Discovering NBA Strategies That Drive Success: An Elastic Net Regression Approach

Aaron Sanders, Drew Kearny, Matt Anderson
Group 10 STAT 654
Texas A&M University, Statistics Department

Thank you Dr. Sharmistha Guha and Mr. Sungee Hong for the semester which has guided this presentation

Outline

1. **Introduction**
- Motivation & Objective
- Challenges
1. **Preprocessing**
- Data Description
- Exploratory Discoveries
1. **Our Model**
- Methodology
- Performance
1. **Discussion**

Introduction

Motivation:

- Identifying factors which lead to team success can help NBA front offices
  - Coaching Strategies
  - Roster Construction/Player Development
- Can allow fans with more predictive power of team success - potentially lucrative...

Project Objectives:

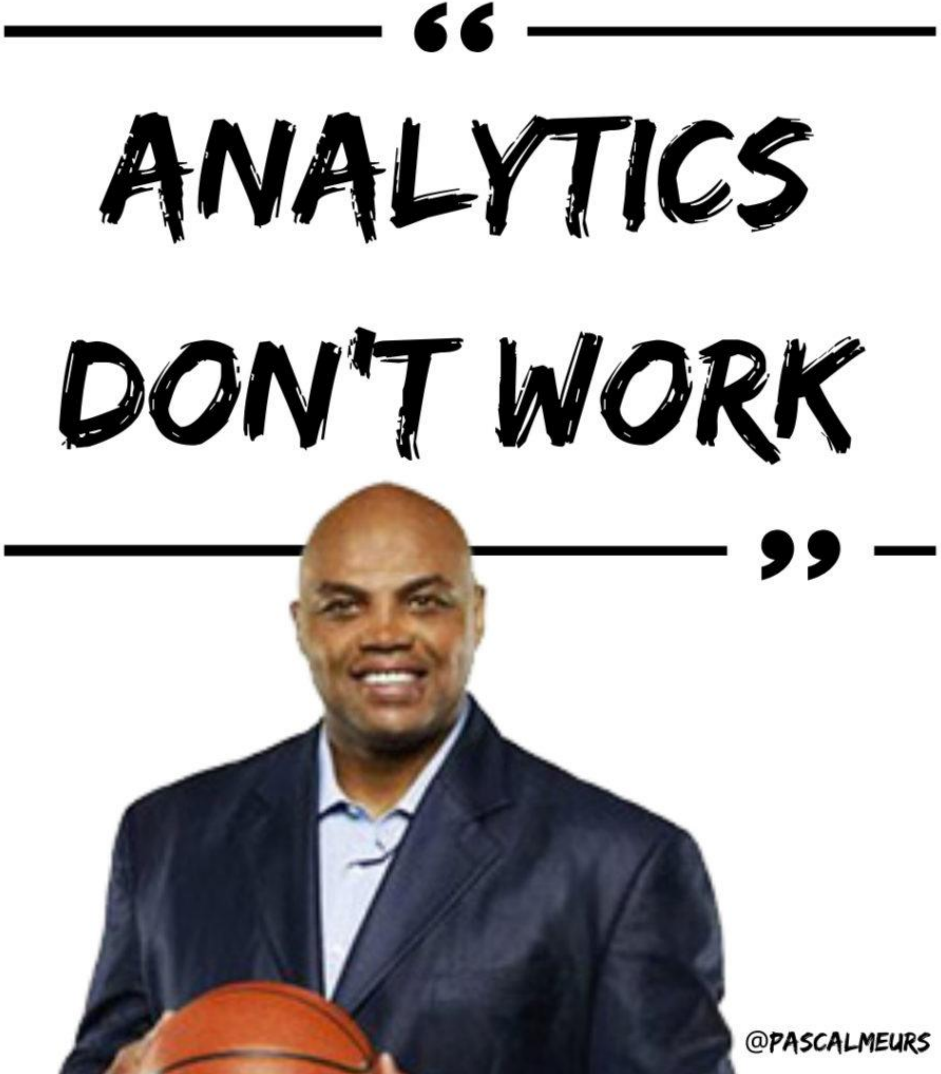
- Identify significant drivers of NBA team success
- Development of an interpretable predictive model
- Validate the model with recent season data

