NBA Statistical Analysis

through a Machine Learning Classifier

Domain Basics

Structure of the NBA and the data set

NBA Single Season	82 regular season games
Number of NBA Teams	30 teams
Temporal Coverage	79 seasons to date, data set covers 41 seasons from 1983 - 2024
Number of Games	49,626 games total in the data set
Team Statistics Only	The model uses team-related data only

Purpose

Data analytics drives our world today, and there are few better examples than the world of sports.

Data:

- Analyze historical trends
- Derive statistics which show impact on winning games

Model:

- A classification model
- Acts as a vessel for deriving latent features behind winning a basketball game

Result:

- Indicate which features may be possible predictors of success, applicable to performance training, sports gambling, predictor models

Data source

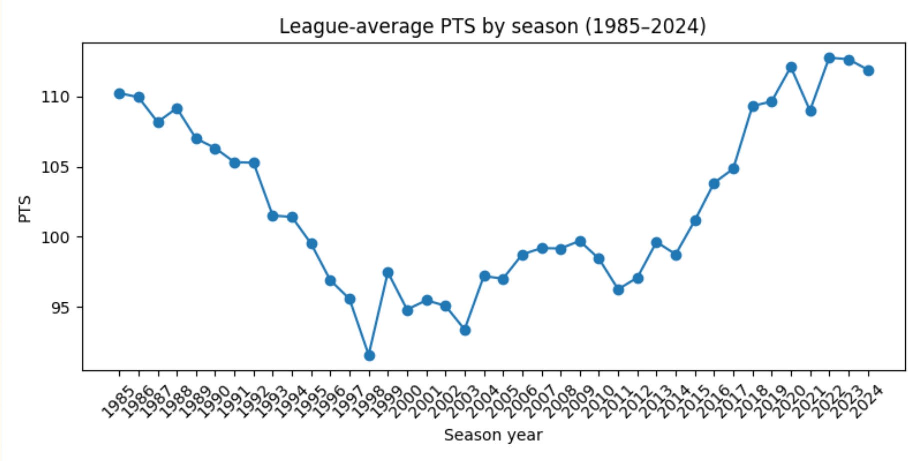
- Data was sourced using the nba_api library as well as contributions from Eoin A Moore and Wyatt Walsh’s Kaggle sets
- Data was cleaned and relevant fields were stitched together and saved in parquet files
- This persistence step was especially important because I was getting rate limited by the nba_api constantly

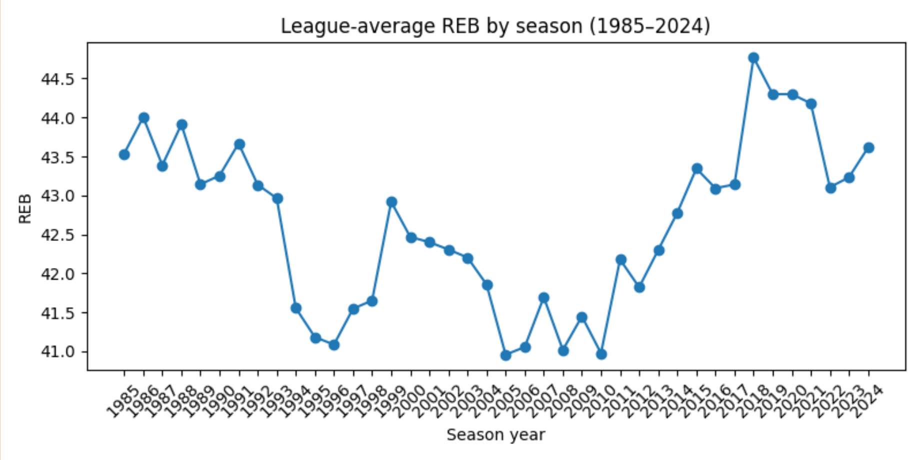
Basketball chaos

Sports are inherently difficult to predict:

