Predicting NBA Players' Three-Point Percentages

FUTURE ANALYTICS STARS

Lauren Manis | December 2023

Project Scope

Predicting each player's three-point percentage at the end of the 2022-23 season, given only their shooting statistics from October/ November 2022.

This project used data from only the one provided source, which contained 12 columns of shooting data for 108 players in the league.

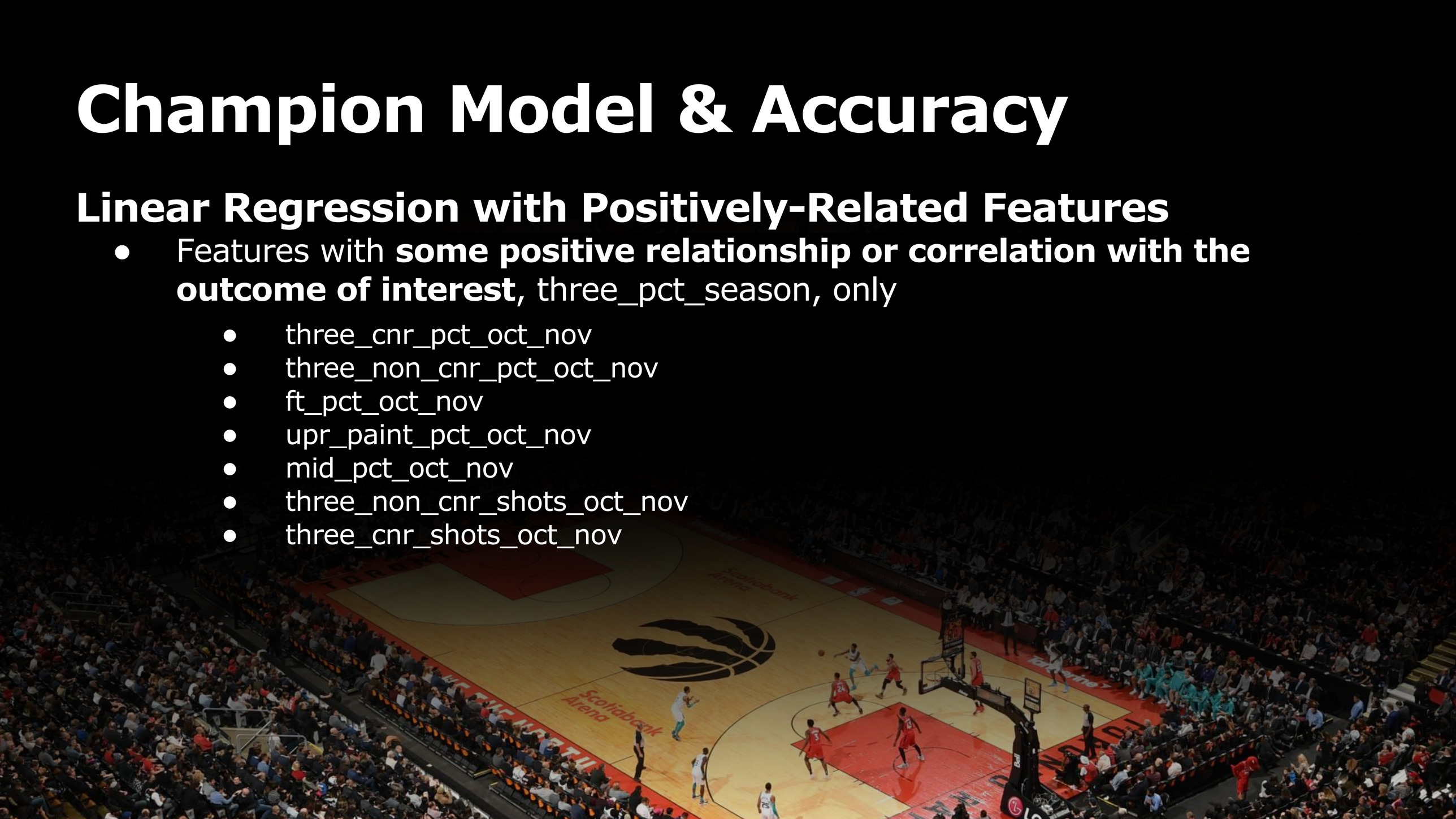
Process

1. **Exploratory Data Analysis**
    1. Looked into correlations between variables, levels of skewness, etc. in order to get familiar with the data
2. **Baseline Simple Linear Regression**
    1. During the EDA phase, it became clear that there would be a positive relationship between many of the variables and the outcome of interest, three_pct_season. Just thinking logically, it makes sense that a player with a high shooting percentage from long and mid-range would also have a high 3-point shooting percentage. Knowing this, it made sense to start looking into **Linear Regressions**
3. **Additional Linear Regression Models**
    1. Tested different combinations of variables in order to find the optimal mix
        1. Accounting for potential collinearity
        2. Removing variables with no predictive power
        3. Etc.
4. **Other Statistical Models**
    1. Fitted the data to Random Forest and Decision Models
5. **Champion Model Selection & Evaluation**
    1. Selected a “champion” model, which was the model with the highest R-Square

Champion Model & Accuracy

Linear Regression with Positively-Related Features

- Features with **some positive relationship or correlation with the outcome of interest**, three_pct_season, only
    - three_cnr_pct_oct_nov
    - three_non_cnr_pct_oct_nov
    - ft_pct_oct_nov
    - upr_paint_pct_oct_nov
    - mid_pct_oct_nov
    - three_non_cnr_shots_oct_nov
    - three_cnr_shots_oct_nov

