Analysis of Elo and Raptor Ratings on NBA Game Outcomes

Introduction

The dataset selected for this project is from the FiveThirtyEight (2022-23 NBA Predictions) data repository, specifically the "nba_elo_latest.csv" dataset. The prediction problem identified is a binary classification task aimed at predicting whether the home team will win a game based on various pre-game statistics. This dataset includes features such as team elo ratings, Raptor ratings, and game-specific statistics. The goal is to build two machine learning models and a baseline model to predict the outcome of NBA games and compare their performance.

Data Engineering

Data Cleaning, Transforming, and Normalizing

Dropping Irrelevant Columns: ('carm-elo1_pre', 'carm-elo2_pre', 'carm-elo_prob1', 'carm-elo_prob2', 'carm-elo1_post', 'carm-elo2_post') were dropped as they are empty.

Handling Missing Values: The 'playoff' column had missing values which were filled with 0, assuming that missing values imply non-playoff games.

Feature Engineering:

- elo_diff: Difference between the home team's pre-game elo rating ('elo1_pre') and the away team's pre-game elo rating ('elo2_pre').
- rap_diff: Difference between the home team's pre-game Raptor rating ('raptor1_pre') and the away team's pre-game Raptor rating ('raptor2_pre').
- home_win: Binary target variable indicating whether the home team won the game.

Normalization: Features were normalized using StandardScaler to ensure uniformity in scale.

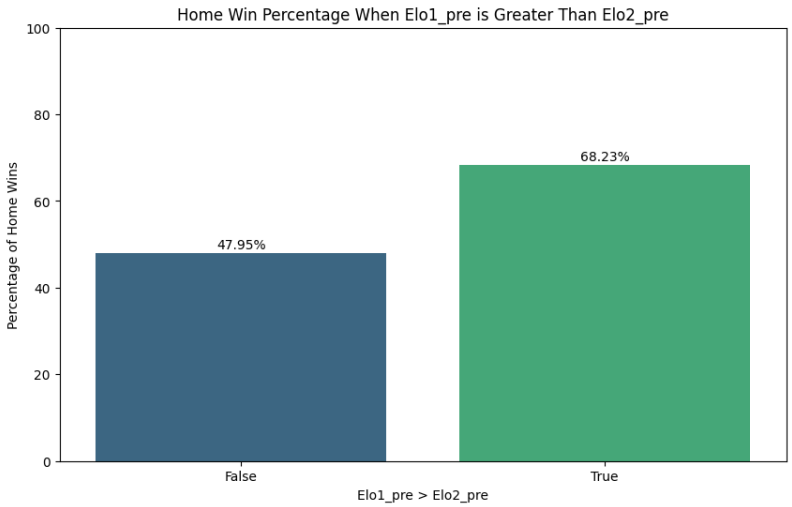


Figure 1 Home Win Percentage When Elo1_pre is Greater Than Elo2_pre

In Figure 1 illustrates the win percentage for home teams based on whether their pre-game elo rating (elo1_pre) is greater than the away team's pre-game elo rating (elo2_pre). This analysis aims to investigate the relationship between the teams' elo ratings and the likelihood of the home team winning the game.

- When elo1_pre <= elo2_pre: The win percentage for home teams is 47.95%. This means that when the home team's pre-game elo rating is less than or equal to the away team's pre-game elo rating, the home team wins approximately 47.95% of the time.
- When elo1_pre > elo2_pre: The win percentage for home teams increases significantly to 68.23%. This indicates that when the home team's pre-game elo rating is higher than the away team's pre-game elo rating, the home team wins approximately 68.23% of the time.

This finding suggests that elo1_pre and elo2_pre are important features for predicting the outcome of NBA games, as higher pre-game elo ratings for the home team correlate with a higher probability of winning.

