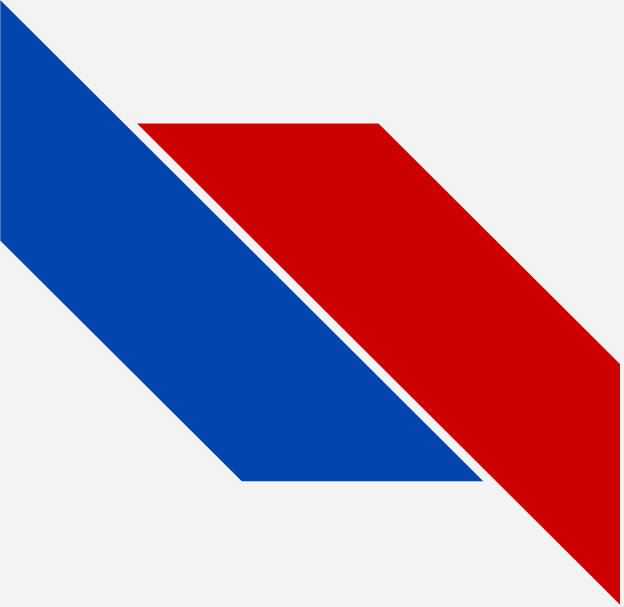


Predicting NBA Champion Based on Regular Season Performance

Jameson Tuck

Introduction

- Every season, a new NBA champion is crowned
- The regular season of the NBA has a sample size of 82 for each team
- How well can we predict the which team will win the NBA championship based on various data collected throughout the regular season?

Objectives

- To build a logistic regression model for predicting the NBA Champion
- To determine what the most significant factor is in predicting which team will win the NBA finals
- Figure out how well we can predict the Champion off of regular season play

About the Data

- We collected data from every teams regular season performance from Basketball Reference
- Obtained every teams basic and advanced statistics from the last 45 NBA seasons
- Data collected from 55,350 games
- Full model has 14 predictor variables

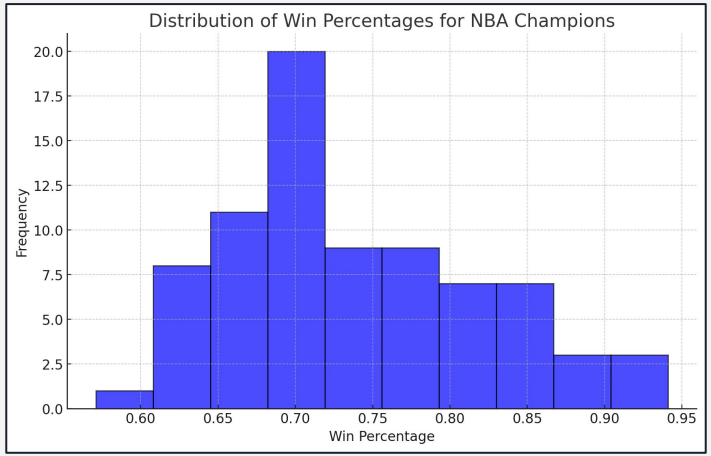
https://www.basketball-reference.com/leagues/NBA_2024.html?sr&utm_source=direct&utm_medium=Share&utm_campaign=ShareTool

About the Data (Continued)

- Based on basic domain knowledge, we were able to isolate 17 predictors of interest (all based on per-game basis)
    - These include: 3-Point Attempts (3PA), 3-Point % (3PP), 2-Point Attempts (2PA), 2-Point % (2PP), Free Throw Attempts (FTA), Free Throw % (FTP), Assists (AST), Steals (STL), Blocks (BLK), Win % (WP), Margin on Victory (MOV), Strength of Schedule (SOS), Offensive Rebound % (ORBP), Turn-Over % (TOVP), Defensive Rebound % (DRP), Opponent Turn-Over % (OTOP), and Average Age (AGE)
- These represent traditional success indicators while avoiding overly derived metrics
- It should be noted that only teams that qualified for the playoffs in the given year were included in our analysis as these are the only teams that have a chance to win the championship

