Expected Effective Field Goal Percentage

A project by Joey Callahan
May, 2021

Background

In recent years it has come to light that traditional individual box score stats do not fully encompass a player's entire contribution to their team as it relates to winning a game. Stats like points, assists, and steals, while still important, are heavily influenced by volume and opportunity.

A relatively new stat that measures a player's efficiency of scoring is effective field goal percentage (eFG%), which takes traditional field goal percentage (FG%) and intuitively weights three-point field goals made (3pt FGM) 1.5-times more heavily than two-point field goals (2pt):

$$eFG\% = \frac{2pt\ FGM + 1.5 * 3pt\ FGM}{FGA}$$

Background

All sports contain some degree of random chance or luck, which, over time, evens out and allows for true talent level to be observed. For example, a pitcher who limits total runs allowed but gives up a high rate of solo homeruns has likely been experiencing good fortune that no one was on base when he's made mistakes. He will eventually make these same mistakes in more damaging situations and allow more runs.

In football, if a quarterback has avoided throwing interceptions but hasn't been accurate in completing passes, it can be expected that he will turn the ball over with more frequency if he continues to throw passes that don't find receivers at a constant rate.

The same may be true of shooting in basketball. At some threshold of volume, a hot streak or a cold streak must reflect a true level of talent, but there may be ways to forecast if performance is sustainable, or if regression toward mean should be expected.

Objective

Create a new metric – expected effective field goal percentage (**xeFG%**) – that predicts the expected eFG% of shot attempts by a player.

A number of features will be considered for use in a predictive model, including shot location and distance, proximity of the nearest defender, and the angle of the defender to the shooter relative to the basket, (to see what degree the defender was in front of the shooter).

Use Cases

Prediction

It is possible xeFG% will be a better predictor of future shooting efficiency than eFG%

Evaluation

At a large enough volume, both eFG% and xeFG% should stabilize, which will allow for evaluation of potential differences between the two metrics.

A player who consistently outperforms his xeFG% is likely one of high skill, who takes difficult shots and makes them with greater regularity than is typical. A player who underperforms his xeFG% is likely one with lower true talent, who misses shots most players make.

Coaching

In observing the potential differences or similarities between players' eFG% and xeFG%, coaches can optimize how a player is used in their offense.

For example, if player's eFG% and xeFG% are similar but below league average, this could indicate poor shot selection, as he may be taking difficult shot attempts and need his offensive role reduced or to reevaluate his style of play.

Data

The dataset comes from the NBA’s briefly publicly available player movement tracking database, and contains over 500 games from the 2015-2016 NBA season.

Cameras track and record the exact coordinates of all 10 players and the ball 25 times per second, creating 2-2.5 million rows of data per game.

The column that would be used to join movement data with a corresponding data frame of box score events is dirty and unreliable, thus making extracting relevant rows difficult.

Due to the uncertainty of matching unlabeled continuous location observations with the correct box score event, the end result was a dataframe of just 43,000 shot attempts.

Modeling

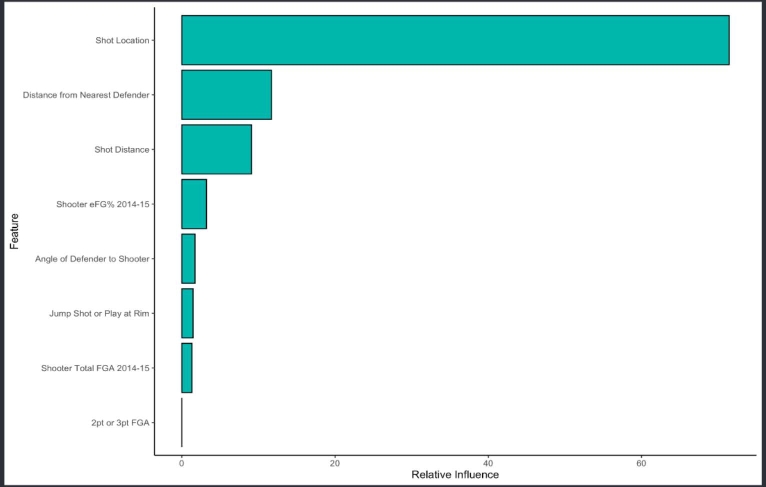
The model with best performance was a gradient boosting machine regression trained on a stratified sample of data containing an even number of 2pt and 3pt shots and even number of each shot type made and missed, created by oversampling the minority classes.

Performance:

- RMSE at shot level: 1.171
- RMSE at player level: 0.408

When predicting the actual value of a shot attempt (0-3 points), the model did not perform well, but when aggregated to the player level, the RMSE dropped notably. The model was predicting fairly well on average, even if not at the level of an individual observation.