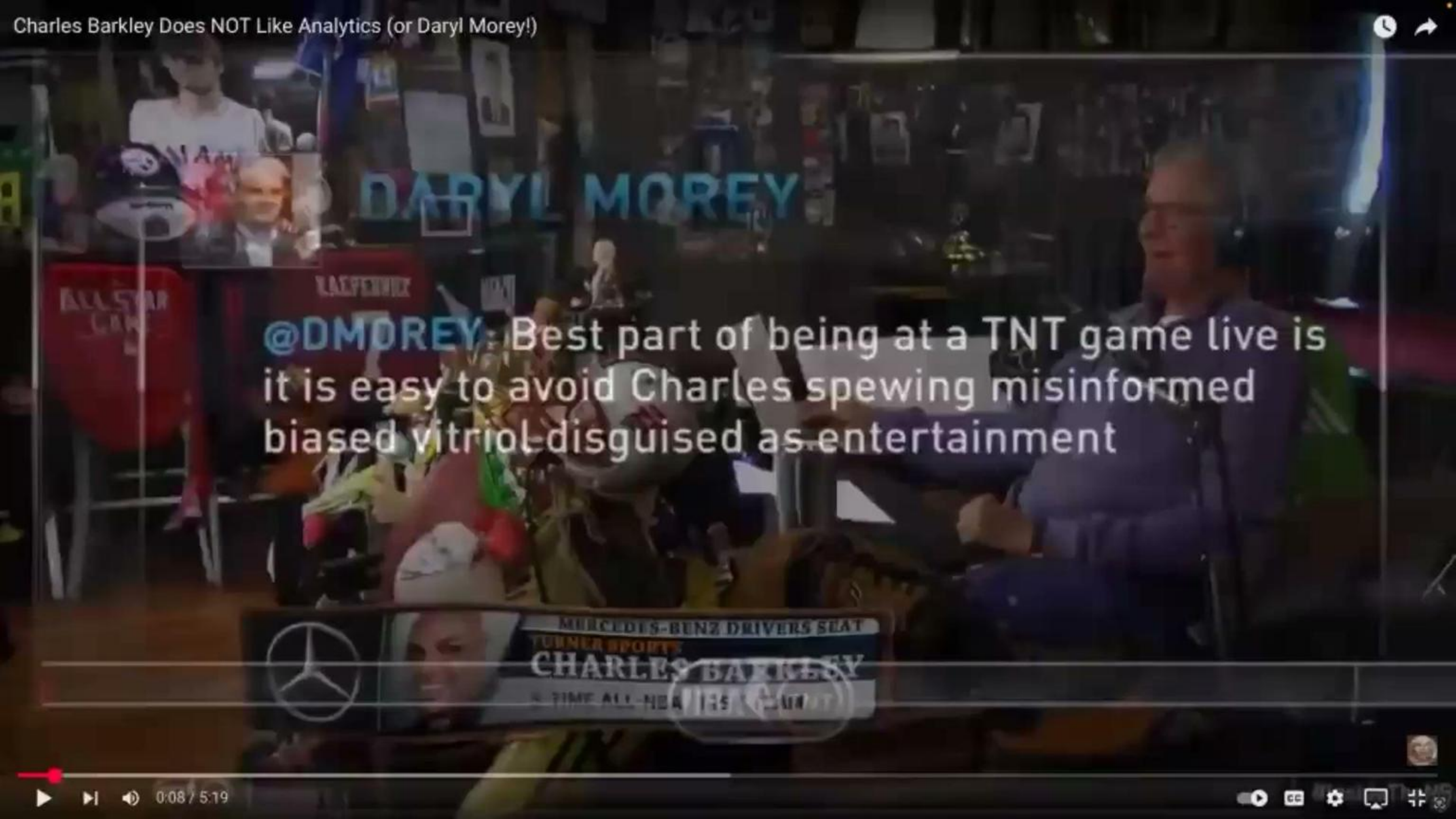
Challenges:

1. Difficult for analytics to measure “intangibles” that can contribute to success

1. Multicollinearity between predictors

1. Translation of season data to playoff success




Charles Barkley Does NOT Like Analytics (or Daryl Morey!)
DARYL MOREY
@DMOREY: Best part of being at a TNT game live is it is easy to avoid Charles spewing misinformed biased vitriol disguised as entertainment
MERCEDES-BENZ DRIVERS SEAT
CHARLES BARKLEY
0:08 / 5:19

