Men’s Basketball Score Predictions

(using Linear Regression)

Tony Ghabour

January 2020

Motivation

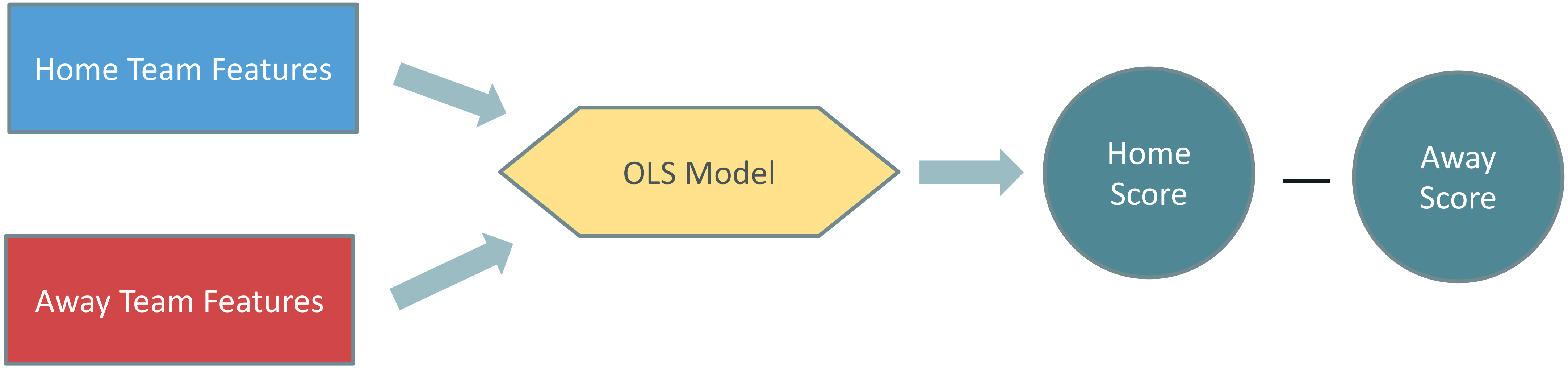
GAMBLING!

Motivation

More specifically…

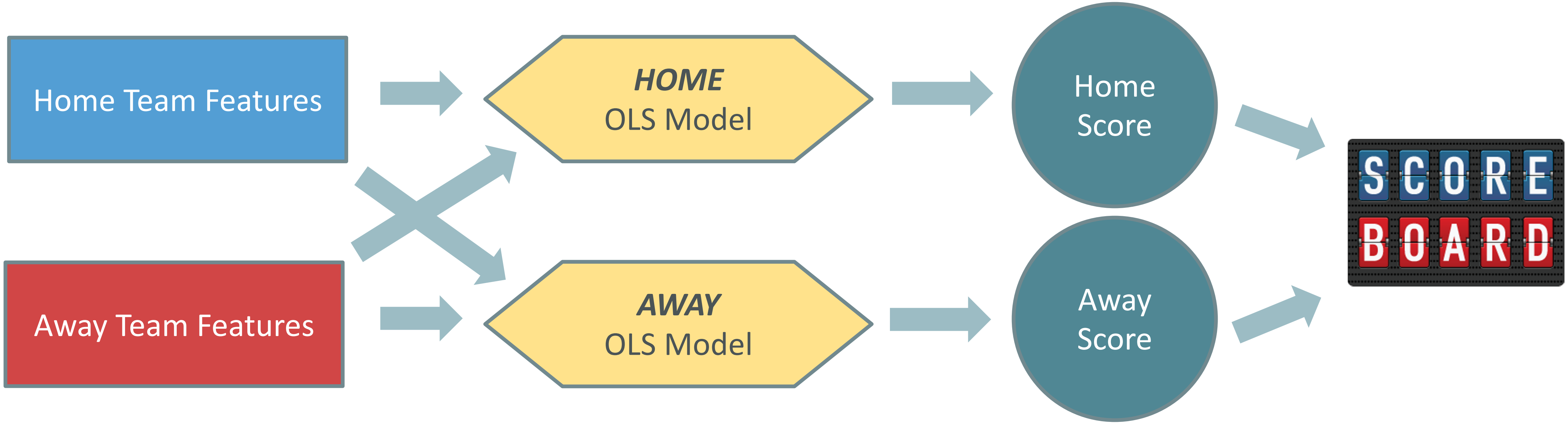
Rely on historical NCAA basketball performance metrics and OLS to "accurately" predict the outcome of a given game.

