AY2025/2026 Semester 1
ES2001: Computational Earth Systems Science
Group 12

Final Report

Which National Basketball Association (NBA) team has the highest chance of winning the championship?

Matriculation Number	Student Name
U2440278B	Safwan Bin Jam Hari
U2440923C	Lee Yi Shuen
U2340547A	Chow Jia Hui
U2440937J	Caitlin Beatrice Albay Valencia
U2440182B	Lew Sun Zhe

Table of contents

1. Introduction

1.1. Background and Context

The National Basketball Association (NBA) is a globally renowned professional basketball league, with each season's game producing detailed statistical data. It consists of 30 teams, divided into the Eastern and Western Conferences. After the regular season, the top 8 teams in each conference advance to the playoffs, where they progress through a series of elimination rounds. As the bracket narrows, the winners of each conference meet in the NBA Finals, and the team that wins this final series is crowned the NBA champion. Given the NBA's global popularity and the extensive datasets available, we chose it as the focus of our project to apply the predictive tools learned in ES2001 to predict the league's champion.

1.2. Overview

To achieve this, our group developed a prediction model that analyses historical NBA data to estimate each team's likelihood of becoming champion in a given season. We start by using linear regression to identify the top five metrics most strongly correlated with win rate. These metrics are then assigned weights based on their R^2 values, to give greater influence to variables that lead to higher winrate. The model then processes the selected team statistics and calculates the team's win probability.

1.3. Scope

Despite over 50 seasons worth of data being available, we narrowed our scope to ensure the use of current and relevant statistics. Rule changes and evolving play styles over the years can alter the impact of different metrics, making recent seasons more appropriate for analysis. We selected a 10-season period from 2015 to 2024 as the primary dataset for training and evaluating our model. We exclude player based data, external factors such as trades and injuries, and playoff specific variables to keep the scope focused and manageable.

While playoffs determine the eventual champion, their smaller sample size and higher variability make playoff data less suitable for analysis. Hence, we chose to use regular season data, with its larger sample size and greater consistency, allowing for better insight into a team's strength.

1.4. Assumptions

To ensure consistency and account for the unpredictable nature of the sport, several assumptions have been established:

a. **Team performance statistics are consistent across each season**
   Season averages are treated as a reliable representation of a team's actual strength.
b. **Teams perform at a similar level in both the regular season and the playoffs**
   Assume that regular season strength can be used to approximate playoff potential, despite the variability introduced by matchups or situational factors.
c. **Win rate is an effective proxy for overall success**
   Since win rate directly reflects team performance, it is used as a basis for determining the metrics
d. **Team metrics directly affect winrate**
   Team statistics such as shooting efficiency and net rating are assumed to have impacts on winning games

2. Data Processing

2.1. Source

We used the official NBA website and 'Basketball Reference' to obtain data sets from seasons 2015 to 2024 covering basic and advanced statistics for each team. We combined both datasets to obtain a greater range of metrics that could be used as indicators for our model.

2.2. Loading data onto MATLAB

The datasets of basic and advanced team statistics respectively were separated by season, each in a separate CSV file. We used a for loop to automate the processing of data. Within the for loop, textread loaded the data from the CSV files, then all the statistics are stored in a struct, where each field represents each of the metrics in the dataset. All the structs for each season are then combined into one table. This for loop was used for both sets of basic and advanced statistics for seasons 2015 to 2024. After which, the basic and advanced tables are then combined into one large table.