Preprocessing

Preprocessing - Data Description

Category	Variables
Team Characteristics	Age
Pace & Style	Pace (possessions per 48 minutes)
	3-point attempt rate (3PAr)
	Free throw rate (FTr)
Shooting Efficiency	True Shooting % (TS%) Effective FG% (eFG_off)
Ball Possession	Turnover % (Offensive) Offensive Rebound % (ORB_off)
	Free Throws per FG Attempt (FT_FGA_off)
Defensive Metrics	Opponent eFG% (eFG_def) Turnovers Forced (TOV_def)
	Defensive Rebound % (DRB_def)
	Opponent FT/FGA (FT_FGA_def) Defensive Rating (DRtg)

Data Source:

- The dataset is sourced from Basketball-Reference.com
- Owned and operated by Sports Reference, LLC, a reputable American sports statistics company

Seasons & Observations:

- 5 NBA seasons: 2019-2020 to 2023-2024 (2024-2025 data added recently and used as a test set)
- 30 teams × 5 seasons = 150 team-season observations

Target Variable:

- Win Percentage (win_pct) = Wins / Games Played

Other:

- Added a variable to track team's leading ppg individual
- Serves as a proxy to measure "star power"

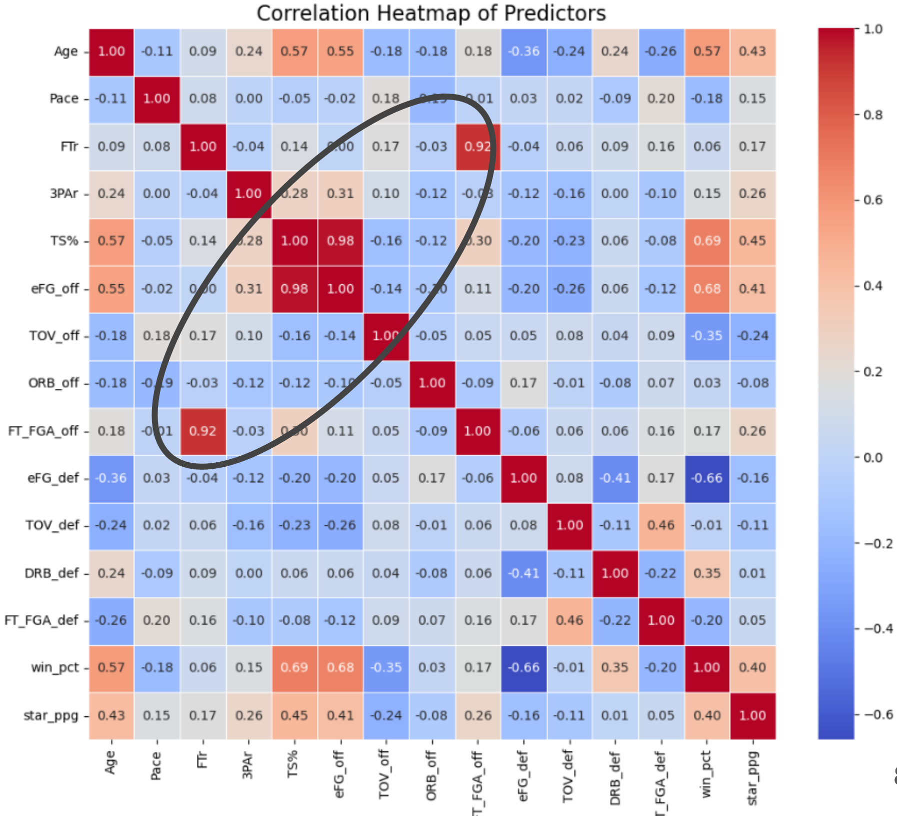
Preprocessing - Exploratory Discoveries

Issue with Multicollinearity:

- Difficulty to determine individual effects impact on response
- Leads to an inflated standard error
- Calls for rectification in predictor choice

Promising Potential with Response:

- Some of the predictors being fairly correlated with the response can reflect potential predictive strength of our model