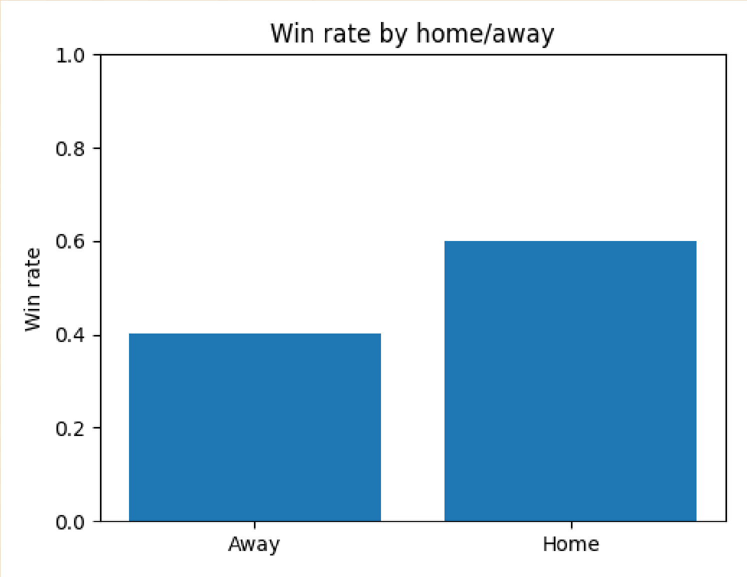
- Basketball especially is an inventory sport, with many games and variance over a season.
- Anyone game is especially hard, with variables including momentum, absences, additions, etc.
- However with large enough sample sizes trends can be found and indicators of success can be uncovered

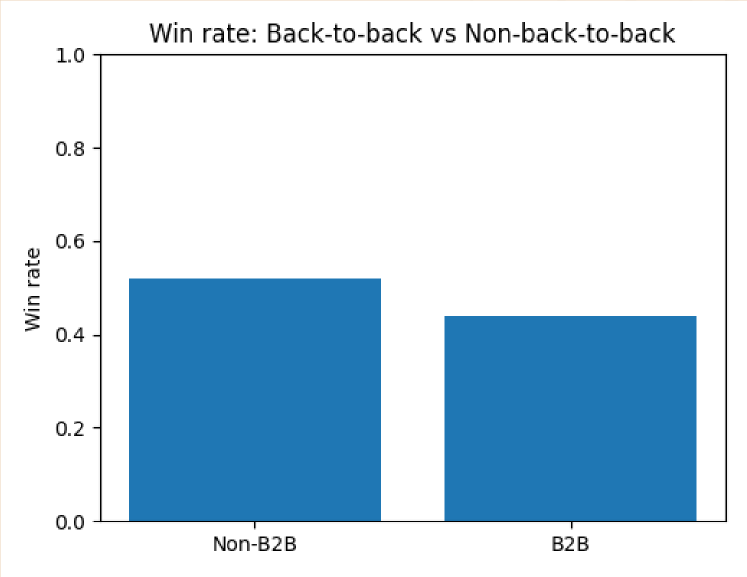
Historical Trends





Historical Trends





Features

The full set of features used in the investigation were as follows, however the model used a deliberately selected subset

Team Stats:

- PTS, OREB (offensive rebounds), DREB (defensive rebounds), AST (total assists), STL (total steals), BLK (total blocks)
- Identity features like team ID and opposing team ID

