Modeling Techniques to Forecast NBA Player Fantasy Points

Abhinav Madabhushi

June 3, 2025

Abstract

With the increasing popularity of fantasy sports, especially the NBA, there is a growing need to predict player performance for the upcoming season. Most models focus on predicting per-game stats, which are extremely volatile. In this paper, I take a different approach by predicting season averages for each player, which are more stable and useful for both fantasy and real-life applications like free agency and trades. I divide the problem into three sub-tasks: predicting stats for existing NBA players using their previous season, predicting rookie stats using college performance and draft pick, and predicting games played using injury history. I use multiple machine learning models including Gradient Boosting, CatBoost, LightGBM, Random Forests, and others. Deep learning models like LSTMs and Transformers were also used for returning players to take advantage of sequential data. Feature engineering played a major role in improving performance, and I introduced a weighting technique based on fantasy points to prioritize better players. Results show that tree-based methods significantly outperform neural networks, and the model can be used to accurately predict fantasy points for any scoring system.

1 Introduction

With the rise in sports culture across the world, the number of people who play fantasy sports has also increased significantly. Due to this, there is also an increasing need to create a model to predict fantasy points, or how well a player would do next year based on their statistics from the current season. There are multiple companies like DraftKings and SportLine, among others, that already have models to predict how well a player will do the following season. As expected, sports are unpredictable, and although there cannot be an exact prediction, we can come close to it by utilizing the power of machine learning. I will be focusing particularly on predicting fantasy points in the NBA.

There have been several attempts to achieve this task in the past, all with slightly different goals. One attempt was to build a separate model for each player to predict per-game statistics [1]. Another approach used machine learning optimization models for lineup optimization using constraints such as salary caps and player positions[3]. A third paper worked on predicting per-game player statistics using the XGBoost model and a new concept of ”synergies” which measures how well players mesh with each other, improving or decreasing player performance [4]. Another paper questions the ”profitability” of using predictions from companies like DraftKings and SportsDuel, and how it is not worth investing in paid services since the predictions are not that different from using naive prediction[2].

Most papers as mentioned try to predict per-game statistics which are extremely unpredictable. I decided to predict something more stable: season averages of a player. By doing this, we can not only help players playing fantasy sports but also get insight into which players the team would want to sign in free agency. There are three main sub-parts to this problem: predicting player statistics for players already in the NBA, predicting player statistics for rookie players using college statistics, and predicting games played by a player using their injury history. There are issues concerning each of these problems which I will address in this paper. Overall, I use feature engineering techniques and multiple models to achieve this task.

2 Methods

The main dataset used for this task was player season averages for each statistic for all the seasons starting from 1960 to 2022 from the balldontlie API [5]. Player-specific data such as position of the player, age of the player in that season, and the team of the player in that season were also extracted. If a player got traded mid-season, I used a weighted average based on games played for each team to calculate the player’s statistics for the whole season. Finally, I created a feature named team change which indicated if the player changed his team from the previous season or not, and if the player is playing in the NBA for the first time this variable would be -1.

I specifically focused on predicting 12 statistics that are most used to calculate the fantasy points of a player [9]. These 12 statistics are points, rebounds, assists, field goals made, field goals attempted, 3-point makes, 3-point attempts, free throws made, free throw attempts, steals, blocks, and turnovers. There was another key variable that I predicted apart from the 12 standard statistics that measures fantasy points: games played. Fantasy sports rely on players who not only average high points but also play most of the games so that they can maximize the points gained from that player. The data was trained on all the years up till 2021, and was tested on the year 2022.

2.1 Predicting Statistics of Players Already in the NBA

Extra variables were created to predict player statistics for players already in the NBA. For each player-season combination, I used the statistics of the previous season as features to predict all the statistics. Taking two years back statistics did not seem to affect the model much, so I decided to just use the statistics of the previous year. Multiple machine learning models were used for this task: Gradient Boosting Regressor, Random Forest Regressor, AdaBoost Regressor, CatBoost Regressor, Support Vector Machines, K-Nearest Neighbours, and Multi-Layer Perceptrons.

Two other deep learning models that are usually used to predict sequential data were also used: Long-Short Term Memory and Transformers. By using these models, we are not restricted to using only the data of the past year, but the data of all the past years of the player. The data was transformed to fit the form needed for an LSTM: (samples, features, timesteps). The timesteps used was 22 seasons, since the maximum number of seasons played by a player in the NBA is 22. For every player who has played less than 22 seasons, the seasons not played by the player were padded with 0s.

