



Shot by Shot: Can we predict NBA shot success?

By: Jonathan Beyene

TABLE OF CONTENTS



01

About the project

Brief overview of the goals of this project

02

Relevant Work

Discussing other relevant work on this topic

03

Data

Describing the dataset we are working with

04

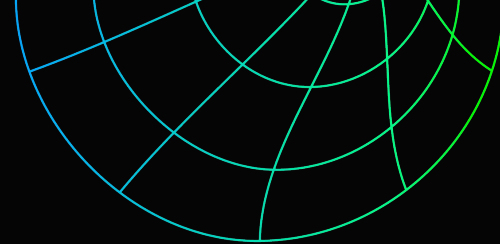
Modeling

Going over decisions made in the modeling process

05

Results

Explanation of overall model performance





01

▶▶▶▶

About the Project





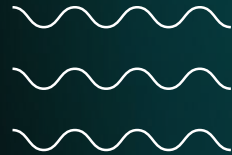
Basketball
is a game
where two
teams
compete to
see who can
score the
most points

It would be
beneficial to
players,
teams, and
organizations
if they were
able to
predict the
outcome of a
shot





In this project we will
be attempting to see
how well we can predict
the outcome of a
basketball shot using
machine learning





02



Relevant Work



Predict The Result of an NBA Shot: Machine Learning Analysis Project



This project was written by Albert Tan where he trained seven different models using data from the 2014-2015 season and focused on features such as 'Shot Clock', 'Dribbles', and 'Closest Defender Distance'. His best model was a Decision Tree Classifier with ADA Boost that had an accuracy of 61.9%.

How I plan to differentiate my project is through different feature selection. This project uses 'Closest Defender Distance' to capture defensive pressure. I was unable to find this data and instead utilized a team's defensive rating to capture this aspect of the game. I also implemented a multitude of different features such as a player's shooting averages and the one-hot encoded action type for the shot.



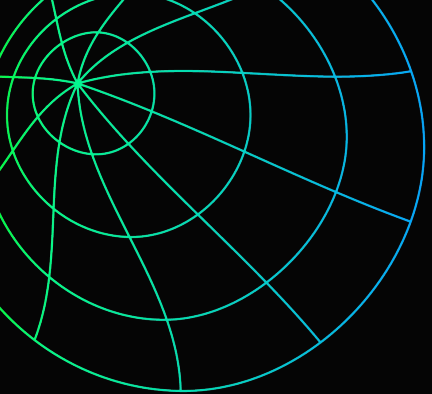
Analysis of Machine Learning Models Predicting Basketball Shot Success



This project uses shooting data from the 2015-2016 NBA season and utilizes Logistic Regression, Neural Networks, and Gradient Boosting as models. The top accuracy achieved by this project came from the Gradient Boosting model with an accuracy of 65.2%.

This project and my own dataset contain some similarities in features, such as utilizing 'ACTION_TYPE', 'MINUTES REMAINING', and 'SECONDS REMAINING', however this biggest differentiator is my attempt to see if utilizing previous season shooting averages and the opposing team's defensive rating can improve the accuracies.

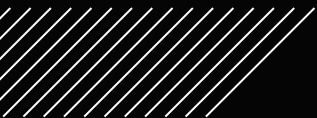





03



Data



Dataset Breakdown



Date range: 2004 to 2021

Shooting

Dataset found on github that scraped shooting data from NBA.com

Player Statistic

"Comprehensive" dataset that contains every NBA player and their associated statistics for each season

Defensive Rating

Contains every teams and their associated team defensive rating

Shooting Dataset

In the beginning, I used a play by play dataset found on kaggle that scraped the data from a website called basketball references. This dataset had shot data contained within it but difficulties occurred when attempting to merge this data with the player statistic dataset due to name formatting



Left: Trae Young
Right: Thaddeus Young

T. Young = T. Young

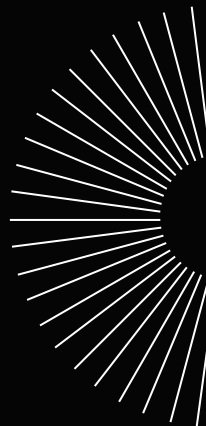


Shooting Dataset



I later pivoted to a new dataset found on github, which scraped its data from NBA.com and contained relevant features such as 'SHOT_DISTANCE', 'LOC_X', 'LOC_Y', and 'ACTION_TYPE'.

'ACTION_TYPE' is a categorical feature that contains the description of the shot action. Since I wanted to contain it as a feature I one-hot encoded all the possible values and this expanded my feature space significantly.

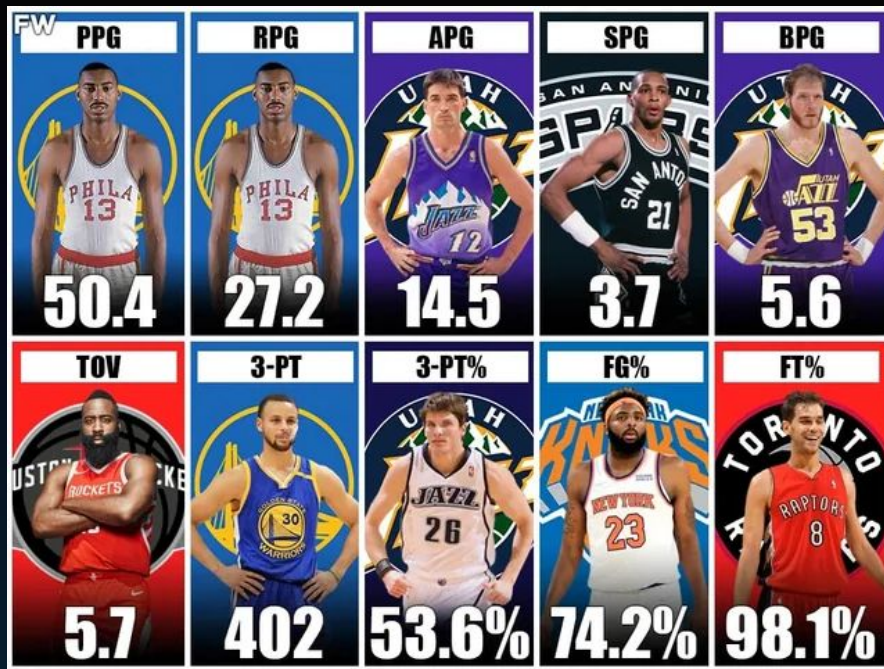


Why do I believe ACTION_TYPE has some importance



Player Statistics

The player statistics dataset is a “comprehensive” dataset that has every single player and their associated season averages for a given season. After exploring this dataset I realized that it truly wasn’t comprehensive. For some players, it was missing a season in which they actually did play in, and for others it is missing entire careers.