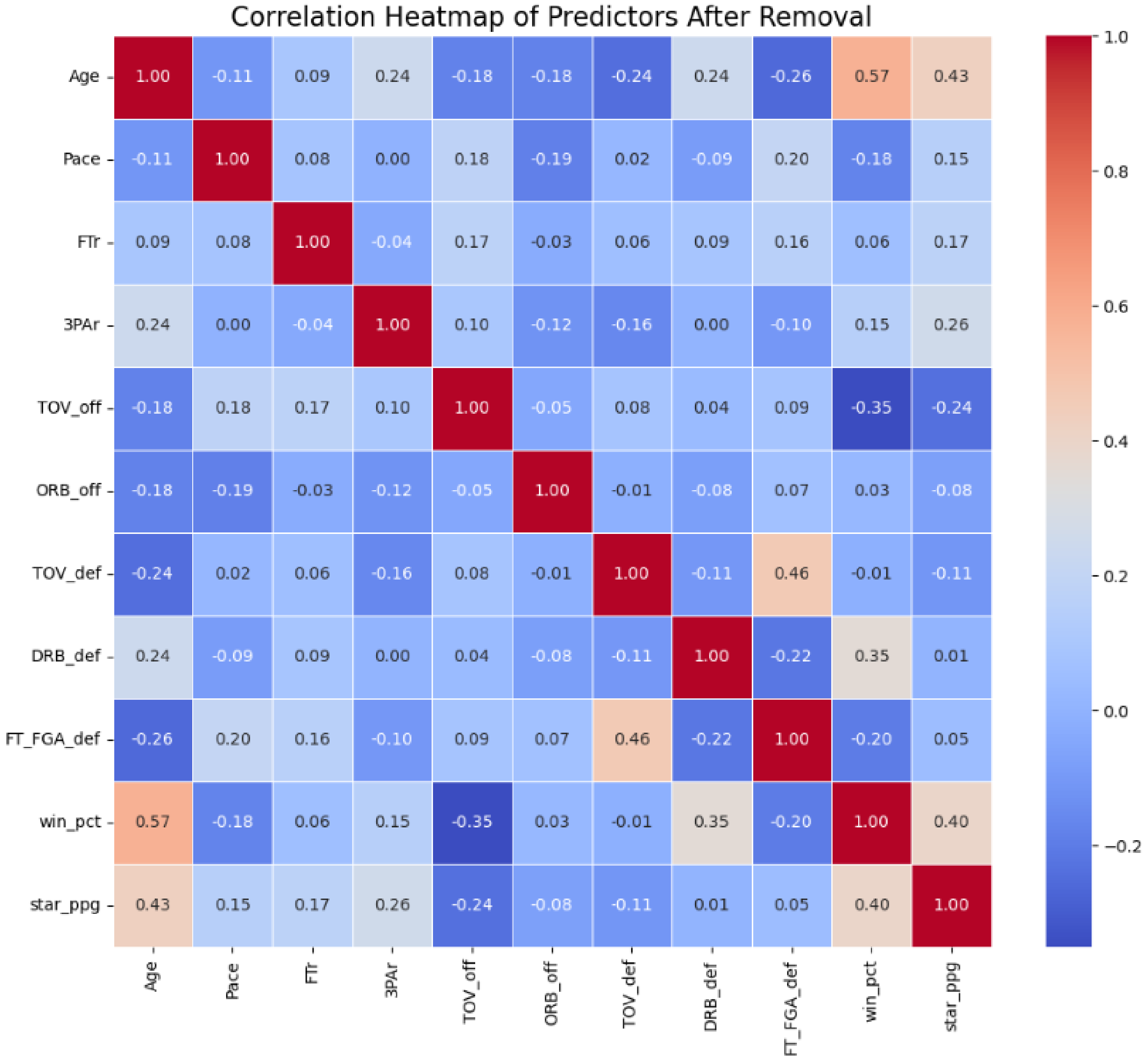
Preprocessing - Exploratory Discoveries

Feature Engineering Decisions:

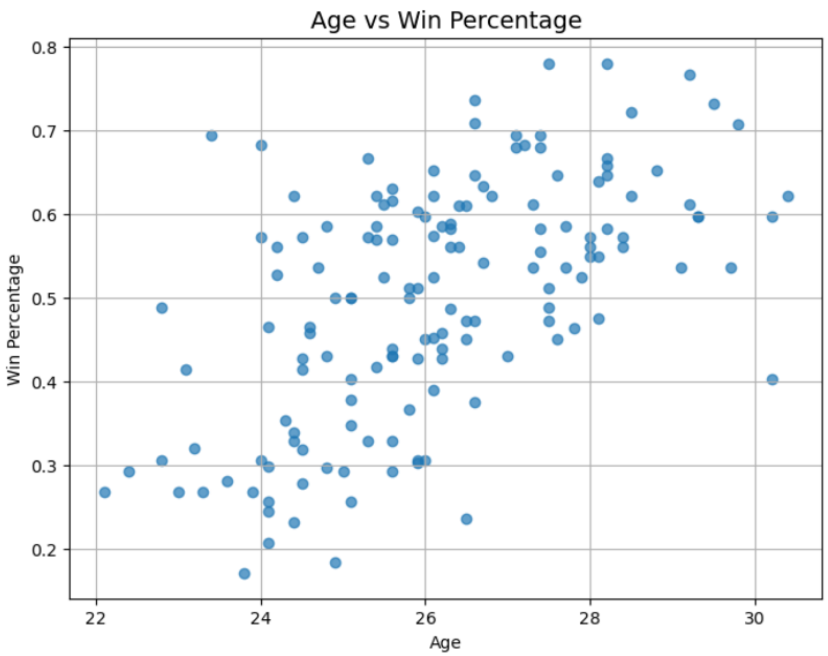
- **Removed Correlated metrics:**
  - Kept FTr, removed FT_FGA_off (high correlation)
  - Removed TS%
- **Excluded composite metrics:**
  - Removed overall efficiency metrics (ORtg, DRtg, eFG)
  - Focused on specific, actionable factors coaches can influence
- **Kept Metrics:**
  - Age to test experience vs. youth tradeoff
  - Kept Pace despite weak correlation with winning

