# Predicting NBA All-Star Selections

By Scott McCracken

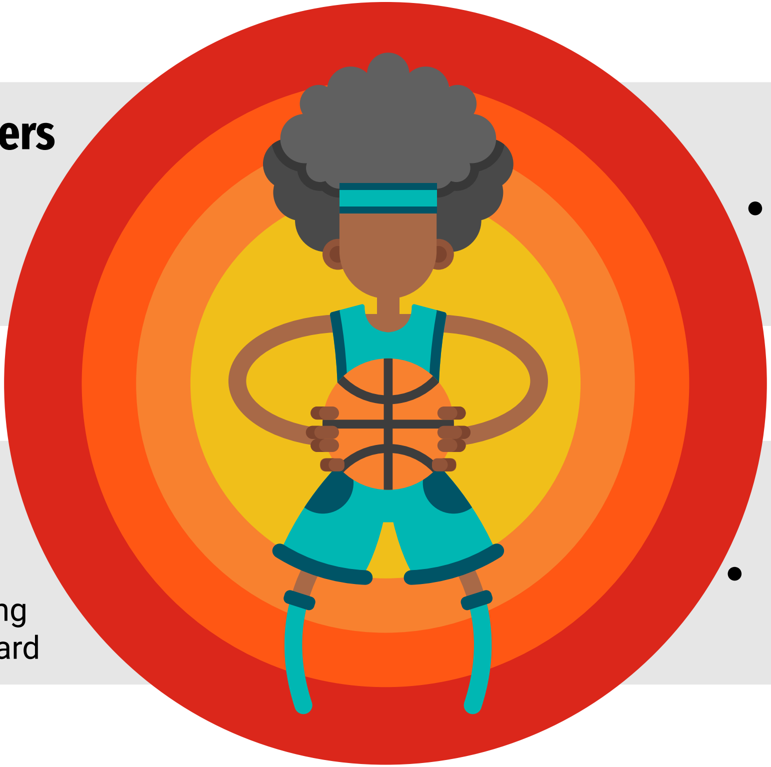
# The Problem:
## All-Star selections have major consequences but can be difficult to anticipate

### Financial Impact on Players

- Reward clauses
- Higher contracts
- Sponsorships

### Historical Legacy

- All-Star teams are a major metric for individual player success in the past



### Online Sports Betting

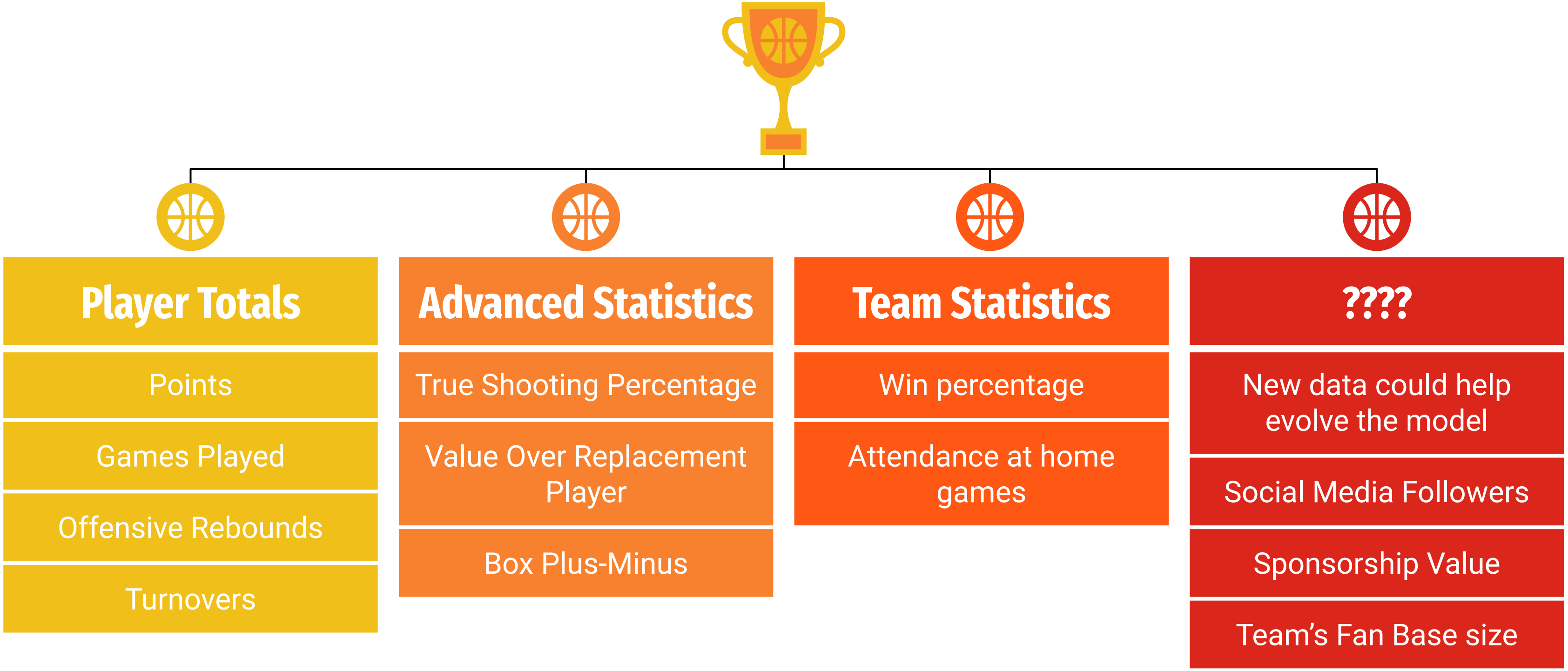
- $100 billion industry
- NBA players will have betting sponsorships moving forward