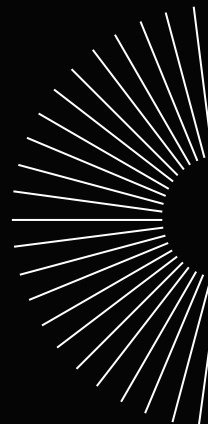
Player Statistics



This missing data proved to be problematic for the merging of these datasets. I took the simplest approach and dropped the missing values, however if I were to re attempt this I would have a different approach.

For players missing a season, I would take the average of all of their previous season stats and use those values to fill the missing data

For players who have entire careers missing, I would create an “average nba” career statistics that would be the average of the whole dataset for the previous season, and use those values as to fill the missing data



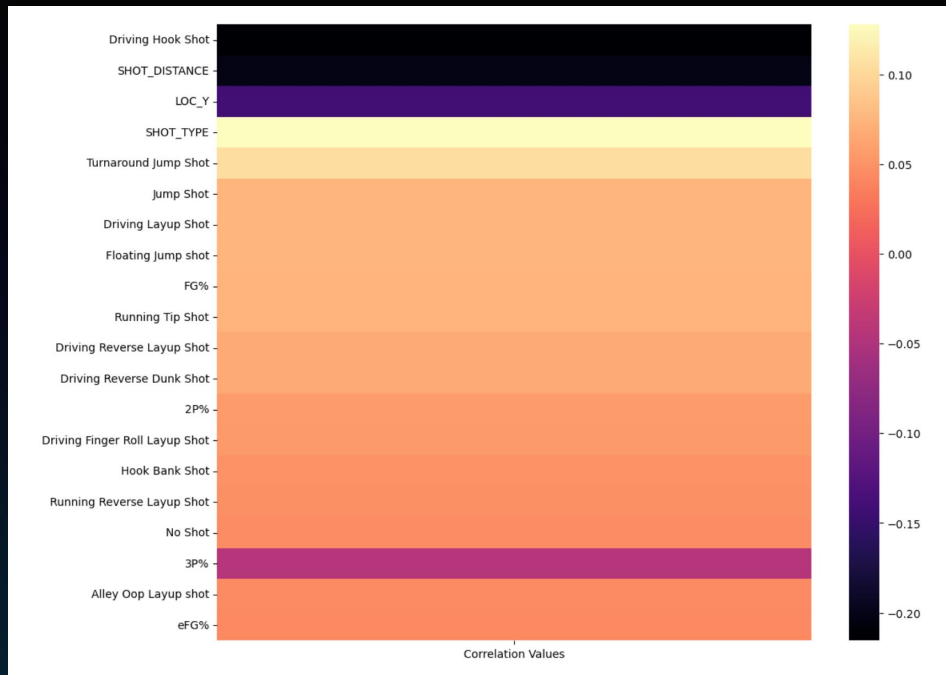
Defensive Ratings

Due to domain knowledge of basketball, I wanted to include some sort of defensive feature as shots are usually contested. After scouring the internet I was unable to find a premade defensive rating dataset. I found the data on NBA.com however it would require me to scrape it off of their website. Since there are only 30 NBA teams and my years for this data span from 2004-2021 I decided to take the simpler route and manually input these values into an excel file as I deemed it to be more time efficient. This dataset contains the teams name, the season, and their defensive rating for that season.