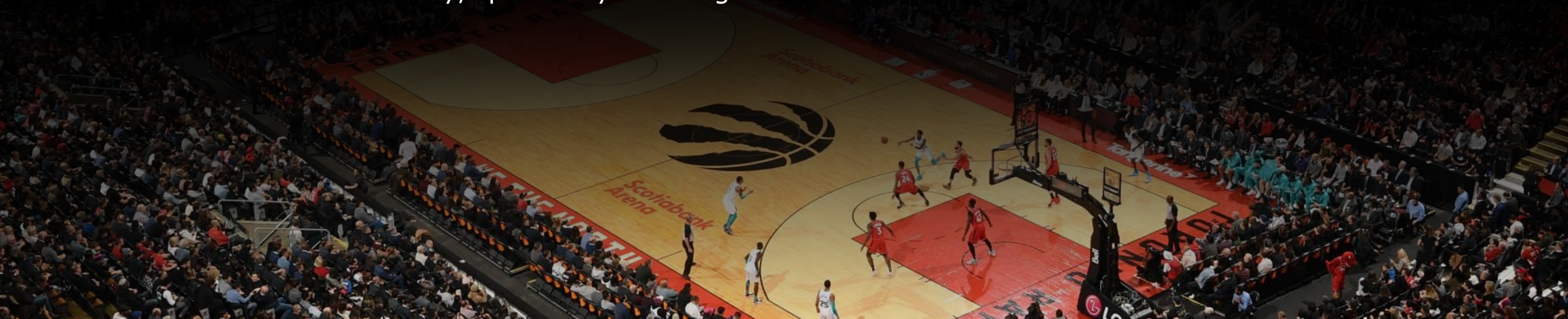


Champion Model & Accuracy

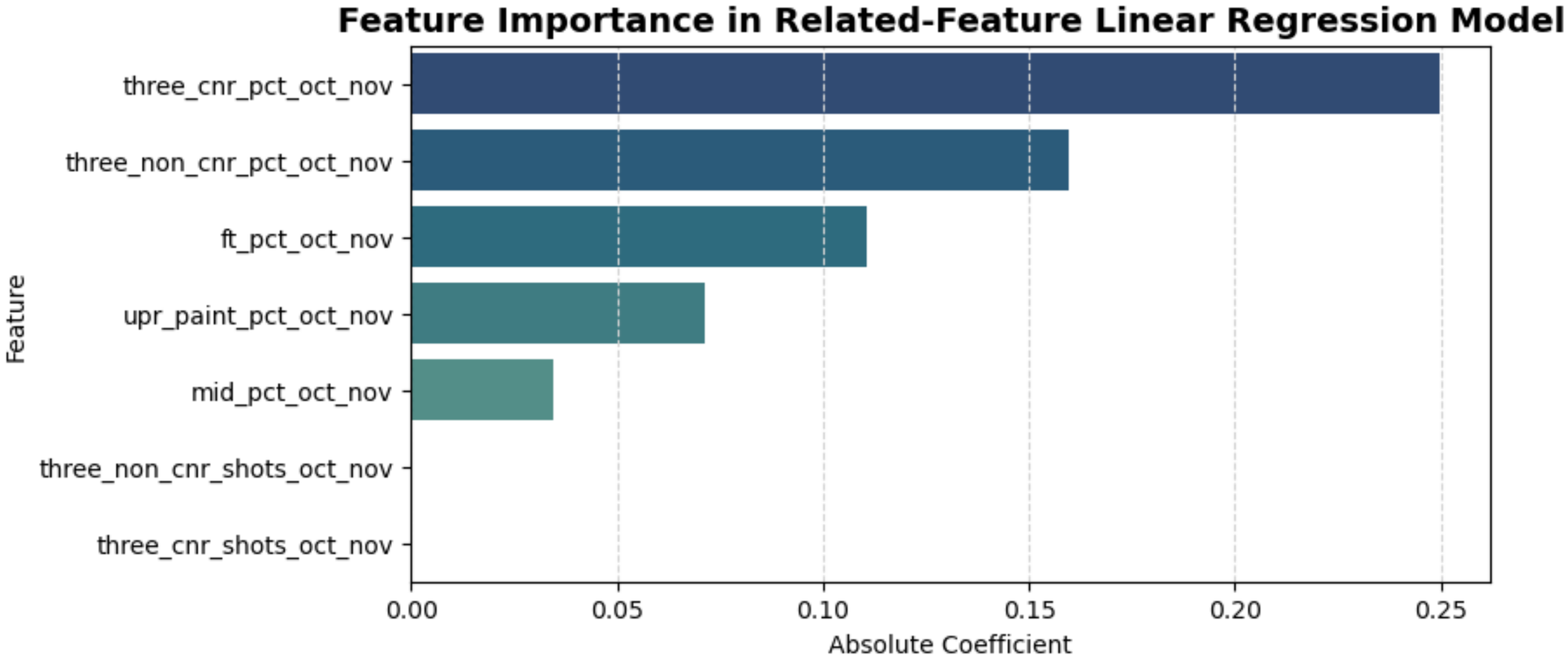
Linear Regression with Positively-Related Features

- Features with some positive relationship or correlation with the outcome of interest, three_pct_season, only

- **R-Square = 0.506**
  - The R-square of 0.51 indicates that this model is only **accounting for about 51% of the variance, or only getting us about 51% of the way to perfectly predicting players' full-season three-point percentages**.
    - R-Square is a very commonly used statistic when evaluated model performance and accuracy, specifically with Regression.



