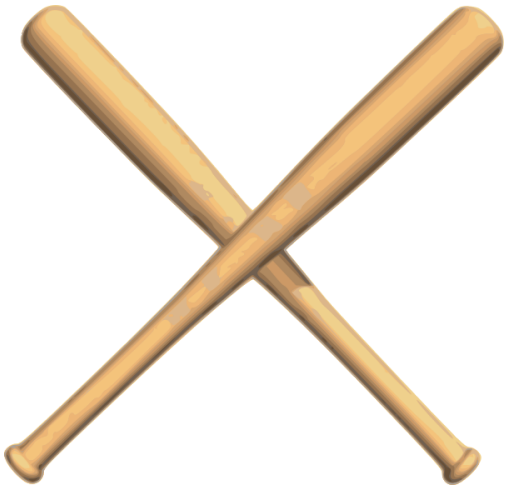
SPRINGBOARD DATA SCIENCE CAREER TRACK

MLB DFS HITTING CAPSTONE
FINAL REPORT



Introduction

Predicting points for hitters is arguably the most challenging task in daily fantasy sports. The best hitters will rise to the top over the course of a 162-game season, but on any single day, it's not unusual for a Bryce Harper or an Aaron Judge to score zero fantasy points.

Those zeroes take you out of the running for the big prizes in large tournaments, and they hurt your chances of finishing in the money in cash games. Our model will help you avoid those big, fat bagels and predict the hitters who will help you turn a profit.

Methodology

Our data comes from 30 days of the 2021 Major League Baseball season, spaced four days apart from May 1 to Aug. 29. We used 96 variables to predict our target variable, DraftKings points. FanGraphs was used as our data source.

Final model: XG Boost Regressor

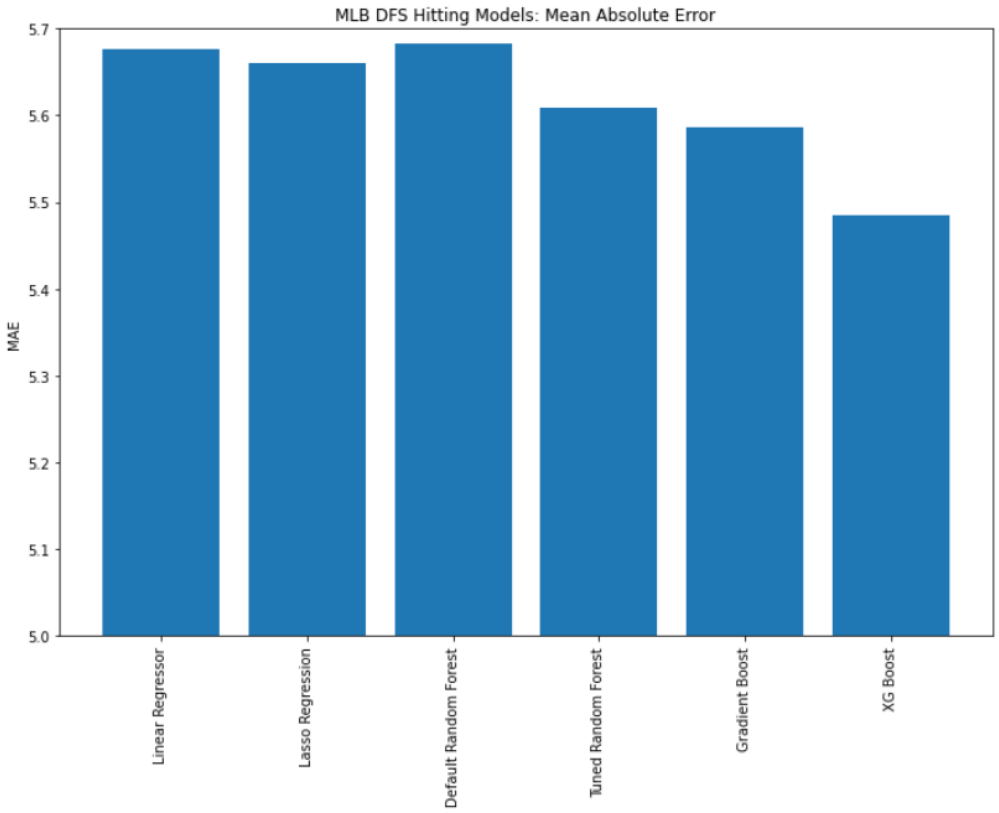
None of our variables were strongly correlated with DraftKings points. We knew we were facing an uphill battle.