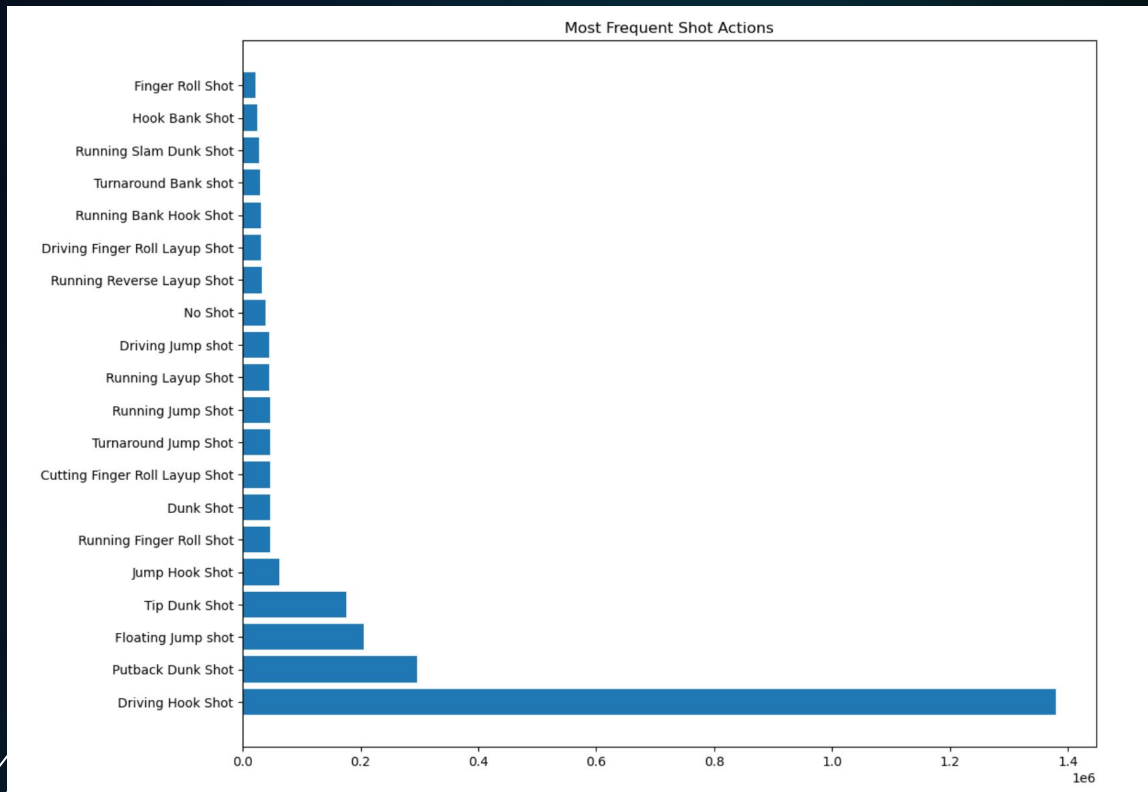


Exploratory Data Analysis

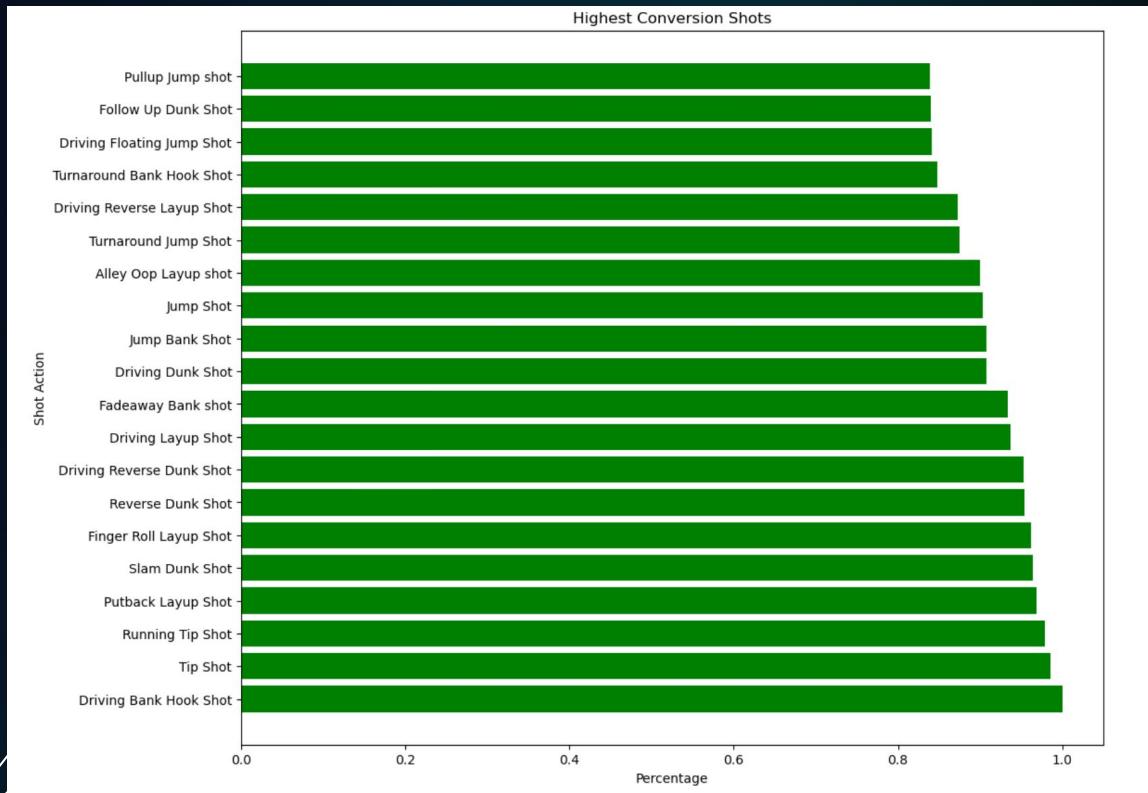
I began the EDA process by looking at the correlation between features and the target variable. After observing the top 20 features with the strongest correlations, I noticed that the implementation of shooting averages seemed to be strongly correlated with the target variable. I was shocked to see that many of the top 20 correlated features were the action type categories turned into features through one-hot encoding.



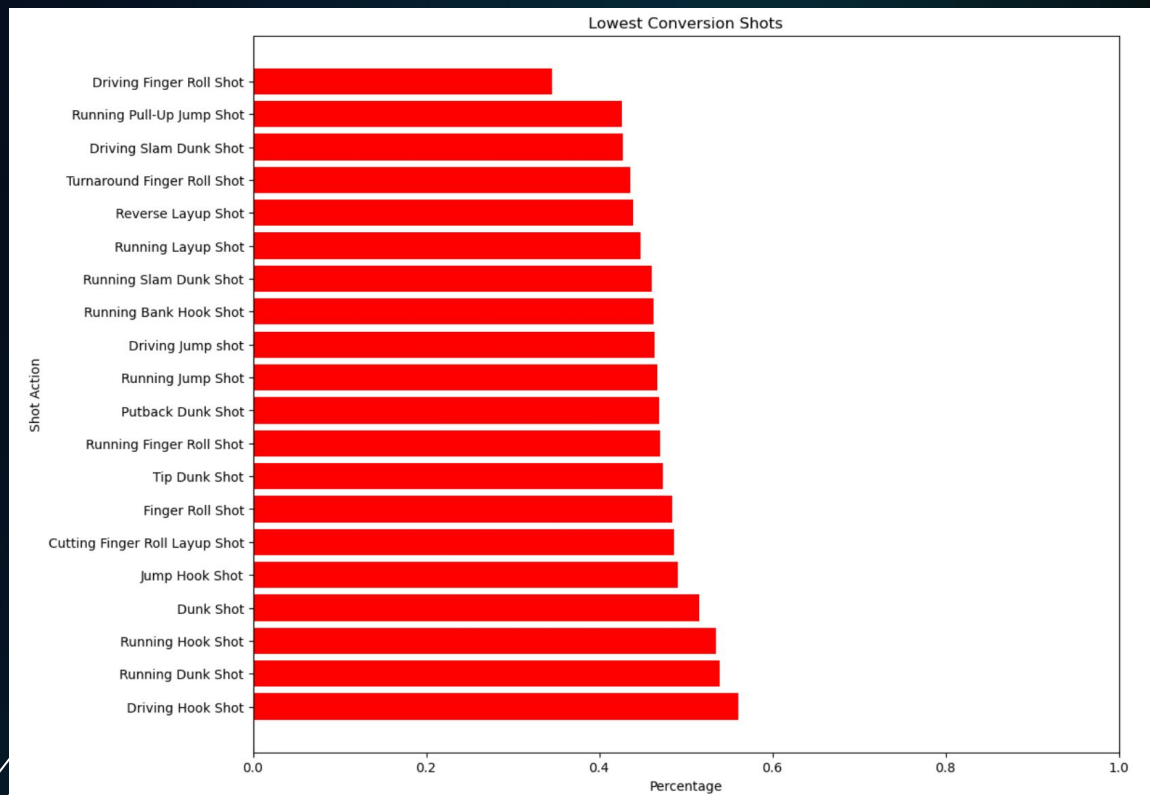
Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis





