Insights into the Tennis Court through Machine Learning: Analysis and Evaluation of the Wimbledon Men's Singles Final

Zhen Gao[1, *], Weixi Sun[2]

[1]School of Computer Science, Qufu Normal University, Shandong, China, 276826

[2]School of Cyber Science and Engineering, Qufu Normal University, Shandong, China, 273165

*Corresponding author: gz_06212003@163.com

Abstract. In today's data-driven world of sports science, in-depth analysis of tennis match data is especially critical for improving athlete performance, refining training strategies, and enhancing the viewing experience. This study starts from all aspects of data and key features of the players during the game, using Stacking integrated machine learning methods, a kind of integrated algorithms through XGBoost, GBDT, CatBoost, and ExtraTrees to build prediction models, aiming to always keep an eye on the dynamics of the game, using the data to reveal the changes in the player's strength and to quantify the impact of the momentum on the player as well as the game's trend. The experimental results show that we have identified factors that influence changes in players' momentum, which can provide real-time feedback to coaches and players so that they can make data-based decisions during matches, opening up new avenues for improving tennis players' scoring ability and maximizing help for players to improve their winning percentage in matches.

Keywords: Momentum, Stacking model, Decision tree integration algorithm, Correlation analysis.

1. Introduction

Tennis is a global sport, and in-depth analysis of its game data is important for improving player performance, optimizing training strategies, and enhancing game viewing. Especially in a top tennis tournament like Wimbledon, every match is a comprehensive test of a player's technical, tactical, and mental skills. With advances in data collection technology and big data analytics methods, researchers are now able to meticulously analyze every move, and every point gained or lost in a game to reveal the key factors in victory and defeat. This research endeavors to quantify the impact of potential energy on tennis matches through comprehensive data analysis. Its primary goal is to unravel the causes of fluctuations and explore their predictability. The overarching objective is to construct a model capable of monitoring potential energy variations during the Wimbledon 2023 men's singles matches, facilitating an in-depth analysis of player performance and intensity during critical junctures.

In sports such as tennis, potential energy emerges from the force and momentum generated during events. It's a recognized psychological element influencing player performance, but quantifying and forecasting it within the realm of sports science has posed challenges.

Traditionally, assessing potential energy in sports science relied on qualitative methods. However, recent research has shifted towards statistical and machine learning techniques for quantifying and predicting potential energy and match outcomes. For example, Morris et al. employed statistical modeling to forecast tennis match results, while Smith and Gombert delved into psychological factors. These approaches, though insightful, often lack precise match-time data. We have developed a data-driven model that takes into account factors such as points and serve wins in the 2023 Wimbledon tournament.

In this study, data were obtained from https://www.comap.com/, which collects data from tennis matches, including player scores, court positions, and various statistics. Initially, we analyzed the dataset to understand feature distributions and relationships. Subsequently, missing values were handled, and models were established based on the acquired feature indicators. An ensemble machine learning approach was employed, utilizing XGBoost, GBDT, ExtraTrees, and CatBoost models. These models were integrated using Stacking to reveal relationships among the indicators.

2. Data preprocessing and preparation

2.1. Data Description

Analyzing the provided dataset reveals comprehensive information about tennis matches. Key columns include match ID, player details, match duration, set/game/point numbers, and sets/games won. The dataset also contains detailed points, serves, and returns data, such as serve speed, direction, and return depth, enabling deeper analysis of match strategies, player traits, and pacing.

We conducted data exploration and visualization to understand feature distributions and relationships. This helped uncover tournament characteristics, player performance, and factors influencing tournament outcomes as shown in Figure 1.