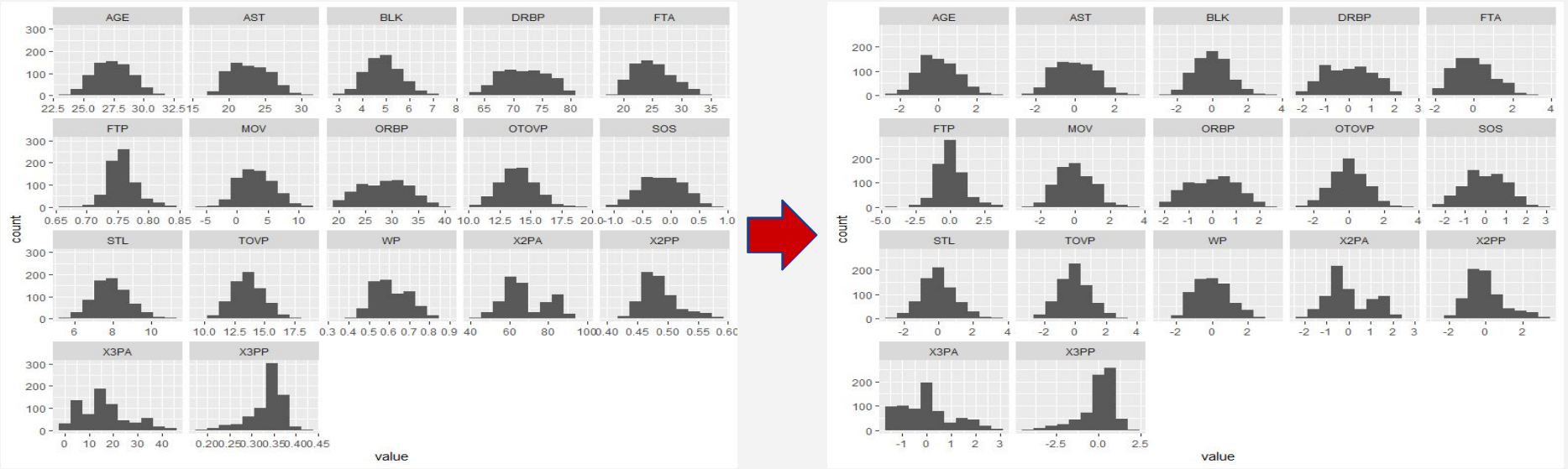
Data Exploration

- Upon looking at the data, the clear correlation to becoming an NBA champion is with winning percentage
- The mean winning percentage is obviously 50% but for an NBA champion it is about 73.5% in our data with its distribution plotted below



Data Exploration (Continued)

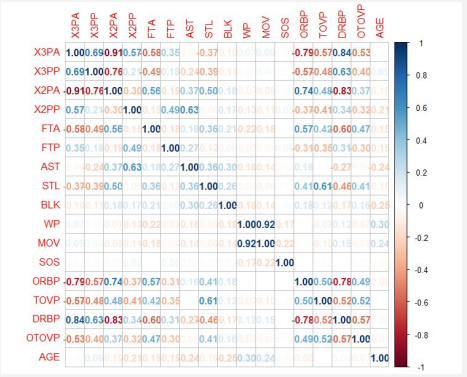
- In order to aid in the interpretability of our model(s), we normalized our predictors so that each would have a similar scale and range of values

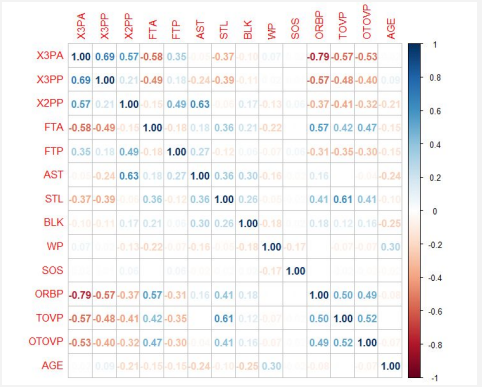


Verifying Assumptions: Multicollinearity

- The matrices display correlations between predictors in our data, comparing the state before and after addressing multicollinearity (removing |correlation| >= 0.8; set level based on subject knowledge)