2.3. Shortlisting of metrics for 1st prediction model

We calculated the R^2 values and correlation coefficient for every metric in our dataset against win percentage.

```
%% Compute target variable
combinedTable.WinRate = combinedTable.W ./ (combinedTable.W + combinedTable.L);
targetVar = 'WinRate';
```

Firstly, we specified our target variable, win rate, by taking the number of wins divided by the total games played.

```
%% Select numeric predictors
numericVars = varfun(@isnumeric, combinedTable, 'OutputFormat', 'uniform');
predictorNames = combinedTable.Properties.VariableNames(numericVars);
predictorNames(strcmp(predictorNames, targetVar)) = [];
X = combinedTable{:, predictorNames};
y = combinedTable.(targetVar);
```

Then, we extracted all numeric columns, except WinRate, to extract the metrics for evaluation.

```
%% Evaluate each metric
results = [];
for i = 1:length(predictorNames)
    xi = X(:, i);
    valid = ~isnan(xi) & ~isnan(y);
    xi_v = xi(valid);
    y_v  = y(valid);

    G = [ones(size(xi_v)), xi_v];
    m = (G' * G) \ (G' * y_v);
    y_pred = G * m;

    SS_tot = sum((y_v - mean(y_v)).^2);
    SS_res = sum((y_v - y_pred).^2);
    R2 = 1 - SS_res / SS_tot;

    C = corrcoef(xi_v, y_v);
    r = C(1,2);
```

Our for loop evaluates every metric, and calculates how much a single metric explains variation in WinRate. Then, R^2 is calculated to find out which metrics are the best indicators of a team winning. Then, we calculate the correlation between the metric and win rate. After running the for loop, we got the following top 5 metrics with the highest R^2 values and strongest correlation coefficients:

1. Win amount (W),
2. Loss amount (L),
3. Margin of Victory (MOV),
4. Net rating (NRtg),

5. Simple Rating System (SRS)

3. Methodology: Predictive Modelling and Mapping Visualisation

3.1 Data Segmentation

We decided to split the data into training and testing datasets for our model so that the model can be trained on historical patterns to predict future win percentages.

- **Training Data:** Nine seasons of historical data (2015-2023) were used to train the model.
- **Testing Data:** The 2024 season was used for testing the model's predictive accuracy by comparing it to actual data, as the 2025 data was incomplete at the time of analysis.
- **Implementation:** The most recent season (2024) in the dataset was first identified. Logical indices were then created: trainIdx selects all seasons prior to 2024 for model training, while testIdx isolates the 2024 season exclusively for testing. The training and testing datasets were subsequently constructed using these indices.

```
%% Predict the most recent season (2024–2025)
latestSeason = max(combinedTable.Season);
trainIdx = combinedTable.Season < latestSeason;   % 2015–2023 for training
testIdx = combinedTable.Season == latestSeason;   % 2024 for testing

% Training data
X_train = combinedTable{trainIdx, selectedMetrics};  % train features
y_train = combinedTable.W(trainIdx) ./ (combinedTable.W(trainIdx) + combinedTable.L(trainIdx));  % train target

% Testing data
X_test  = combinedTable{testIdx, selectedMetrics};   % test features
y_test  = combinedTable.W(testIdx) ./ (combinedTable.W(testIdx) + combinedTable.L(testIdx));   % test target
teams_test = combinedTable.Team(testIdx);           % team names for testing
```

3.2 Variable Selection

The input variables for the model were the top five metrics identified in section 2.2 and the output variable was the team's actual win percentage.

3.3 Linear Regression Equation

The relationship between the input metrics and the win percentage is expressed by the following linear equation:

$$y = \beta_0 + \beta_1 W + \beta_2 L + \beta_3 MOV + \beta_4 NRtg + \beta_5 SRS$$

where:

- y = Predicted win percentage
- β_0 = Intercept
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ = Coefficients (weights) for each respective metric

3.4 Coefficient Estimation and Interpretation

The linear regression equation is first implemented in matrix form as:

$$d = Gm$$

where:

- d is the vector of the observed win percentages.
- G is the design matrix containing a column of ones (for the intercept) and the values of the five selected metrics.

$$G = \begin{bmatrix} 1 & W_1 & L_1 & MOV_1 & NRtg_1 & SRS_1 \\ 1 & W_2 & L_2 & MOV_2 & NRtg_2 & SRS_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & W_n & L_n & MOV_n & NRtg_n & SRS_n \end{bmatrix}$$

- m is the vector of the coefficients $[\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5]^T$.