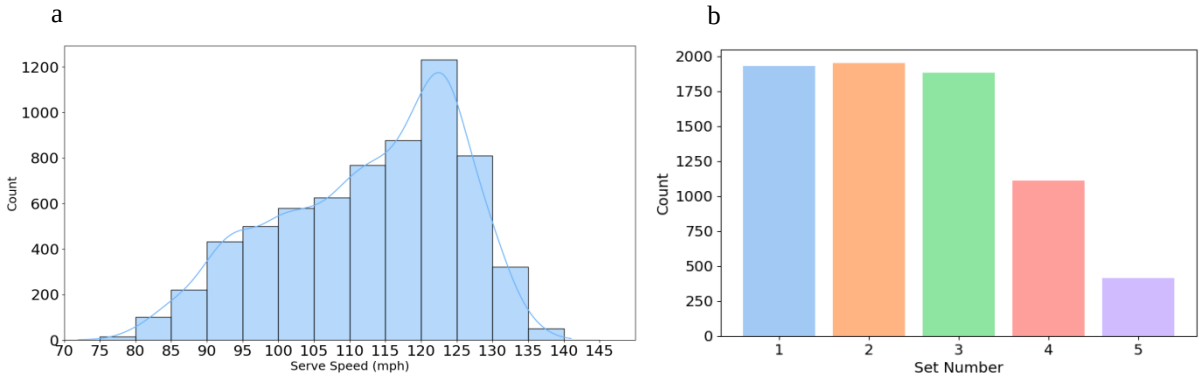

Figure 1. Visualization of key data

2.2. Missing value handling

Handling missing values is critical in the data preprocessing phase to ensure accurate analysis and effective model training. In real-world datasets, missing values are common due to various factors, including data collection errors, information loss, or incomplete records. Failing to address them can lead to biased conclusions and adversely affect machine learning model performance.

Python's pandas library offers valuable tools for addressing missing values. The DataFrame.isnull() function is commonly used to detect missing values, helping assess data completeness. Data visualization techniques, such as heat maps, can visually represent the distribution of missing values, making it easy to identify problematic areas and guide handling strategies. Heat maps use color variations to illustrate the density of missing values, simplifying the process of identifying issues and making informed decisions. As shown in Figure 2.

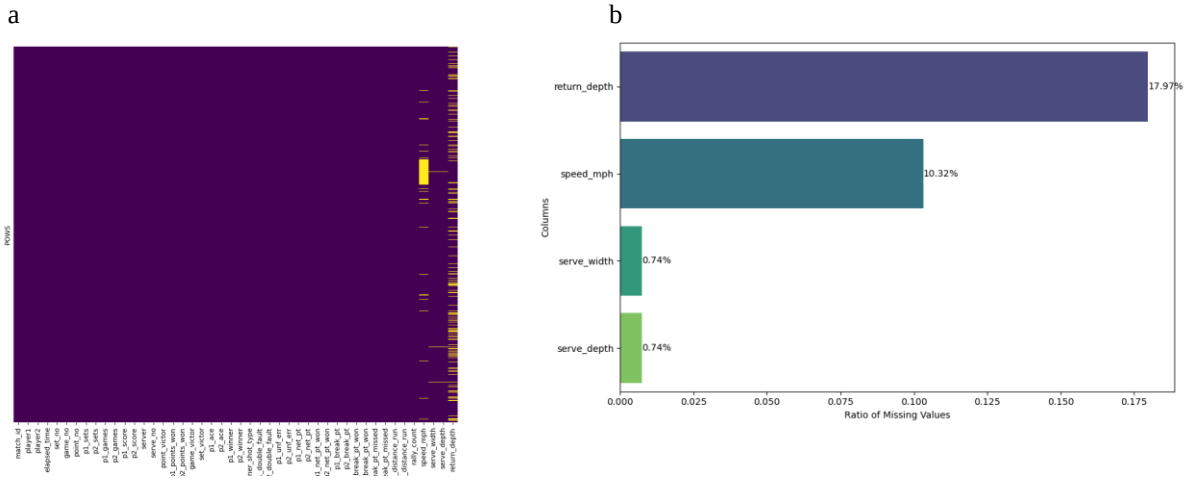

Figure 2. Visualization of missing data

After examining the visualization graph, we note the existence of missing values in columns such as speed_mph, serve_width, serve_depth, and return_depth. The heatmap illustrates that these missing

values are evenly distributed and do not make up a significant portion of the dataset. To handle this, we utilized the K-nearest neighbors (KNN) method to impute missing values in numerical columns, while rows with missing values in non-numeric columns were excluded.