2.2 Predicting Rookie Statistics

Feature Engineering was done to predict the statistics for Rookie players. Since rookie players do not have a previous year played, I decided to use the statistics of their last season in college as the features instead [6]. Along with this, I used the overall draft pick of the player as a predictor [8]. If the player was undrafted, then I made the overall pick as -1. After this feature engineering, I used the same machine learning models as mentioned above. I could not use LSTMs or Transformers for this task since there is only 1 previous season (1 timestamp).

2.3 Predicting Games Played

There was a lot of focus on predicting games played for this project, especially since predicting injuries and games played for the season heavily impacts fantasy points of the player. Feature engineering was done here so that we could accurately predict games played for the next season. Initially, I extracted the most common words from the injury notes [7] of the players and used those as features. Eventually, I extracted the most frequent body parts and injury types for NBA players extracted through ChatGPT. For each player-season, I extracted the number of injuries for the particular body parts and injury types up till that season to assess how likely it is for a player to get injured the coming season. Other features included whether the player was injured at the start of the season and the number of days since the last injury.

2.4 Combining into a Single Model

All the above features were combined into a single model to predict the statistics and games played for any given player, whether a rookie or not. For this task, an additional column named rookie was added to indicate whether the player was a rookie. For rookie players, the previous season features were the statistics of the last year of their college basketball career. The overall draft pick was added to all players regardless of whether the player was a rookie or not.

All the machine learning models were used, including a new addition for each tree-based model: weighting. Since for fantasy points, we need the error for players with higher fantasy points to be lower, I used the weighting option in the tree-based models and set the weight for each player as proportional to the fantasy points in the previous season. This way, the model will make better predictions for better players at the cost of making worse predictions for worse players.

3 Results

3.1 Data Analysis

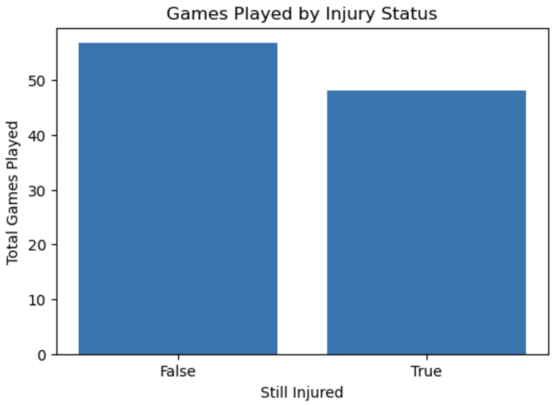

Figure 1: Games played vs Injury Status

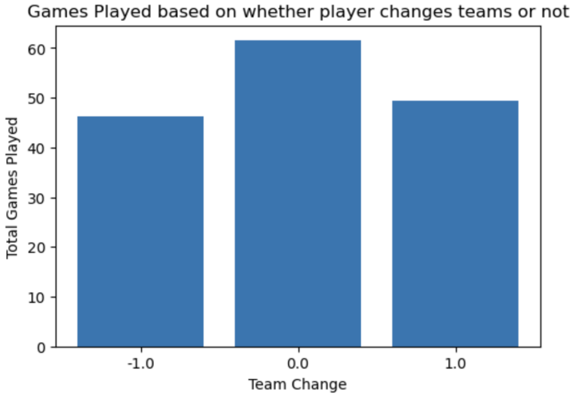

Figure 2: Games played vs Team Change

Out of all the features created to predict games played, only the variable Injury Status, which depicts whether a player is injured or not at the start of the injury, and the variable team change, which depicts whether a player has changed team from last season or not, has a significant correlation with games played. As shown in Figure 1, a player injured at the start of the season plays 10 more games on average than players who are not injured at the start of the playoffs. We can also see that players in the same team end up playing more games than players who change teams, which is evident by Figure 2.

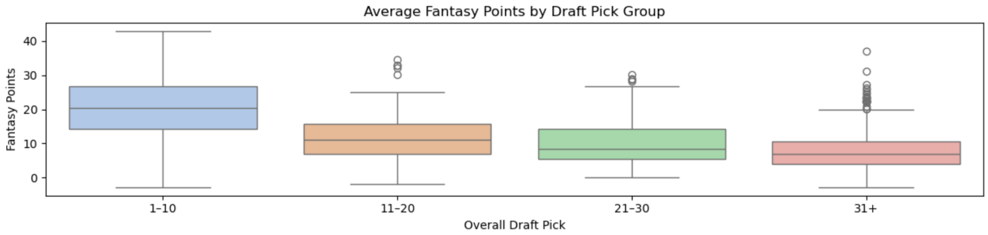

Figure 3: Draft Pick vs Fantasy Points

Overall draft pick was also an extremely good predictor to predict any statistic. The fantasy points metric plotted in the graph above is from ESPN, and the formula is:

$$\text{FP} = \text{PTS} + \text{3PM} - \text{FGA} + 2\,\text{FGM} - \text{FTA} + \text{FTM} + \text{REB} + 2\,\text{AST} + 4\,\text{STL} + 4\,\text{BLK} - 2\,\text{TOV}.$$

As you can see in Figure 3, the fantasy points, and every other statistic, is higher for higher draft pick players, which is an obvious conclusion, and which is why it is used in the model as a proxy for player scouting data.