As a baseline, we started with the simplest form of Linear Regression. From there, we went to Random Forest. That was an improvement, but Random Forest consists of several decision trees that all work independently. We knew we ultimately would need a model in which each individual sub-model learns from, or is boosted by, the previous one.

Mean Absolute Error

Mean Absolute Error is the average difference, positive or negative, between a player's predicted points and his actual points.

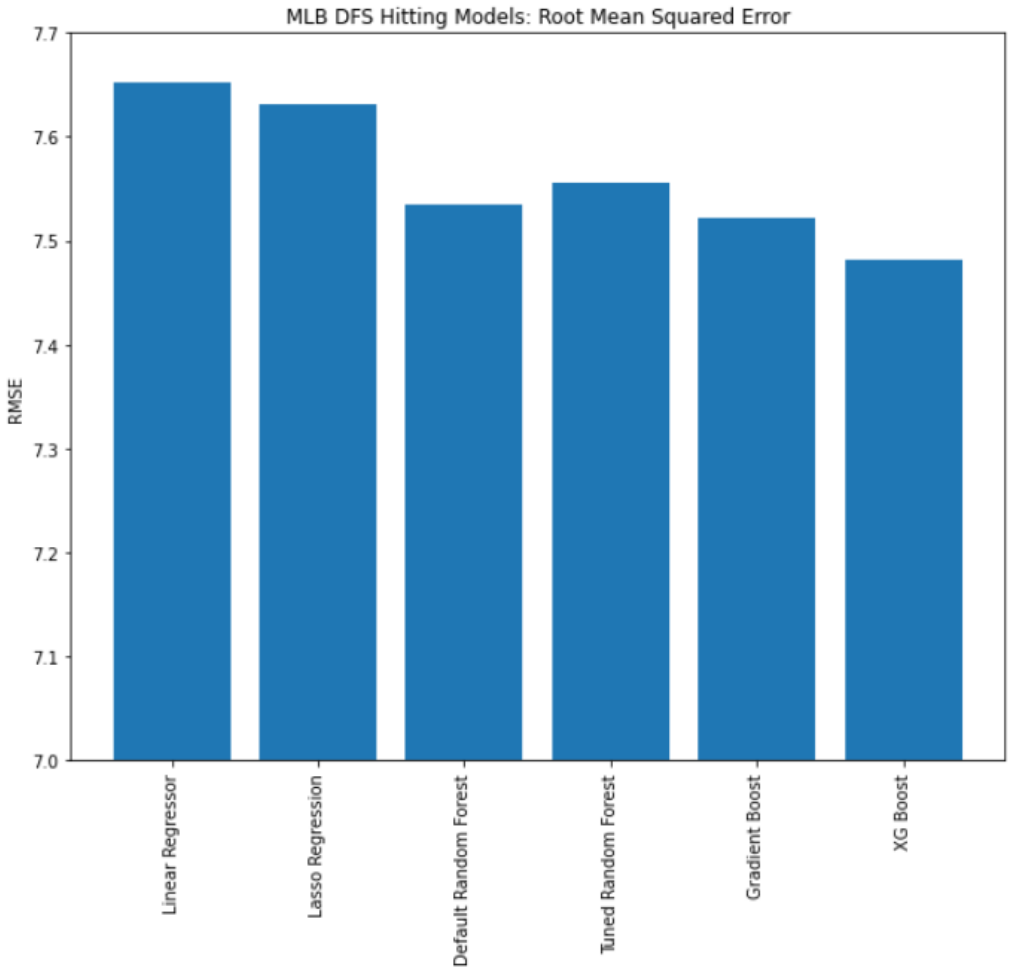
Please note that the following bar chart for Mean Absolute Error starts at 5.0 rather than zero. The MAE for our first model, Linear Regressor, was 5.676. For our final model, XG Boost, it was 5.489.



