NBA Fantasy Predictor

James Sobrino

Problem Statement

My friends are significantly better at Fantasy Basketball than me (there is also a $20 buy in for our league and I'm tired of donating every year).

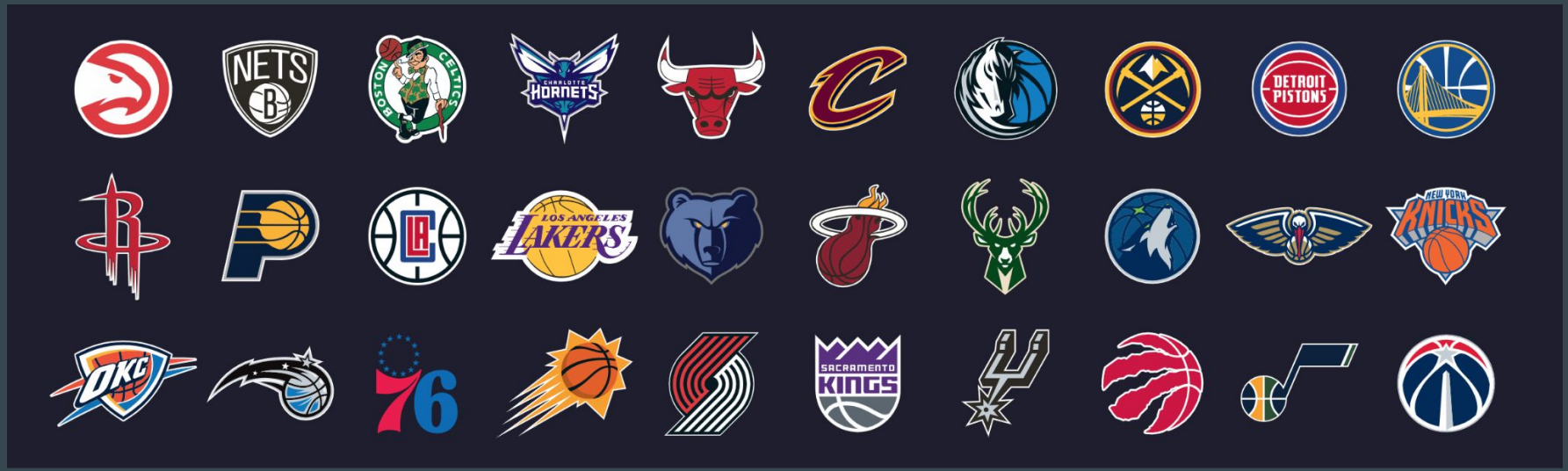
Fantasy Sports requires a lot knowledge and a lot of attention to consistently have the best players on your roster.

Can I create a model that can accurately predict fantasy output for any player in a given week?

NBA Background

There are 30 teams in the NBA with up to 15 players per team

82 games played per year (72 in 2020-2021 season)

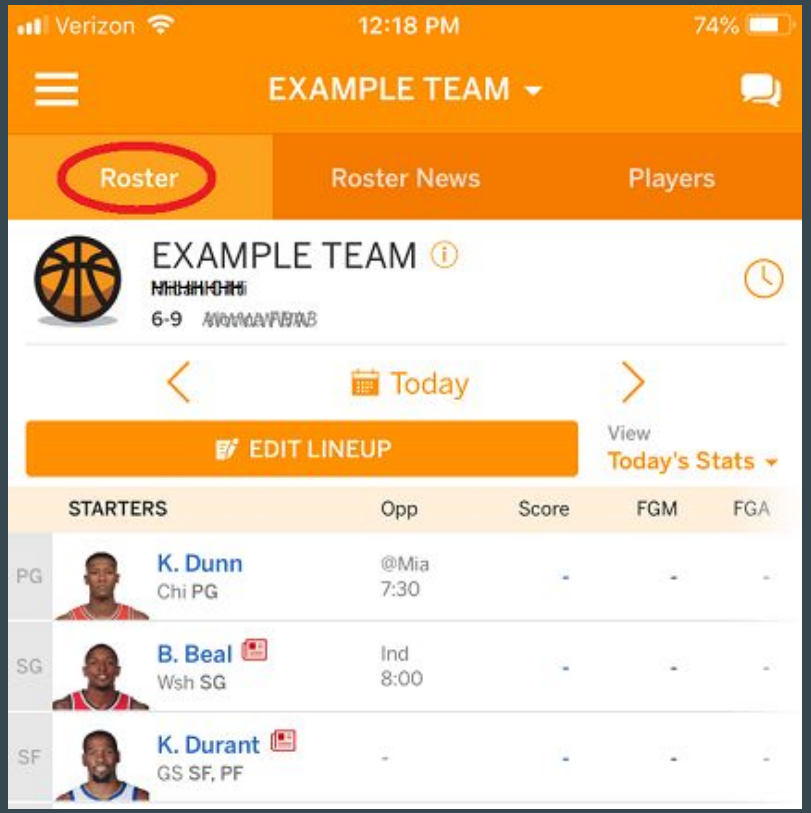


Fantasy Sports Background (League Specific)

There are many types of Fantasy Sports, we will be focusing on the NBA.

