Forecasting NBA Scores

Final project presentation

David Goldman

Background

The NBA betting markets are known to be *highly efficient* in producing accurate betting lines (forecasts), yet there is surprisingly very little public information as how these forecasts are actually created

This project aims to gain a better understanding of the mechanics involved in forecasting NBA scores

Current Research Question

- Can NBA betting lines be *reasonably accurately* recreated through the *simple* Data Science technique of Linear Regression?

Future Research Questions

- Can NBA betting lines then be matched further through more <u>*advanced*</u> Data Science techniques (i.e. machine learning, referee analytics, advanced statistical modeling techniques, etc.)
- Can NBA betting lines then actually be *beaten* (surpassed in accuracy) through betting into the cases where the advanced model *differs* from the betting line (and thus earn a positive ROI over the long-term)?

The Hypothesis

- NBA basketball scores can roughly forecasted using ***simple*** Data Science techniques

- Given that NBA betting lines are ***highly efficient***, a <u>successful</u> first attempt should be just to *roughly replicate* the market betting lines (within a reasonable margin)

The Data

- **Basketball-reference.com** – Best for the overall variety of data
- **Bigdataball.com** – Best for exportable box scores
- **NBA.com** – Best for team-level advanced analytics
- **Pinnacle.com** – Industry-leading sportsbook for current betting lines
- **Rotoword.com** – Best for current player news

Additional Reference

Within the widely respected book, “**Basketball on Paper**” the author identifies the “*four factors of basketball success*” and their associated weights of importance:

- **Shooting (40%) - eFG%**
- **Turnovers (25%) - TOV%**
- **Rebounds (20%) - ORB%**
- **Free Throws (15%) - FT / FGA**

… which served as the starting point for this analysis

- One more metric, “Pace” was also included in the model

Linear regression yielded consistent results in both train and test

Train results

- R^2 = 0.8928

Coefficients

- Intercept = -66.31
- eFG_pct = 144.07 (p-value = 0.000)
- TOV_pct = -130.92 (p-value = 0.000)
- ORB_pct = 48.65 (p-value = 0.000)
- FT_divby_FGA = 31.99 (p-value = 0.000)
- PACE = 0.99 (p-value = 0.000)

Now that we have our betas, let's forecast x-values for any given NBA game

- Each team will have offensive and defensive stats for each independent variable

- We'll treat teams as essentially two different teams for each of their Home and Away stats

- An average between Team A offense vs Team B opponents stats should suffice for each independent variable

…Let's clarify with an example!

2/23 LA Clippers (away) vs Golden State Warriors (home)

- PTS = -66.31 + 144.07 * eFG_pct + -130.92 * TOV_pct + 48.65 * ORB_pct + 31.99 * FT_divby_FGA + 0.99 * PACE

- GSW (home) eFG_pct: 59.5%
- LAC (away) *opponent's* eFG_pct: 51.5%
- Taking the average yields an expected eFG_pct of 55.4% for GSW

- This same methodology can be applied to TOV_pct, ORB_pct, FT_divby_FGA and Pace

- Crunching the math for each team and each independent variable yields an expected score of **LAC 108 GSW 116**

Final considerations

- “Plain vanilla games” vs games where more attention is required – i.e. where an impact player is not playing… or a team is playing back-to-back games where fatigue is a factor… etc.

- Measuring the ongoing accuracy of the model: Logging the model forecasts each day against the betting lines and the actual scores