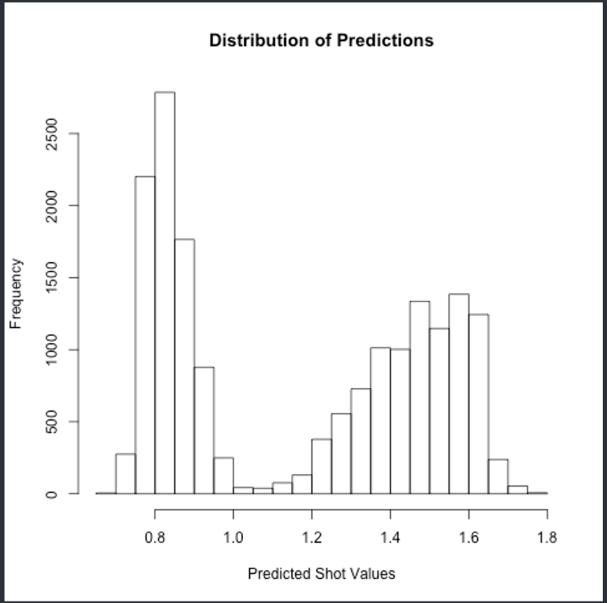
The figure at right shows the relative influence of the eight features used as model inputs. Shot location, a categorical variable with six different locations in the half court was far and away the most influential, followed defender proximity and shot distance.

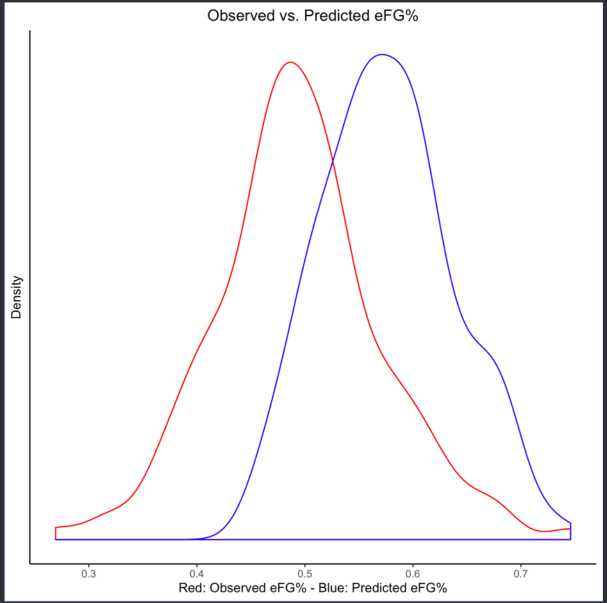


Results

Results

Left: the distribution of predictions passes the eye test. The highest concentration of predictions occurs at values less than one, signifying the likelihood of a missed shot, and the range of predictions end before 2, which would indicate a 3pt FGA with .667 chance of being made, or a 2pt FGA with certain success.

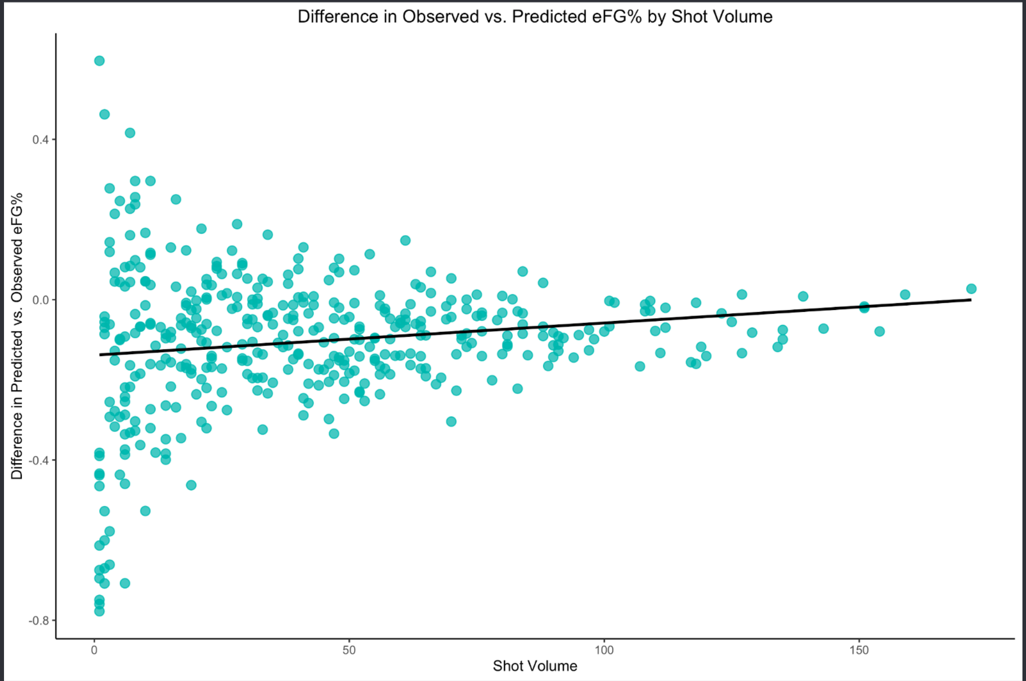




Unfortunately, however, the model inflated the expected value of shot attempts across all shots. The figure on the right shows the observed eFG% of all shots in the test set was .504, while the predicted eFG% was .585

Analysis

- Due to the extraction and cleaning issues that resulted in a relatively small set of data for training and predicting, it is not possible to determine if a stabilizing volume threshold was reached by any player in the sample, or what that threshold might be.