Teams are matched up in head to head competitions.

Players fantasy points are added up after each night for a week, the highest total points wins.



Dataset

https://www.kaggle.com/nathanlauga/nba-games

This dataset contains data from all games in the 2004-2020 seasons, the player's individual box scores, team data, game outcome and csv's of player and team IDs

Used data from 2015+ seasons (220,000 rows) because the data set was so large

Data Wrangling

Dataset needed very little cleaning

Used SQL to join player data and game data

Created a Fantasy Points target column by combining boxscore data and transforming necessary columns. Ex: Points + FG made - FG attempted + 2*Steals

GAME_DATE_EST	PLAYER_NAME	PLAYER_ID	GAME_ID	SEASON	FGM	FGA	FG_PCT	FG3M	FG3A	FTSY_PTS
2015-10-02	C.J. Wilcox	203912	11500001	2015	1.0	2.0	0.500	1.0	2.0	2.0
2015-10-02	Jameer Nelson	2749	11500001	2015	3.0	8.0	0.375	1.0	3.0	16.5
2015-10-02	Randy Foye	200751	11500001	2015	2.0	4.0	0.500	2.0	2.0	3.0
2015-10-02	Wilson Chandler	201163	11500001	2015	5.0	12.0	0.417	1.0	3.0	18.0
2015-10-02	JJ Hickson	201581	11500001	2015	1.0	7.0	0.143	0.0	0.0	-1.0

Modeling

I used the rolling average(5) of a player's box score stats to generate variables

Tested a couple of models: Linear Regression, Tree/Forest Regressors, 2 types of Recurrent Neural Networks

Recurrent Neural Networks performed significantly better at predicting individual players

Tested both GRU and an LSTM activation layer, LSTM consistently outperformed the GRU model slightly

LSTM Model

The model used MSE for loss and MAE as the metric

Depending on the player the val MAE ranged from 2 to 25

The model had a lower MAE for players that tended to earn a lot of fantasy points, but had a very high MAE for players that either were very inconsistent or earned very little fantasy points

Next Steps

After training a model for a single player for a single game the next step was to have it predict fantasy points for a week.

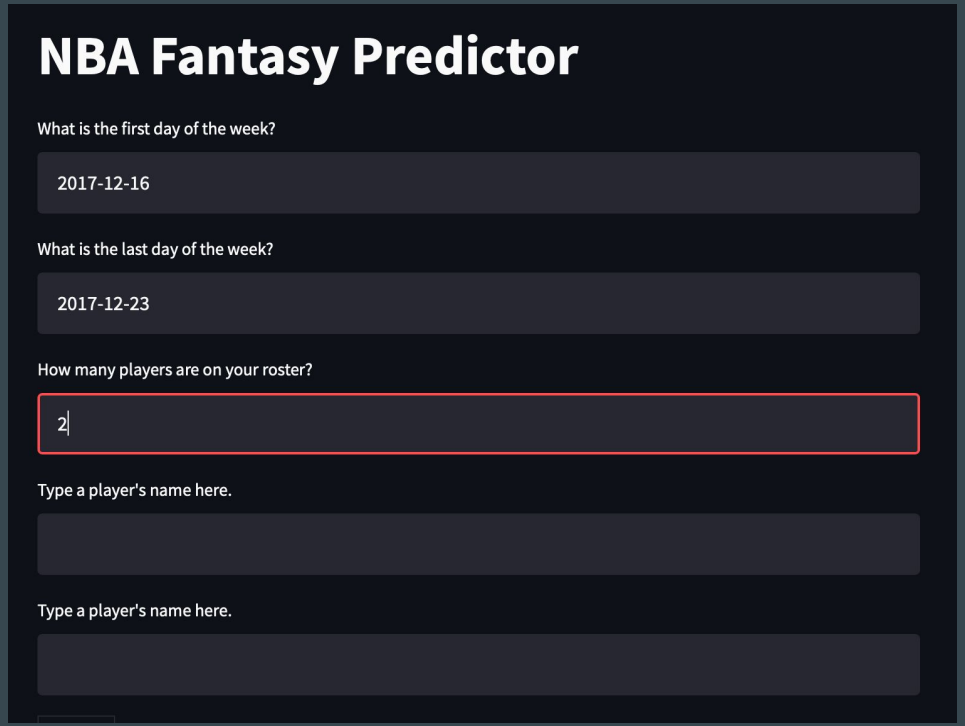
Once this was accomplished I had the model predict a list of players provided by an input function.

```
What is the first day of the week? 2019-01-08
What is the last day of the week? 2019-01-15
How many players are on your roster? 6
Type a player's name here. Stephen Curry
Type a player's name here. LeBron James
Type a player's name here. Kevin Durant
Type a player's name here. Paul George
Type a player's name here. Damian Lillard
Type a player's name here. Kyrie Irving
```

Streamlit (shoutout Noah)

With a functionable model I decided to make a quick app so I wouldn't have to open up jupyter every time I wanted to check a player.