- This adjustment improves the model's stability and interpretability by reducing redundancy and lowering potential issues with inflated variance in coefficient estimates



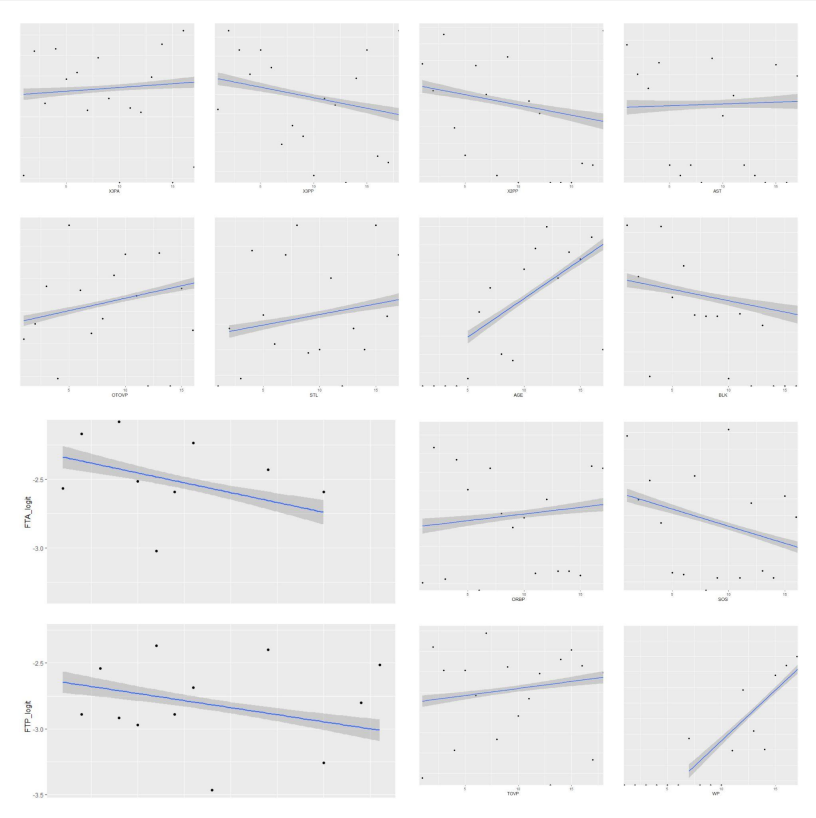


Verifying Assumptions: Linearity

Linearity assumption: logit of the probability of the outcome has a linear relationship with the predictors

- Predictors satisfying linearity: Age, Win Percentage, Strength of Schedule, Margin of Victory, Free Throw Percentage
- Predictors deviating from linearity: Offensive Rebound Percentage, Turnover Percentage, Three Point Percentage, Free Throw Attempts, Steals

Overall, these deviations seem to indicate a weak relationship rather than a non-linear one, so we will simply proceed with caution