The model coefficients were then estimated using the least squares method, which minimises the sum of squared differences between the observed and predicted values. The formula is given by:

$$m = (G^T G)^{-1} G^T d$$

```
%% Train the regression model on past data
X_design_train = [ones(size(X_train, 1), 1), X_train];
beta = inv(X_design_train' * X_design_train) * X_design_train' * y_train;
```

The resulting coefficients were:

Metric	**Coefficient (β)**
Intercept (β_0)	0.45909
W (β_1)	0.00651
L (β_2)	-0.00549
MOV (β_3)	0.01462
NRtg (β_4)	-0.01822
SRS (β_5)	0.00529

Interpretation of Coefficients:

- **Positive Coefficients ($\beta_1, \beta_3, \beta_5$):** An increase in Wins, Margin of Victory, or SRS leads to an increase in the predicted win percentage.
- **Negative Coefficients (β_2, β_4):** An increase in Losses decreases the predicted win percentage, as expected. The negative coefficient for Net Rating suggests a complex

interaction with the other metrics in the model, particularly MOV and SRS, which may capture similar aspects of team strength.

- The strongest predictor was MOV, indicating that teams which consistently win by larger margins are statistically more likely to achieve a higher season win percentage.

3.5 Model Performance and Validation

The model's performance was evaluated by comparing its predictions for the 2024 season against the actual win percentages.

Prediction Implementation: To generate win percentage predictions for the test data, the following code was executed:

```
%% Predict Win% for the latest season
X_design_test = [ones(size(X_test, 1), 1), X_test];
predicted_WinPct_test = X_design_test * beta;
```

The design matrix X_design_test was first constructed by prepending a column of ones to the test features X_test, accounting for the model intercept. The predicted win percentages predicted_WinPct_test were then computed through matrix multiplication with the previously derived coefficient vector beta.

- **R-squared (R^2) value:** The model achieved an exceptional R^2 value of 0.9983 on the test data.
    - R^2 was manually calculated using the following code:

```
%% Compute model accuracy for the latest season
R2_2024 = 1 - sum((y_test - predicted_WinPct_test).^2) / sum((y_test - mean(y_test)).^2);
fprintf('Model 1 R² (on 2024–2025 season): %.4f\n', R2_2024);
```

- This near-perfect score indicates that the linear regression model, using the five selected metrics, explains 99.83% of the variance in team win percentages. It demonstrates a very strong fit and high predictive accuracy for the 2024 season.

A sample of the predicted versus actual win percentages for select teams in the 2024 season is shown below:

```
=== Predicted vs Actual Win% for 2024–2025 Season ===
                Team                 Actual_WinPct    Predicted_WinPct
    ____________________________    _____________    ________________

    {'ATLANTA HAWKS'          }        0.4878           0.48495
    {'BOSTON CELTICS'         }        0.7439           0.74477
    {'BROOKLYN NETS'          }       0.31707            0.3132
    {'CHARLOTTE HORNETS'      }       0.23171           0.22504
    {'CHICAGO BULLS'          }       0.47561           0.47173
    {'CLEVELAND CAVALIERS'    }       0.78049           0.78984
    {'DALLAS MAVERICKS'       }       0.47561            0.4773
    {'DENVER NUGGETS'         }       0.60976           0.61569
    {'DETROIT PISTONS'        }       0.53659            0.5392
    {'GOLDEN STATE WARRIORS'  }       0.58537           0.59184
    {'HOUSTON ROCKETS'        }       0.63415           0.64311
    {'INDIANA PACERS'         }       0.60976           0.61027
    {'LA CLIPPERS'            }       0.60976           0.61516
    {'LOS ANGELES LAKERS'     }       0.60976            0.6125
    {'MEMPHIS GRIZZLIES'      }       0.58537           0.59551
```

The close alignment between the actual and predicted values across all teams confirms the model's robustness and effectiveness in forecasting team performance based on the chosen metrics.

3.6 Mapping

After calculating the win probabilities of every team, our map will show markers representing each team's home arena, and these markers will be colour-coded based on their chance of winning.

