

NBA GAME PREDICTION
David Hamilton Michael Yang Steffen Roehrsheim

NBA GAME PREDICTION
David Hamilton Michael Yang Steffen Roehrsheim

PROJECT OBJECTIVES

A

Point Spread Prediction

Regression methods to model the points spread between teams as close as possible, .

A Point Spread Prediction

Regression methods to model the points spread between teams as close as possible, .

A Point Spread Prediction

Regression methods to model the points spread between teams as close as possible, .

A Point Spread Prediction

Regression methods to model the points spread between teams as close as possible, .

B Classification

Predicting which team is going to win the match up, and if possible try to predict the market line.

B Classification

Predicting which team is going to win the match up, and if possible try to predict the market line.

B Classification

Predicting which team is going to win the match up, and if possible try to predict the market line.

FEATURES
Team stats from stats.nba.com on a game by game basis.

PROCESSING
We choose the expanding window averaging for the features, starting from

FEATURES
Team stats from stats.nba.com on a game by game basis.

PROCESSING
We choose the expanding window averaging for the features, starting from

PROCESSING

We choose the expanding window averaging for the features, starting from game one.

PROCESSING

We choose the expanding window averaging for the features, starting from game one.

The diagram illustrates the data set split. A large circle is divided into two main sections. The left section is labeled 'TIMEFRAME' and contains the text: 'Taking the last five season of data, 2012-2017. Using the first 3 for train, 1 validate, 1 test.' The right section is labeled 'Format' and contains the text: 'For each training example, we take the home & away team and stack it as an vector.' The circle is further divided into three colored segments: a large orange segment on the left, a smaller yellow segment on the right, and a small red segment at the bottom. The text 'DATA SET' is written in large, bold, grey letters across the center of the circle.

The diagram illustrates the data set split. A large circle is divided into two main sections. The left section is labeled 'TIMEFRAME' and contains the text: 'Taking the last five season of data, 2012-2017. Using the first 3 for train, 1 validate, 1 test.' The right section is labeled 'Format' and contains the text: 'For each training example, we take the home & away team and stack it as an vector.' The circle is also divided into four colored segments: orange (top-left), yellow (top-right), green (bottom-left), and red (bottom-right). The word 'DATA SET' is written in large, bold, black letters across the center of the circle.

**DATA
SET**

ADVANCE STATS

For improving proformance, we also included stats like Offense rating, Opponent's Defensive Rebounds

ADVANCE STATS

For improving proformance, we also included stats like Offense rating, Opponent's Defensive Rebounds

Format

For each training example, we take the home & away team and stack it as an vector.

Format

For each training example, we take the home & away team and stack it as an vector.

METHODS

Linear Regression

Ridge Regression, Lasso Regression, Elastic-net

Hard Classification

Logistic Regression, SVM, Kernel-SVM

Bayesian Methods

Naive Bayesian, Bayesian Ridge Regression

Tree Models

Gradient Boosting Trees, Random Forest

K-Neighbors

K-Neighbors Regression

Method	Accuracy (%)
Linear Regression	68.0%
Hard Classification	68.7%
Bayesian Methods	67.7%
Tree Models	68.8%
K-Neighbors	68.7%

Model	MAE	RMSE	R^2
Linear Regression	0.0001	0.0001	0.9999
Ridge Regression, Lasso Regression, Elastic-net	0.0001	0.0001	0.9999

Model	MAE	RMSE	R^2
Linear Regression	0.0001	0.0001	0.9999
Ridge Regression, Lasso Regression, Elastic-net	0.0001	0.0001	0.9999

The diagram illustrates the relationship between Hard Classification and Soft Classification. A large light blue rounded rectangle is divided into two horizontal sections. The top section is labeled "Hard Classification" in a dark blue rounded rectangle. The bottom section is labeled "Soft Classification" in a light blue rounded rectangle. Below the "Soft Classification" section, the text "Logistic Regression, SVM, Kernel-SVM" is displayed. To the right of the main rectangle, there are three vertical bars of different colors: orange, green, and teal.

The diagram illustrates the relationship between Hard Classification and Soft Classification. A large light blue rounded rectangle is divided into two horizontal sections. The top section is labeled "Hard Classification" in a dark blue rounded rectangle. The bottom section is labeled "Soft Classification" in a light blue rounded rectangle. Below the "Soft Classification" section, the text "Logistic Regression, SVM, Kernel-SVM" is displayed. To the right of the main rectangle, there are three vertical bars of different colors: orange, green, and teal.