2.3. Outlier handling

The 3 Sigma Rule and Deviation Point Rejection stand as effective techniques for identifying and managing outliers during data preprocessing and exploratory data analysis (EDA). These methods help reveal data distribution characteristics and pinpoint unusual deviations that can impact the analysis and model training process.

The 3 Sigma Rule, based on statistical principles, provides a means to detect outliers, assuming data follows a normal distribution. It identifies data points located more than three standard deviations from the mean as outliers, offering a statistically grounded approach to recognizing potential anomalies.

Deviation Point Rejection, on the other hand, is a straightforward yet effective method for identifying and removing data points that significantly deviate from the typical data distribution. When combined with the 3 Sigma Rule, it allows for the identification of outliers as data points deviating by more than three standard deviations from the mean, enabling further analysis or necessary processing.

When we applied the 3 Sigma Rule to the speed_mph column, we found certain data points that exceeded three standard deviations from the mean, which are typically flagged as outliers in statistical analysis. However, it's important to acknowledge that statistical methods come with their limitations, and a comprehensive understanding of the real-world context and the data is crucial. As shown in Figure 3.

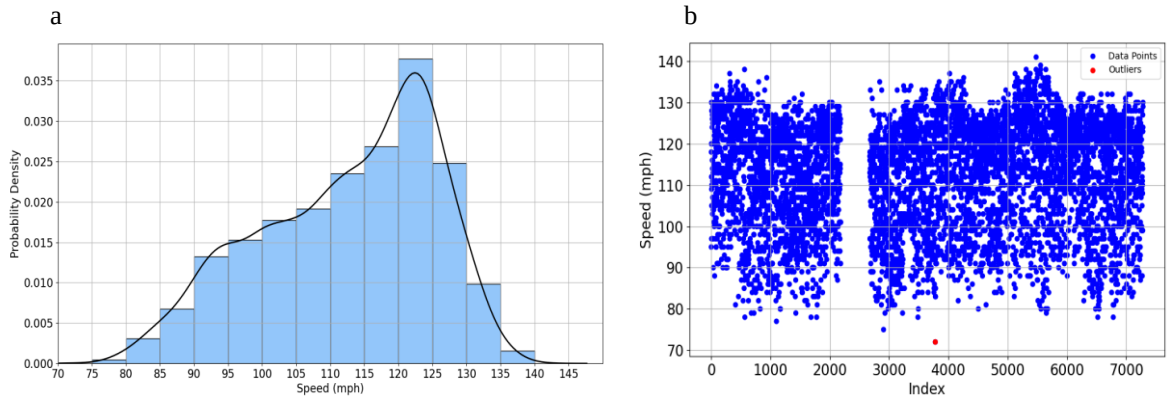

Figure 3. Outlier Detection in Speed Data

Upon closer inspection of the anomalies detected in the speed_mph column, taking into account specific race conditions and variations in player abilities, we concluded that these data points fall within the expected range. These variations may be influenced by factors such as race conditions or individual player characteristics, which can impact the recorded results. By adopting this holistic approach, we ensure that our decision-making process is well-informed and that we retain valuable data while upholding the quality, validity, and reliability of our data analysis.

2.4. The Establishment of the Scoring Model

The task is to create a tennis match-scoring model that assesses a player's performance at specific moments. To achieve this, we'll extract relevant data features and build an evaluation system that considers factors like scoring ability, consistency, key moment performance, and opponent comparisons.

2.5. Structure of the Evaluation System

To assess a tennis player's scoring performance, we have developed a comprehensive multi-indicator evaluation system. The key performance metrics considered in this system cover various

aspects of the game, ranging from basic scoring to critical moments. Here's a concise description of each metric: Set-specific Games Won, Scoring Lead Progress within a Game, Serve Identification, No-touch Point Forehand/Backhand, and so on.

A challenge we encountered in building a comprehensive player performance evaluation system was the inconsistency in metric scales. To standardize these metrics and make them comparable, we applied min-max normalization. This technique scales data to a range of 0 to 1, preserving relative relationships and enhancing model interpretability. The specific formula for min-max normalization is:

$$X_{norm} = \frac{(X - X_{min})}{(X_{max} - X_{min})} \tag{1}$$

In this transformation:

X represents raw data values,

X_{min} represents the minimum value of the indicator in the dataset,

X_{max} represents the maximum value of the indicator in the dataset.