We first used axesm to set the map projection as Mercator, and geoshow to plot the base map of the US. We then added .mat files of US state borders using geoshow to the same figure. Canadian province borders are needed as the Toronto Raptors are based in Toronto. Thus, we downloaded a shapefile of Canadian province borders (*Awips Map Database Catalog - Canada Province*, 2017). Then, we used textread to load a .csv file containing the locations of each NBA team's home stadium.

```
%% 3) Plot with color-coded markers based on win probability
% Create colormap: red (low) -> yellow (medium) -> green (high)
cmap = [linspace(0.8,0,100)', linspace(0,1,100)', zeros(100,1)];
colormap(cmap);
```

Figure 1: Setting colours for colour map

For the colour coded markers, we set the range of colours from red to green: the highest win probability reflects a bright green marker while the lowest one reflects a bright red marker. Firstly, we created a 100x3 matrix, cmap, where each row is an RGB triple in the form [R G B]. linspace(0.8, 0, 100)' is used to create a column vector of 100 values under the red column, evenly spaced from 0.8 (for bright red) to 0 (to least amount of red). Similarly, linspace(0, 1, 100)' creates a column vector under the green column, with 100 values evenly spaced from 0 (least amount of green) to 1 (bright green). zeros(100, 1) makes the values in the blue column always 0, as there is no blue component in the colours. As a result, this will form a smooth colour gradient from bright red [0.8, 0, 0] in row 1, a mix of red and green [0.4, 0.5, 0] in the middle rows, and bright green [0, 1, 0] as the highest value.

```
for v = 1:length(team_abbrevs)
    % Determine color based on win probability
    color_idx = round(predicted_WinPct_test(v) * 99) + 1;
    marker_color = cmap(color_idx, :);

    % Plot points
    plotm(lat(v), lon(v), 'o', 'MarkerSize', 10, ...
        'MarkerFaceColor', marker_color, ...
        'MarkerEdgeColor', 'k', 'LineWidth', 0.5);
```