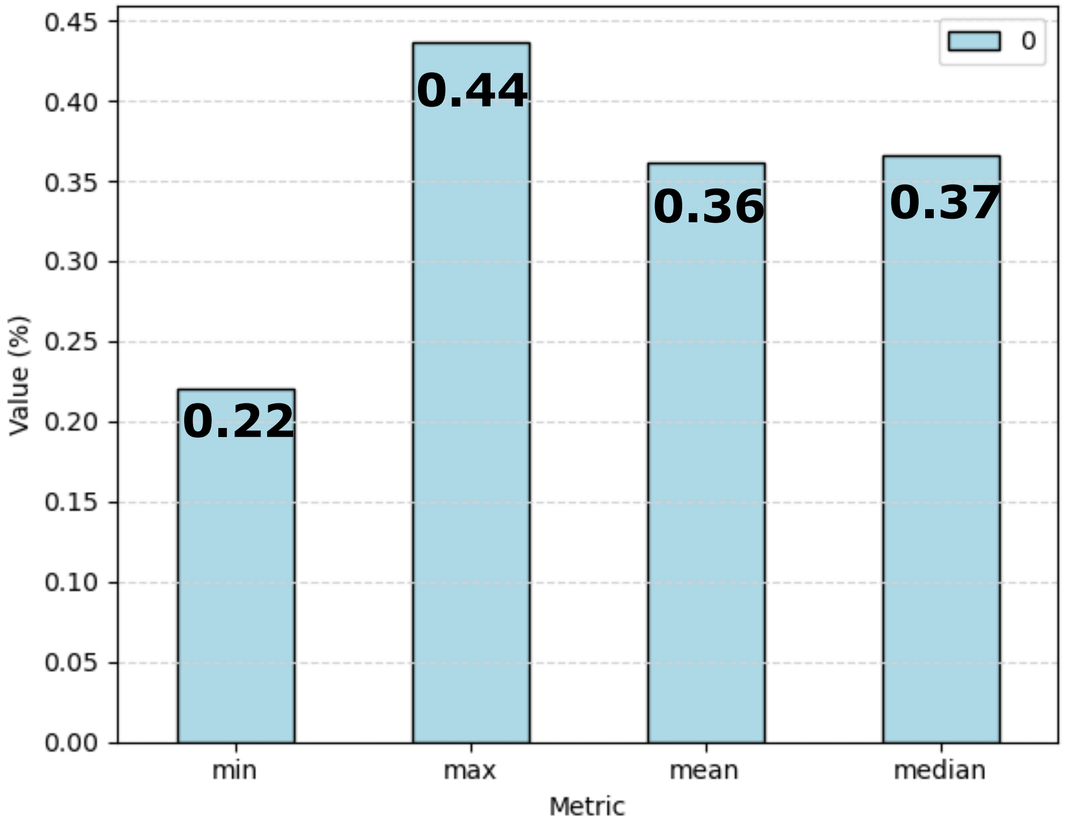
Champion Model Predictions



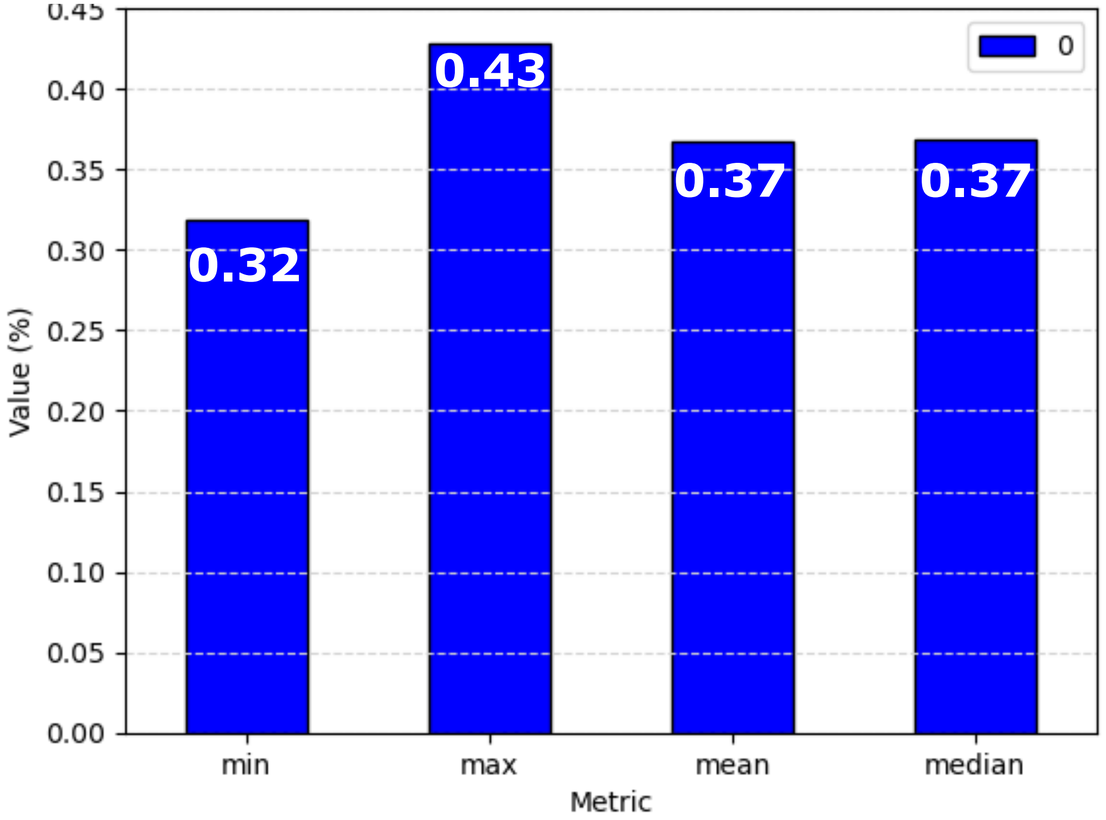
The **most important or impactful features in the model are three_cnr_pct_oct_nov and three_non_cnr_pct_oct_nov**. It absolutely makes sense that players with high 3-point percentages in October and November would go on to have higher 3-point percentages across the entire season.