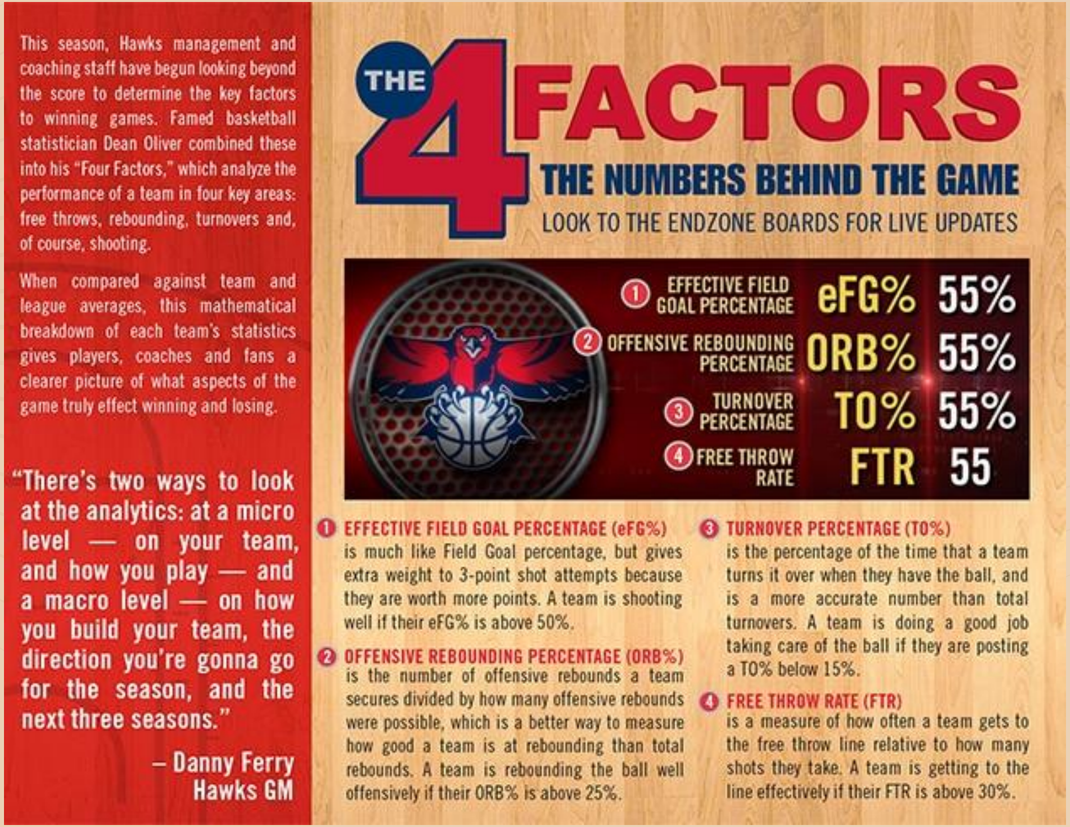
Context:

- HGA (Home Game Advantage), Back-to-Back Game
- Last game outcome (win or loss for previous game)

Efficiency Stats:

- EFG% (Effective Field Goal Percentage), TOV% (turnover rate), FTR (free-throw rate), TS% (True Shooting Percentage)

