Proceedings of the 3rd International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/34/20240767

Exploration of various predictive models for the National Basketball Association

Will Bao[1,6,†], Hanyin Liang[2,†], Ashley Liu[3,†], Kevin Yang[4,†], Terry Zhang[5,†]

[1]Auburn High School, Auburn, 36830, US
[2]East Lost Angeles College, LA, 91754, US
[3]Great Neck North High School, Great Neck, 11023, US
[4]Trinity School, New York City, 10024, US
[5]Ward Melville High School, East Setauket, NY, 11733, US

[6]willbao2025@gmail.com
[†]All the authors contributed equally to this work and should be considered as co-first authors.

Abstract. This research paper explores the application of various predictive models for the National Basketball Association (NBA), with the aim of predicting game outcomes and margins of victory for the 2015- 16 season. The research focuses on three, distinct models, all created using the statistical analysis software R–the mean score difference model, the weighted mean wins model, and the team strength model. Each model employs different methodologies and data analysis techniques to predict game outcomes and team strengths. Our study found that the team strength model and mean score difference model performed best due to their use of point margins as their primary variable, while the weighted mean wins model was less accurate due to its use of wins and losses. None of the models, however, were able to match the predictions from sports gambling organizations, called the "bookies," showing that much more work is needed, not only to surpass the bookie models but to inch closer from prediction to actuality.

Keywords: NBA, point margin prediction, data analyses, modeling with R

1. Introduction

It's what makes face-painted kids roar, erupts, and teeter on the edge of their seats; it's what realizes and shatters dreams and careers; it's what compels tens of millions of Americans to put their hard-earned money on the line. It's the big question–who's going to win?

Every year, die-hard fans and vigilant analysts seek to predict the answer to that question. Their methods may range from wild guesses from the knowledge they swear to have to more accurate, intentional calculations from crunching numbers. Whatever it may be, the outcome of a sports game has long fascinated this vibrant crowd. Further, the rapid ascendancy of the sports gambling industry has made knowing the future even more sought after–now with real, green bills at stake [1-3].

Through this study, we aim to explore this fascination through basketball, hopefully with the same excitement and vitality its viewers effuse. Given bookie point spreads and other game data from the National Basketball Association (NBA), we developed and explored three distinct models to predict

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/34/20240767

not only who wins or loses, but also the margins of victory between two teams. While our models did not prove to be absolutely perfect, they did demonstrate the multiplicity of approaches that can be taken for this area of research. However, these approaches certainly require further analysis and exploration before they can be considered viable alternatives [4-7].

2. Data Preparation

2.1. Data Selection

We were given a total of 18 different datasets, each containing similar data from around twenty-five seasons since 1993. To simplify our exploration, we focused on the three most recent seasons in the datasets, from 2015 to 2017. We first needed to select the most favorable dataset that provided complete, accurate, and well-structured information. To do this, we created an R script titled "findbestdataset.R," which tested each dataset's contents through a series of passes: naming conventions, number of rows, missing data, and complete cases. With each pass, we eliminated several datasets until we were left with datasets 2 and 18, out of which we selected the former for its slightly better aesthetics.

2.2. Data Cleaning

While a simple goal, a single script was used to clean '2.csv', a complex process. First, a copy of '2.csv' called 'df_filtered1' was created, containing data only for the seasons '2014-15', '2015-16', and '2016-17'. Cleaning involved identifying and correcting inaccurate data. To identify flaws, 'df_filtered1' was compared to '7.csv' via a copy named 'df_filtered2'. Further cleaning addressed differences between 'df_filtered1' and 'df_filtered2', attributing disparities to data inaccuracies.

The process began with cleaning columns, where formatting differences such as in scores were addressed. Then, matching team name formatting between 'df_filtered1' and 'df_filtered2' was undertaken. It was discovered that 'df_filtered1' team names were typically stored as a substring of 'df_filtered2' team names. Thus, an algorithm was made to assign team names from 'df_filtered2' to their corresponding substrings in 'df_filtered1', with exceptions to the algorithm being handled individually.