Champion Model Predictions

Actual three_pct_season Summary Statistics



Predicted three_pct_season Summary Statistics



The **summary metrics from the predictions are pretty close to the summary metrics found from the actual data** for three_pct_season, showing that the model's predictions are indeed plausible.

Champion Model Predictions

Top 10 Predicted Three-Point Percentages

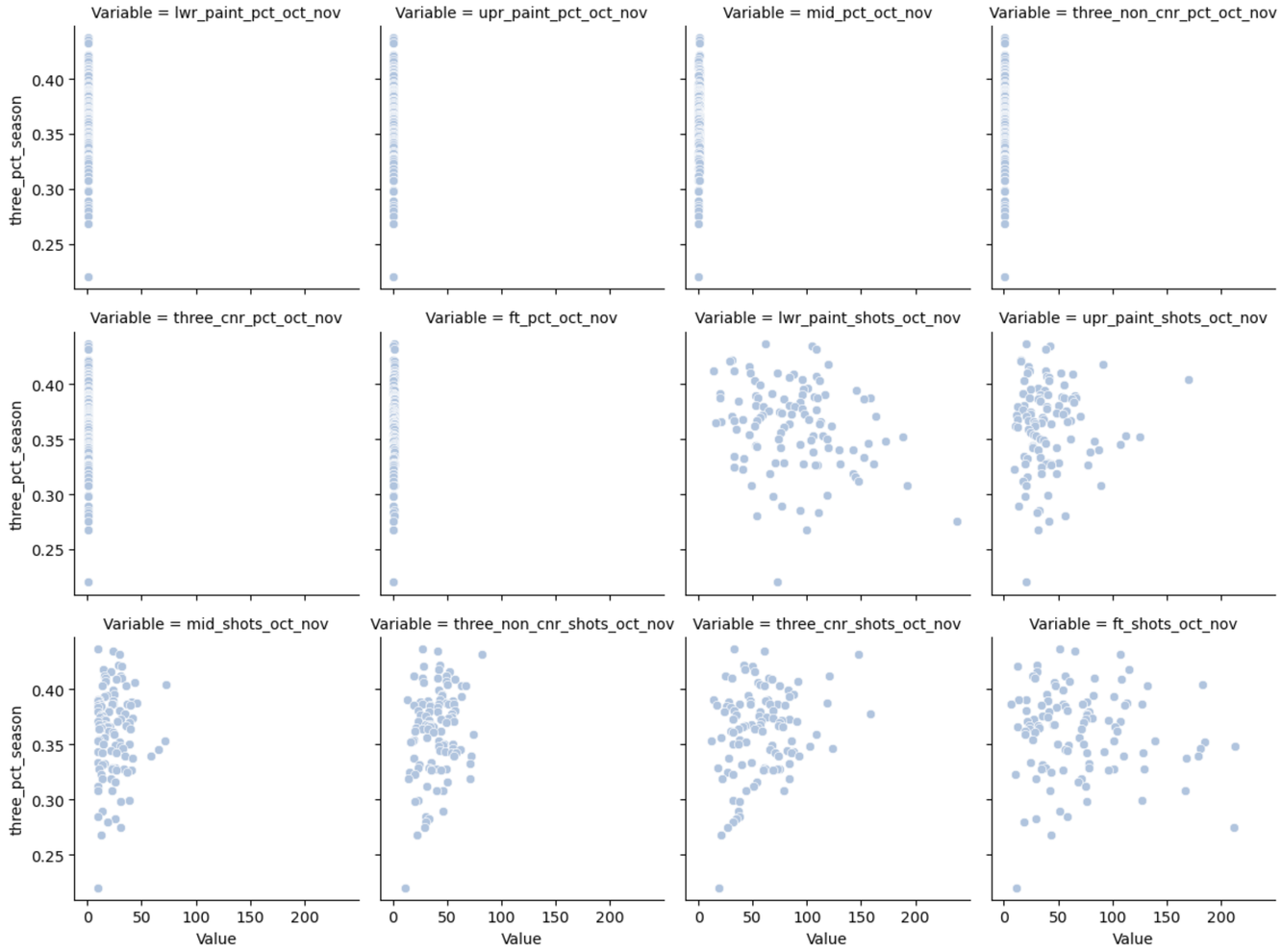
	Name	predicted_three_pct_season	three_pct_season
12	Malcolm Brogdon	0.428374	0.437
4	Malik Beasley	0.402283	0.359
9	Bojan Bogdanovic	0.399224	0.409
6	AJ Griffin	0.389037	0.387
1	Patrick Williams	0.387497	0.410
2	Jevon Carter	0.386318	0.421
19	Brook Lopez	0.380301	0.375
14	Andrew Wiggins	0.379889	0.396
20	Tyrese Haliburton	0.378278	0.407
7	Bobby Portis	0.370320	0.373

Many of the projected three point percentages are relatively close (within a few percentage points of) **to the true values**.

Within this group of players, the one with the highest actual three_pct_season was Malcolm Brogdon, and he is also the player with the highest predicted three_pct_season value, which is a positive sign.

Champion Model Predictions

Feature Correlations with three_pct_season



three_cnr_pct_oct_nov	0.508241
ft_pct_oct_nov	0.347510
upr_paint_pct_oct_nov	0.290882
three_non_cnr_pct_oct_nov	0.271946
lwr_paint_shots_oct_nov	0.249411
three_non_cnr_shots_oct_nov	0.230933
three_cnr_shots_oct_nov	0.227750
mid_pct_oct_nov	0.164970
ft_shots_oct_nov	0.083265
mid_shots_oct_nov	0.068135
lwr_paint_pct_oct_nov	0.062074
upr_paint_shots_oct_nov	0.052534

Model Improvement & Next Steps

- The most obvious way to improve the model would be to **use a larger data set, both in terms of rows and columns**.
    - Modeling and making predictions with only 108 data points, or in this case, players, is not ideal and certainly, using data on every player in the league over the last several seasons would allow for better and more accurate projections.

- Additionally, other information such as **player's historic three-point and overall shooting percentages, positions, and other statistics** would also likely enable better predictions.

- Another approach could be to **apply some transformations to the data, or hyper-parameter tune** the random forest or decision tree models for better accuracy.