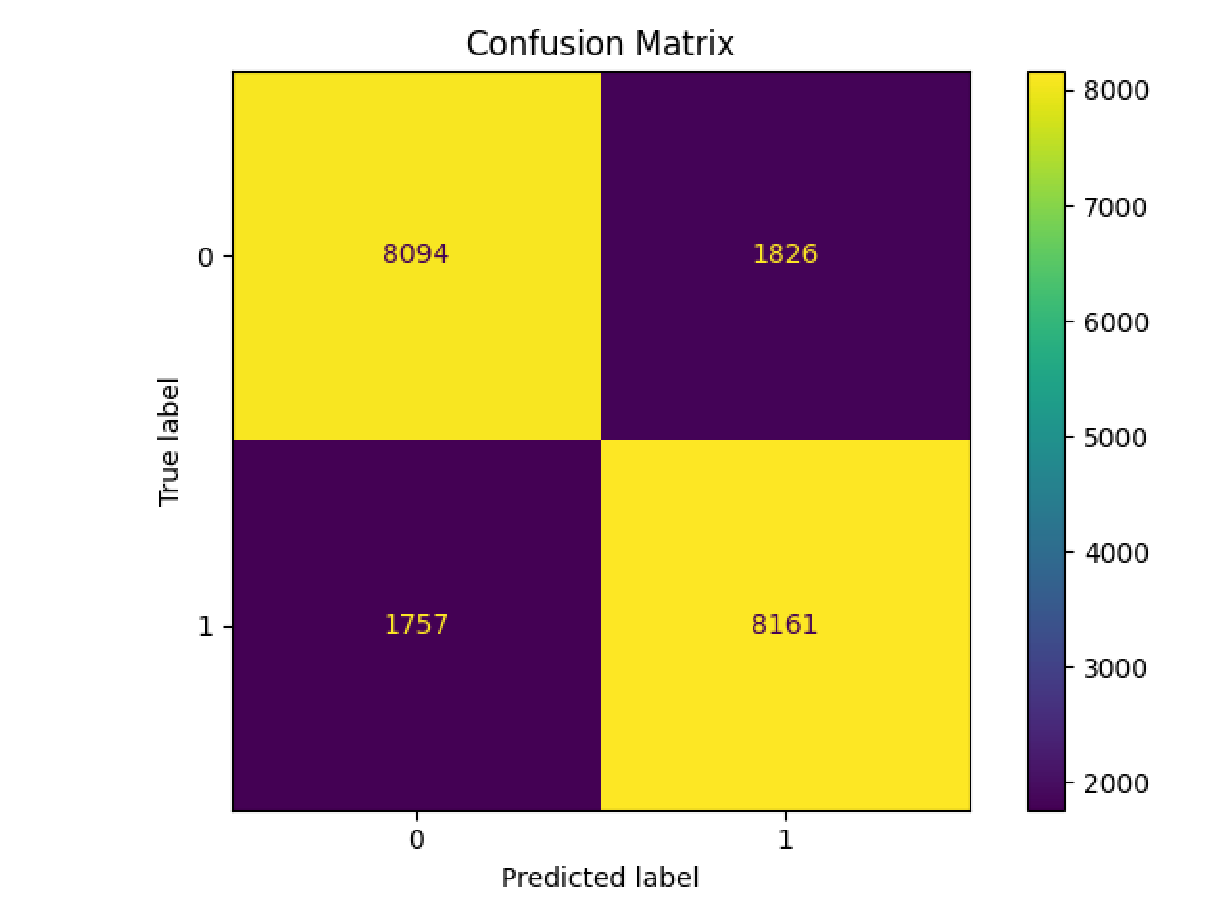
Developed by statistician Dean Oliver



The Model

Gradient Boosted Decision Tree

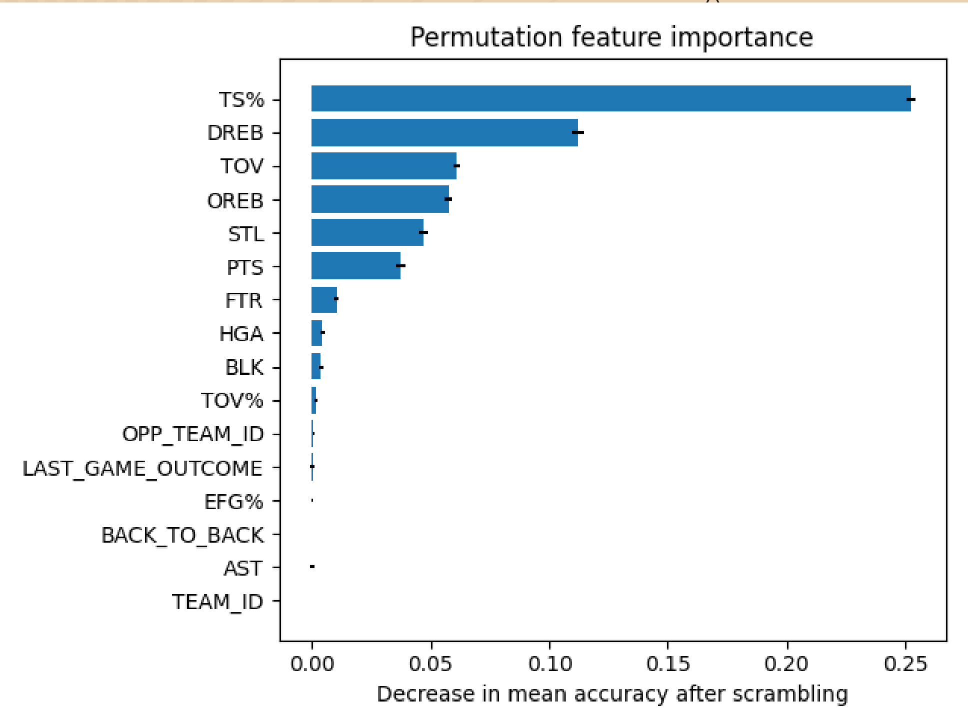
- HistGradientBoostingClassifier from scikit-learn
- Comparable to XGBoost
- Test train split done 80-20% without a temporal split
- Model is interested in classifying a game as a winner or loser by seeing only one teams stats with no reference to other team's performance
- Then upon successful classification latent features are derived to investigate which are being used as predictors of success