Original Objective



Score differential would allow user to make predictions regarding the "spread" and "money line."

REVISED Objective



Independent score predictions permit "over/under" picks in addition to spread and money line.

Strategy

Top-down approach:

1. Collect numerous features
2. Use polynomial terms to artificially create more features
3. Start with overfitted baseline (intentional)
4. Rely on regularization to identify/emphasize important features

>

Raw Data

sportradar

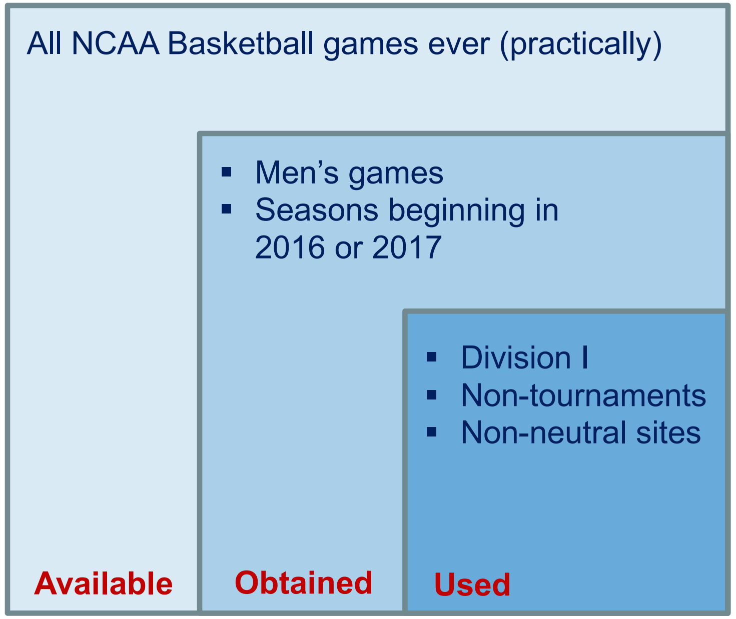
(hosted by)

- Huge data dump of historical statistics
- Offense vs. defense
- Home vs. Away
- Additional team & game-specific info (e.g. conference, venue data)

- Scraped roster-specific data
- Player height and weight

Pre-processing

Narrow domain
(types of games)