Now, there was one problem left to fix: reordering the rows and columns of the two datasets to match. The column titles had differences in capitalization, which was dealt with to enable their reordering. Meanwhile, the rows were rectified by dividing games by seasons in chronological order, then by teams in alphabetical order, and finally by date in chronological order.

With all differences in data formatting resolved, only the integrity of the data itself was in question. Cross-checking 'df_filtered1' and 'df_filtered2' yielded two rows of inaccurate data, which, after being corrected, produced an improved version of '2.csv'.

3. Main models

3.1. Model 1: Mean Score Difference

In our first model, we evaluate the outcome of employing a team's mean margins of victory or defeat in predicting a score margin. We calculate a game's score difference by subtracting the first team's score from the second team's (whichever unit is the first is arbitrary). Here, a team's mean score difference (MSD) is considered an indicator of a team's strength, and it is assumed that this strength remains constant throughout a season (or seasons). Since the score difference is already measured in units of points, the actual score difference of a game between these two teams can be estimated by subtracting the mean score difference of one team from the mean score difference of another.

To calculate each team's cumulative MSD, an additional column storing the score difference for each game was created in our dataset (the product of section 1.3). We then constructed a list to record the sum of score differences for games that a specific team played, along with the total number of games the team participated in. These two values are then updated per game for each team with a loop.

DOI: 10.54254/2753-8818/34/20240767

Afterward, the mean score difference corresponding to the first team in each row is calculated (each game has two corresponding rows, with the placement of teams and their related data swapped; the first team in the row will be designated as teamA) by dividing the sum of its score difference by the game count. These means now appear in a data frame column.

Our model searches for both teams' most recently updated MSD from this column. For every game, the model generates two temporary datasets that contain all rows in which the score difference column exists for our target teams (teamA & teamB) before the current game. We then obtain the valid cumulative MSD for the target team thus far by selecting the last value in the team's temporary dataset.

Once the MSDs for both teams are selected, we can predict the score difference for the current game by subtracting teamB's MSD from teamA's. To ensure that each team has a relevant and accurate MSD, the model only generates predictions for the 2016-17 season despite updating the MSDs over the 2015-16 season. Running the previously laid out algorithm for forecasts regarding the 2016-17 season, the standard deviation of our difference from the actual point margins is just 0.747 above that of professional bookie point spreads, 12.550.

Looking to improve the accuracy of our model, we considered home court advantage as another variable. As discussed in the introduction, a team playing on their home court generally holds an advantage over a team playing as a visitor. First, we needed to quantify this "home court" advantage. We found the average number of additional points that a team scores at home by calculating the difference between the sum of all score margins from every home team game compared to the same sum from every away team game and dividing this by the number of games played. This difference proved to be 2.83 points. To account for this, we would increase the host team's MSD as a strength measure and decrease the other team's MSD. For instance, if two teams have comparable MSD, after adding the location points, the point margin between these two teams will increase from 0 to |2.83|. When team1 is at home, we subtract 2.83 from the predicted point margin – as the score difference is obtained by subtracting the first team's score from the second team's, and team 1 is expected to have a higher score than their MSD; when team1 is a visitor team, we add 2.83 to the predicted point margin, since team 2 is predicted to have a higher score than their MSD, which means team1 will lose more points or win fewer points. This lowers the standard deviation of our difference from the actual point margins to 12.93– just 0.391 away from that of the bookie point spreads. The scatterplot displaying each score difference as a function of our predictions follows a linear equation with a slope of 0.836– ideally, the regression would demonstrate a slope of 1.