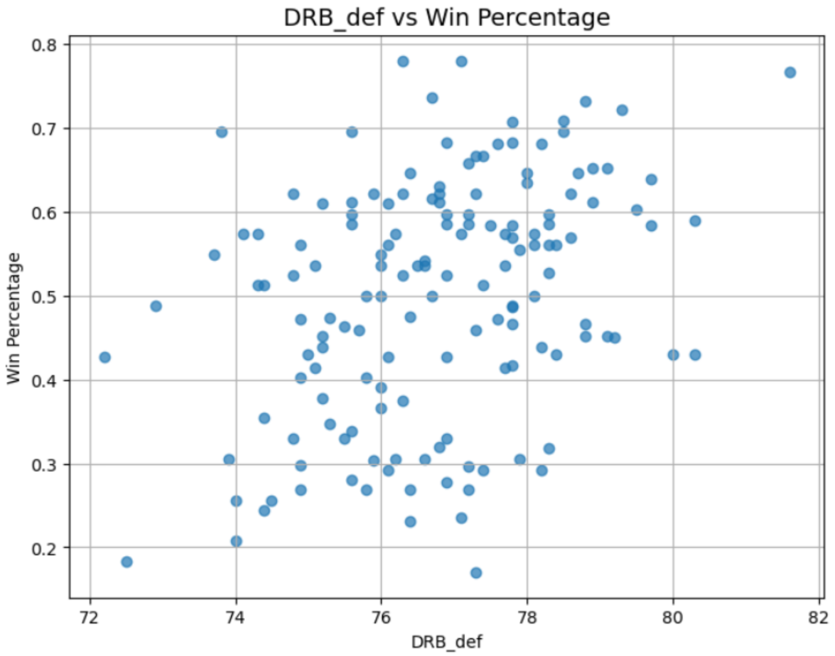
Final Modeling Predictor Set:

- **Balance of offensive and defensive factors**
- **Focus on actionable elements for coaching staff**
- **10 strategic factors teams can potentially influence:**
  - Age: Strategy choices between youth and experience
  - Pace: Offensive tempo
  - FTr: How often teams draw fouls per FG attempt
  - 3PAr: Proportion of shots from 3-point range
  - TOV_off: Ball security
  - ORB_off: Second-chance opportunities
  - TOV_def: Defensive disruption
  - DRB_def: Ability to end opponent possessions
  - FT_FGA_def: Defensive discipline
  - star_ppg : Measures star player power


Correlation Heatmap of Predictors After Removal
Age 1.00 -0.11 0.09 0.24 -0.18 -0.18 -0.24 0.24 -0.26 0.57 0.43
Pace -0.11 1.00 0.08 0.00 0.18 -0.19 0.02 -0.09 0.20 -0.18 0.15
FTr 0.09 0.08 1.00 -0.04 0.17 -0.03 0.06 0.09 0.16 0.06 0.17
3PAr 0.24 0.00 -0.04 1.00 0.10 -0.12 -0.16 0.00 -0.10 0.15 0.26
TOV_off -0.18 0.18 0.17 0.10 1.00 -0.05 0.08 0.04 0.09 -0.35 -0.24
ORB_off -0.18 -0.19 -0.03 -0.12 -0.05 1.00 -0.01 -0.08 0.07 0.03 -0.08
TOV_def -0.24 0.02 0.06 -0.16 0.08 -0.01 1.00 -0.11 0.46 -0.01 -0.11
DRB_def 0.24 -0.09 0.09 0.00 0.04 -0.08 -0.11 1.00 -0.22 0.35 0.01
FT_FGA_def -0.26 0.20 0.16 -0.10 0.09 0.07 0.46 -0.22 1.00 -0.20 0.05
win_pct 0.57 -0.18 0.06 0.15 -0.35 0.03 -0.01 0.35 -0.20 1.00 0.40
star_ppg 0.43 0.15 0.17 0.26 -0.24 -0.08 -0.11 0.01 0.05 0.40 1.00
Age Pace FTr 3PAr TOV_off ORB_off TOV_def DRB_def FT_FGA_def win_pct star_ppg
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Preprocessing - Exploratory Discoveries