Through this process, the minimum value of each indicator becomes 0, the maximum becomes 1, and others fall proportionally in between.

With normalized metrics, we can fairly compare the importance of different indicators and assign suitable weights. These weights can be determined using statistical analysis, expert knowledge, or machine learning algorithms.

3. Machine Learning Decodes Wimbledon Final

In this experiment, we select four efficient machine learning models: Gradient Boosted Decision Tree (GBDT), XGBoost, ExtraTrees, and CatBoost. These models are ensemble learning methods based on decision trees, which enhance prediction accuracy by constructing multiple decision trees and combining their predictions. They excel in handling complex nonlinear relationships and interaction effects, making them particularly suitable for scenarios with multiple indicators within our evaluation system.

(1) GBDT, or Gradient Boosting Decision Tree, is an iterative algorithm that constructs new trees by minimizing errors from the previous round's tree predictions.

(2) XGBoost, an efficient implementation of GBDT, further enhances the algorithm with optimizations and extensions. It introduces regularization terms to control model complexity and employs a more efficient tree-splitting algorithm.

(3) ExtraTrees, a variant of the Random Forest algorithm, is an integrated learning method that, similar to Random Forest, enhances model performance by aggregating predictions from multiple decision trees.

(4) CatBoost is a relatively recent machine learning algorithm based on Gradient Boosted Decision Trees (GBDT). Developed by researchers at Yandex, it excels in addressing classification problems, especially when dealing with datasets containing categorical features.

By inputting normalized metric data into these models and employing appropriate evaluation metrics (e.g., accuracy, AUC, etc.) to assess model performance, we can effectively evaluate the significance of these metrics in predicting player scores. Improved predictions not only highlight the importance of these metrics in understanding game dynamics but also offer data-driven insights for refining training and game strategies.

Building upon the earlier information, we incorporated a weighted fusion model into our methodology. Weighted fusion involves calculating the weighted average of predictions generated by the three models we created, aiming to improve prediction accuracy. When working with diverse datasets, weighted fusion emerges as a simpler and more efficient method for model integration. To determine the weights for this fusion, we utilized the performance of each model on a validation set, allowing for automatic adjustments to achieve optimal predictions. After comparing it to alternative

integration techniques like stacking, we concluded that weighted fusion is the most effective approach for our current dataset. As shown in Figure 4.

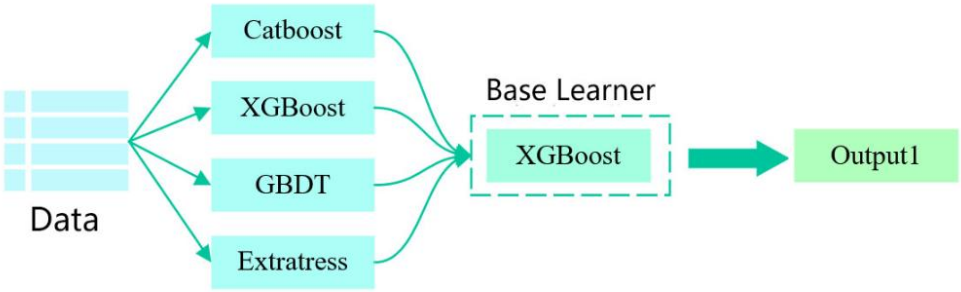

Figure 4. Stacking Model

3.1. Optimizing Prediction Models

To enhance the efficiency and accuracy of our selected model, we employed the following strategies:

(1) Cross-validation: We applied cross-validation, a statistical technique that divides the dataset into overlapping subsets, thereby enhancing the model's generalization during training. We adopted a five-fold cross-validation method, dividing the dataset into five parts, with four for training and one for testing. This approach mitigated the risk of overfitting.

(2) Grid search optimization: To further enhance the performance of the tree model, we leveraged grid search hyperparameter optimization. This process systematically explored various parameter combinations to identify the optimal set, including parameters such as learning rate, tree depth, and the number of trees. Through grid search, we identified the most effective parameter combinations.