Method	Percentage
Bayesian Methods	67.7%
Naive Bayesian, Bayesian Ridge Regression	

Method	Percentage
Bayesian Methods	67.7%
Naive Bayesian, Bayesian Ridge Regression	

Tree Models

Gradient Boosting Trees, Random Forest

Tree Models

Gradient Boosting Trees, Random Forest

K-Neighbors

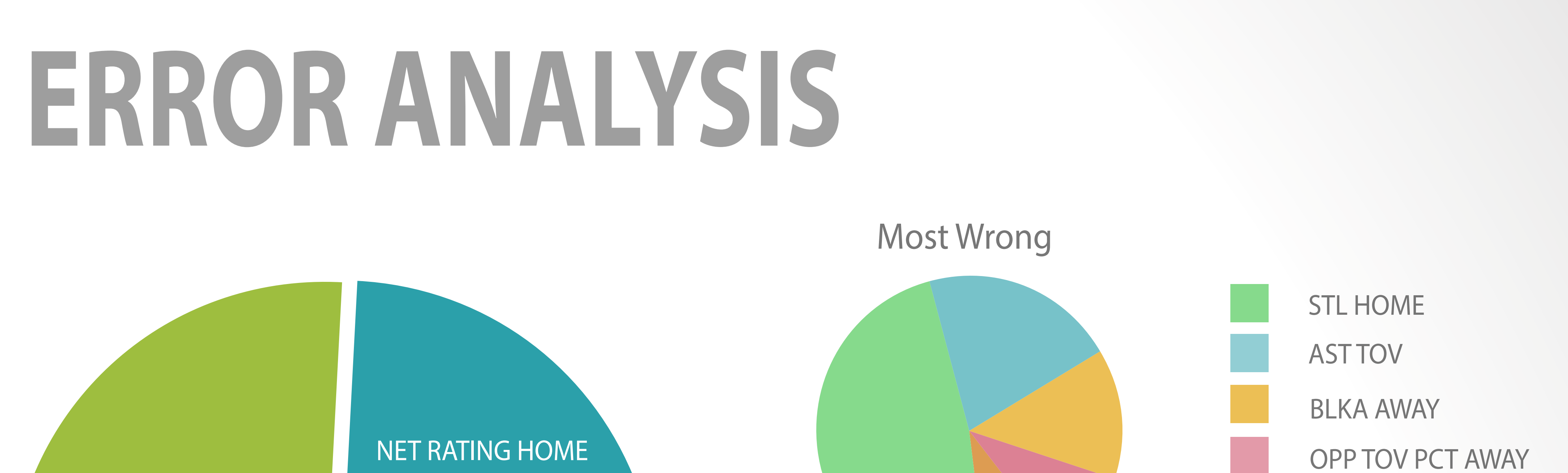
K-Neighbors

ERROR ANALYSIS

NET RATING HOME

Most Wrong

- STL HOME
- AST TOV
- BLKA AWAY
- OPP TOV PCT AWAY

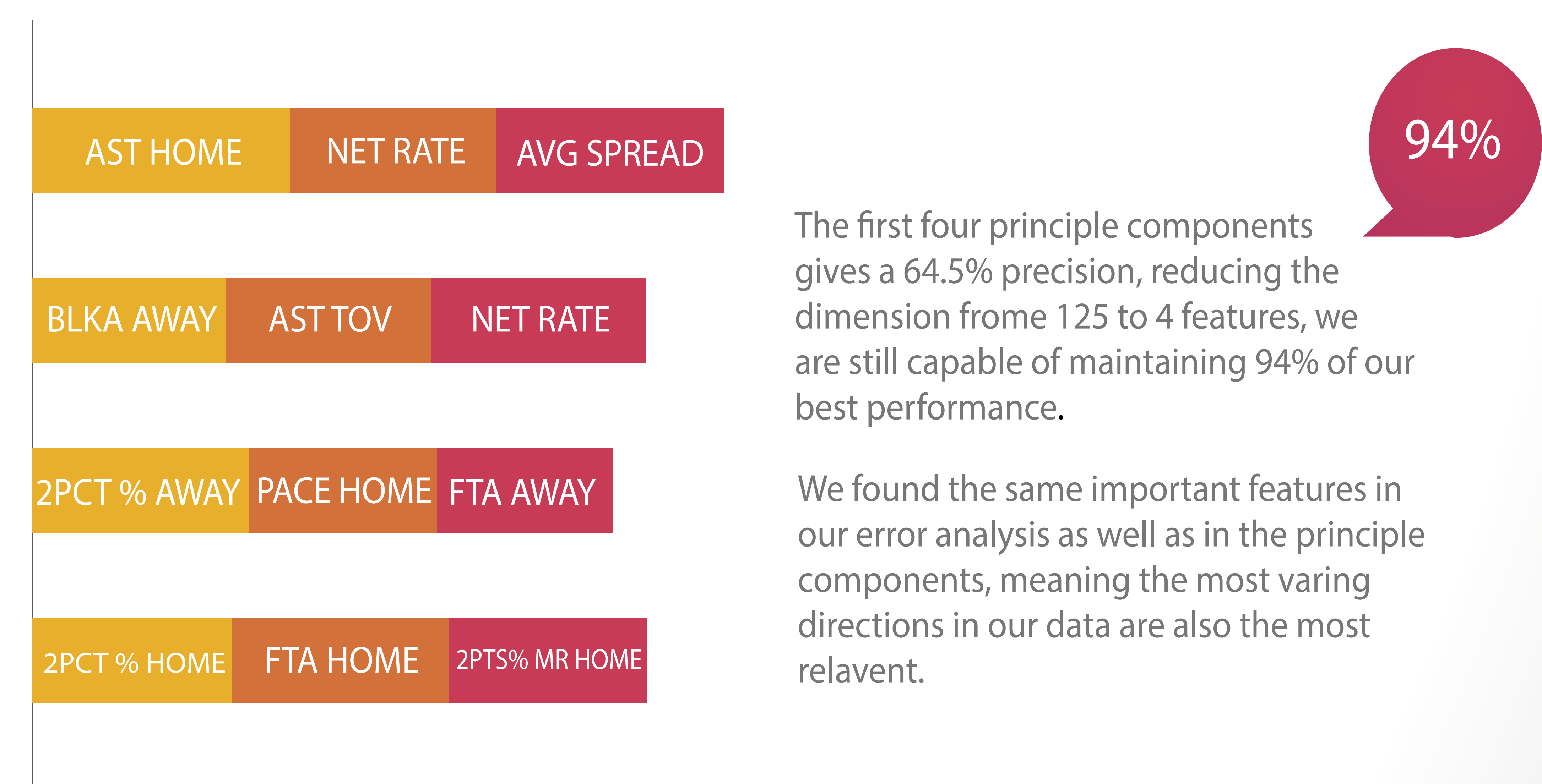


PCA ANALYSIS

From looking into the teams we consistently predicted wrong, we isolated a couple of features, running only with and without those features gives roughly the same performance. The similar performance indicates there's a strong correlation in the features, thus we are looking at PCA.

PCA ANALYSIS

From looking into the teams we consistently predicted wrong, we isolated a couple of features, running only with and without those features gives roughly the same performance. The similar performance indicates there's a strong correlation in the features, thus we are looking at PCA.



AST HOME NET RATE AVG SPREAD

BLKA AWAY AST TOV NET RATE

2PCT % AWAY PACE HOME FTA AWAY

2PCT % HOME FTA HOME 2PTS% MR HOME

94%

The first four principle components gives a 64.5% precision, reducing the dimension from 125 to 4 features, we are still capable of maintaining 94% of our best performance.

We found the same important features in our error analysis as well as in the principle components, meaning the most varying directions in our data are also the most relevant.

AST HOME NET RATE AVG SPREAD

BLKA AWAY AST TOV NET RATE

2PCT % AWAY PACE HOME FTA AWAY

2PCT % HOME FTA HOME 2PTS% MR HOME

94%

The first four principle components gives a 64.5% precision, reducing the dimension from 125 to 4 features, we are still capable of maintaining 94% of our best performance.

We found the same important features in our error analysis as well as in the principle components, meaning the most varying directions in our data are also the most relevant.

FUTURE IMPORVEMENTS



Include player statistics, of perticular interest here would be to match up players against players to predict the would perform against each other.

- A Include player statistics, of particular interest here would be to match up players against players to predict the would perform against each other.

- A Include player statistics, of particular interest here would be to match up players against players to predict the would perform against each other.

We found that kernelizing is not helping our out of sample performance, so the non-linearity of this problem is not captured by common kernels, we need to find smart ways of combining features to achieve better performance.

We found that kernelizing is not helping our out of sample performance, so the non-linearity of this problem is not captured by common kernels, we need to find smart ways of combining features to achieve better performance.