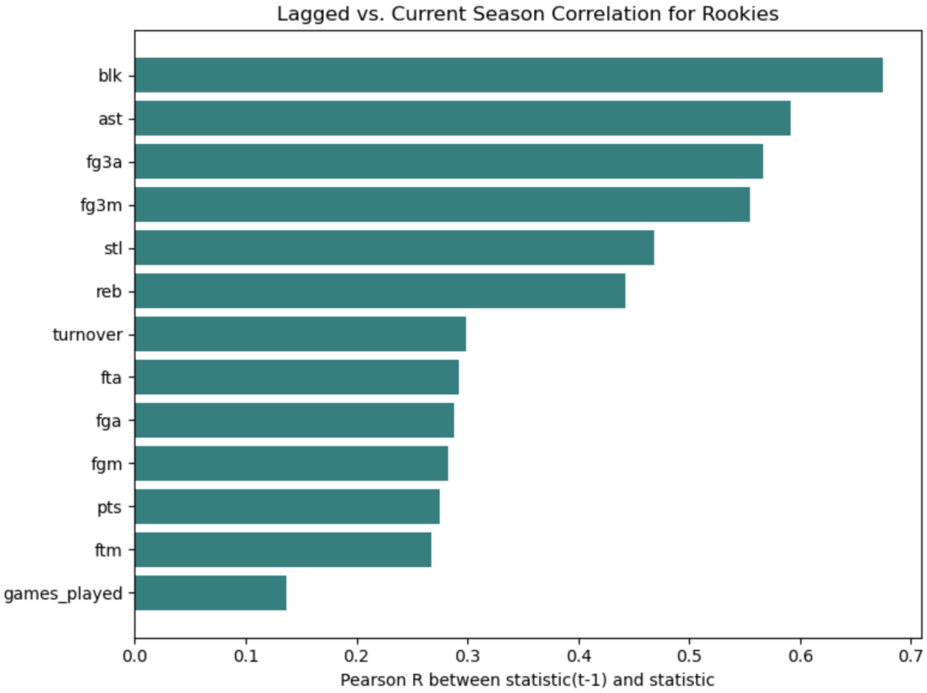

Figure 4: Lagged vs. Current Season Correlation for Rookies

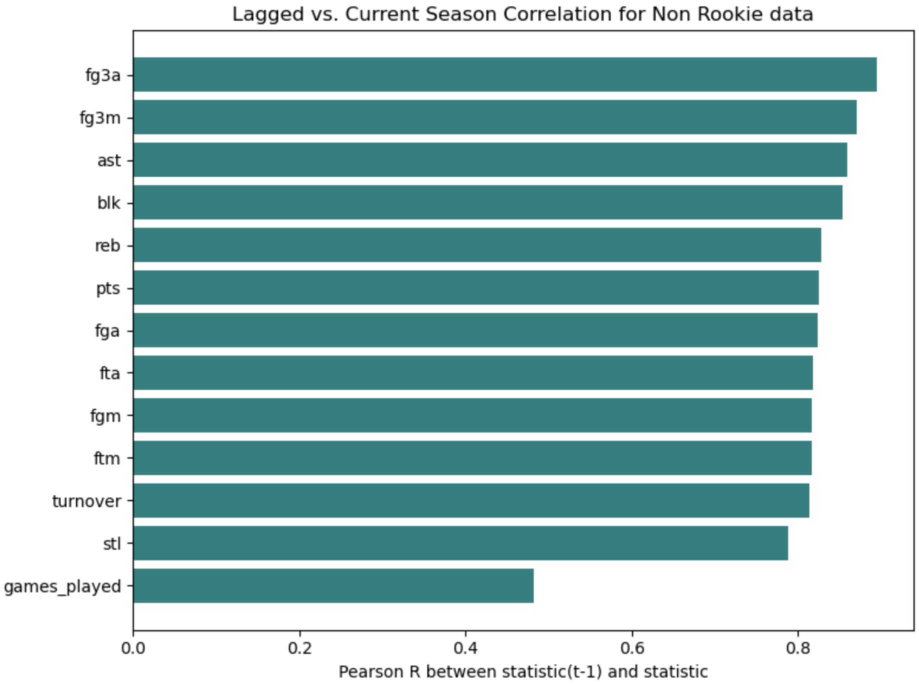

Figure 5: Lagged vs. Current Season Correlation for Non-Rookies

Lastly, as you can see in Figure 4 and 5, there is a stark difference between the correlation of each statistic with the previous season for rookies and non-rookies. For non-rookies, the correlation for every statistic is around 0.85 with the exception of games played, which has a correlation of 0.5. This is in strong contrast with statistics for rookies, where only blocks, assists, 3-pointers attempts, 3-pointers made, steals, and rebounds have a correlation above 0.45, and the rest of the statistics have a correlation below 0.3 with their previous season. We can come to an interesting conclusion from this: blocks, assists, steals, rebounds, and 3 point shooting ability translates to the NBA, but points and all related statistics do not translate to the NBA from college. This implies that scoring in the NBA is much tougher due to the increase in defensive ability, and also because there can be only 1 or 2 star players in each team, which would mean that only some players can score the bulk of the points. For all other statistics, ball dominance is not required, and these statistics can be accumulated without the ball, which is what good role players do. Hence, it is easy to conclude that role players ability transfers from college to the NBA, but being a ball-dominant star player is extremely difficult.

3.2 Model Fitting

Table 1: Test R^2 Scores for Different Models on Each Target

Target	**LSTM**	**Transf.**	**GB**	**RF**	**Ada**	**Cat**	**ET**	**LGB**	**HGBR**	**KNN**	**SVM**
games_played	-2.4596	-0.7584	0.346	0.325	0.286	0.369	0.289	0.356	0.363	0.195	0.325
pts	0.6922	0.7165	0.781	0.770	0.556	0.782	0.740	0.784	0.786	0.625	0.590
fgm	0.6594	0.7033	0.773	0.763	0.569	0.774	0.730	0.773	0.775	0.598	0.585
fga	0.6370	0.7119	0.786	0.775	0.663	0.784	0.748	0.776	0.785	0.607	0.618
fg3m	0.5798	0.6664	0.714	0.691	0.571	0.703	0.647	0.707	0.708	0.531	0.537