Our Model

Our Model - Methodology

Started with Ordinary Least Squares (OLS) regression:

- Simple, interpretable model for coaches and front office staff
- Great for the level of inference we wanted to achieve
- Avoids "black box" methods that obscure relationships

Mathematically We minimize the squared prediction error: $\min_\beta \|y - X\beta\|^2$

- **y** = vector of win percentages
- **X** = matrix of predictor variables
- **β** = vector of coefficients

OLS proved problematic with our basketball data:

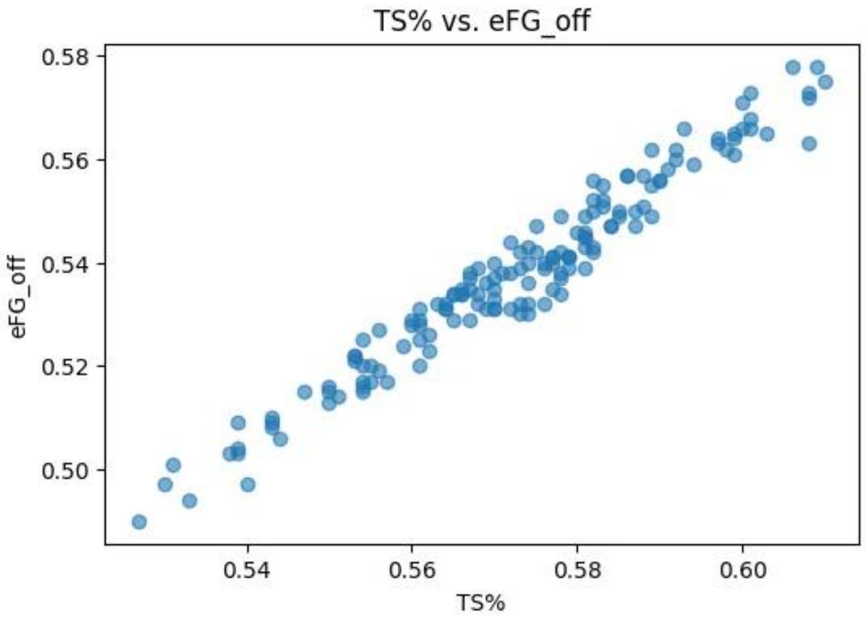
- Coefficient estimates with incorrect signs (e.g., suggesting good shooting hurts winning)
- Implausibly large coefficient magnitudes
- Unstable results when small changes made to input data

Need: A method that maintains interpretability while handling correlations

Our Model - Methodology

Why Correlation Matters

- Traditional regression (OLS) struggles with correlated predictors:
  - Unstable coefficient estimates
  - Inflated standard errors
  - Difficulty isolating variable effects
- Basketball example:
  - Teams shooting well from 3PT often have good overall FG%
  - Metrics like True Shooting % and Effective FG measure close to the same thing



Our Model - Methodology

Our Solution: Elastic Net Regression

Combines strengths of Ridge and Lasso regression

Mathematical formulation: $\min_\beta\{\|y - X\beta\|^2 + \lambda[(1-\alpha)\|\beta\|_2^2/2 + \alpha\|\beta\|_1]\}$

- λ controls the overall strength of regularization
- α determines the mix between L1 and L2 penalties ($0 \leq \alpha \leq 1$)

Benefits for basketball analytics:

- Group selection for correlated variables
- Feature selection capabilities
- Stable coefficient estimates
- Adjustable regularization strength