Model Selection: Full Model

- Full model shows that many of these variables that we believed to be impacting are not statistically significant to model
- Some had a negative correlation to winning the finals, for instance blocks and free throw attempts
- Our most significant statistic, as anticipated, was winning percentage
- Resulted in an AIC of 155.83

```
> summary(full_model)

Call:
glm(formula = IsChampion ~ ., family = binomial, data = data_train_model)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.93375  -0.28519  -0.11549  -0.05219   2.67053

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.433548   0.531886  -8.336  < 2e-16 ***
X3PA        -0.001251   0.629742  -0.002    0.998
X3PP         0.113645   0.336154   0.338    0.735
X2PP         0.656197   0.610166   1.075    0.282
FTA         -0.478593   0.344797  -1.388    0.165
FTP         -0.334474   0.291031  -1.149    0.250
AST         -0.220532   0.504743  -0.437    0.662
STL          0.161411   0.384794   0.419    0.675
BLK         -0.424808   0.272604  -1.558    0.119
WP           2.023953   0.374420   5.406 6.46e-08 ***
SOS         -0.059699   0.228715  -0.261    0.794
ORBP         0.739818   0.485188   1.525    0.127
TOVP         0.097143   0.397508   0.244    0.807
OTOVP       -0.013144   0.316232  -0.042    0.967
AGE          0.254245   0.290568   0.875    0.382
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$Logit(\pi(x)) = \beta_0 + \beta_1 * 3PA + \beta_2 * 3PP + \beta_3 * 2PP + \beta_4 * FTA + \beta_5 * FTP + \beta_6 * AST + \beta_7 * STL + \beta_8 * BLK + \beta_9 * WP + \beta_{10} * SOS + \beta_{11} * ORBP + \beta_{12} * TOVP + \beta_{13} * OTOVP + \beta_{14} * AGE$$

Model Selection: Forward Model

- The next model that was fit used forward selection
- As may be expected based on the full model summary, only the winning percentage was deemed significant enough to add to the model
- This model achieved a lower AIC (138.73) than the full model (155.83)
- It should be noted that we also ran backwards selection, which resulted in the same model with winning percentage being the only predictor

```
> summary(forward_model)

Call:
glm(formula = IsChampion ~ WP, family = binomial, data = data_train_model)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.21005  -0.31900  -0.13950  -0.06955   2.62115

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.1050     0.4456  -9.213  < 2e-16 ***
WP            2.0118     0.3252   6.187 6.14e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 200.42  on 419  degrees of freedom
Residual deviance: 134.73  on 418  degrees of freedom
AIC: 138.73

Number of Fisher Scoring iterations: 7
```

$$Logit\big(\pi(x)\big) = \beta_0 + \beta_1 * WP$$

Model Selection: Custom Model

- Because of the wide gap between the full and forward models, we wanted to see if the incorporation of some predictors might help model performance even if they are not significant
- The predictors selected were the 4 most significant predictors in the full model
- This model's AIC of 141.14 falls between each of the other models

```
> summary(custom_model)

Call:
glm(formula = IsChampion ~ FTA + BLK + WP + ORBP, family = binomial,
    data = data_train_model)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8747  -0.3054  -0.1283  -0.0623   2.6696

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.2468     0.4789  -8.867  < 2e-16 ***
FTA          -0.3726     0.3021  -1.233    0.217
BLK          -0.2551     0.2365  -1.079    0.281
WP            1.9720     0.3329   5.925 3.13e-09 ***
ORBP          0.4152     0.2807   1.479    0.139
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 200.42  on 419  degrees of freedom
Residual deviance: 131.14  on 415  degrees of freedom
AIC: 141.14

Number of Fisher Scoring iterations: 7
```

$$Logit\big(\pi(x)\big) = \beta_0 + \beta_1 * FTA + \beta_2 * BLK + \beta_3 * WP + \beta_4 * ORBP$$

Model Selection: Model Comparison

- Based on the Analysis of Deviance shown below, the added information provided by the additional predictor variables in the custom and full models is not significant when compared to the increased model complexity
- Because of this, the forward model (the simplest model) appears to be most optimal
- We will select the forward model as our primary model, but we will additionally evaluate the custom and full models to allow for more observations to be drawn

```
> anova(forward_model, custom_model, full_model, test="Chisq")
Analysis of Deviance Table

