Analyzing NBA Stats

Antonio Hila

Which Stats Tell Us The Most About Team Wins

I looked at nba team stats from the last 6 years to try and find which ones would be the most influential to team success (Wins)

I predict offensive statistics such as points and 3 point shooting will have the highest correlation to wins

My Data

Monstars

Starters	FGM-A	3PM-A	FTM-A	REB	AST	STL	BLK	TO	PTS	PTS BY DUNK
Pound (Barkley)	16-16	5-5	0-0	0	0	6	0	1	37	37
Bang (Johnson)	3-3	0-0	0-0	0	0	2	0	0	6	6
Nawt (Bogues)	0-0	0-0	0-0	0	6	4	0	0	0	0
Bupkus (Ewing)	15-16	4-4	0-0	0	0	2	0	0	34	34
Blanko (Bradley)	0-0	0-0	0-0	0	0	0	0	0	0	0
Totals	34-35	9-9	0-0	0	6	14	0	1	77	77

Pulled Data from **Basketball-reference.com**

- Primarily Box Score stats with some important advanced statistics (True shooting %, Assist%, etc)
- And Defensive Field Goal% to encapsulate the teams defense beyond just steals and blocks

I found that the 3 most influential statistics were

- **Defensive Field Goal %**
- **True Shooting %**
- **Turnovers**

The Process

Data Gathering and Cleaning

- Pulled data from all different areas of basketball reference for 180 Teams and 2700+ players over 6 seasons
- Organized all player data by year to add on team data

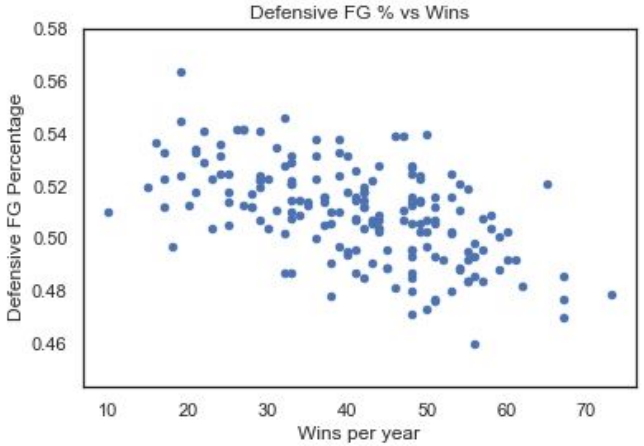
Exploratory Data Analysis

- Created scatter plots for all the data against wins
- Generated correlation heat maps

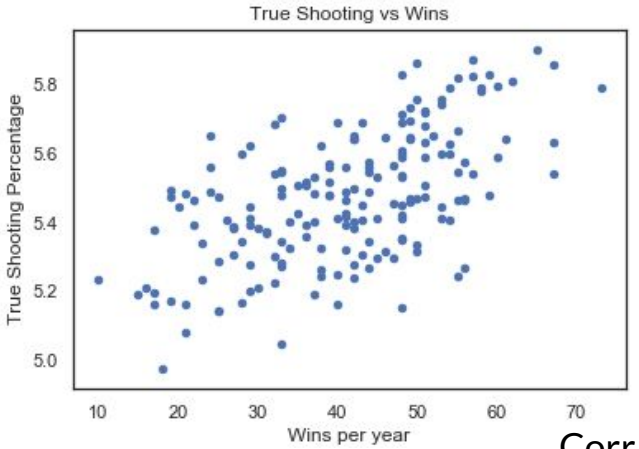
Modeling

- Ran multiple linear regression models while constantly changing features to get the best model

EDA Results

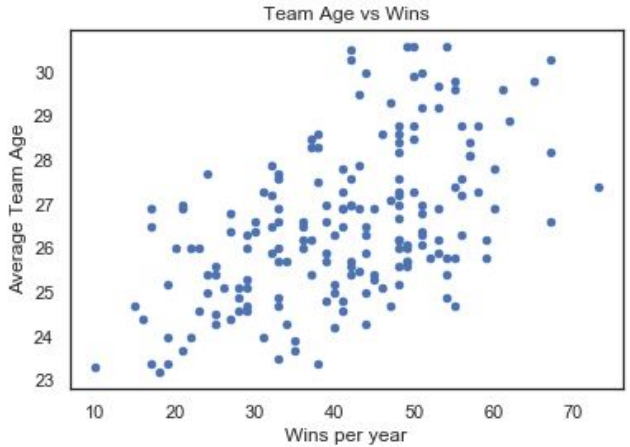


Corr = -0.55740



Corr = 0.58688

Corr = .52273



Modeling the Data

- Used RFE, Lasso and K-best Tests to narrow down and find best features to keep in final model
- Lasso Method worked best
- Original model had a ton of Collinearity so even if it has the best number it is not the most predictive it is just the most overfit model

Accuracy of final model using Cross Validation = 83.4%

Final Model Predictions were only .4 STD away from the mean

	# of Features	Training RMSE	Testing RMSE
All Features (Baseline)	42	2.47	3.92
Selected Features (EDA/VIF/Models)	17	3.89	6.17
Selected Features (w/interactions)	10	3.41	5.61