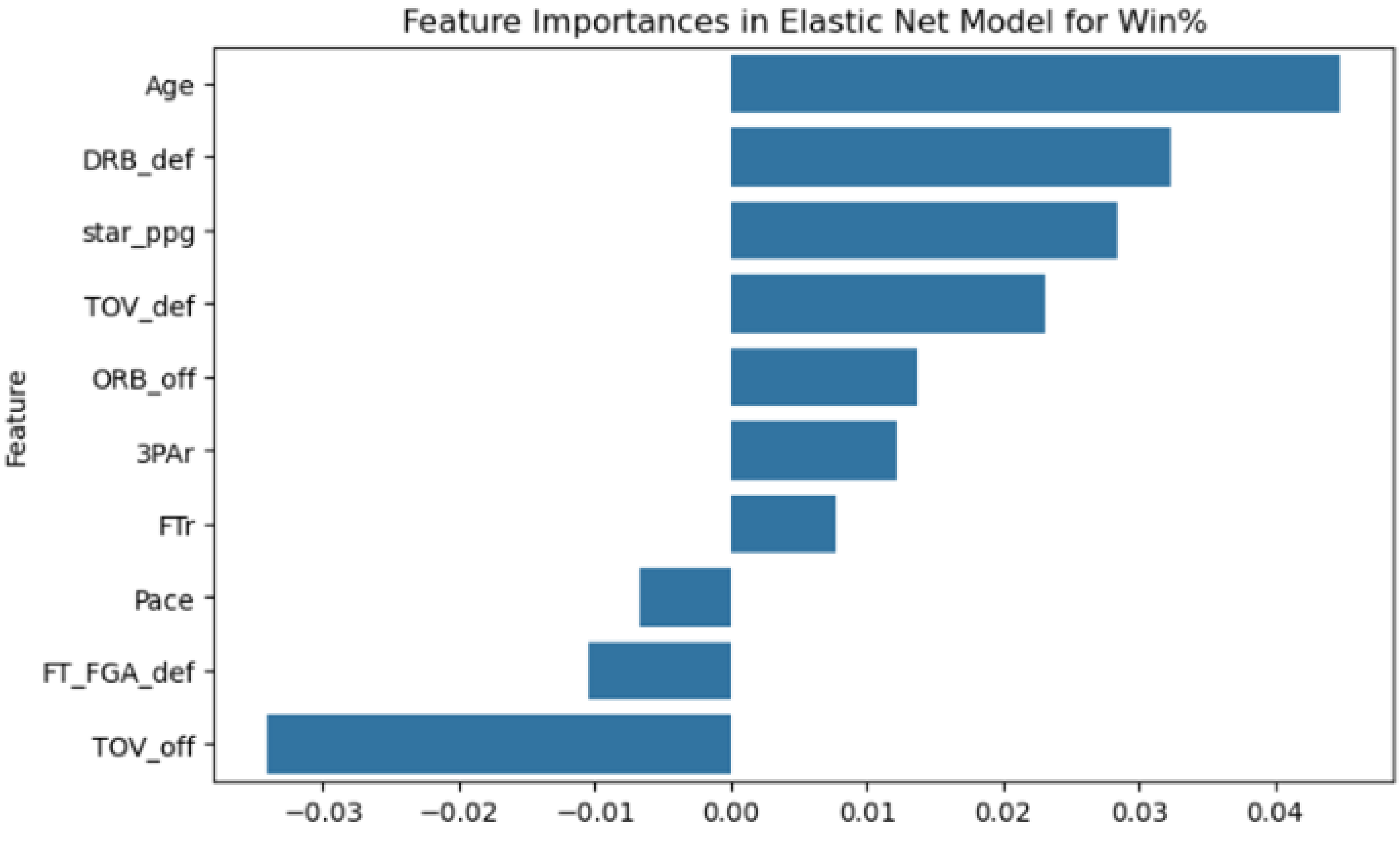
Our Model - Performance

Model Development:

- Hyperparameter(λ and α) selection via 5-fold cross-validation
- Tested range of alpha values (10^{-6} to 10^{6})
- Tested range of L1 ratios (0.01 to 0.99)
- Optimal parameters:
  - Alpha (λ) = .123285
  - Moderate regularization strength Strong enough to address correlation without over constraining
  - L1 ratio (α) = 0.01 (strong preference for Ridge regression)
  - Focuses on addressing coefficient variance rather than aggressive feature selection

Elastic Feature Selection:

- L1 component confirmed all 10 selected metrics contribute to winning (none zeroed out).
- Shrank coefficients of less impactful factors like Pace (-0.0067) while preserving stronger predictors .
- Stabilized estimates between correlated basketball metrics.