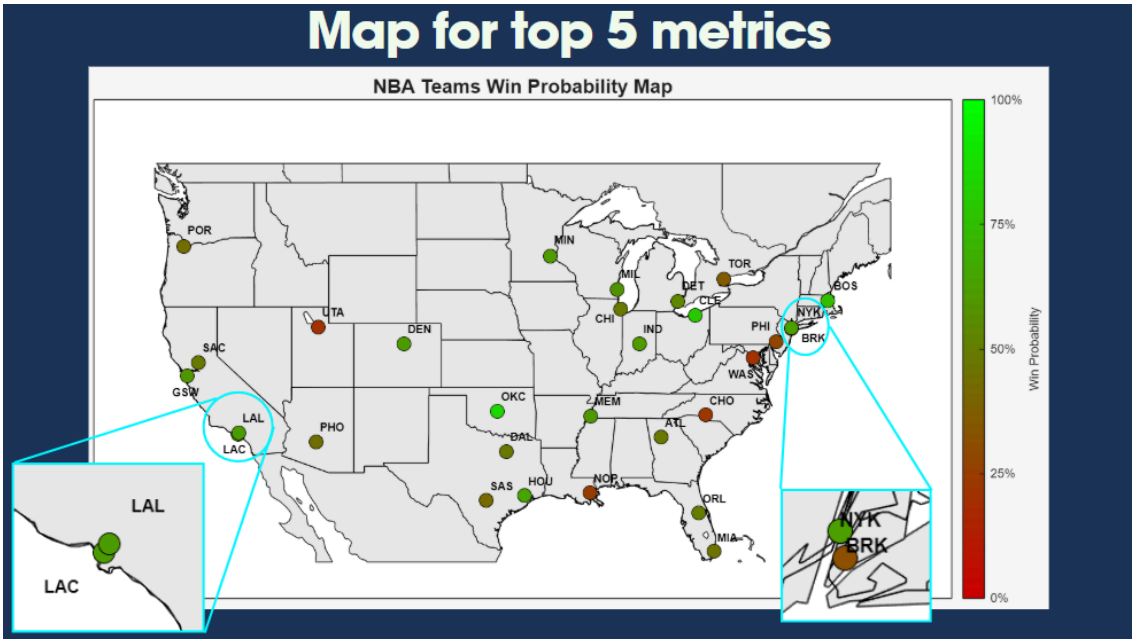
Figure 2: Determine colour of markers

We then assign a color to each marker based on the team's win probability (Fig. 2). Firstly, as the predicted win probabilities range from 0 to 1, we multiply it by 99 to map it to [0 99]. We then used the round function to make sure that every index is an integer. Since the indices in cmap are from 1 to 100, we added 1 to shift the values from 0-99 to 1-100. We plotted the markers using plotm.

Since the text labels would cover the markers if we used the markers' latitudes and longitudes to plot the location of the labels, we offset every label's latitude and longitude by 0.5 and 0.3 respectively. However, after plotting the map, some labels were obstructed by the borders, so we manually shifted those labels by a larger amount. Afterwards, we added a colour bar to our map to show the colours at each probability.

As both Los Angeles and New York City had more than one team representing the city, causing the markers in those locations to overlap, we created zoomed-in maps for both

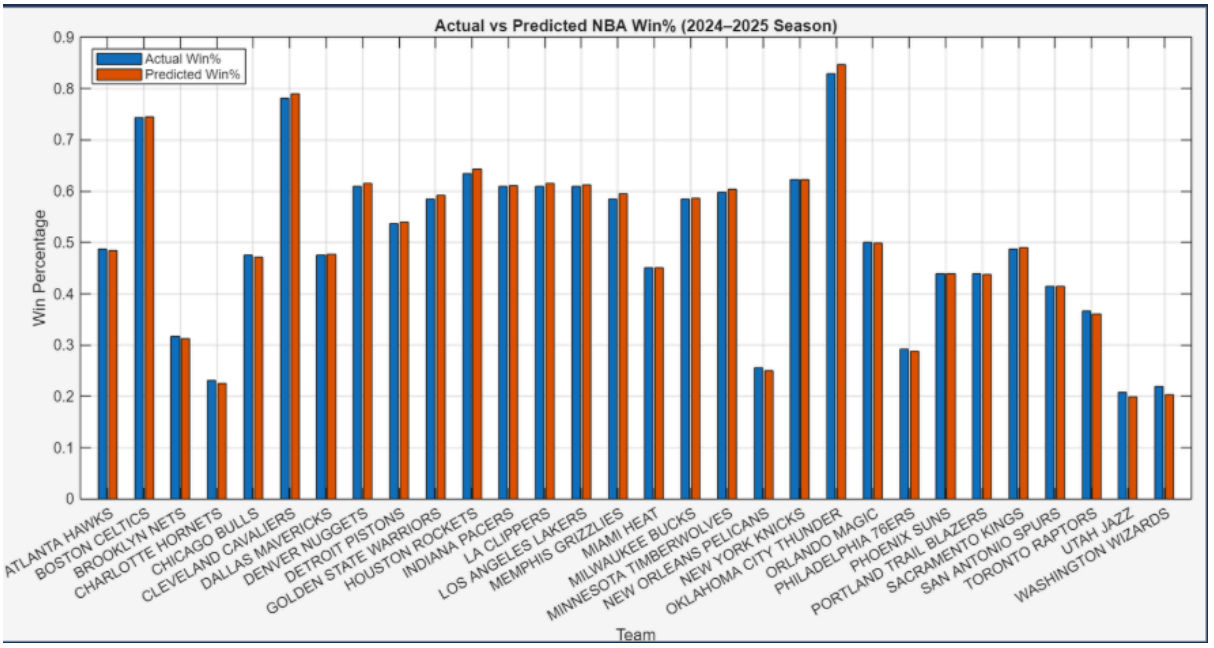
locations where the markers could be clearly seen for each team. Since we created 3 models, we repeated the above mapping steps for every model.