Model 1: IsChampion ~ WP
Model 2: IsChampion ~ FTA + BLK + WP + ORBP
Model 3: IsChampion ~ X3PA + X3PP + X2PP + FTA + FTP + AST + STL + BLK +
    WP + SOS + ORBP + TOVP + OTOVP + AGE
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       418     134.73
2       415     131.14  3   3.5852   0.3099
3       405     125.83 10   5.3131   0.8693
```

Model Performance on Test Set: Forward

Based on Team with Highest Predicted Probability to Win Championship in a Given Year

	Actual: Champion	**Actual:** Not Champion
Predicted: Champion	6 Teams	12 Teams
Predicted: Not Champion	12 Teams	254 Teams

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = 0.9155$$

$$\text{Precision} = \frac{TP}{TP + FP} = 0.3333$$

$$\text{Recall} = \frac{TP}{TP + FN} = 0.3333$$

Model Performance on Test Set: Custom

Based on Team with Highest Predicted Probability to Win Championship in a Given Year

	Actual: Champion	**Actual:** Not Champion
Predicted: Champion	9 Teams	9 Teams
Predicted: Not Champion	9 Teams	257 Teams

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = 0.9366$$

$$\text{Precision} = \frac{TP}{TP + FP} = 0.5000$$

$$\text{Recall} = \frac{TP}{TP + FN} = 0.5000$$

Model Performance on Test Set: Full

Based on Team with Highest Predicted Probability to Win Championship in a Given Year

	Actual: Champion	**Actual:** Not Champion
Predicted: Champion	11 Teams	7 Teams
Predicted: Not Champion	7 Teams	259 Teams

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = 0.9507$$

$$\text{Precision} = \frac{TP}{TP + FP} = 0.6111$$

$$\text{Recall} = \frac{TP}{TP + FN} = 0.6111$$

Model Performance on Test Set: MRR

- The Mean Reciprocal Rank (MRR) is a metric that accounts for predictions that were close but not perfect
- We would prefer that the champion predicted by our model is always among the top few teams as opposed to either correctly predicting the team or ranking the actual champion very low

$$\text{MRR} = \frac{\sum \frac{1}{\textit{Rank Given To Champion By Model}}}{\textit{Number of Years Predicted by Model}}$$

	Forward Model	Custom Model	Full Model
MRR	0.5926	0.6782	0.7384

Reflections

- **AIC/Model Complexity:** While AIC favored simpler models, these likely underfit to the data, failing to capture the critical patterns necessary for accurate predictions
- **Importance of Insignificant Predictors:** Predictors deemed statistically insignificant in simpler models actually contained valuable information, which contributed to better performance on the test set when included
- **Test Set Performance:** Models incorporating these additional predictors achieved better generalization, highlighting the limitations of relying solely on AIC or p-values when selecting features
- **Future Direction**
  - Future iterations should balance expert domain knowledge with statistical significance tests to identify impactful predictors
  - Incorporating a more nuanced feature selection process and domain-informed transformations could improve both model interpretability and predictive power

Future Research

- **Incorporating Eras:** Future research could leverage the distinction between the three unique eras of the NBA as eras reflect shifts in playing styles and team archetypes (such as the emphasis on the three point shot in the modern NBA)
    - The paper below implemented K-means clustering on the same data to identify these 3 distinct eras
        - Rocha da Silva, João Vítor and Rodrigues, Paulo Canas. 'The Three Eras of the NBA Regular Seasons: Historical Trend and Success Factors'. 1 Jan. 2021 : 263 – 275.
- **Feature Engineering by Era:** Future iterations of the model could incorporate era-informed features, such as weighted importance of 3-point attempts, percentages, and defensive metrics
- **Generalization Across Eras:** Investigating whether certain predictors remain consistent across eras if they need to be tailored could lead to better insights involving team dynamics

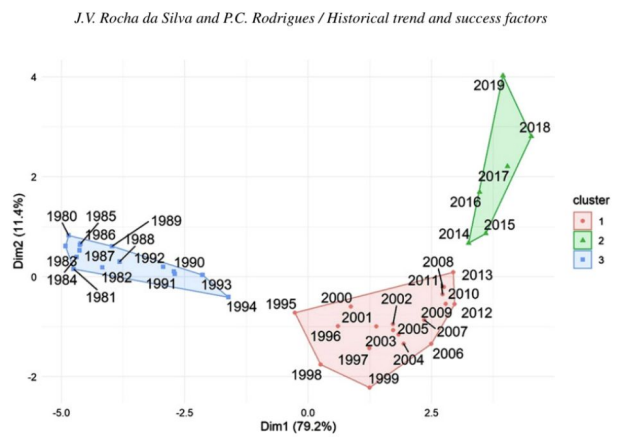


Fig. 5. K-means cluster analysis for all NBA seasons, from 1980 to 2019.

Thank You For Listening