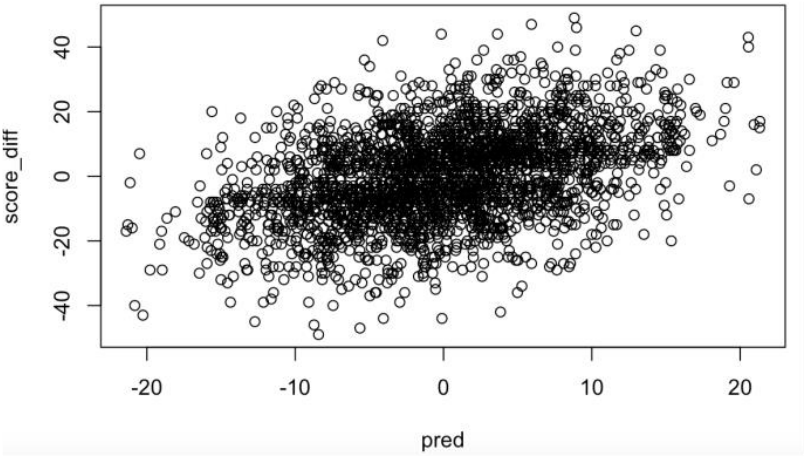

Figure 1. Score differences plotted over our predictions and bookie point spreads.

DOI: 10.54254/2753-8818/34/20240767

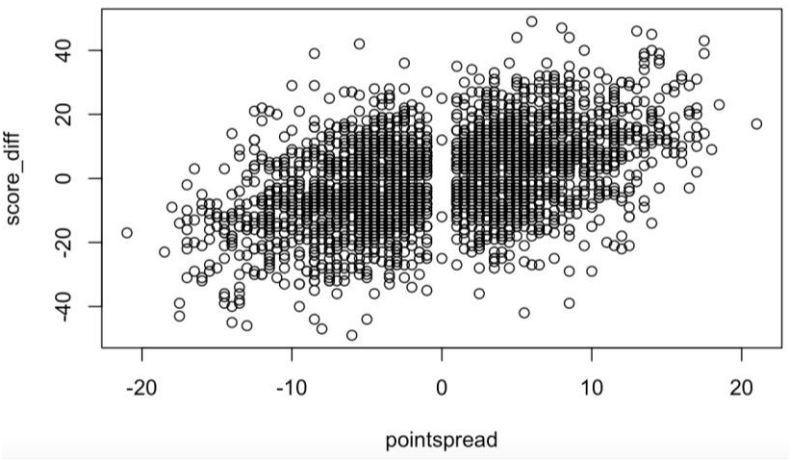

Figure 2. The behavior of our predictions closely resembles bookie predictions.

Although this model is somewhat successful, it is far too simple to be considered reliable. Figure 3 demonstrates that the mean score difference remains static over time, even though a team's skill often fluctuates. Furthermore, although recent games more accurately illustrate a team's strength, the MSD system places equal weight on all games. As the prediction summary in Figure 4 shows, there is a slight variance among predictions compared to the actual score difference summary. In addition, we ran a rough calculation to measure the home court advantage across two seasons, and thus, this value may not be accurate. Therefore, this model can be improved to weigh recent games more or to multiply predictions by a coefficient to expand their range.

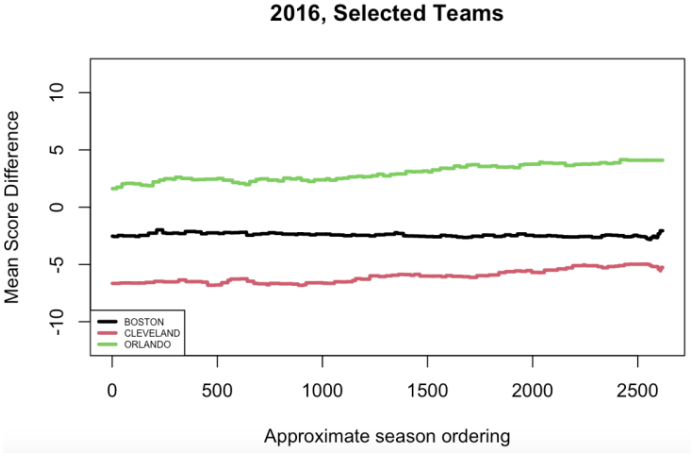

Figure 3. MSDs of Boston, Cleveland, and Orlando plotted over the 2016 season

Table 1. Summaries for true score differences and MSD predictions

Min.	1st Qu.	Median	Mean	3rd Qu	Max
-49	-9	0	0	9	49
-21.392901	-4.817385	0.010188	0.007549	4.835026	21.328667