fg3a	0.5932	0.6876	0.752	0.718	0.627	0.746	0.681	0.743	0.746	0.548	0.586
ftm	0.6437	0.6751	0.769	0.753	0.316	0.767	0.720	0.760	0.768	0.634	0.547
fta	0.6486	0.6835	0.763	0.751	0.360	0.766	0.720	0.758	0.762	0.631	0.549
reb	0.5902	0.6269	0.681	0.671	0.524	0.677	0.634	0.662	0.677	0.508	0.533
ast	0.6439	0.6554	0.759	0.735	0.592	0.767	0.680	0.759	0.762	0.587	0.594
stl	0.3467	0.3050	0.570	0.538	0.308	0.566	0.476	0.555	0.567	0.385	0.482
blk	0.3361	0.4694	0.595	0.599	-0.320	0.600	0.546	0.617	0.613	0.391	0.477
turnover	0.5333	0.5177	0.705	0.684	0.382	0.707	0.643	0.699	0.701	0.535	0.558

As seen in table 1, all the tree-based methods did the best. Also, the gradient boosting, Catboost, light gradient boosting, and Histogram-based gradient boosting did the best out of all the tree-based methods. We can see that LSTMs and Transformers performed pretty poorly, especially on games played. From this, we can conclude that the data size, limited by our player-season combinations, is too small, which is why the tree-based methods end up doing better.

Table 2: Top 5 Predictors for Each Statistic

Statistic	**Rank 1**	**Rank 2**	**Rank 3**	**Rank 4**	**Rank 5**
games_played	games_played(t-1)	pts(t-1)	team_change	player_id	min(t-1)
pts	pts(t-1)	fgm(t-1)	team_change	fga(t-1)	Rookie
fgm	fgm(t-1)	team_change	pts(t-1)	Overall Pick	Rookie
fga	fga(t-1)	team_change	pts(t-1)	fgm(t-1)	Rookie
fg3m	fg3m(t-1)	fg3a(t-1)	team_change	Rookie	player_id
fg3a	fg3a(t-1)	fg3m(t-1)	Rookie	team_change	player_id
ftm	ftm(t-1)	Rookie	team_change	games_played(t-1)	pts(t-1)
fta	fta(t-1)	Rookie	team_change	ftm(t-1)	games_played(t-1)
reb	reb(t-1)	team_change	Rookie	fta(t-1)	Overall Pick
ast	ast(t-1)	team_change	Rookie	Pos	games_played(t-1)
stl	stl(t-1)	games_played(t-1)	Rookie	team_change	player_id
blk	blk(t-1)	team_change	Rookie	Pos	games_played(t-1)
turnover	turnover(t-1)	team_change	Rookie	games_played(t-1)	Overall Pick

As you can see above, the (t-1) variable for each statistic is its best predictor as expected. Also, team change, rookie, and overall pick seem to be in the top 5 predictors for most statistics, showing their importance.