By implementing these techniques, we substantially improved the model's accuracy and generalization capabilities while optimizing the efficiency and performance of our algorithm. As shown in Table 1.

Table 1 optimal parameters for each model

mold	GBDT	XGBoost	ExtraTrees	CatBoost
learning_rate	0.1	0.05	NULL	0.05
n_estimators	140	120	140	200
max_depth	8	8	12	8

Here are the accuracy versus F1 values for both the test and validation sets with the model using the optimal parameters, along with the accuracy curves during the training of each model. As shown in Table 2 and Figure 5.

Table 2 optimal parameters for each model

	GBDT	XGBoost	ExtraTrees	CatBoost
Accuracy	1.000	0.999	1.000	0.8533
F1 value	0.993	0.832	0.988	0.7854
validation	0.6933	0.6882	0.6773	0.6992

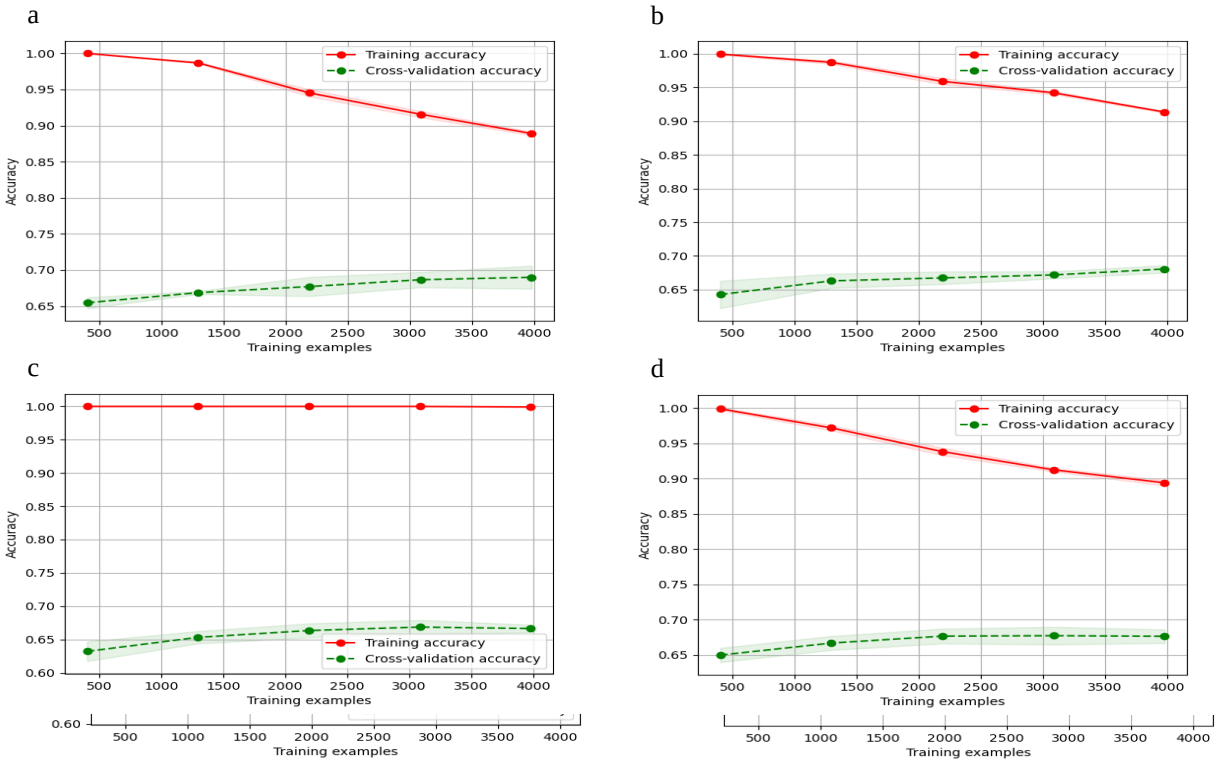

Figure 5. Model Performance Comparison Chart

In the improved model, we attained a 69% accuracy rate, showcasing the model's ability to predict a player's scoring performance in real matches.