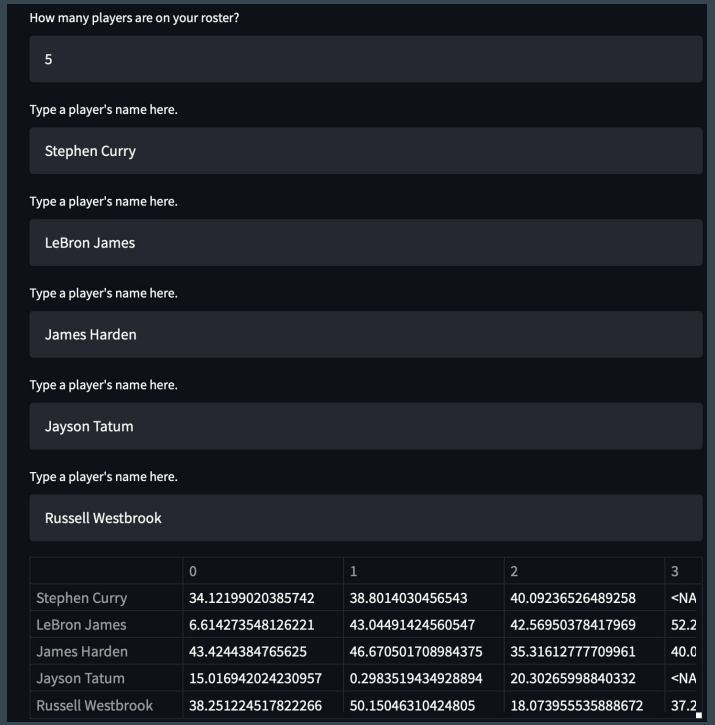
Streamlit is a python library for making apps. I was able to copy most of my code with some minor tweaks.



Streamlit cont.

I'd show you a demo of how it works, but it would take too long, here is what it looks like after it is finished.

Below is the expanded dataframe.



	0	1	2	3
Stephen Curry	34.12199020385742	38.8014030456543	40.09236526489258	<NA
LeBron James	6.614273548126221	43.04491424560547	42.56950378417969	52.2
James Harden	43.4244384765625	46.670501708984375	35.31612777709961	40.0
Jayson Tatum	15.016942024230957	0.2983519434928894	20.30265998840332	<NA
Russell Westbrook	38.251224517822266	50.15046310424805	18.073955535888672	37.2

	0	1	2	3	4
Stephen Curry	34.12199020385742	38.8014030456543	40.09236526489258	<NA>	<NA>
LeBron James	6.614273548126221	43.04491424560547	42.56950378417969	52.25846481323242	<NA>
James Harden	43.4244384765625	46.670501708984375	35.31612777709961	40.02192306518555	23.93157196044922
Jayson Tatum	15.016942024230957	0.2983519434928894	20.30265998840332	<NA>	<NA>
Russell Westbrook	38.251224517822266	50.15046310424805	18.073955535888672	37.261470794677734	<NA>

Conclusion

My friends are going to beat me in Fantasy again.

The model is unable to form accurate predictions for mid-tier players, these are arguably the most important to predict properly because these are the players that I would have to decide whether to play or not.

What Next?

Scrape more data- the data collected is a good starting point, but there are more advanced analytics that could provide a better model.

This model does not take into account the team the player plays against nor does it take in whether it is a home or an away game. (Steph Curry does not play well in Staples Center)

I would also want to add in player specific data such as points scored off of pick and roll, and how team defenses fare against the pick and roll.

Special Thanks to...

Noah and Patrick for putting up with my frantic messages begging for meetings.

Thomaz and Kennedy for hanging out while I cried over my capstone.

Kennedy one more time for pointing me towards the dataset.

Sources

Streamlit - https://streamlit.io

Kaggle dataset - https://www.kaggle.com/nathanlauga/nba-games

Images:

https://support.espn.com/hc/en-us/articles/360000958652-Setting-Your-Lineup

https://github.com/ChrisKatsaras/React-NBA-Logos