Root Mean Squared Error

Root Mean Squared Error is a model metric that penalizes larger errors. The point distribution for MLB hitters is weighted heavily toward zero, but there are a few high-scoring outliers. Accurately predicting and rostering those outliers is the key to winning at daily fantasy baseball. We wanted a model that would try to predict those outliers, which is why we prioritized RMSE over MAE when fine-tuning our model.

The following bar chart for RMSE starts at 7.0. The improvements in RMSE were similar in scale to that of MAE, with a 7.653 score for Linear Regressor and a 7.482 score for XG Boost.

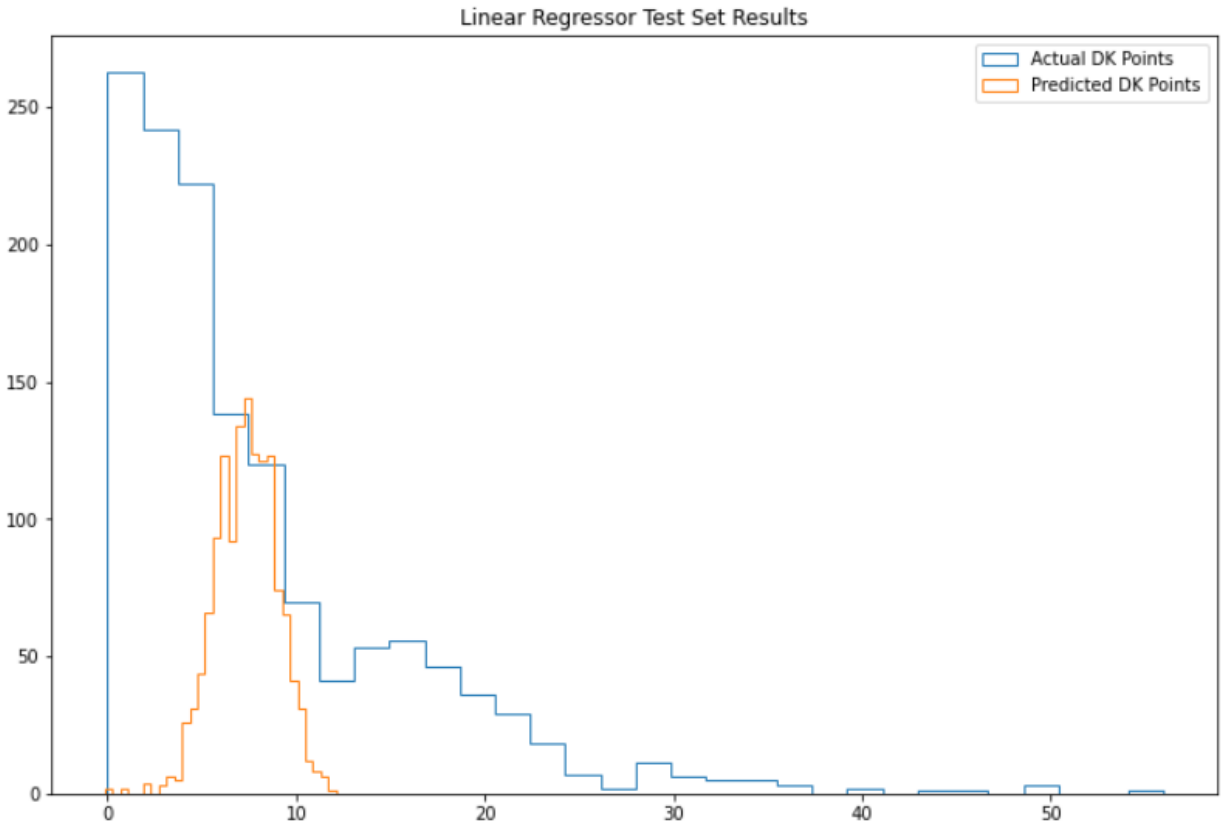


Linear Regressor predictions

The following histograms compare the distribution of model predictions to actual DraftKings points.