3.2. Exploring Significant Contributions with SHAP Values

SHAP (Shapley Additive exPlanations) values are a method for interpreting the predictive results of a model, which gives the degree of contribution of each feature to the model output. By ranking the SHAP values of the features, we can learn which features have the greatest impact on the model's prediction results, which can help us identify important features, and potentially guide subsequent feature selection, model tuning, or problem interpretation.

Upon closer examination of the model's variables, it became clear that at least half of them made highly significant contributions to the predictions. The heatmap generated is presented in Figure 6.

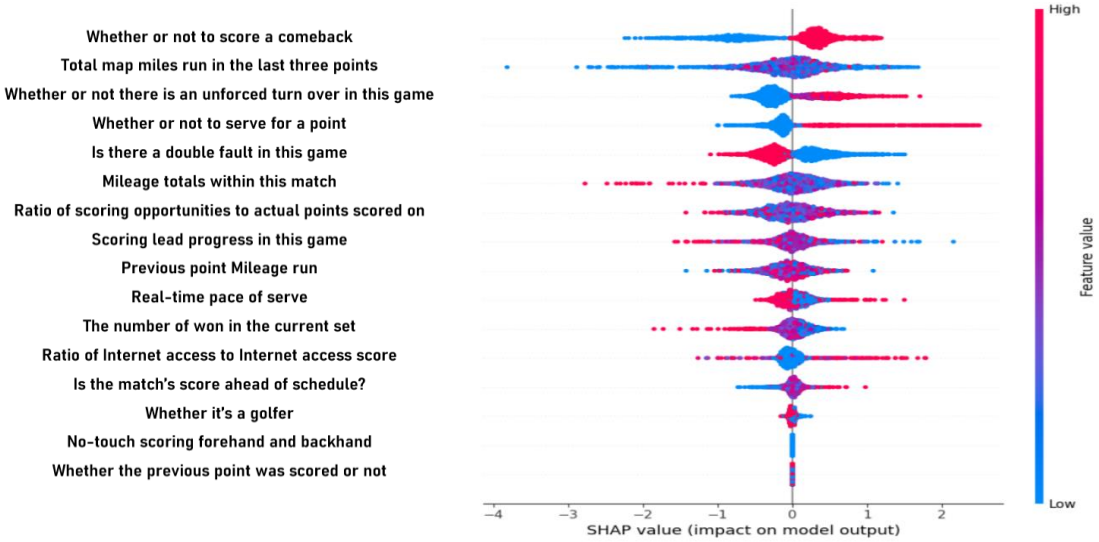

Figure 6. Contribution ranking of indicators based on SHAP

Using the above graph, we can find out which features in the dataset are effective in quantifying the impact of momentum on game scores.

3.3. Quantification and Correlation analysis of momentum

In tennis, momentum plays a pivotal role in shaping match outcomes. It signifies shifts in a player's control and dominance throughout a match, often achieved by scoring consecutive points, winning critical moments, or showcasing superior skills. Momentum not only bolsters a player's confidence but also exerts psychological pressure on the opponent, potentially influencing the course of the match, leading to reversals or further consolidating leads.

To quantify momentum for match analysis, we first conduct an in-depth quantitative examination of match data. We have previously established an evaluation system to gauge tennis players' scoring performance during a match, enabling us to quantify a player's capabilities by modeling their scoring potential at any given moment. Given the relevance of this evaluation system to defining momentum, we selected ten highly significant features to construct a quantitative momentum system.

Subsequently, we proceed to train the player's momentum utilizing the Stacking Integration Model structure, an efficient machine learning technique that enhances prediction accuracy by amalgamating various base models. Through this approach, we can build an optimal model capable of accurately forecasting changes in a player's momentum throughout a match.

Unlike conventional classification tasks, our objective here is to predict the likelihood of a player scoring at the current point, rather than merely determining their ability to score. This probability value serves as a quantitative measure of the player's "momentum." It is presented in Figure 7. Such quantification not only offers a more precise comprehension of momentum fluctuations during the game but also provides real-time, data-driven insights for coaches and players, facilitating improved decision-making during matches. To quantify this concept and evaluate its connection to a player's actual score, we utilize the Spearman correlation coefficient as our analytical tool.