3.2. Model 2: Weighted Mean Wins Model

The data required further cleaning before implementing models 2 and 2.5, including removing duplicate games. Instead of each game occupying two rows from the perspective of either team, each game is summarized in one row. We also added an individual strength column for every team. This

DOI: 10.54254/2753-8818/34/20240767

strength is recalculated after every game and is unique to each game a team has played. This cleaned dataset only accesses 2015- 16 and 2016- 17 data.

Model 2 ignores scoring altogether and only focuses on each team's wins. To compare the different sections, this model assigns a power rating to each group; the higher the number, the more influential the team. This strength is recalculated after every game and is unique to each game a team has played. This is done for every game chronologically, such that the win-loss ratio (WLR) is updated after every game. For example, if a team wins their first match of the season, their WLR after the game would be 1. If they lose the three following games, their WLR would decrease to .5, .33, and .25 as they lose each game.

Each team's column is filled with wins and losses to achieve this effect. Each row will have a 1 (a win) or a 0 (a loss), with rows where they did not play games set to - 1. Then, we calculate the cumulative mean using R, ignoring the - 1 values. The cumulative mean function adds up all wins and divides the sum by the total number of games played by each team previously to a row, thus giving a win-loss ratio. We can assess the group's overall strength by plotting each team's WLR over time. According to our analysis, the Golden State Warriors were the strongest during the 2015- 16 season, closely followed by Cleveland.

However, this model has its limitations. Teams' performance can change drastically during a season due to various factors. For example, a group that plays well at the beginning of the season can perform worse after an injury. So, in general, a team's more recent wins should affect their strength score more than their wins in the past. A recency weight was introduced in Model 2.5 to accommodate this idea, giving more importance to recent games. This was done by changing the cumulative mean function to apply a weight to each score. A win in the previous model would have counted as 1, regardless of when it occurred. However, a win from a long time ago is estimated as less than 1, reducing its significance. Meanwhile, a victory that occurred more recently will have a more significant influence on the cumulative mean. Initial testing of this model used an x^2 weight and compared results to the bookie scores. Two other variants of this model used an exponential function and a sqrt(2x) function, with results expressed as a standard deviation difference from the point spread.

These models only provide a metric for comparing team strengths, not predicting score differences. Although finding a coefficient that converts our strength measure to a point measure is feasible, the models currently don't incorporate other influential variables like home court advantage or margin of victory. So, while a team's win by 30 points reflects a different level of dominance than a win by 1 point, our model doesn't consider this yet. Nonetheless, by comparing the strengths of two competing teams, this model can reasonably identify the stronger team and quantify the degree of their advantage, thus allowing it to predict which team is more likely to win.

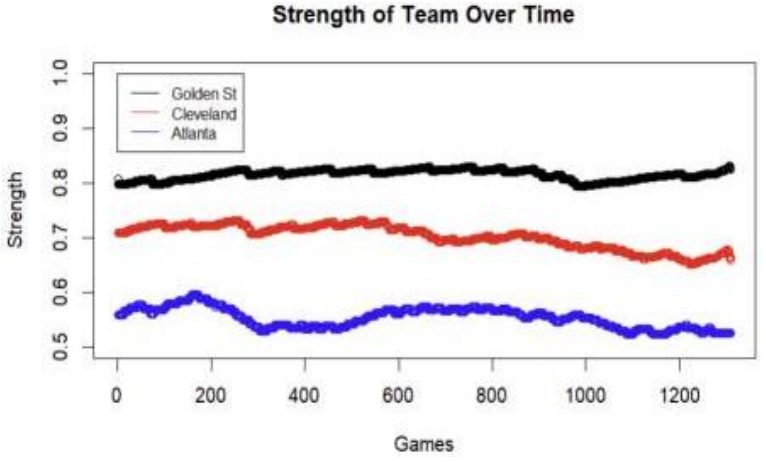

Figure 4. Strength of team over time with linear function.