### Popularity Contest

- Fans make up 50% of the vote

# The Solution

## Train a model to predict All-Star selections based on data



### Player Totals

- Points
- Games Played
- Offensive Rebounds
- Turnovers

### Advanced Statistics

- True Shooting Percentage
- Value Over Replacement Player
- Box Plus-Minus

### Team Statistics

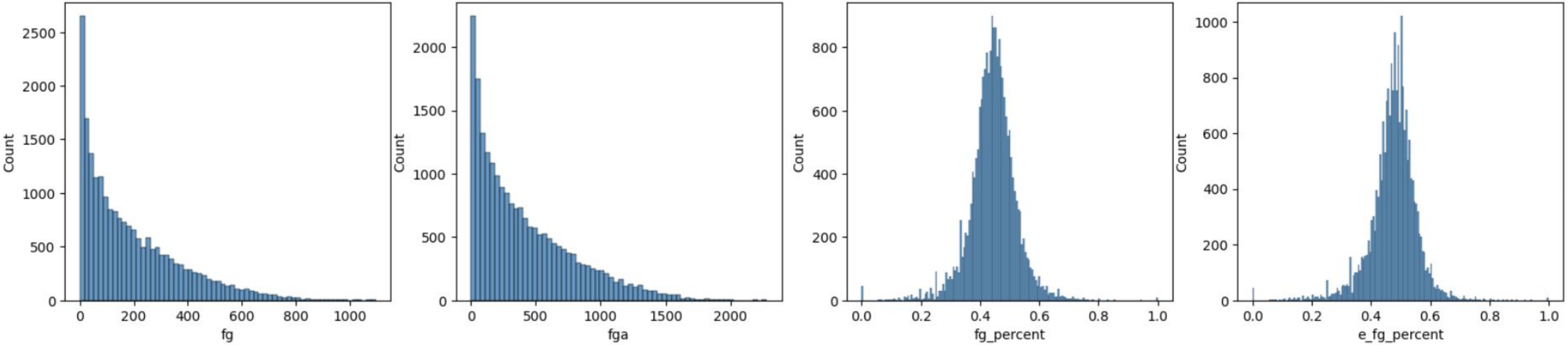
- Win percentage
- Attendance at home games

### ????

- New data could help evolve the model
- Social Media Followers
- Sponsorship Value
- Team’s Fan Base size

# The Data

## Over 50 features and 20,000 records dating back to 1979, when the three-point line was first implemented.



Sumitro Datta formatted this data from Basketball-Reference.com in csv files on Kaggle: https://www.kaggle.com/datasets/sumitrodatta/nba-aba-baa-stats