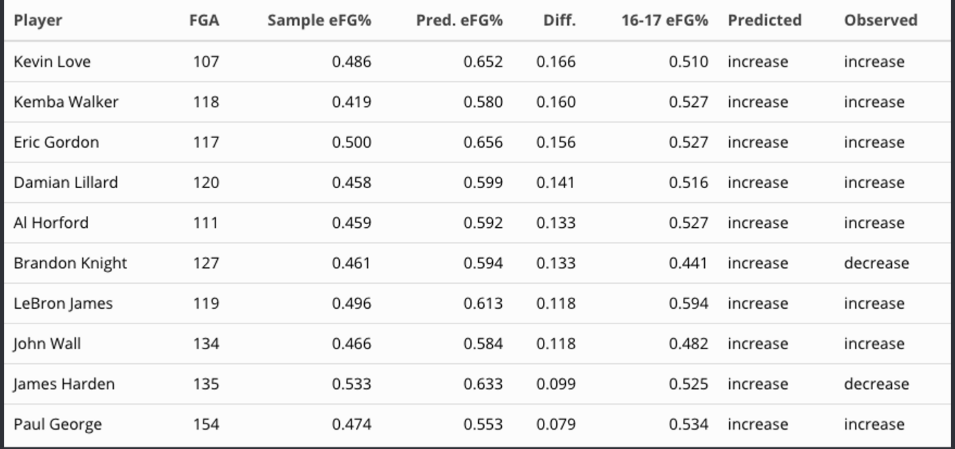


The plot illustrates the higher variance between predicted and observed eFG% values at lower shot volume

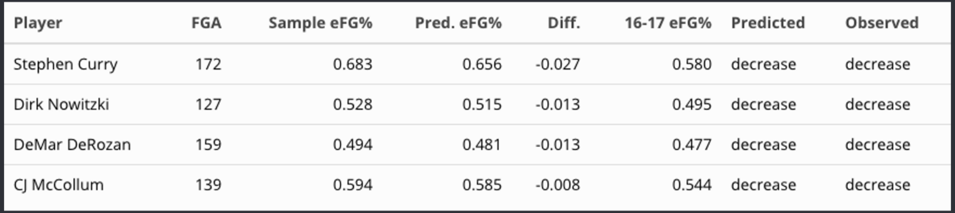
Analysis

Given the uncertainty created by the size of the data, the following conclusions should be considered with appropriate skepticism.

Filtering for players with at least 100 FGA attempts in the prediction set, of the players with the top 15 largest predicted increases in eFG%, 12 saw their eFG% increase during the 2016-2017 season. Due to the model generally over estimating xeFG%, only four players with 100 FGA were predicted to see a drop in eFG%, but all four did experience a decrease the following year.



Player	FGA	Sample eFG%	Pred. eFG%	Diff.	16-17 eFG%	Predicted	Observed
Kevin Love	107	0.486	0.652	0.166	0.510	increase	increase
Kemba Walker	118	0.419	0.580	0.160	0.527	increase	increase
Eric Gordon	117	0.500	0.656	0.156	0.527	increase	increase
Damian Lillard	120	0.458	0.599	0.141	0.516	increase	increase
Al Horford	111	0.459	0.592	0.133	0.527	increase	increase
Brandon Knight	127	0.461	0.594	0.133	0.441	increase	decrease
LeBron James	119	0.496	0.613	0.118	0.594	increase	increase
John Wall	134	0.466	0.584	0.118	0.482	increase	increase
James Harden	135	0.533	0.633	0.099	0.525	increase	decrease
Paul George	154	0.474	0.553	0.079	0.534	increase	increase



Player	FGA	Sample eFG%	Pred. eFG%	Diff.	16-17 eFG%	Predicted	Observed
Stephen Curry	172	0.683	0.656	-0.027	0.580	decrease	decrease
Dirk Nowitzki	127	0.528	0.515	-0.013	0.495	decrease	decrease
DeMar DeRozan	159	0.494	0.481	-0.013	0.477	decrease	decrease
CJ McCollum	139	0.594	0.585	-0.008	0.544	decrease	decrease

It is possible that xeFG% could become a valuable forecasting tool with some upgrades.

Future Research

Obtaining a larger and cleaner dataset will be critical to future research. Given the richness of this data and the value it could hold to an NBA team, I believe this must already exist, but is proprietary to those affiliated with the league.

With more data and more reliable data, future research should target understanding the effect of volume on xeFG% to better interpret and apply its use cases.

After achieving a better understanding of volume’s effect on xeFG% and better data theoretically improving model performance, the ultimate goal of this research should introduce another new metric that aims to quantify the impact a player has on the xeFG% of all shot attempts while on the floor. With location and distance from defender, this could capture elements of the game like court spacing and defensive rotations that are not easily measurable at present.