04



Modeling



Types of Models

Logistic Regression

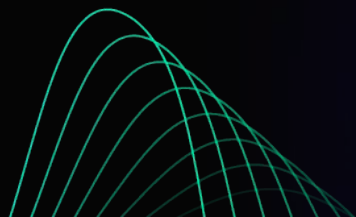
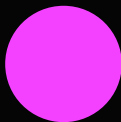
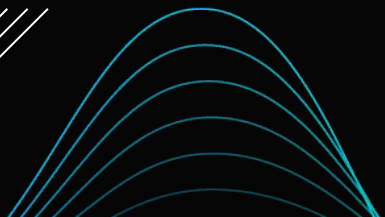
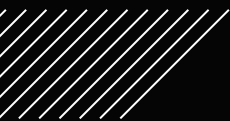
Logistic Regression is a classification model that can be used to make binary predictions. In the case of my project, we can model a shot going in as a 1 and a shot missing as a 0, and structure the problem into a binary setting

Decision Trees

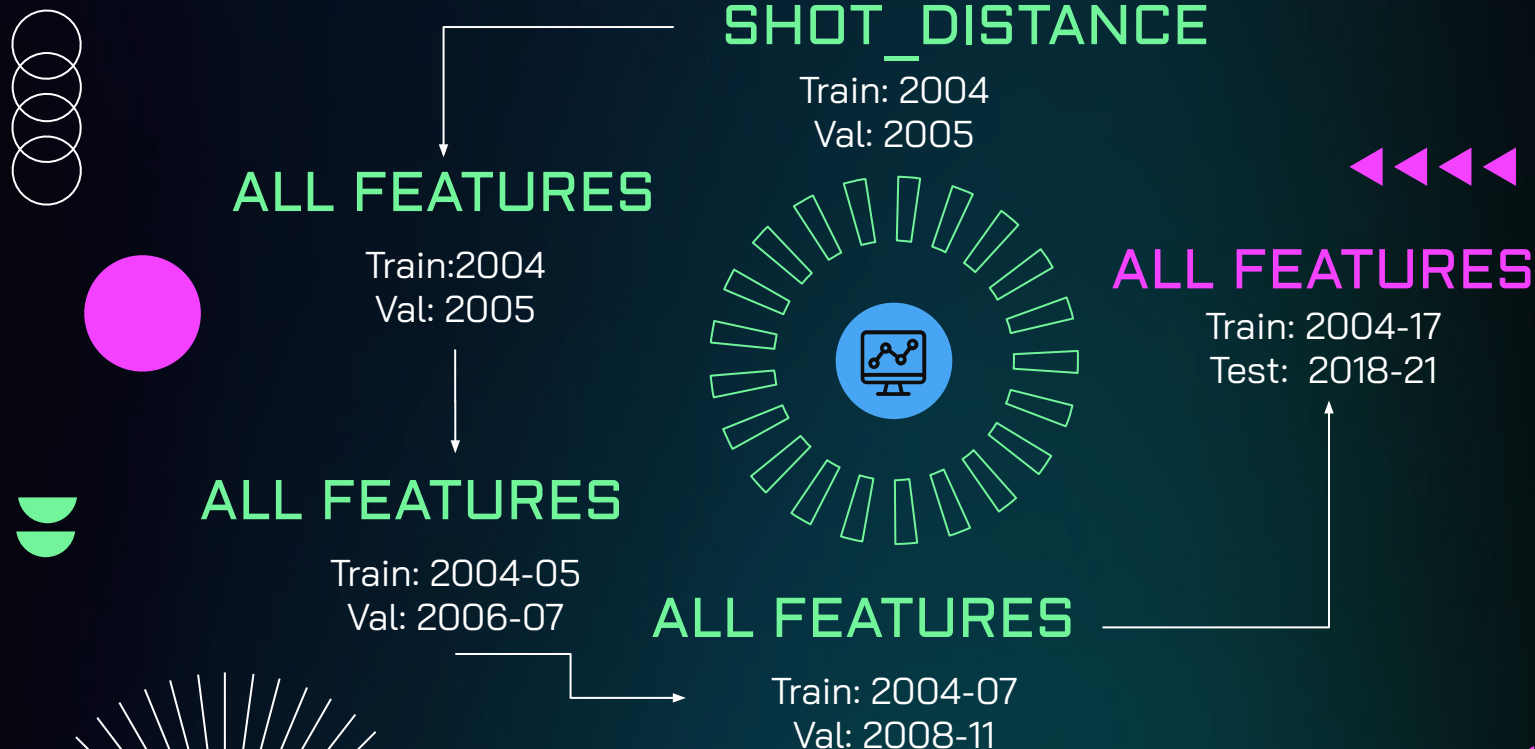
A Decision Tree can be used in a classification setting and splits nodes by a specific decision rule in order to classify a datapoint.

XGBoost

XGBoost is an ensemble method utilizing Decision Trees, where each new Decision Tree is fit on the residual error of the previous tree.