Feature Importances in Elastic Net Model for Win%
Age
DRB_def
star_ppg
TOV_def
ORB_off
3PAr
FTr
Pace
FT_FGA_def
TOV_off
Feature
−0.03
−0.02
−0.01
0.00
0.01
0.02
0.03
0.04

Team	Predicted Win%
Boston Celtics	0.705013
Golden State Warriors	0.642912
Los Angeles Clippers	0.623210
Milwaukee Bucks	0.603312
Oklahoma City Thunder	0.600677
New York Knicks	0.575982
Sacramento Kings	0.574266
Minnesota Timberwolves	0.567079
Los Angeles Lakers	0.543736
Cleveland Cavaliers	0.539127
Miami Heat	0.536789
Orlando Magic	0.535760
Houston Rockets	0.516976
Denver Nuggets	0.514182
Phoenix Suns	0.510366
Philadelphia 76ers	0.505384
Detroit Pistons	0.486740
Atlanta Hawks	0.454891
Dallas Mavericks	0.454007
San Antonio Spurs	0.451621
Indiana Pacers	0.437774
Brooklyn Nets	0.415650
New Orleans Pelicans	0.412547
Memphis Grizzlies	0.408349
Charlotte Hornets	0.402365
Chicago Bulls	0.387966
Toronto Raptors	0.350721
Portland Trail Blazers	0.322164
Utah Jazz	0.236025
Washington Wizards	0.212718

Real World Test on the Recent 2024-2025 Season

Key Takeaways:

1. 75% success rate in predicting which team entered the playoffs, and the errors were typically low seed teams

1. More than 67% of predicted win percentages were within a .1 margin

1. Predictions thrown off by oddity teams that have young players who play at a fast pace yet manage to perform well

Team	Actual Win%
Oklahoma City Thunder	0.829268
Cleveland Cavaliers	0.780488
Boston Celtics	0.743902
Houston Rockets	0.634146
New York Knicks	0.621951
Los Angeles Clippers	0.609756
Denver Nuggets	0.609756
Indiana Pacers	0.609756
Los Angeles Lakers	0.609756
Minnesota Timberwolves	0.597561
Memphis Grizzlies	0.585366
Golden State Warriors	0.585366
Milwaukee Bucks	0.585366
Detroit Pistons	0.536585
Orlando Magic	0.500000
Atlanta Hawks	0.487805
Sacramento Kings	0.487805
Dallas Mavericks	0.475610
Chicago Bulls	0.475610
Miami Heat	0.451220
Portland Trail Blazers	0.439024
Phoenix Suns	0.439024
San Antonio Spurs	0.414634
Toronto Raptors	0.365854
Brooklyn Nets	0.317073
Philadelphia 76ers	0.292683
New Orleans Pelicans	0.256098
Charlotte Hornets	0.231707
Washington Wizards	0.219512
Utah Jazz	0.207317

Discussion

Discussion - Strategy Insights

- Factors that maintained your OWN team’s possession were heavily weighted
  - Defensive Rebounding
  - Ball security on offense
- Star power should not be underestimated
  - successful teams generally need a strong scoring leader
- If a team wants to push towards winning, they want more veteran-aged players. However, there’s nuance to this since those players will retire sooner than younger players

Discussion - Limitations

Data Size		
- Only 30 NBA teams - Potential for focusing on game by game analytics, but we thought single games might have too much noise to model well - Can use more seasons for data, but our third point addresses why this may not work		

Discussion - Limitations

Data Size	Model Complexity	
- Only 30 NBA teams - Potential for focusing on game by game analytics, but we thought single games might have too much noise to model well - Can use more seasons for data, but our third point addresses why this may not work	- Could afford to use more predictors, but data for them can be hard to generate and collinearity remains an issue - More complex models, such as XGBoost or DNNs, but data size limitation may mitigate their usefulness	

Discussion - Limitations

Data Size	Model Complexity	Dynamic League
- Only 30 NBA teams - Potential for focusing on game by game analytics, but we thought single games might have too much noise to model well - Can use more seasons for data, but our third point addresses why this may not work	- Could afford to use more predictors, but data for them can be hard to generate and collinearity remains an issue - More complex models, such as XGBoost or DNNs, but data size limitation may mitigate their usefulness	- Trades can taint in-season data - Rules are constantly changing, forcing teams to adapt playstyle - In season success does not necessarily translate to playoff success, although it's closely related

Discussion - Future Work

- More predictors to evaluate
  - Potential web scraping
  - Unique predictors that offer value and don’t affect multicollinearity

- Explore possibility of using game data and stronger models to deal with the extra noise - instead of a general strategy analysis, shift our focus to a more win predictive approach

Thank You!

Questions?