity rather than learning generalizable trends. Due to this reason, I decided to not use linear regression for my final trade simulator due to the unexpectedly high accuracy.

Decision Tree:

The last and final algorithm used for this project is the decision tree algorithm. This is a supervised machine learning model used for predicting the target ranking based on the performing feature. This algorithm was chosen because of the simplicity and interpretability of the clear tree-style view and how predictions are made based on the input feature. This decision tree algorithm predicts NBA player rankings based on key performance metrics.

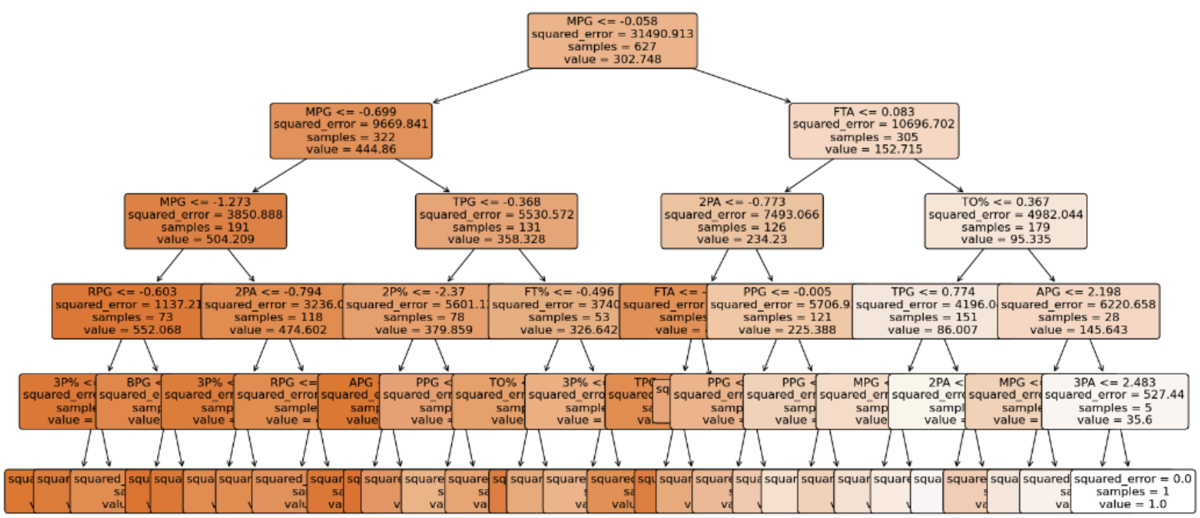


Figure 10: Visual representation of decision tree metrics.

```
Relaxed Accuracy (±10): 44.16%
Mean Squared Error (MSE): 6961.99771167048
Mean Absolute Error (MAE): 58.67963386727689
R-squared (R²): 0.6546039733880566
```

Figure 11: Accuracy, MSE, MAE, R^2 of decision tree algorithm

The visual representation of the decision tree reveals that a player's ranking is strongly correlated with their minutes per game(MPG) since that is the first deciding factor of the entire tree. This correlation is logical because a high-ranking player needs sufficient court time to demonstrate their skills and make a significant impact.

It is important to highlight that StandardScaler from sklearn.preprocessing was applied to both the training and testing data for this decision tree. In basketball, features such as PPG(points per game) and turnover(TO%) have vastly different numerical scales. By using StandardScaler, all features are standardized to ensure they are on the same scale, preventing any single feature from dominating the decision-making process due to its magnitude. When tested without the StandardScaler, the entire decision tree algorithm was tested based on PPG since offensive points hold the highest numerical value in all of the features.

Trading Simulator:

This part of the project implements an NBA trading simulator that uses the decision tree regressor algorithm to estimate players' trade values based on their performance metrics. Player data from the dataset is used to train the model which predicts a trade value for each player based on feature. The trade values are then normalized by inverting their rank, ensuring that higher trade values correspond to better players. A custom labeling function assigns categories to players based on their trade value(superstar, starter, role player, bench). Using a dictionary of player data, the simulator evaluates trades between two teams by summing the trade values of players on each team and determining if the trade is fair(trade value difference <= 5) or favors one side. Additionally, a function that displays detailed information about the players involved in a trade is implemented.

```
Team 1:
  LeBron James: Rank = 15, Label = Starter
  Anthony Davis: Rank = 15, Label = Starter
Total Team 1 Value : 30

Team 2:
  Stephen Curry: Rank = 1, Label = Superstar
  Draymond Green: Rank = 209, Label = Bench
Total Team 2 Value : 210

Trade Result:
Team 1 wins the trade!
```

Figure 12: Final result of the trading simulator

This simulation uses the decision tree algorithm's 'Unique_rank' and 'NAME', which is the ranking of the decision tree algorithm tested and the name corresponding to it. It calculates the team value based on the ranking found and the lower the value the better. I ended up choosing the decision tree algorithm to run this trading simulator because the decision tree algorithm had a higher relax accuracy than Knn and it had an easier visual representation of the feature that impacted their decision.

Conclusion:

In conclusion, we tested multiple machine learning algorithms to predict the rankings for the current NBA 2024 season and compared the results to professionally ranked data. While the predictions showed promise, they were not as accurate as desired and could be improved. The predictions were more accurate for top players with significantly different statistics. However, for average players with similar statistics or limited playtime, the algorithm struggled to differentiate them properly, often grouping them together. Some improvements that could be made include implementing a better standard scaling method and enhancing the decision tree algorithm's ability to differentiate between features. Additionally, some players were ranked highly purely due to their playtime, even though their performance did not justify such rankings. Given the diversity of basketball statistics and the varying roles of players who excel in different areas, a more tailored algorithm that evaluates different types of players separately could be beneficial. With these changes, the machine learning model could achieve higher accuracy and provide more detailed rankings for the trade simulator.