This distribution is shown in 30 bins, meaning it's broken down into 30 equal categories.

The actual distribution of DraftKings points, our target variable, is right-tailed. The distribution of predictions for Linear Regressor is left-tailed.

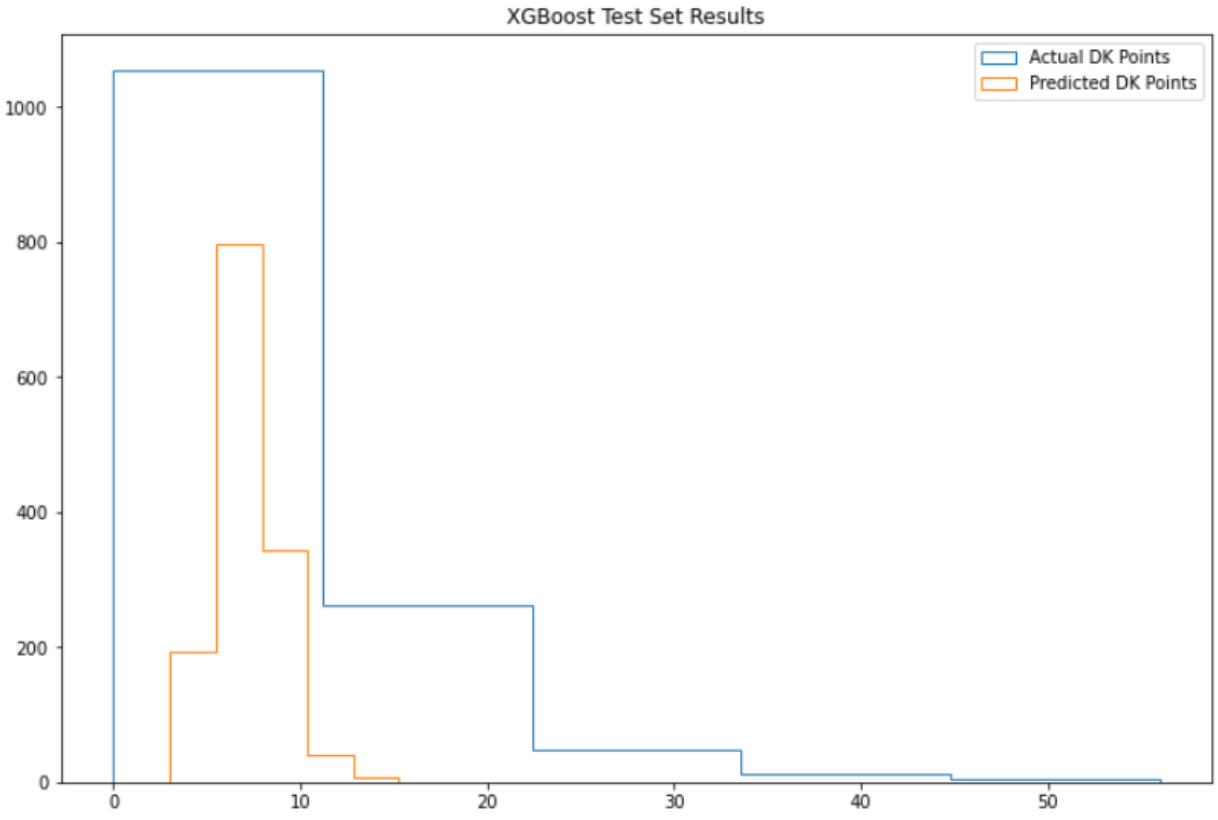


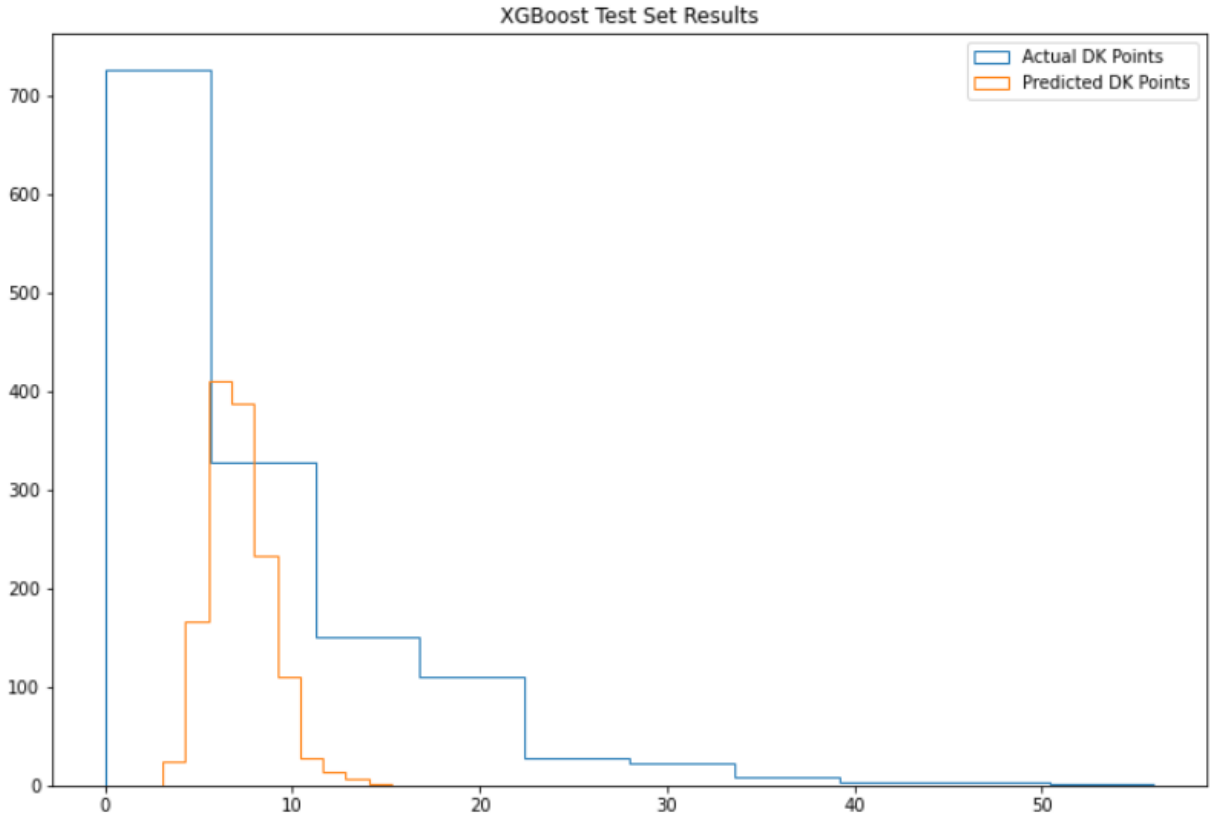
XG Boost predictions

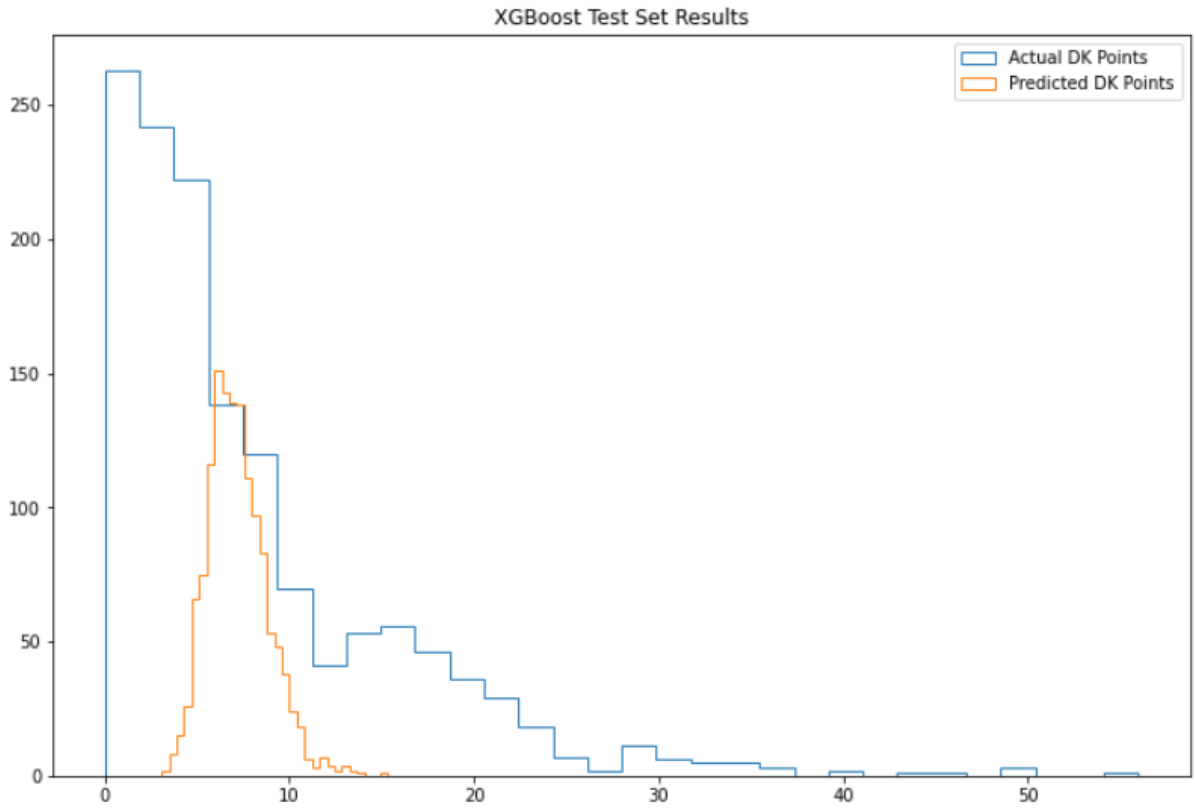
For XG Boost, we will show three histograms. The first contains five bins, the second 10 bins and the third 30 bins.

None of our models, not even XG Boost, will predict those high outlying numbers. The maximum number of actual DraftKings points in our dataset is 56, and XG Boost's maximum prediction is 15.3.

Notice, however, that at each bin level our prediction distribution shadows that of the actual scores. It's much more right-tailed than our baseline model. It's almost like our predictions are a child of the actual score distribution.