Not to scale

High-level Feature Selection

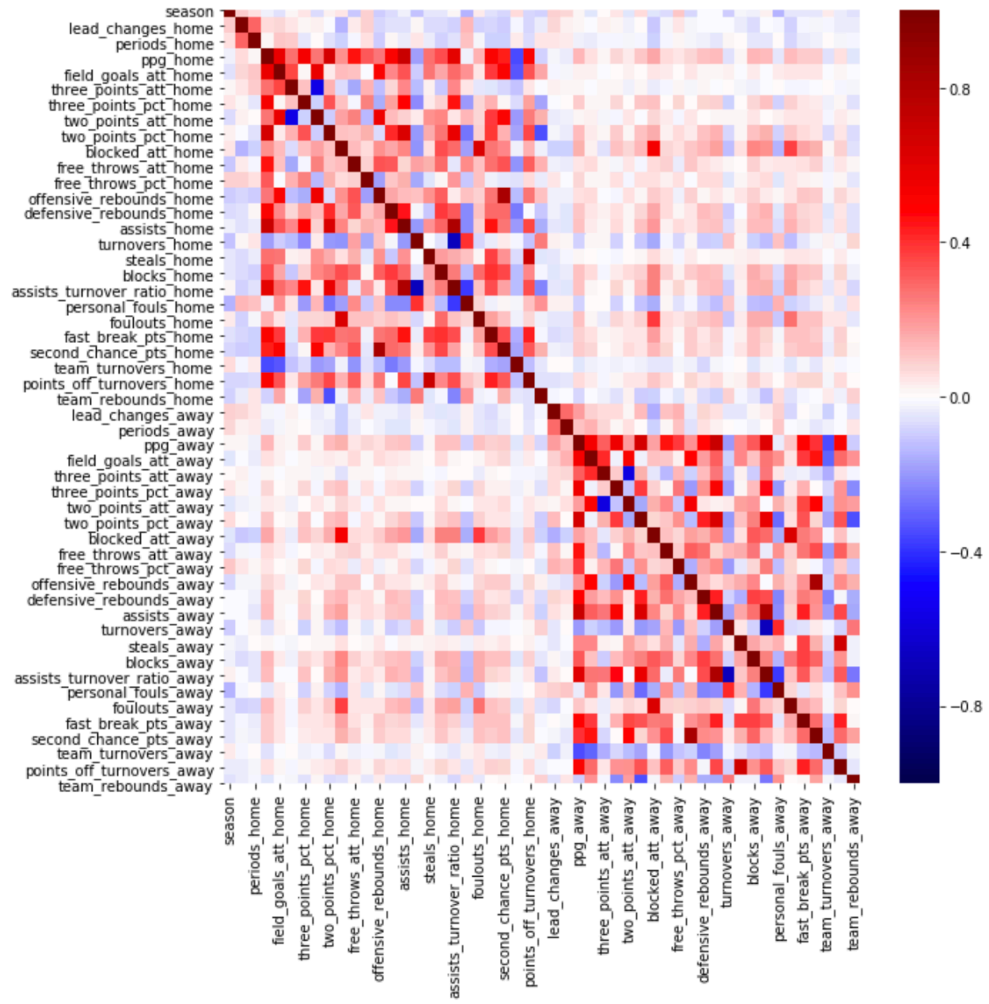
- Venue details
- Team colors
- Mascots
- Logos
- etc…

Transform for Modeling

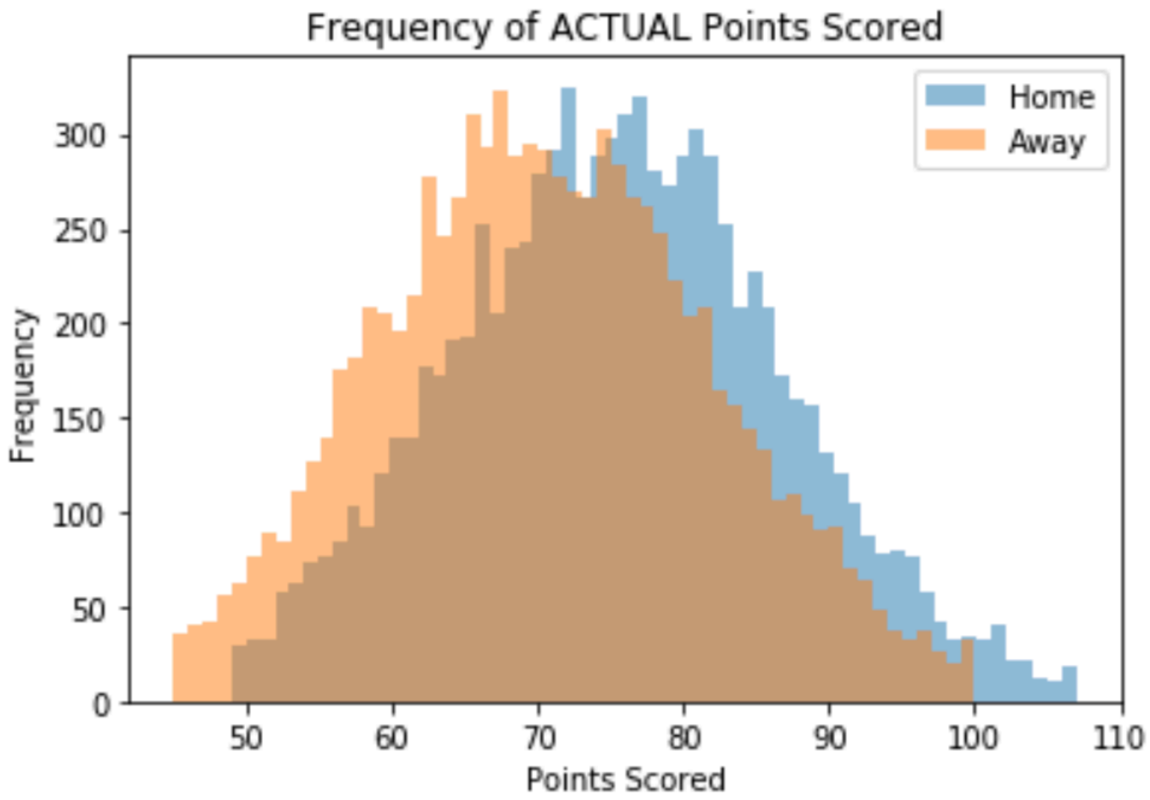
- Raw data = game-level
- Model(s) require historical team averages preceding prediction