DOI: 10.54254/2753-8818/34/20240767

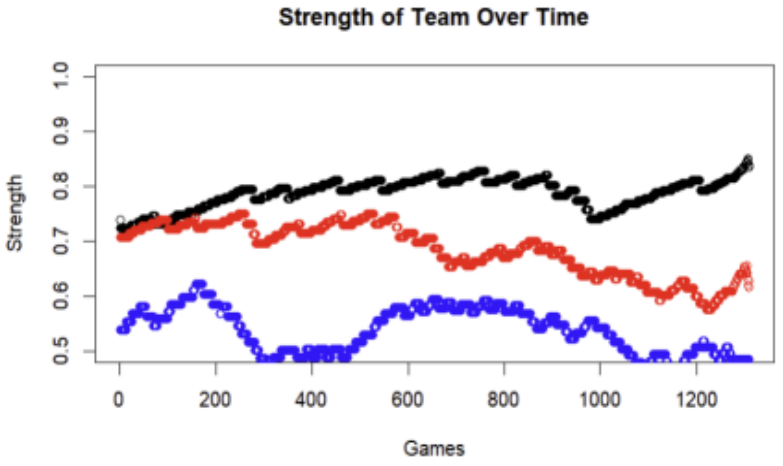

Figure 5. Strength of team over time with x^2function.

3.3. Model 3: Team Strength Model

The team strength model was adapted from Professor Emerson and yielded the best results out of all three approaches. The model revolves around a team strength variable that is updated with every new game, unique for all 30 teams in the league. Thus, the more information available, the more accurate the model becomes. The team strength is in terms of point margins, meaning that the forces of any two teams can be directly compared, and their difference reflects the margin of victory in a hypothetical matchup between them.

3.3.1. Testing measures. To test the efficacy of this model, we measured each iteration's results with a "rdif" value. To find an iteration's "rdif," subtract the standard deviation of the model's predictions from the actual score difference from the standard deviation of the bookie point spreads from the fundamental score difference. The standard deviation indicates the amount of variation, in this case, of the predictions from the real-world outcome. Thus, the lower the standard deviation, the more accurate a model is. Likewise, the less negative the "rdif" score is, the more precise the model.

3.3.2. First iteration. The first iteration of the model is the base model given by Professor Emerson. To begin, three columns are initialized–a "scorediff" column for the actual game score differential provided by the score of team2 minus team1, a "pred" column to store the model's prediction for the score differential, and lastly, a "residual" column, for the difference between the actual "scorediff" and the model's prediction. In addition, a new column, "S_[team name]" is created for each team active in the league, containing that team's strength value.

The model works chronologically, beginning with the first game played in the 2014- 15 season. For any two given teams in fun, the team strength values from their last match are copied down. A prediction is made with these team strengths by subtracting one from another, stored in the aforementioned "pred" column. Then, this prediction is compared to the actual "scorediff" once the hypothetical game is played, and the difference between them is included in the "residual" column. At this point, the team strength values for both teams are updated based on the residual to reflect the game's outcome. If team one wins, for instance, the residual will be added to its team strength and subtracted from team two's team strength value, and vice versa.

This first iteration produced a standard deviation of 13.20241 points compared to the bookies' standard deviation of 12.03156 for all three seasons, yielding a "rdif" score of - 1. 170856. This could certainly be improved.

3.3.3. Second iteration. The second iteration of the model tested for the ideal "residual coefficient," denoted by the variable "cresidual" in the script. The residual coefficient is a weight applied to the

DOI: 10.54254/2753-8818/34/20240767

residual after each game, essentially controlling how significant that game was to a team's strength value. If the coefficient is comparatively high, each game's results matter more and lead to more volatile trends in team strengths; small fluctuations or "noise" in the game data may produce pronounced effects, and more significant factors such as injuries may have greater consequences. If the coefficient is comparatively low, then each game's results matter less, leading to more consistent and steady trends in team strengths; however, while less prone to overfitting, the team strengths may "lag," needing a period to adjust to a team's true power.