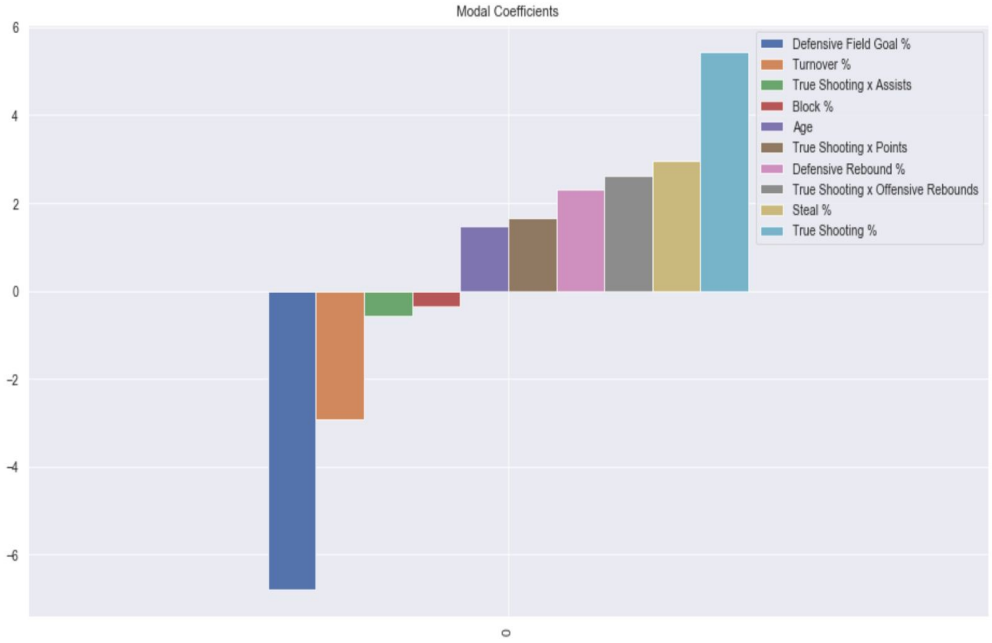
Analyzing the Models Best Features

True Shooting %

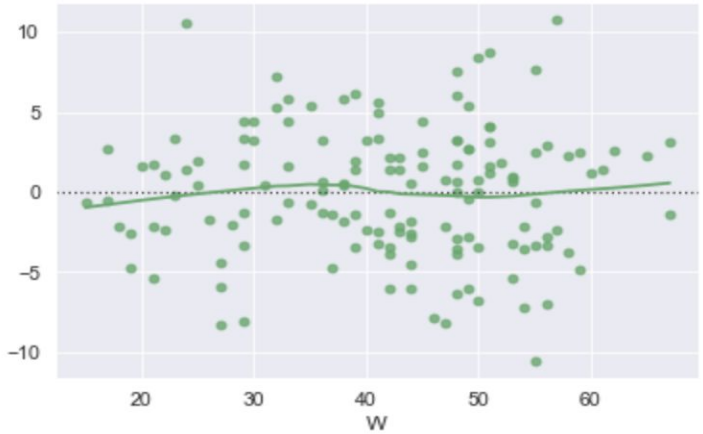
- Efficiency mattered more than scoring in volume when it came to the final data

Defensive Field Goal %

- Highest relation to the final results of the model
- Efficiency on defense as a whole is more important than individual plays (Steals/Blocks)



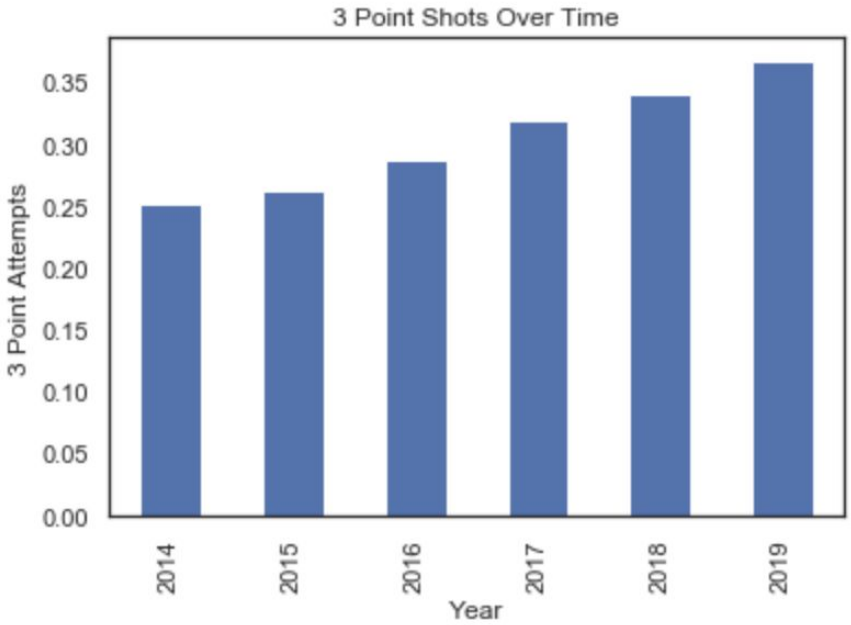
Final Thoughts on the Model



- Was I Right
  - No and Yes
  - Top value was a defensive statistic
  - 3 of the next 4 highest rated features

- **Efficiency is key!**
  - No matter what side of the ball you're on efficiency and possessions had a large influence on the final model

Problems with the data



- 3 Pt Shooting rate is at an all time high