The “System”

Multicollinearity and endogeny



Similarly Distributed Targets



Analysis & Model Evaluation

Model	Regularization*	R^2	RSME
HOME MODELS			
“Vanilla OLS”	None	0.237	9.928
Linear Regression	Lasso	0.236	9.937
Linear Regression	**Ridge**	**0.238**	**9.923**
Polynomial (deg 2)	Lasso	0.162	10.407
Polynomial (deg 3)	Lasso	0.168	10.370
AWAY MODELS			
“Vanilla OLS”	**None**	**0.236**	**9.778**
Linear Regression	Lasso	0.233	9.800
Linear Regression	Ridge	0.236	9.779
Polynomial (deg 2)	Lasso	0.155	10.284
Polynomial (deg 3)	Lasso	0.166	10.216

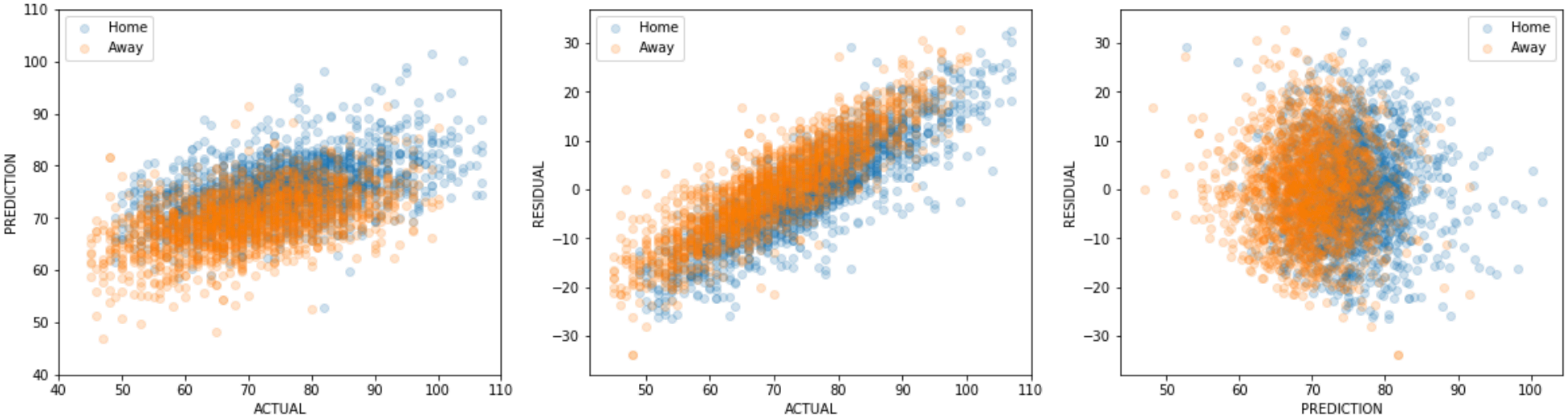
* Regularization hyperparameter λ = 1 in all cases, as applicable.

