NBA Fantasy Basketball Optimizer

Rachel Price
price.1803@osu.edu
Department of Physics, The Ohio State University

INTRODUCTION

Fantasy sports have been steadily growing in popularity in the United States since their conception. Analysts across the world have attempted to create models that will predict game outcomes, total points scored, individual player performances, and thousands of other predictions.

While there are a great many factors that affect a player/team's performance for a given game/season, we can use a few of the most prominent and influential factors to train our model.

Oe can use deep learning to gain greater insight

AIM

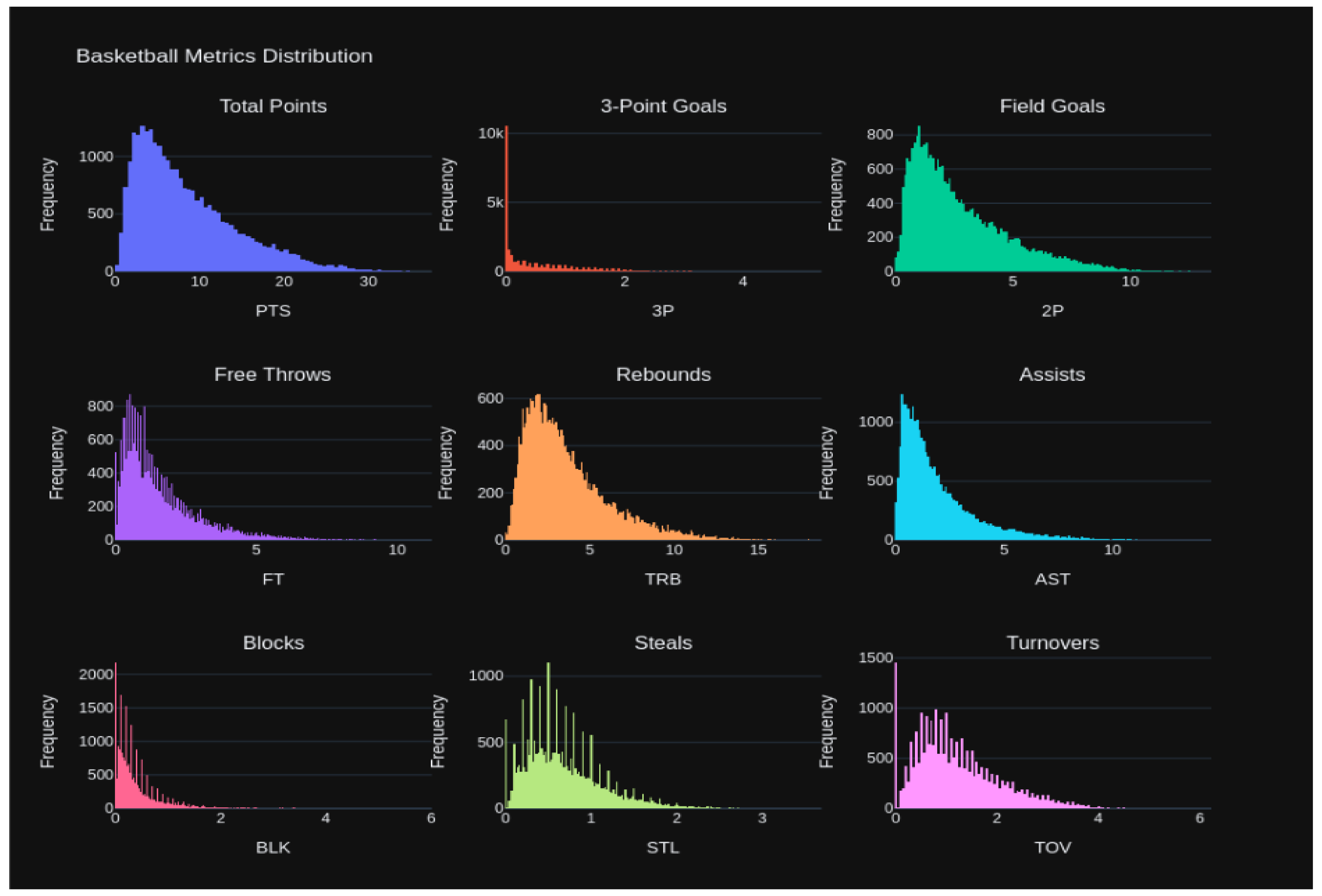
Our model has two objectives:

1. Create a hypertuned RandomForestRegression model that is trained on over 50 years of past NBA players' metrics and can predict current players' performances in the upcoming season with a certain level of accuracy.
2. Find the optimal fantasy league lineup given a variety of constraints.