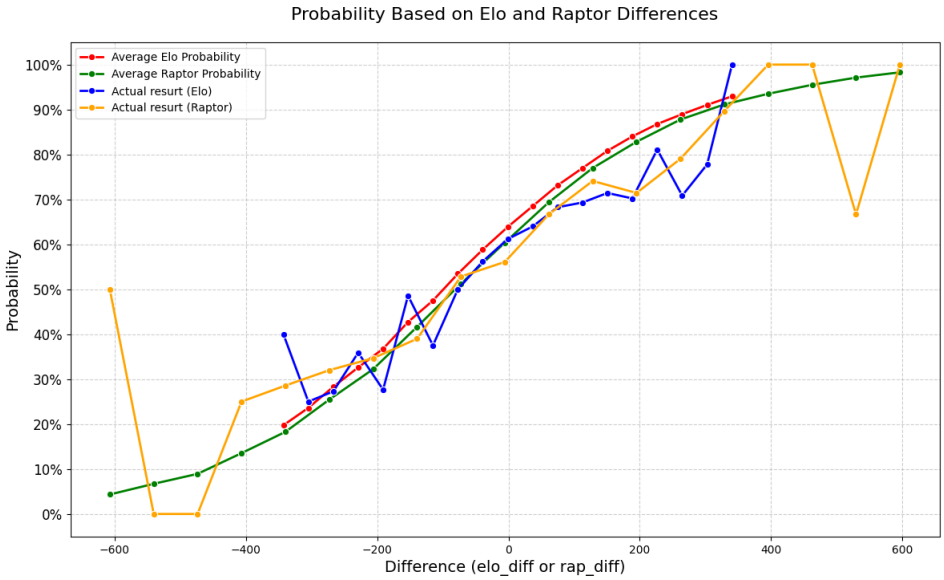


Figure 2: Probability of Home Team Win Based on Elo and Raptor Differences

In Figure 2 visualizes the relationship between the differences in pre-game ratings (Elo and Raptor) and the probability of the home team winning the game. The x-axis represents the difference between the home team's rating and the away team's rating (elo_diff or rap_diff), while the y-axis represents the probability of a home team win.

- Average Elo Probability (Red Line): This line shows the average predicted probability of a home team win based on the Elo rating difference. As the Elo rating difference increases, the probability of the home team winning also increases.
- Actual Result (Elo) (Blue Line): This line plots the actual win percentage of home teams at various Elo rating differences. It represents the empirical win rate corresponding to each rating difference.

- Average Raptor Probability (Green Line): This line displays the average predicted probability of a home team win based on the Raptor rating difference. Similar to the Elo probability, the home team's win probability rises with an increasing Raptor rating difference.
- Actual Result (Raptor) (Yellow Line): This line shows the actual win percentage of home teams at different Raptor rating differences. It reflects the real-world outcomes relative to the Raptor rating differences.

The close alignment of the actual results with the predicted probabilities indicates that both Elo and Raptor ratings are effective predictors of game outcomes. The smoother average probability lines (red and green) highlight the expected trend more clearly, while the actual results (blue and yellow) provide empirical validation.

Model Selection

Chosen Models and Justification

Logistic Regression: Selected for its simplicity and interpretability in binary classification tasks. It provides probabilistic outcomes and is computationally efficient. Logistic Regression is a good starting point for binary classification problems and serves as a benchmark.

Random Forest Classifier: Random Forests are well-suited for classification problems where the relationship between the features and the target variable is complex and non-linear.

Baseline Model

A simple baseline model was implemented, predicting the majority class observed in the training dataset for every instance. This model provides a reference point to compare the performance of the more sophisticated models.

Feature Selection

Variable Input Selection

Features were selected based on their relevance to the prediction task. **'elo_prob1'** (elo-based win probability for the home team), **'raptor_prob1'** (Raptor-based win probability for the home team). **'elo_diff'**, and **'rap_diff'** had correlations with the game outcome ('home_win').

$$['elo\_diff'] = ['elo1\_pre'] - ['elo2\_pre']$$

$$['rap\_diff'] = ['raptor1\_pre'] - ['raptor2\_pre']$$

$$['home'_{win} = IF(['score1'] > ['score2') ==> True$$

$$= ELSE ==> False$$

Columns related to post-game statistics such as 'elo1_post', 'elo2_post', 'score1', 'score2', 'quality', 'importance', and 'total_rating'. They are not available before the game and thus not useful for prediction.