4. Results Discussion

4.1 Model 1

From our map, Oklahoma City Thunder (OKC), has the highest winning probability, followed by, Cleveland Cavaliers (CLE), and Boston Celtics (BOS). However, since OKC had the highest win amount and Net Rating, it was obvious that they would be the champion.



4.2 Surprise Finding in Model 2

Once we have answered our research question in model 1, we felt that it was too straightforward and uninsightful because it is obvious that the metrics that are most directly related to winning are going to have a high correlation to win percentage and a very high R^2. We wanted to challenge ourselves by finding a more interesting result. Hence we chose to explore more diverse metrics in our 2nd prediction model as shown in the figure below.

🏆 Metrics ranked by predictive strength:

Metric	R2	Correlation
{'WIN_PCT'}	1	1
{'W'}	0.96877	0.98426
{'L'}	0.9558	-0.97765
{'MOV'}	0.93082	0.96479
{'NRtg'}	0.93032	0.96453
{'SRS'}	0.92834	0.9635
{'PW'}	0.90508	0.95136
{'PlusMinus'}	0.89168	0.94429
{'PL'}	0.88996	-0.94338
{'ORtg'}	0.38993	0.62444
{'ThreeP_PCT'}	0.37232	0.61018
{'TS'}	0.36294	0.60244
{'FG_PCT'}	0.35228	0.59353
{'OffeFG'}	0.33761	0.58104
{'BLKA'}	0.2985	-0.54635
{'Age'}	0.27454	0.52397
{'DRtg'}	0.27017	-0.51978

Figure showing top 17 metrics

We chose:

#10 ranked metric: offensive rating for a holistic measure of offensive performance,

#11 ranked metric: 3 point percentage to account for shooting ability

#15 ranked metric: Percentage of team's blocked field goal attempts to account for offensive physicality

#16 age

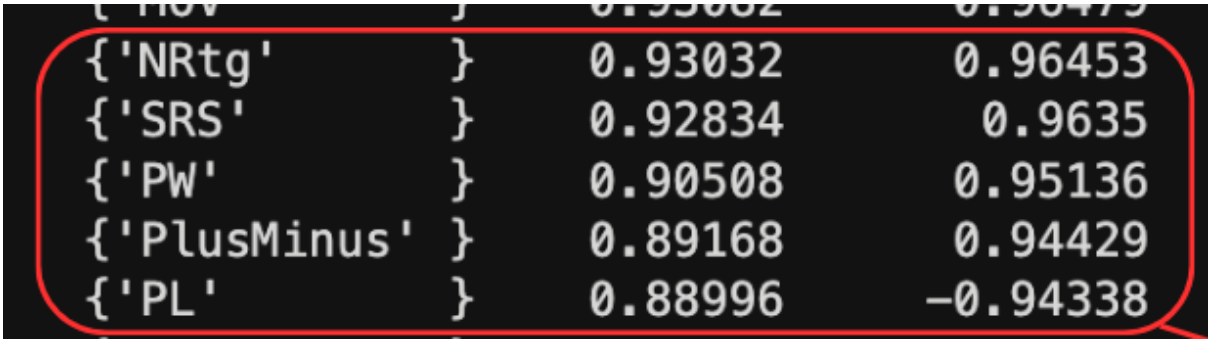
#17: defensive rating for a holistic account for defensive performance

The results for model 2 were surprisingly accurate as well, as it had predicted the Oklahoma City Thunder to win as well for the 2024-2025 season as seen in the figure below. Not only that but its R2 value is also very high at 0.9559. Meaning it can be used to explain our data very well.

```
Model 2 R² (on 2024–2025 season): 0.9559
Mean Absolute Error (on 2024–2025 season): 0.0262

🏆 Predicted NBA Champion (2024–2025): OKLAHOMA CITY THUNDER
```

While doing some exploration, we decided to build model 3, this time picking metrics from 4–9th place shown in the figure below.