DATA SET

Over 50 years of NBA metrics were obtained from various Kaggle datasets. These datasets contained a variety of player statistics (age, height, weight, draft round pick, seasons of experience, position, and team), along with actual season metrics per game (games played, minutes played, rebounds, assists, steals, blocks, turnovers, as well as all-attempts/successful-attempts for: overall shooting, 3-point goals, 2-point goals, and free throws).

After cleaning the dataset, we have a total of 26,489 rows and 30 columns. The distributions of the nine metrics the model will attempt to predict are below.



RANDOM FOREST REGRESSION MODEL

A RandomForestRegression model is a collection of decision trees that aim to overcome the two main drawbacks of decision trees: overfitting, sensitivity to the distribution of training data. This model trains many randomized trees (randomized through bootstrapping and using a small selection of the features) and averages the results. The model then passes each desired sample through all the trees and aggregate the results through voting.

OTHER MODELS TESTED

Two other models were tested: a linear regression model and a time-series forecasting model.

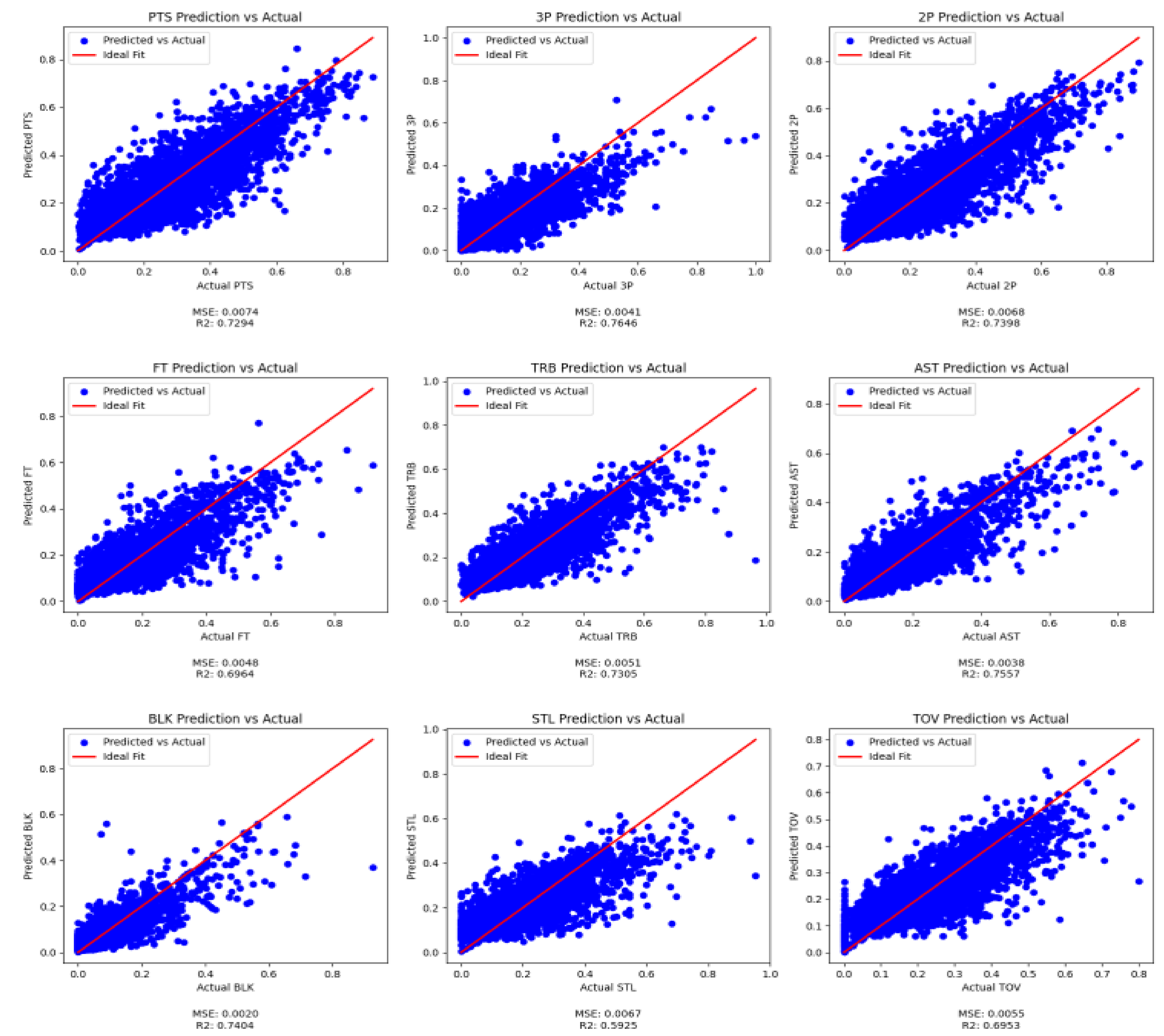
Each of the three models were trained using our training dataset, and then used to predict our test dataset, and then both the mean squared error, as well as residual-squared values were recorded, and the best model was chosen to be hypertuned.