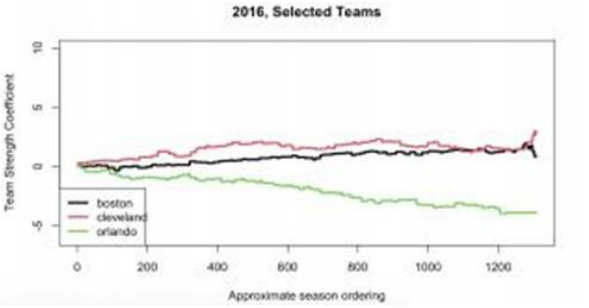

Figure 6. Graphs of three teams and their team strength over time. (low residual coefficient of 0.03)

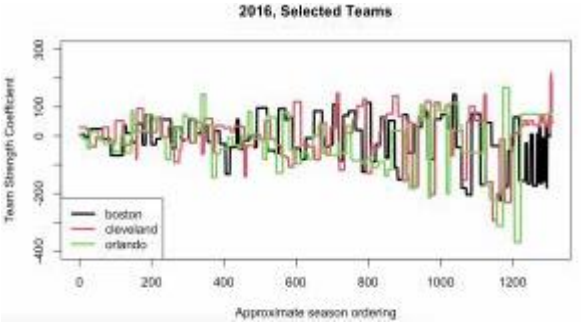

Figure 7. Graphs of three teams and their team strength over time. (high coefficient of 1.0)

The goal for the second iteration of the model was to find the ideal residual coefficient. To do this, we ran a for loop that iterated through one hundred potential coefficients, ranging from 0.01 to 1, and monitored the results. Out of them, a residual coefficient of 0.03 yielded the best results: an improved "rdif" of -0.7636477. The testing for a residual coefficient increased accuracy significantly, but the model thus far was only one-dimensional.

3.3.4. Third iteration. The model's third iteration incorporated each game's location as a factor to account for home-court advantage. As mentioned before, teams playing in their home arena tend to perform better for various reasons, from familiarity with the court to the support from the fans. Thus, the location of a game proved to be a critical factor in our exploration of this model.

First, we needed to quantify this "home court" advantage. To do this, we found the number of additional points a team scores at home.

$$((sum(z\$score1[z\$location == \text{“H”}]) + sum(z\$score2[z\$location = \text{“V”}])) - (sum(z\$score1[z\$location = \text{“V”}]) + sum(z\$score2[z\$location == \text{“H”}]))) / 3936$$

We found that teams, on average, scored 2.82 points more when playing at home. With this number in mind, we then created a function titled "location" that parsed each game and, based on its status as "H," home, or "V," away, updated a variable called "ans" to be either 2.8, or -2.8. This variable "ans" was then added to the prediction, meaning that if team one is at home, then ans is negative, and the strength of the team2 is reduced; likewise, if team two is at home, then ans is positive, and the power

DOI: 10.54254/2753-8818/34/20240767

of team 2 is increased. Thus, the game's location affects the model's prediction by altering team strengths, helping the home team, and hurting the away team.

The third iteration of the model produced the best results we have seen so far. The standard deviation of our model from the actual score differential was only 12.46316, yielding a "rdif" value of just -0.4316024, under half a point away from the bookie's standard deviation of 12.03156.

3.3.5. Potential Model Improvements. The team strength model could be improved in a couple of ways, though one is most hopeful. Aside from its location, the model treats every single game the same; they all affect the team's strength value identically. However, this does not accurately reflect the significance of all the games in the season. For instance, more recent games show the current strength of a team going into an upcoming matchup, while more distant games may not be the most optimal indicators. A team may be on a hot streak, winning more games than usual, or in a dangerous slump, where they keep losing games they should win. For this reason, a "time weight" could further improve results from this model, perhaps weighing recent games more on some sort of curve, as Model 2 does. This could be pretty difficult, though, as the model currently evaluates each game as it is played, not knowing how many more will come.