XGBoost Test Set Results
Actual DK Points
Predicted DK Points
700
600
500
400
300
200
100
0
0
10
20
30
40
50



14 is the magic number

So what's the use of a model that won't make a prediction higher than 15.3 when we need to fill our lineup with numbers in the 20s and 30s?

Usually when a player scores in the 20s, 30s or beyond, he hits a home run. On DraftKings, a solo home run is worth 14 points (10 points for the homer itself, two points for the run scored and two points for the RBI). Therefore, when a player scores 14 points, there's a reasonable chance that he has hit a home run.

If all eight of your hitters score 14 points, you're not guaranteed to win the big bundle in a large-field tournament, but there's a strong chance that you'll finish in the money. In cash games, you're in a good position if most of your hitters cross the 14-point threshold.

Cash games are the contests where you try to double your money by finishing in the top 40 percent of the field or, say, win $18 in a head-to-head with a $10 entry,

Using Mean Absolute Error

Our XG Boost model reduced our Mean Absolute Error to a number under 5.5. So if we look at players predicted to score at least 8.5 points, of course there's a possibility that they could be a dud with 3 points. But there's a similarly reasonable possibility that they score 14.

Our test set contains 1,381 samples. Of those samples, 273 players were predicted to score 8.5 or more points. That's 20 percent. When this model is deployed live, we won't have to sift through that many players. In a 10-game slate involving 20 teams, there would be 160 eligible hitters. Twenty percent of 160 is 32. With a list of only 32 interesting players, lineup construction becomes much more manageable.