Weighted gradient boosting model was also used. The formula for the weight is given below:

$$\bar{fp} = \frac{1}{N}\sum_{j=1}^{N} fp_j, \qquad w_i = \frac{fp_i}{\bar{fp}}.$$

Using this formula for weight, the mean absolute error of the fantasy points computed from weighted gradient boosting model came down from 5.15 to 5.06 for the top 150 players in the league.

4 Discussion/Limitations

The model for predicting each statistic achieved a R^2 of up to 0.9 for players already in the NBA. However, predicting rookie statistic was a challenge especially due to limited data. Having scouting data for each player could have helped us better predict rookie statistics. Another downfall of our model is its inability to accurately predict points and related statistics for rookie data. Nevertheless, statistics not related to points had a good R^2 of around 0.5. Predicting games played is also heavily tied to accurately predicting the other statistics, since a player playing better would indicate that he would play more games for the team. Injury history of players coming into the NBA would also help in the task of better predicting games played for rookies.

Another issue was the availability of good quality injury data. Injury notes for each player-season combination was obtained from a kaggle dataset, from which injury features were extracted. The fact that none of the injury features were correlated with games played tells us that the injury data used did not have good enough details to be able to predict games played for the next season. Although it is a difficult task to predict games played, since we also need to take into factors other than injury such being given less minutes by the coach or being prone to getting suspended more, better data with regards to physical tests conducted by each team can give us more insight into how fit a player is and can help us predicting how likely the player is to get injured.

5 Conclusion

Using the predictions for each of the different statistics and games played, one can easily calculate the predicted fantasy points for the player using the formula for the fantasy points. Since each platform has its own formula for fantasy points, individually predicting each statistic will help us not only find the predicted fantasy points for any given formula, but will also help us analyze each statistic individually. More specifically, we can look into the R^2 for each statistic and add any relevant features that we think would be relevant to predicting that statistic better. Moreover, since we are using tree-based models which are interpretable, we can individually explain which features are the best to predict each statistic, helping us better tune our models.

Being able to predict how well a player will perform has implications not only in fantasy sports, but also in making better free agency decisions and draft pick selections. By accurately predicting how rookies will perform in the NBA given scouting data can significantly help teams draft the right personnel for their team. Also, accurately predicting non-rookie statistics and games played can help teams during free agency and mid-season trades.

References

[1] Marathe, A., Banerjee, T., & Zhang, Y. (2019). *An innovative method for accurate NBA player performance prediction*. International Journal of Computer Applications, 177(34), 1–5.

[2] Tang, B., Zhang, Y., & Wang, H. (2023). *Analyzing the Profitability of NBA Daily Fantasy Sports: A Causal Inference Perspective*. Journal of Sports Analytics, 9(1), 85–105. https://doi.org/10.1007/s12197-023-09646-7

[3] Yi, S., Lien, C., & Ding, H. (2020). *Machine Learning Optimization Model to Predict Fantasy Basketball Teams*. Proceedings of the International Conference on Machine Learning and Data Engineering (iCMLDE), 2020, 15–19.

[4] Zheng, Y., Li, Y., & Xu, H. (2020). *NBA Player Comprehensive Score Prediction based on Synergy and Machine Learning*. 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), 1040–1044.

[5] Balldontlie API. *Season-by-season NBA player statistics from 1960–2023*. Retrieved from `https://www.balldontlie.io/`

[6] Basketball Reference. *College and NBA statistics, positions, and team affiliations for players*. Retrieved from `https://www.basketball-reference.com/`

[7] Kaggle Injury Dataset. *NBA player injury list and injury descriptions (1960–2023)*. Retrieved from `https://www.kaggle.com/datasets`

[8] NBA.com. *Historical draft pick data for NBA players*. Retrieved from `https://www.nba.com/stats/draft/history/`

[9] ESPN Fantasy Basketball Scoring. *Default points league scoring explained*. Retrieved from `https://www.espn.com/fantasy/basketball/story/_/id/30296896/espn-fantasy-default-points-league-scoring-explained`