Confusion Matrix
8094
1826
1757
8161
True label
Predicted label
0
1
0
1
8000
7000
6000
5000
4000
3000
2000

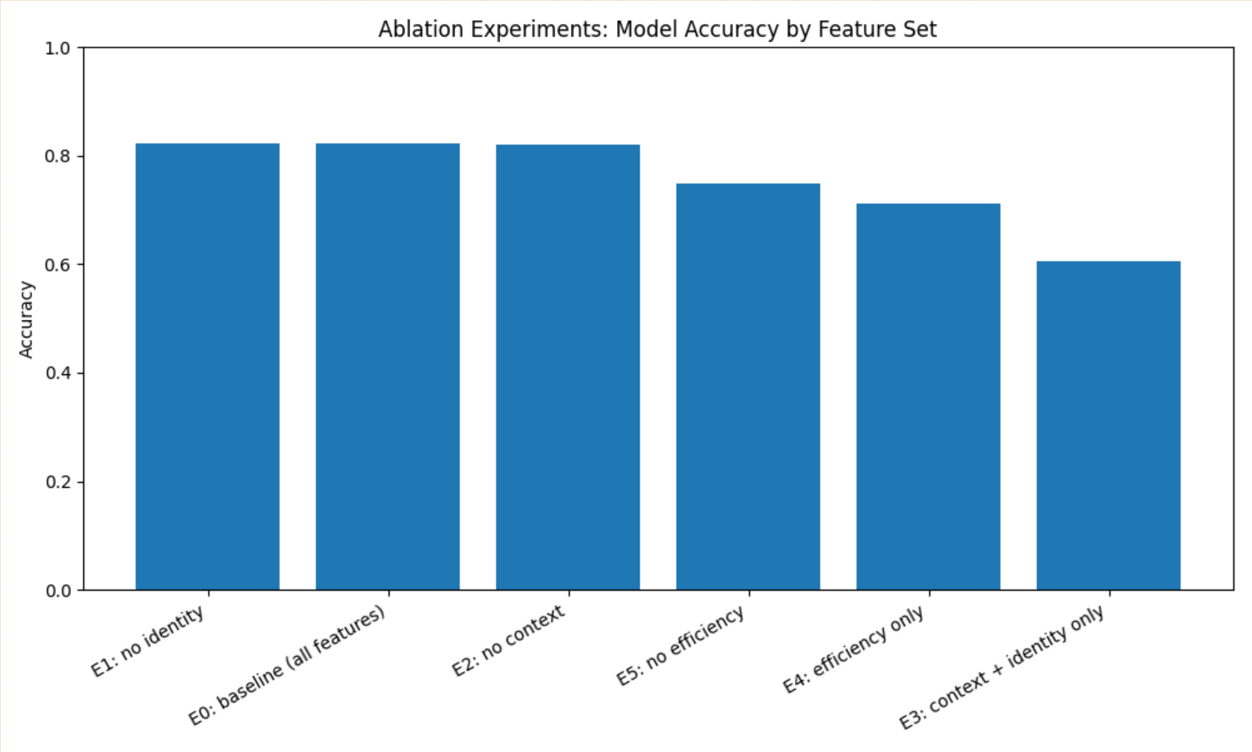
Feature Importance

- Unlike some other tree-based models, HistGradientBoostingClassifier does not have built in feature importance.
- Instead permutation importance is used.
- Single features at a time are scrambled randomly to kill their signal and the decrease in a model's score is measured.