Of course, not all 32 of those players will score 14 points. The idea is to create multiple lineups with different combinations of these players.

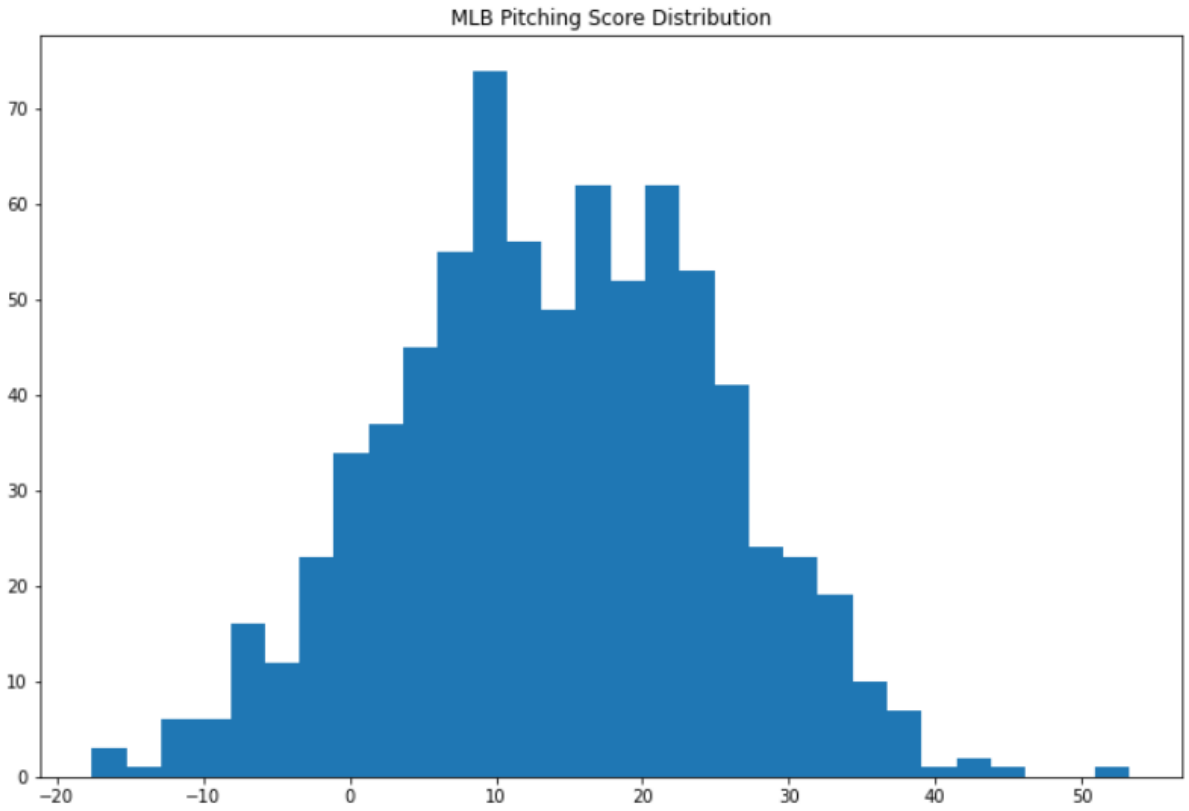
Conclusion

A DraftKings lineup consists of 10 players, eight of which are hitters. The other two spots are filled by pitchers.

Our next step will be to create a pitching model. Pitching is a more reliable source of points than hitting, as the 30-bin histogram below shows.

This is 775 starting pitcher scores from the same 30 dates in 2021 that we used for our hitting model.

Keep in mind that pitching scores can be negative. Pitchers can lose points for allowing hits and earned runs. Still, zero is well to the left of the peak of this distribution. With hitting, zero is always part of the most frequent distribution group.



A profitable DraftKings MLB lineup is anchored by pitching points. That's our next step, but now that we have a hitting model, the hard part is finished.