4. Conclusion

After our study, we are left with one final question–Which model was the best? The true answer is that we don't know; each model presents its unique strengths, limitations, and potential. While some models performed better than others in our tests, note that they all are relatively basic models, and some methods may have more significant potential for implementation in a more advanced or accurate system. The mean score difference model and team strength model produced great results using point margins and location but could not incorporate a time-weighted curve for the significance of certain games in a season. Thus, they can be used to predict the exact difference in scores. The weighted mean wins model, on the other hand, included a "recency weight" that valued recent games more than older ones. This allowed the model to reflect better the current trends of a team's performance, which is more relevant to upcoming games than patterns from the beginning of the season. However, by operating in units of weighted wins instead of points, this model proves more challenging to adapt in terms of score differentials. Nonetheless, it can be used as a solid indication of the point margin of a game. Perhaps if we could combine aspects of each of the three original approaches into a singular model, our results could improve even further. Thus, moving forward, we encourage further exploration and analysis of the models–they are good but far from perfect. It is also important to note that the models were created based on the data available, namely the end scores of a game and its location. The models may be improved further if more data is used, such as individual player statistics or more advanced statistical metrics.

Acknowledgement

All group members contributed to the writing, editing, and reviewing of the final research paper. We extend our gratitude to Professor Emerson, Michael, Amy, and John for their guidance.

References

[1] Smith J T, Sasina N V, Kryshev A I, et al. A review and test of predictive models for the bioaccumulation of radiostrontium in fish[J]. Journal of Environmental Radioactivity, 2009, 100 (11): 950-954.DOI: 10.1016/j.jenvrad.2009.07.005.

[2] Zou K H, O'Malley A J, Mauri L. Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models[J]. Circulation, 2007, 115(5):654-657. DOI:10.1161/CIRCULATIONAHA.105.594929.

[3] Zhao W, Chow D, Yan H, et al. Influential factors and predictive models of indoor clothing insulation of rural residents: A case study in China's cold climate zone[J]. Building and environment, 2022(May): 216. DOI: 10.1016/j.buildenv.2022.109014.

DOI: 10.54254/2753-8818/34/20240767

[4] Ezzati A, Davatzikos C, Wolk D A, et al. Application of predictive models in boosting power of Alzheimer's disease clinical trials: A post hoc analysis of phase 3 solanezumab trials [J]. Alzheimer's & Dementia: Translational Research & Clinical Interventions, 2022, 8. DOI: 10.1002/trc2.12223.

[5] Rajasulochana P, Gummadi S N. A probiotic based product using multi-strain Bacillus species and predictive models for shrimp growth following probiotic intervention [J]. Aquaculture, 2022, 551:737869-. DOI: 10.1016/j.aquaculture.2021.737869.

[6] Morcos F. Evolutionary landscapes derived from amino acid sequence space serve as predictive models of functional phenotypes in protein design[J]. Biophysical Journal, 2022, 121(3):46a-.

[7] A Y M Q, B L P M, A E B. Predictive models of drought tolerance indices based on physiological, morphological and biochemical markers for the selection of cotton (Gossypium hirsutum L.) varieties [J]. Journal of Integrative Agriculture, 2022, 21(5):1310-1320.

Appendix

(1) Margin of victory – the margin of basketball victory is the difference in points between the winning and losing teams at the end of a game. We refer to a team's cumulative margin of victory or defeat as their mean score difference (MSD). Section 2.1 covers the relationship of mean score differences between teams and a game's margin of victory. Section 2.3 uses a team strength rating system based on score differences for each game.

(2) Home court advantage – Home court advantage refers to a team's perceived benefit or advantage when playing on their home court or field, often attributed to familiarity, crowd support, and reduced travel. The models in Section 2.1 and Section 2.3 both attempts to factor in this metric.

(3) Win-loss record – win-loss record is a summary of a team's or individual's performance in a competition, indicating the number of games or matches won and lost. Section 2.2 investigates a strength modeling method involving a team's win-loss record, where recent wins are weighed more heavily.