# Data Wrangling

1. **Similar Leagues and Rules**
   Only use seasons after 1979

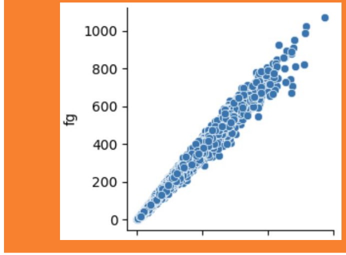
2. **Impute averages when appropriate**
   Attendance at home games, or team win record for players traded mid-season

3. **Drop Irrelevant, Problematic Data**
   Such as players with 0 minutes played

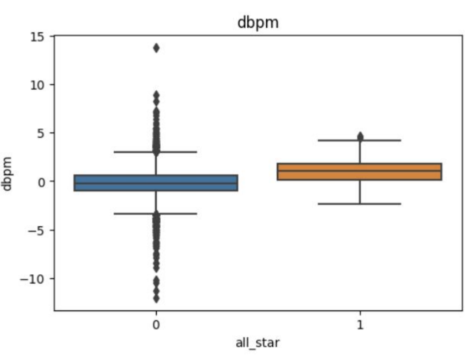
4. **Impute Zero when appropriate**
   Such as a three-point percentage for a center who never shot a three-pointer

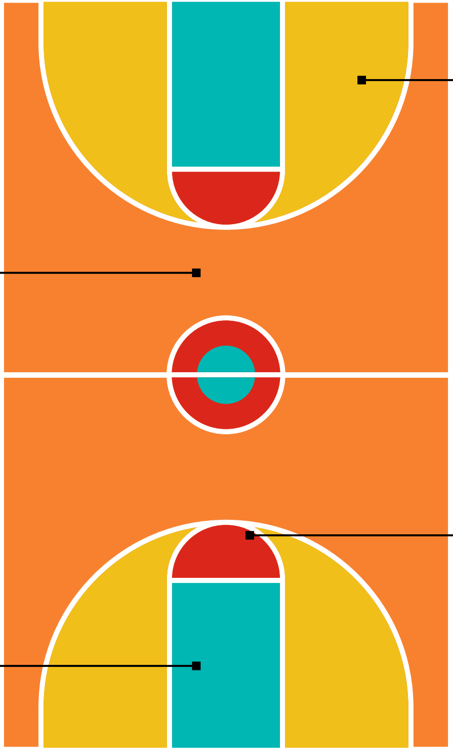
# Exploratory Data Analysis

## Correlations

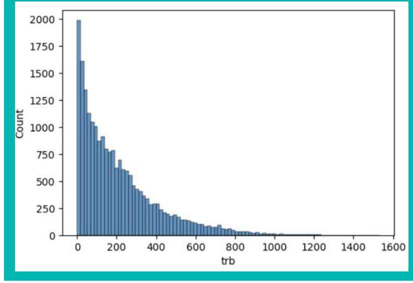


## Outliers





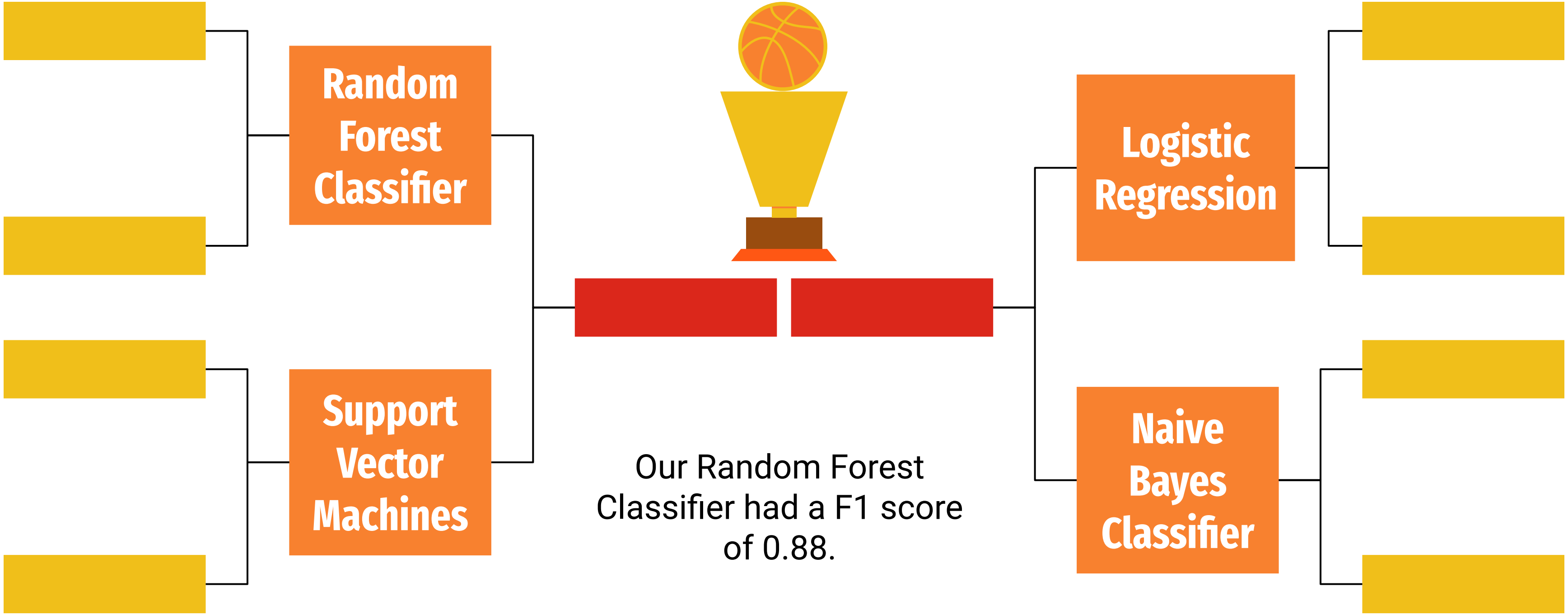
## Skewed Data



## Imbalanced Classes

97% of players are not All-Stars

# Model Selection



Random Forest Classifier