Feature Importance

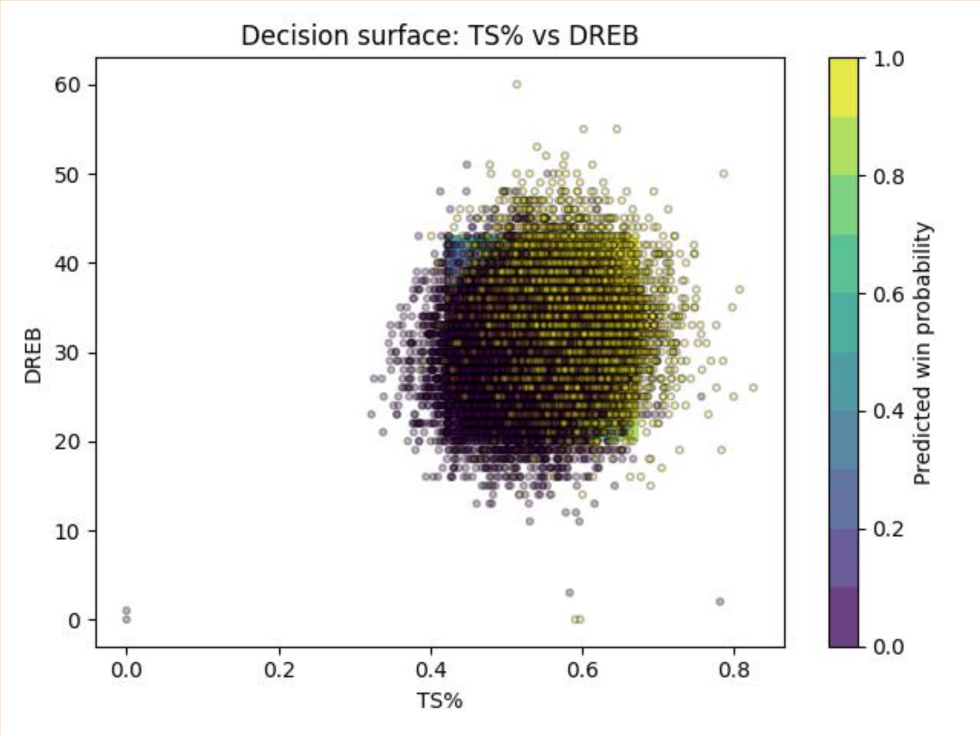
Feature	Description
TS% (True Shooting Percentage)	25.25% decrease in accuracy
	Overall scoring efficiency triumphed over everything
DREB (Defensive Rebounding)	11.21% decrease in accuracy
	Shockingly important, limiting opponent second-chance opportunity was a major win predictor
TOV, OREB, STL, PTS (~3 – 7%)	Suprisingly total points had much less impact
	Turn-overs, offensive rebounds, and steals had greater impact, furthering “possession control is vital”



Removing identity features	Had a slightly positive effect, possibly due to noise interference
Removing context features	Very small drop, possible signal but weak as presently designed
Removing efficiency features	Largest effect with an 8% drop, indicating its signal strength
Efficiency features alone	Using just the Four Factor EFG%, TOV%, FTR, TS% held 71% on its own
Context and identity alone	Showed it had some signal when isolated, better than chance (coin flip)