Results

	actual_home	actual_away	predictions_home	predictions_away	actual_margin	predicted_margin
count	1783.000000	1783.000000	1783.000000	1783.000000	1783.00000	1783.000000
mean	75.418957	70.096467	75.469455	69.771305	5.32249	5.698150
std	11.372038	11.191562	5.582934	5.432878	13.53649	7.428717
min	49.000000	45.000000	52.751343	46.948691	-40.00000	-17.872835
25%	67.000000	62.000000	71.958748	66.484072	-4.00000	0.890124
50%	75.000000	70.000000	75.259291	69.816039	5.00000	5.128873
75%	83.000000	78.000000	78.845928	73.331635	14.00000	9.749419
max	107.000000	100.000000	101.567423	91.541813	54.00000	39.654323

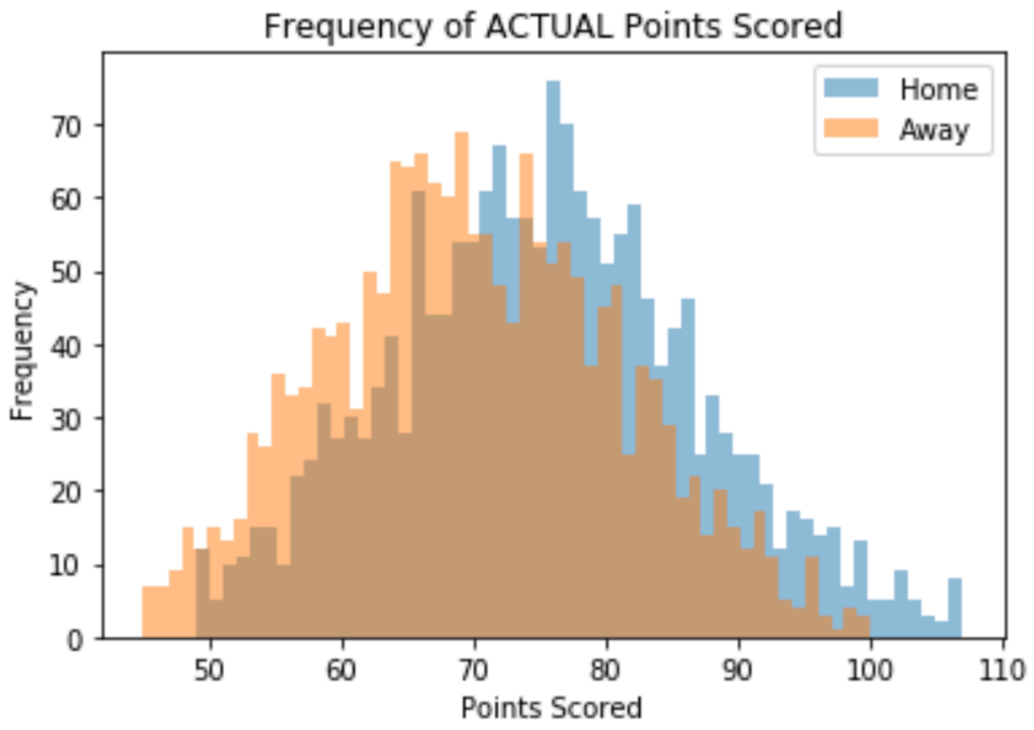
Based on 20% “holdout” test set

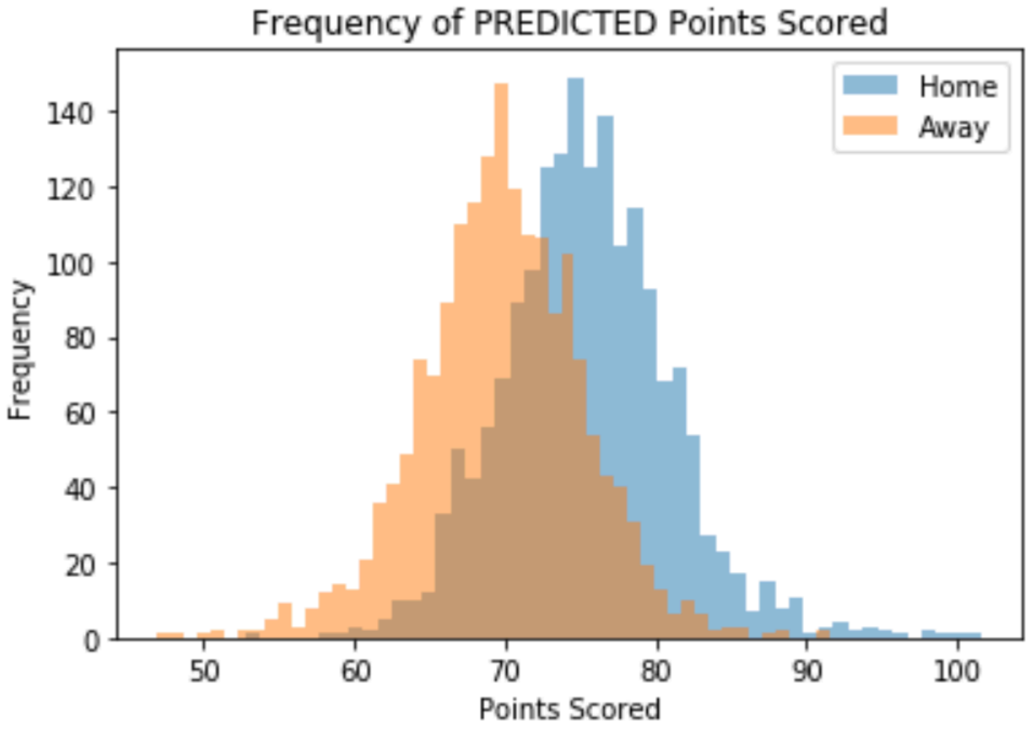
Results



Based on 20% “holdout” test set

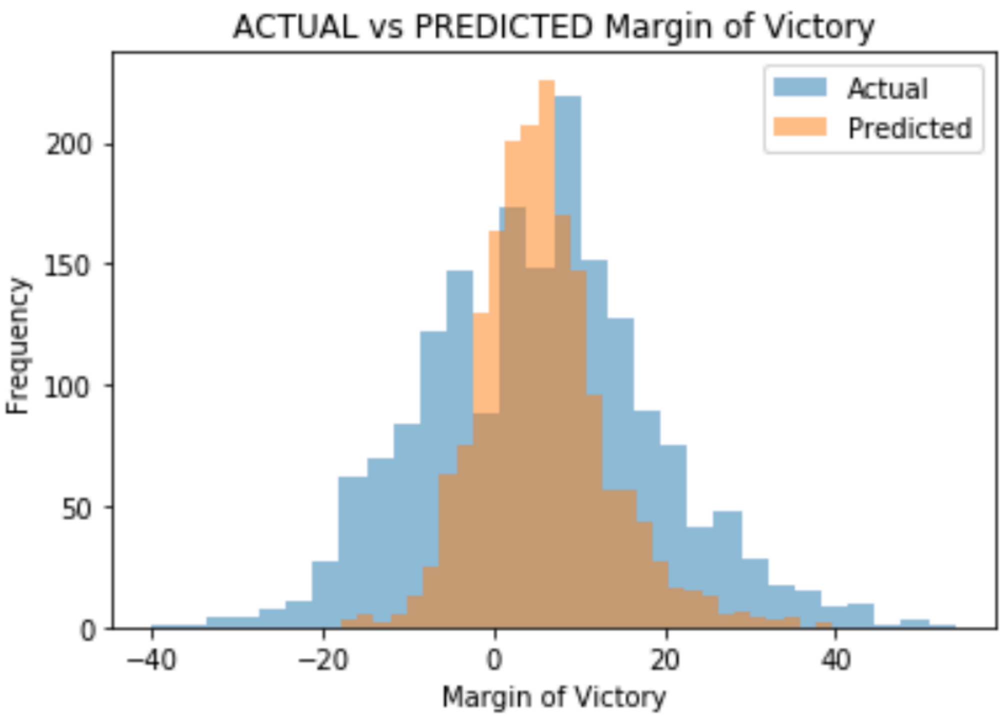
Results

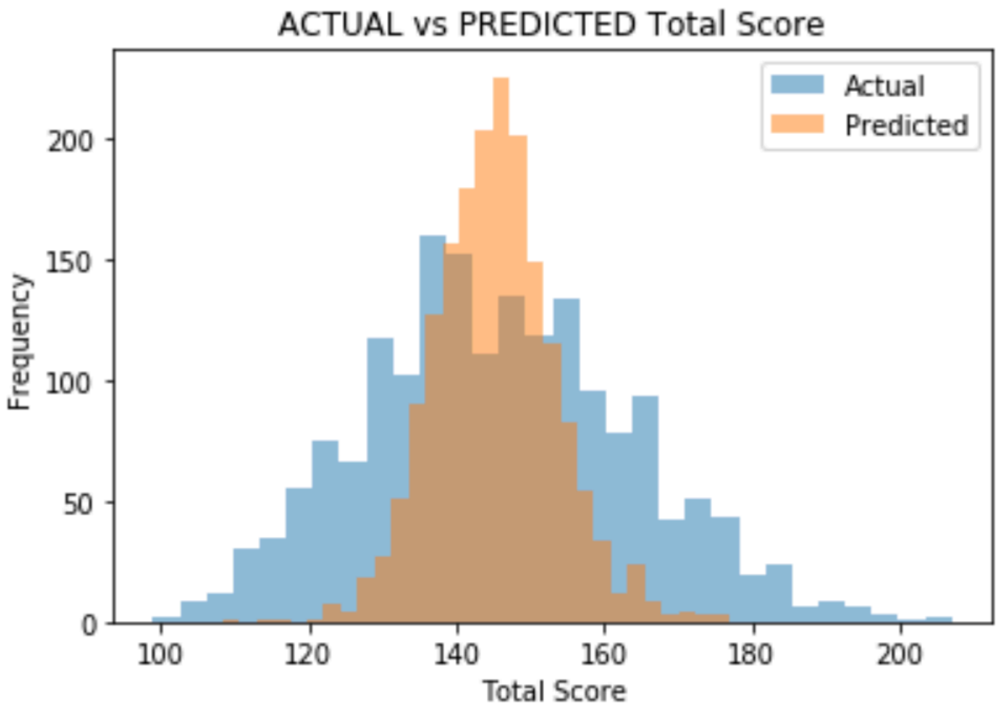




Based on 20% “holdout” test set

Results





Based on 20% “holdout” test set

Conclusions

- **DO NOT USE FOR GAMBLING**
- Ambitious objective
- Great deal of inherent variability in underlying system
  - Example: 2018 Tournament > #1 Overall seed lost in 1st round
- Not enough to match aggregate distributions, <u>matchups matter</u>
- Linear models *may* be inadequate for stated objective

Future Work

- Start with fewer, more targeted features
  - Offensive Rebound % = OR / (OR + Opp. DR)
  - Free Throw Rate = FT attempts / FG attempts
- Incorporate K-fold cross-validation
- Consider categorical variables
  - Conference, in-conference matchup
- Try more robust/appropriate models:
  - Weighted Least Squares (WLS)
  - General Linear Models (GLM)
  - Ensemble Models (combinations)