Incremental Year Tuning



Hyperparameter tuning

For the hyperparameter tuning, I explored the below feature spaces using a randomized grid search and utilized the following parameters:

Logistic Regression

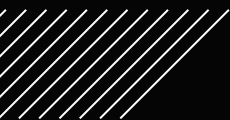
Penalty: L1
Solver = Saga
Max_iter = 10000
C = 0.001

Decision Tree

Max_leaf_nodes: 20
Max_features = log2
Criterion = log_loss
Ccp_alpha = 0.001

XGBoost

N_estimators = 500
Max_leaves = 10
Learning_rate = 0.1





05

Results

Final Testing Accuracies

$$1 - 45.58\% = 54.42\%$$

Naive Benchmark



64.00%

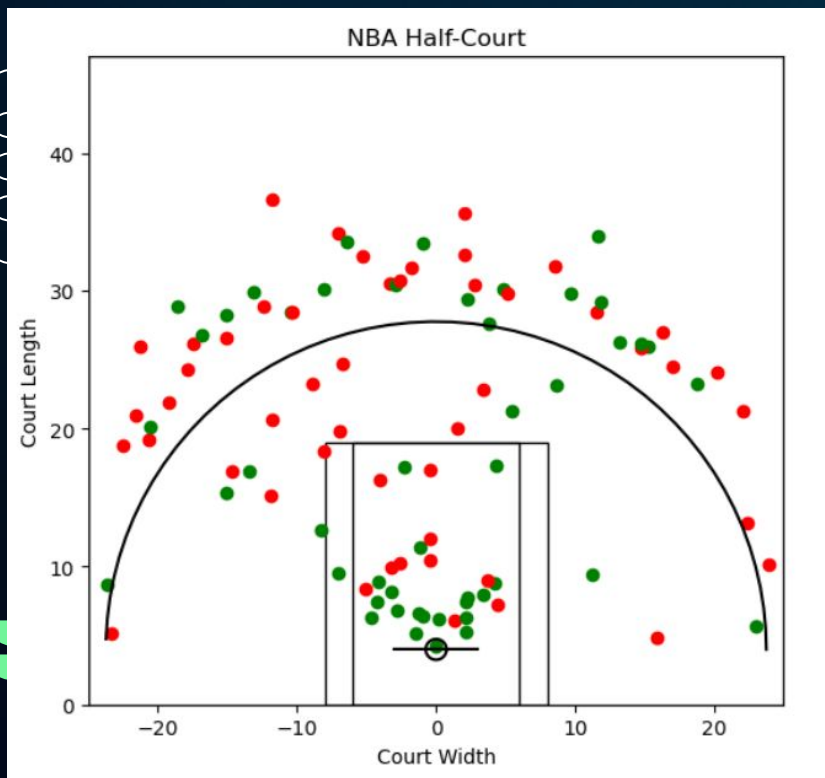
Logistic
Regression

59.46%

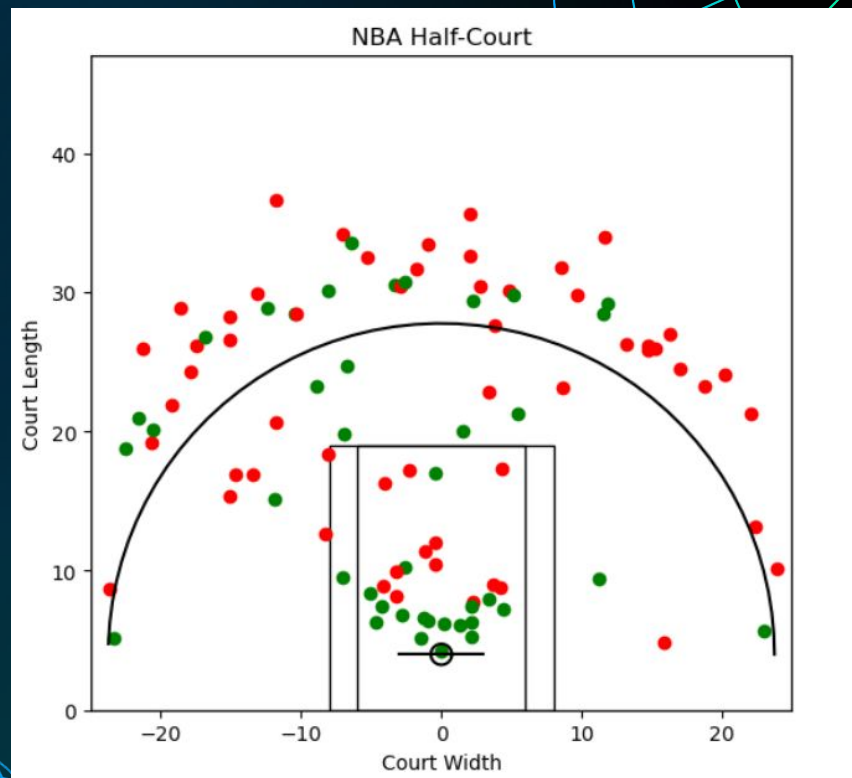
Decision Tree

65.28%

XGBoost

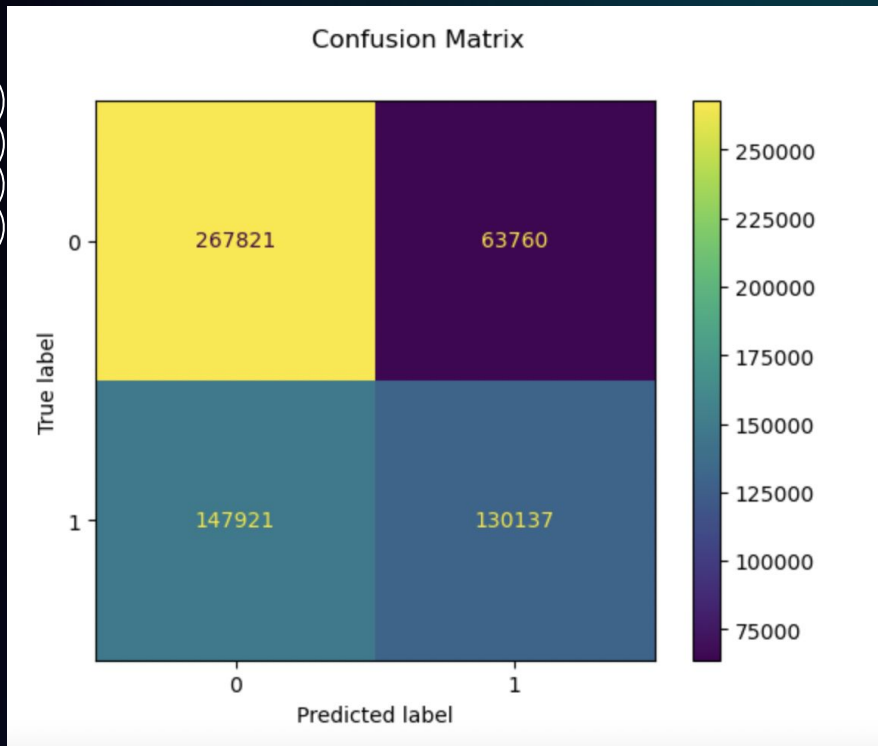


Stephen Curry 2021 Shots



XGBoost Prediction

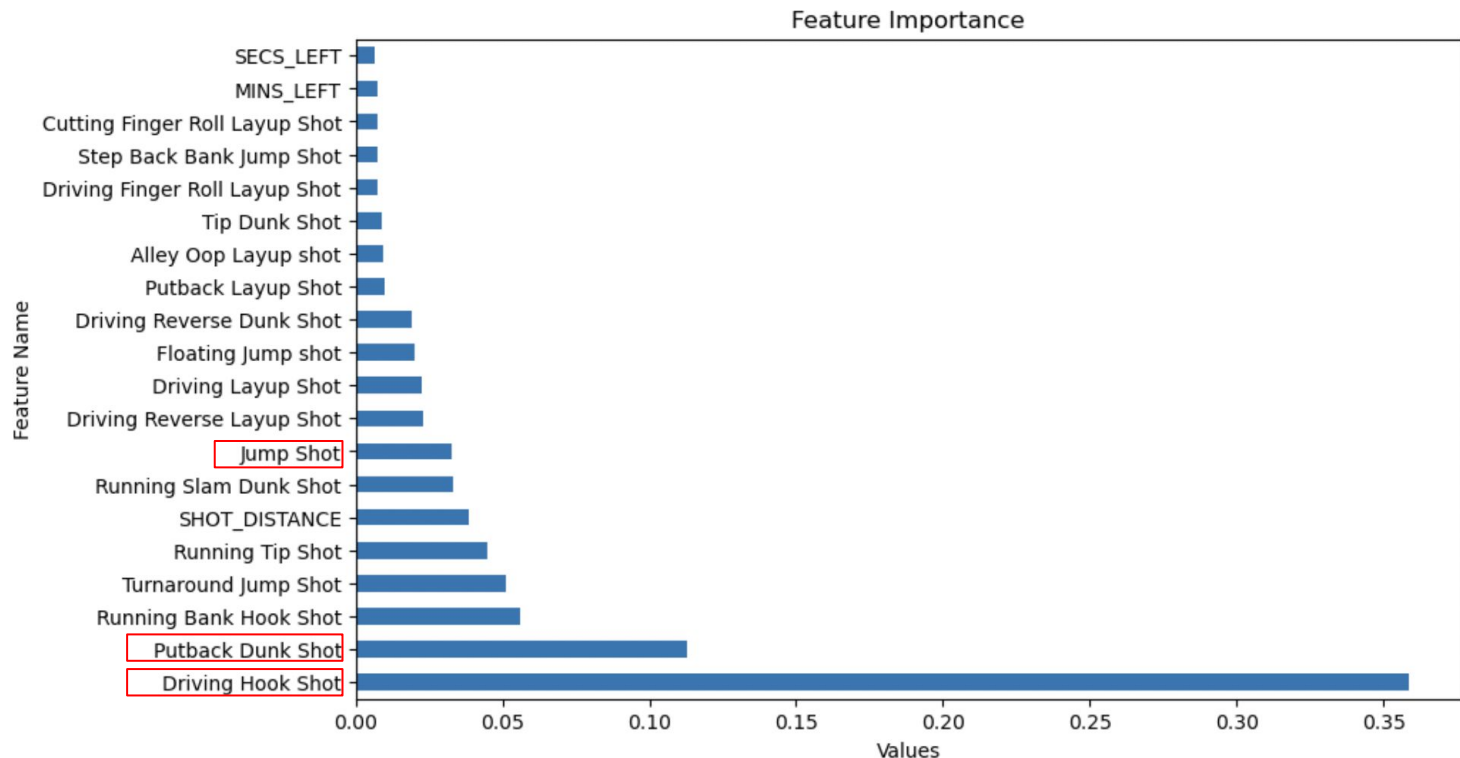
False Negatives



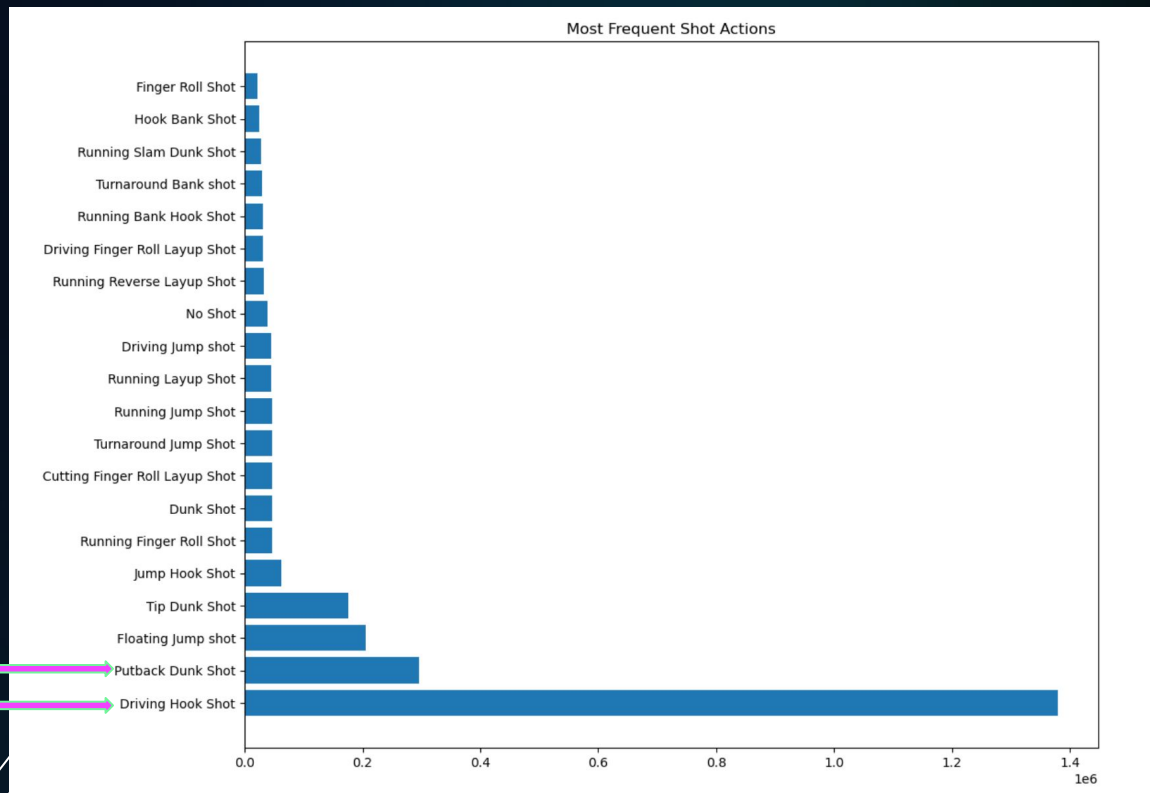
Through visual inspection of the visualization of my XGBoost model, it seems that the model tends miss classify a shot being a miss when in reality it is a make more often than classifying a shot as a make when in reality it is a miss