Cross-Validation and Hyperparameter Tuning

Cross-Validation Method

Stratified K-Fold Cross-Validation was used to maintain the distribution of the target variable across folds. This method ensures each fold is representative of the overall dataset and helps in evaluating the models' performance more reliably.

Hyperparameter Tuning

Logistic Regression: Tuning parameters included regularization strength (C). Grid search was employed to find the optimal values.

Random Forest Classifier: Parameters such as the number of estimators, maximum depth, and minimum samples split were tuned using Grid Search CV. This approach helps in identifying the best hyperparameters for maximizing the model's performance.

Results

Performance Metrics and Comparison

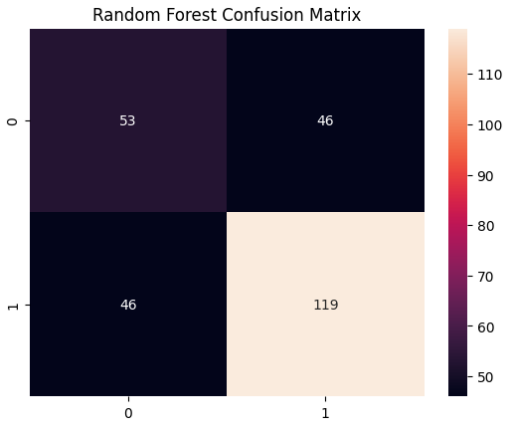
Model	Accuracy	Precision	Recall	F1 Score
Random Forest Classifier	0.6515	0.72	0.72	0.72
Logistic Regression	0.6894	0.74	0.78	0.76
Baseline Model	0.6250			

The Logistic Regression model slightly outperformed the Random Forest Classifier in terms of accuracy, precision, recall, and F1 score. The baseline model's performance was the lowest, underscoring the necessity of more sophisticated models.

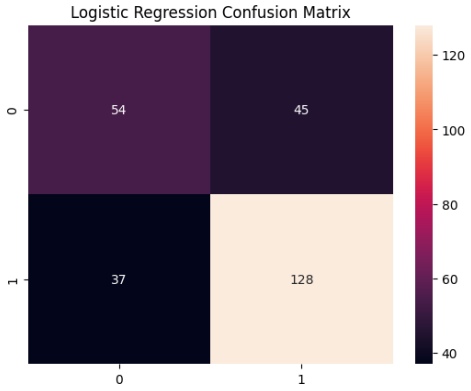
Confusion Matrix and Classification Report

Confusion matrices for both models were generated to visualize the distribution of true positives, true negatives, false positives, and false negatives. The classification reports provided detailed metrics for each class (home win or loss).

Random Forest Confusion Matrix:



Logistic Regression Confusion Matrix:



Logistic Regression had a better balance between precision and recall for predicting home wins.

Conclusion

Advantages:

The use of Logistic Regression provided a simple and interpretable model, which performed well despite its simplicity.

The Random Forest model, although slightly less accurate, offered robust performance and the ability to handle complex interactions between features.

Cross-validation and hyperparameter tuning ensured that the models were well-optimized and generalizable.

Feature engineering significantly improved the models' performance by highlighting the key differences between teams.

Disadvantages:

Logistic Regression, while interpretable, may not capture complex non-linear relationships as effectively as Random Forests.

The Random Forest model, despite its performance, was more computationally intensive and time-consuming to train, especially during hyperparameter tuning.

The baseline model's simplicity highlighted the need for more sophisticated models, but it also underscored the challenge of achieving high accuracy in sports prediction tasks.

Interpretation of Results

The results indicate that for this particular classification task, Logistic Regression provided a slightly better balance between performance metrics compared to the Random Forest Classifier. The baseline model's performance, as expected, was significantly lower, demonstrating the necessity of using advanced machine learning models for predictive accuracy. The pairplot analysis and bar plots provided valuable insights into the relationships between the features and the target variable, further justifying the chosen features for model training.