Support Vector Machines

Logistic Regression

Naive Bayes Classifier

Our Random Forest Classifier had a F1 score of 0.88.

# Scaling, Resampling, and Hyperparameter Tuning

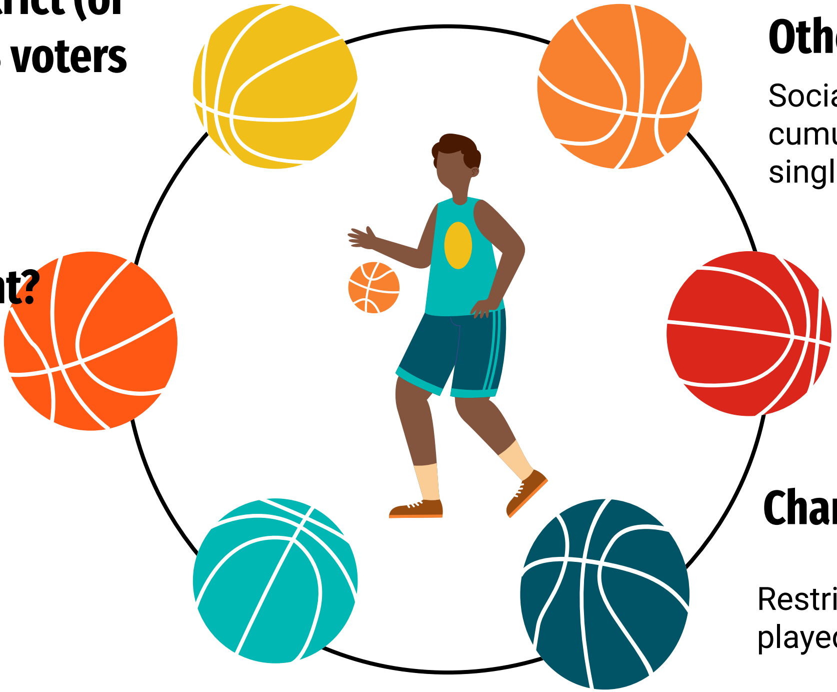
## Standard Scaler

## SMOTE and Tomek Links

## Grid and Random Searches

A Random Forest with default parameters with 300 estimators performed the best.

# Takeaways and Future Research



## The model was more strict (or pessimistic) than 2023 voters

## Other Useful Data?

Social media, sponsorships, cumulative player data vs a single season

## Are older seasons relevant?

As the game evolves, would the model be improved by favoring recent data?

## Changes to selection process

Restrictions around games played, for example.