Unsurprisingly, OKC were once again the predicted champions for the 2024-2025 season, but the real surprise came in its R2 value shown in the figure below

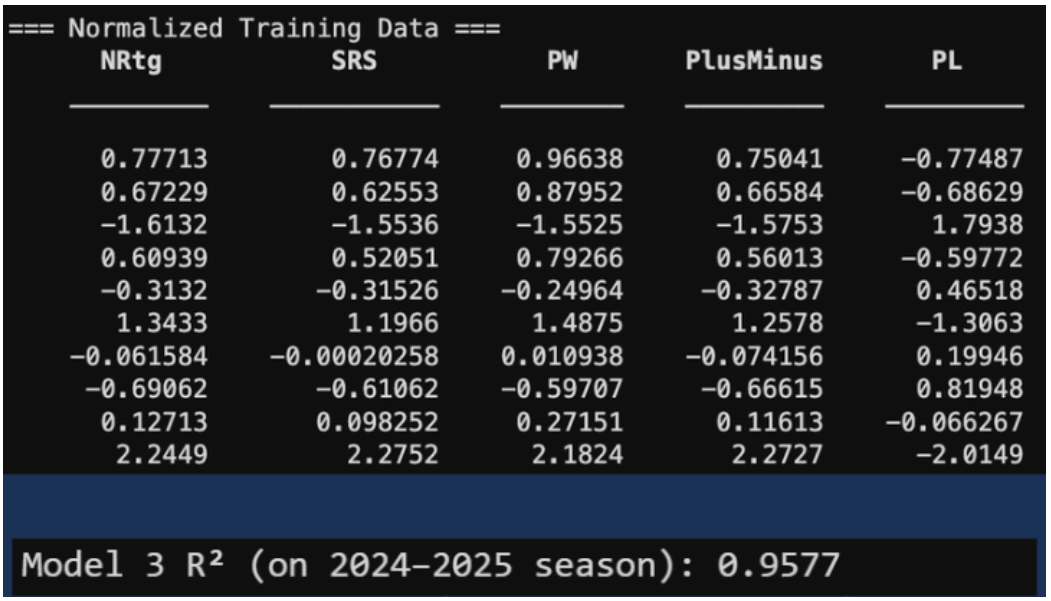
```
Model 3 R² (on 2024–2025 season): 0.9577
Mean Absolute Error (on 2024–2025 season): 0.0266
```

We are surprised because looking at the figure with top 17 metrics, we see that the metrics in model 2 all have an absolute correlation coefficient of less than 0.63, while that of model 3

all have very high correlation coefficients of above 0.90, and yet, both models manage to produce very close R2 values within the range of 0.002.

4.3 Reasons for Surprise

Our first hypothesis for the surprise in model 2 was that the values in the datasets we dealt with were not normalised, which made the weightage distribution unfair since metrics that are calculated with higher absolute values will skew the weightage more in its favour. Hence, we decided to test this hypothesis out by normalising our data and then running the new normalised model 3. The R2 value we got was still the exact same as shown in the figure below hence we concluded that this is not the reason for the surprise.



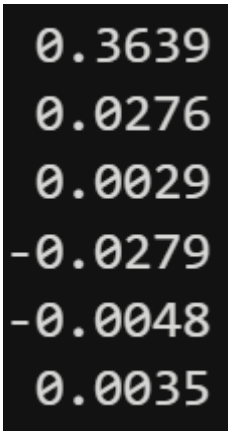
=== Normalized Training Data ===

NRtg	SRS	PW	PlusMinus	PL
0.77713	0.76774	0.96638	0.75041	-0.77487
0.67229	0.62553	0.87952	0.66584	-0.68629
-1.6132	-1.5536	-1.5525	-1.5753	1.7938
0.60939	0.52051	0.79266	0.56013	-0.59772
-0.3132	-0.31526	-0.24964	-0.32787	0.46518
1.3433	1.1966	1.4875	1.2578	-1.3063
-0.061584	-0.00020258	0.010938	-0.074156	0.19946
-0.69062	-0.61062	-0.59707	-0.66615	0.81948
0.12713	0.098252	0.27151	0.11613	-0.066267
2.2449	2.2752	2.1824	2.2727	-2.0149