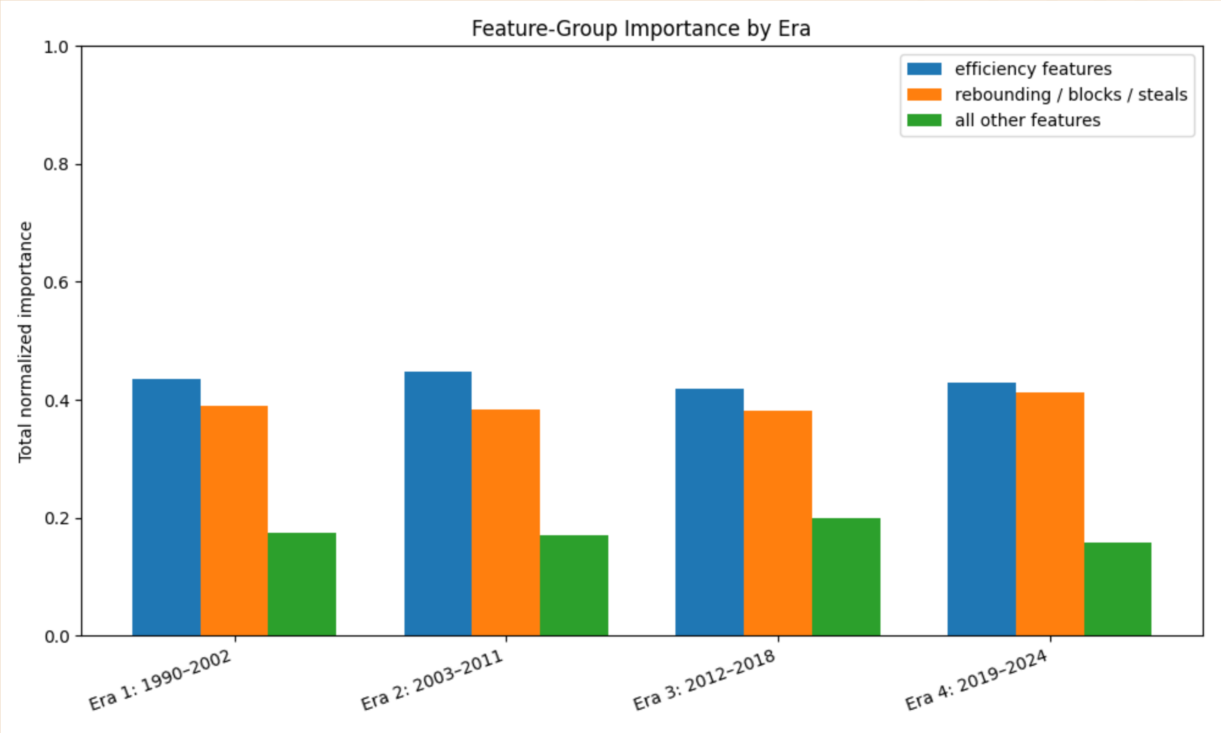
Model’s learned geometry

Based on predicted probability of a win



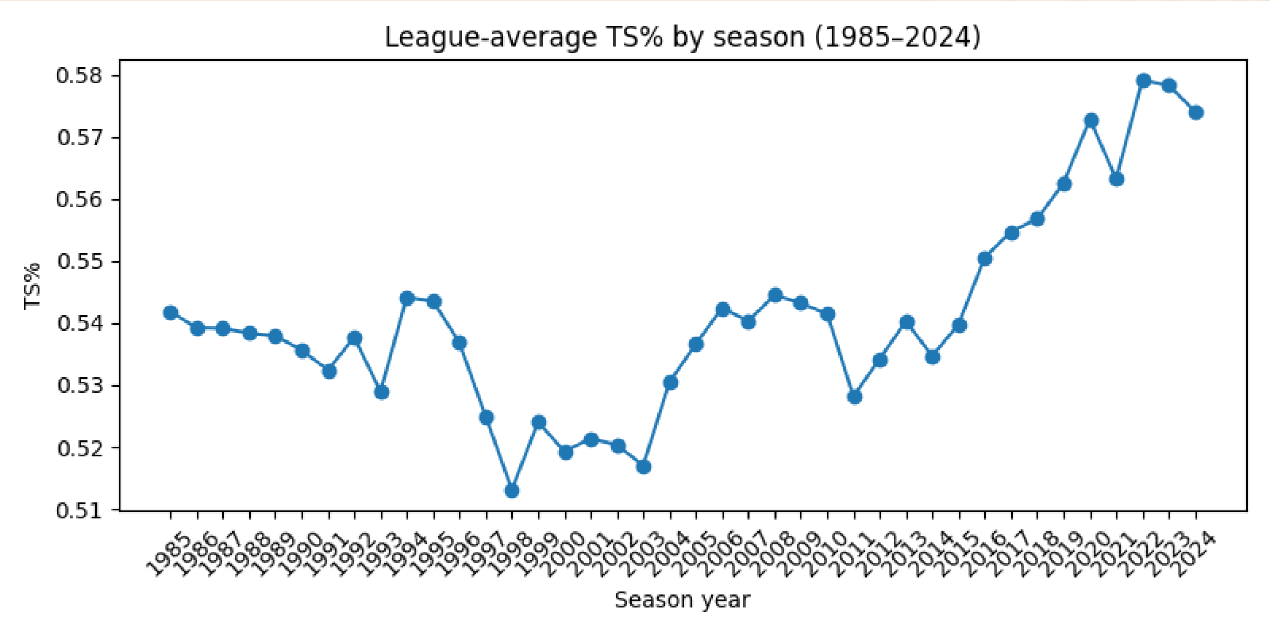
Era-based feature importance

- Rebounding impact surprisingly at its highest currently
- Conventional wisdom would think rebounding might have had more impact in the physical eras of the 90's