This aligns with the confusion matrix representation of my testing set as the amount of false negatives is much higher.

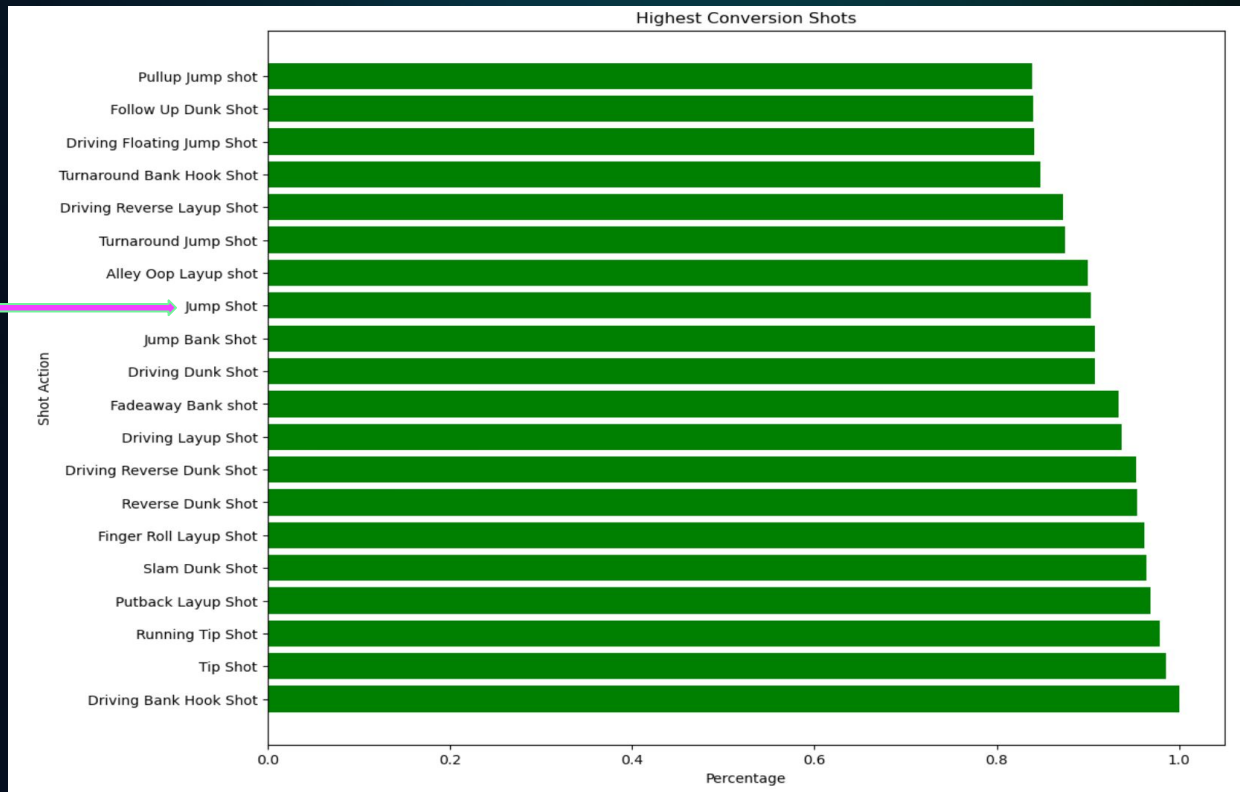
Feature Importance



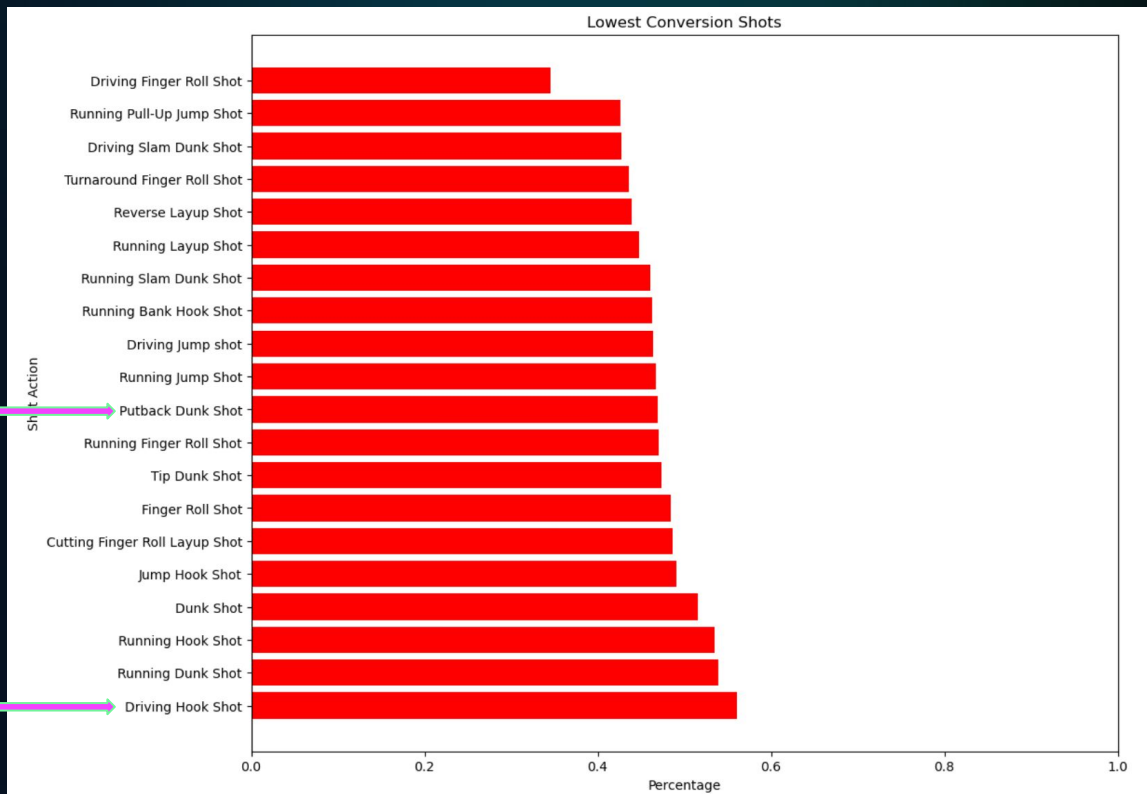
Back to EDA



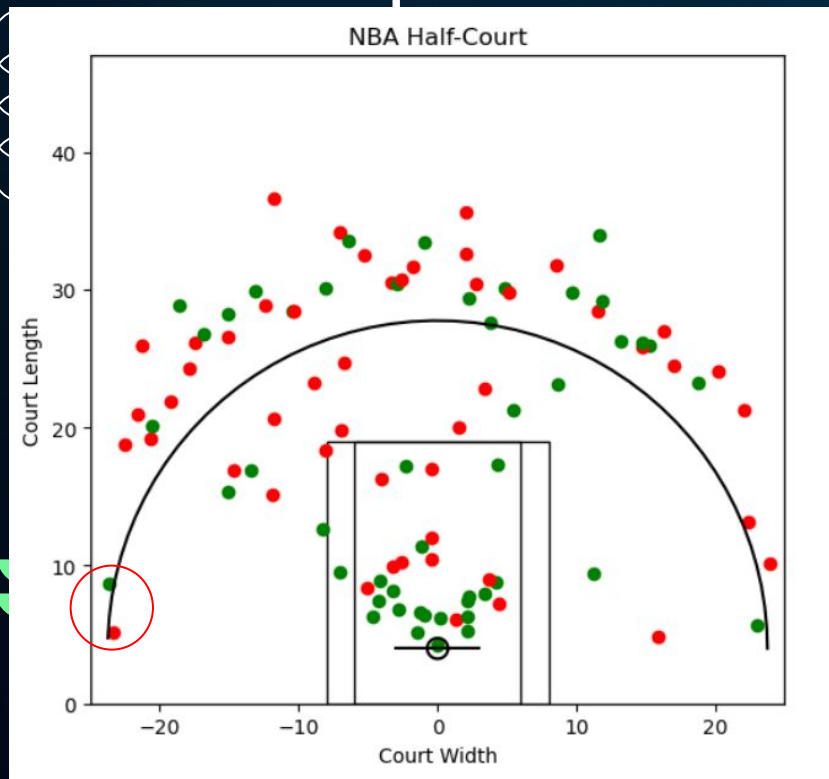
Back to EDA



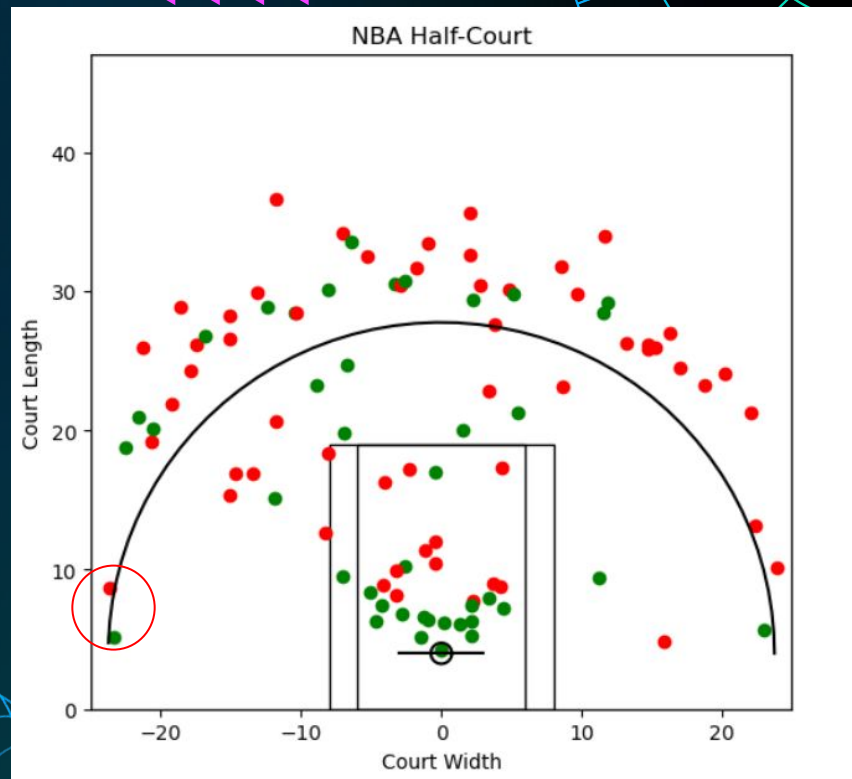
Back to EDA



Next Steps



Stephen Curry 2021 Shots



XGBoost Prediction

Is Basketball Confusing?



After going through this process, it has made me question my basketball domain knowledge. It seems that factors I initially thought would have some importance (defense and a players shooting ability) didn't really play a big role.

If I were to try to improve this project, I would try to be more specific in my feature selection. In my project, I utilized a players season averages, but I now would instead look for data for shooting averages for specific areas on the court (Ex: Corner Three, Top of the key three, Elbow Jump Shot etc...).

Al Horford has a 44.6% Three point percentage on the 2022-23 season. He also has a 55% Corner three point percentage.

Is Basketball Confusing?



For the defensive aspect of basketball, using a teams defensive rating may have been to general of an approach. I would now look for a dataset that included the closest defender along with their defender id, and use their personal defensive rating instead of the teams defensive rating.

Left: Mike Conley

Right: Rudy Gobert (3 time DPOY)

Thanks for Listening!