Both the MSE and R2 scores of our three models are shown below.

	Linear Regression	RandomForestRegression	Time-Series Forecasting
MSE	10.058	9.964	12.594
R2	0.732	0.736	0.713

RESULTS

After hypertuning our RandomForestRegression model, it was determined the ideal parameters are: It was found that the following parameters were optimal for our specific dataset: n_estimators=200, max_depth=20, min_samples_split=2, min_samples_leaf=2, max_features='sqrt', bootstrap=False. We then fit our data using a model with these parameters and used said model to predict the metrics for active NBA players' next season. The residual plots, as well as their associated MSE and R2 scores for each of the nine metrics are recorded below.



RESULTS (continued)

While our model is not perfect, it is more effective than pure guesswork. After training the model, we applied it to our desired NBA players and subsequently ran an optimization model to find the ideal ten-player fantasy lineup, considering the following limitations: only ten players per roster, no more than two players from the same team, a requirement of five frontcourt and five backcourt players, and a maximum total salary cap of 100 million USD (based on the salary evaluations available on the NBA fantasy website).

Our ten selected players are as follows:

1. LeBron James
2. Al Horford
3. Tim Hardaway Jr.
4. Kelly Oubre Jr.
5. De'Anthony Melton
6. Ja Morant
7. Jordan Poole
8. Coby White
9. Cole Anthony
10. Onyeka Okongwu

CONCLUSION

Utilizing a dataset containing individual NBA players' metrics across over 50 seasons, our hypertuned RandomForestRegression model processed this data and predicted active NBA players' metrics for the upcoming 2025 season and optimized these predictions to determine the ideal NBA fantasy lineup.

The RandomForestRegression model achieved the highest R2 score, meaning it was able to explain the greatest proportion of variance within the target values, when compared to our other models. This is likely because random forests are less sensitive to outliers and noise and can assess feature importance more effectively. Any underfitting was fixed by standardizing the features and target values, which created more variation in our dataset.

Our model accounts for roughly 75% of variation in our dataset.

FUTURE WORK

In the future, it would be interesting to rerun the simulation using different types of time-series forecasting models and train the model exclusively on non-active players and test solely on active players. This would require access to additional resources containing more metrics for earlier seasons. Our dataset shrank significantly and excluded any statistics prior to 1974 due to insufficient data. If more time were available, it would be possible to manually research missing statistics and record them individually, which would provide enough data to train the model without using active players.

REFERENCES

[1] AmirHossein Mirzaei. "NBA players scored points prediction". https://www.kaggle.com/code/amirhosseinmirzaie/nbaplayers-scored-points-prediction/comments, 2023.
[2] DimitreOliveira. "Deep Learning for Time Series Forecasting". https://www.kaggle.com/code/dimitreoliveira/deeplearning-for-time-series-forecasting, 2018.
[3] Duy Le. "NBA EDA predict point using RandomForestRegressor". https://www.kaggle.com/code/duyle2201/nbaedapredictpointusingrandomforestregressor/notebookModelingdata, 2024.
[4] Sarpeco. "Gathering And Cleaning Data For Salaries". https://www.kaggle.com/code/sarpeco/gatheringand-cleaning-data-for-salaries/notebook, 2021.
[5] Omri Goldstein. "NBA Players stats since 1950". https://www.kaggle.com/datasets/drgilermo/nbaplayers-stats/data?select=SeasonsStats.csv, 2017.
[6] Vivo Vinco. "2023-2024 NBA Player Stats". https://www.kaggle.com/datasets/vivovinco/2023-2024- nba-player-stats/data, 2024.
[7] Wyatt Walsh. "NBA Database". https://www.kaggle.com/datasets/wyattowalsh/basketball, 2023.
[8] Sumit Redekar. "NBA player stats from 2018 to 2021". https://www.kaggle.com/datasets/sumitredekar/nbastats-2018-2021, 2021.