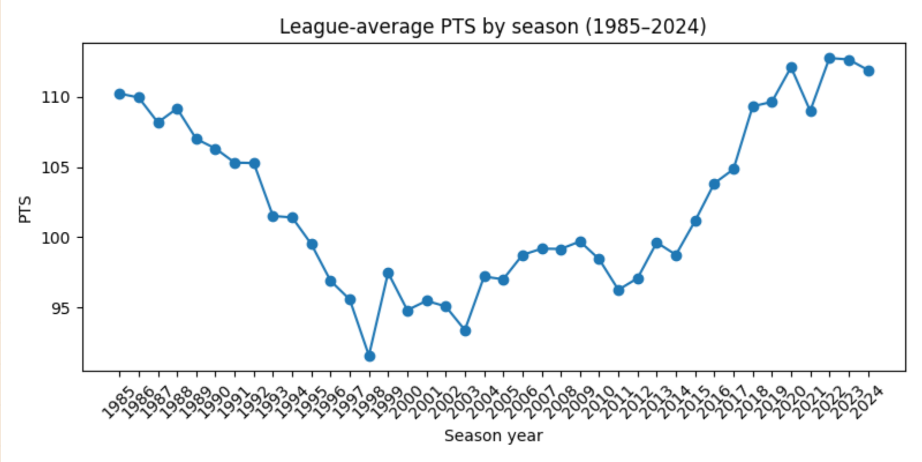


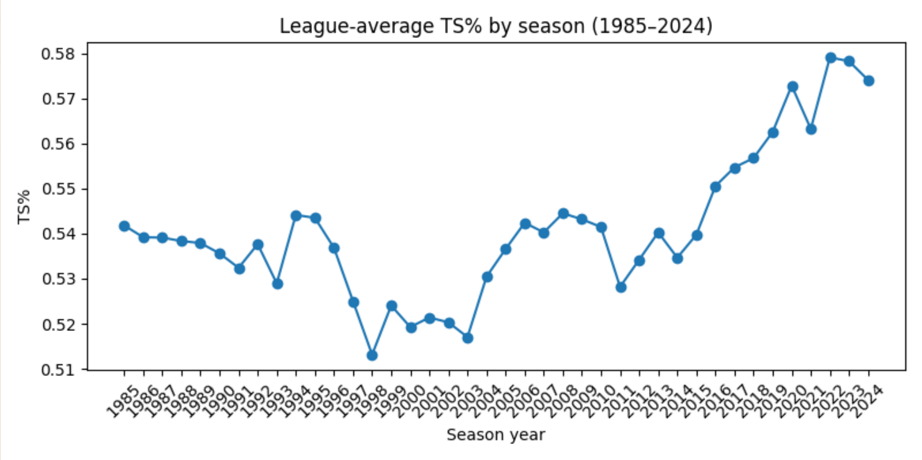
Era-based feature importance

- However TS% has skyrocketed in its importance in the game
- Effeciency stats were shown to always dominate, but that may be true now more than ever in the modern game



Comparing Points and TS%





Conclusions

Dean Oliver purported the Four Factors as being the best indicators of basketball success, and this small analysis supports that notion.

Efficiency stats were far and away the model's choice for successful classification.

Interestingly though, Oliver put weights to his Four Factors giving them the following order of importance: TS%, TOV, REB, FT
Our model however found rebounding to be the stronger indicator and perhaps that is grounds for a more in-depth examination of his theory.

Ultimately this is a quirky approach to the world of advanced analytics used by professional sporting teams.

Sports teams, and the world at large, has become increasingly data driven, and the way the NBA game is played is ever increasingly data-reliant.

Uncovering features most important to basketball success can become a mighty insight for team’s building a roster, predictive and sports gambling models, and simply better understanding what the game values.

Thank you