Model 3 R^2 (on 2024–2025 season): 0.9577

The next possible reason is that model 3 contains redundant information, just like model 1, which is true because net rating and plus minus used in model 1 are essentially the same thing, just that net rating is adjusted for the team's pace of play, but it still conveys how much the team is able to outscore their opponents. Metrics that we have selected for model 2 on the other hand is designed to be diverse and capture different aspects of an NBA game, making it more holistic and hence, even though each metric in model 2 has a smaller correlation coefficient than that of model 3, it is able to generate a very high R2 value as it captures a wider, more representative aspect of the game.

The final reason we think that led to the surprise is hidden in the weightage of each metric in model 2 that we see in the figure below.



These weightages are assigned in order from top to bottom: constant in linear equation, offensive rating, 3 point percentage, defensive rating, block attempts and age. We can see that 2 metrics clearly hold much more significant weightages than the other 3: offensive and defensive rating. The weightages for these 2 metrics is no coincidence because they are related to one of the top 5 metrics, net rating, by the formula: Net Rating=Offensive Rating−Defensive Rating. Net rating is a metric that has a very high correlation coefficient to winning percentage, at ~0.965. Therefore, we think that even though the individual metrics in model 2 have relatively small coefficient correlation, by assigning a very high weightage to offensive and defensive rating and much smaller weightages to other metrics, it is almost as if the outcome of the model is largely determined by net rating, rendering other metrics insignificant in comparison. Thus, this explains why the R2 value of our model 2 is so high as net rating is very highly correlated with winning.

5. Limitations and Solutions

5.1 Metrics chosen are not accurate

Our metrics chosen may not be a good proxy to estimate win rate, as we only gathered the final stats summary for each team per season, which is an average of the stats for every game the team has played in the season. Thus, using the final summary may not take into account some anomalous data in certain games. To improve our model, we could consider team performance at different points during a season.

5.2 Changing league dynamics

We also failed to take into account changing league dynamics, which means that the stats for the more elite teams in 2015 and 2024 could be very different. Hence, this could have a bias

on our model. An improvement would be to apply time-decay weighting when calculating the weightages of each metric.

6. Conclusion

By identifying the metrics most strongly correlated with win rate, assigning appropriate weights, and developing a model to calculate championship probabilities, we were able to compute and rank each team's likelihood of winning the title.

An important insight from this project is variables we overlook or consider insignificant can have a much stronger influence on team success than anticipated. This demonstrated that data analysis can also help us uncover unexpected relationships between metrics that challenge our assumptions.

Although our model offers a structured approach to predicting success, it is limited by the regular season scope, assumptions adopted, and the unpredictability of sports. With more time and richer datasets, the project can be expanded in many creative directions. For instance, we could investigate light-hearted but intriguing ideas such as whether teams with more bald players win more games or whether higher salaries translate to higher success. Ultimately, this project reinforced that the possibilities in sports analytics are endless. With the combination of statistical tools and an open mind, even simple datasets can provide meaningful insights.

7. References

2024-25 NBA season summary. (n.d.). Basketball-Reference.Com. Retrieved 7 November 2025, from https://www.basketball-reference.com/leagues/NBA_2025.html

Awips map database catalog—Canada province. (2017, April 26). https://web.archive.org/web/20170426004503/http://www.nws.noaa.gov/geodata/catalog/national/html/province.htm

Elle-Blogs. (2025, April 8). *Understanding the nba playoffs: A beginner's guide*. Bout Time Pub & Grub. https://www.bouttimepub.com/understanding-the-nba-playoffs-a-beginners-guide/

National Basketball Association. (2024). *Teams – traditional statistics (2024-25 season)*. NBA.com. https://www.nba.com/stats/teams/traditional?DateFrom=&DateTo=&PerMode=PerGame&Season=2024-25&dir=A&sort=W_PCT