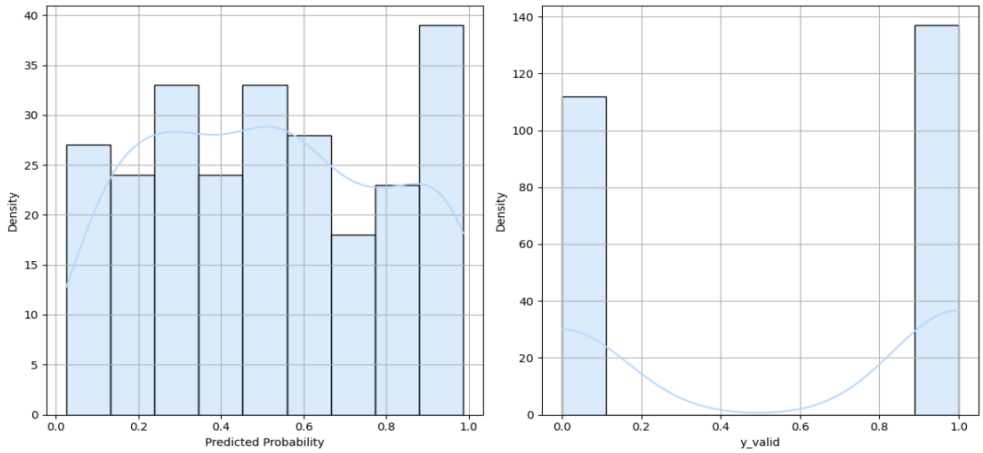

Figure 7. Density distribution graphs

The Spearman correlation coefficient is a non-parametric statistical method used to measure the strength and direction of monotonic relationships between two variables. It is based on the ranks (or orders) of the variables rather than their raw data values, does not rely on the assumption of data normality, and is well-suited for assessing the monotonic correlation between two variables, making it applicable to any type of data distribution.

Spearman's correlation coefficient is calculated by dividing the covariance between two variables by the product of their respective standard deviations.

$$\rho = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \tag{2}$$

Of which.

X_i and Y_i are the observed values of the two variables, respectively.

$\bar{X}$ and $\bar{Y}$ are the average of these observations, respectively.

Upon calculation, we obtained a Spearman's correlation coefficient of 0.479, signifying a moderately positive correlation between momentum and a player's actual score. In simpler terms, when the model-predicted momentum value rises, the player's probability of scoring also increases. This positive correlation holds statistical significance, implying that the connection between momentum and a player's score is unlikely to be a random phenomenon.

4. Conclusions

This study aims to develop a comprehensive model for assessing player performance in tennis matches, beginning with the collection and preprocessing of match data to extract key features such as scoring ability, consistency, and performance during critical moments. Through min-max normalization and assigning different weights based on statistical analysis, expert authority, and machine learning algorithm performance, a robust evaluation system is constructed. Decision tree integration algorithms like GBDT, XGBoost, ExtraTrees, and CatBoost are employed to predict players' scores at specific game moments, demonstrating efficient learning capabilities and accuracy. Further analysis reveals a moderate positive correlation between momentum and score, validating its effectiveness in predicting players' scores. Synthesizing these findings, strategic recommendations are provided for leveraging momentum against opponents, encompassing recognizing momentum influence, developing mental resilience, establishing effective coping strategies, enhancing physical fitness and technique, and analyzing opponents' momentum.

The methodology of this study not only reveal the potential value of data in quantifying player performance but also provide tools and theoretical support for applying these insights in actual matches. Our study provides real-time feedback for coaches and players to make data-based decisions during matches, opening up new avenues for improving tennis players' scoring ability.

References

[1] Dietl, Nesseler. Momentum in tennis: Controlling the match [J]. UZH Business Working Paper Series, 2017, (365).

[2] Fitzpatrick, Stone, Choppin, Kelley. How are elite tennis matches won at Wimbledon? A comparison of close and one-sided contests [J]. European Journal of Sport Science, 2024, 24(2): 190-9.

[3] Gao, Kowalczyk. Random forest model identifies serve strength as a key predictor of tennis match outcome [J]. Journal of Sports Analytics, 2021, 7: 255-62.

[4] Ghosh, Sadhu, Biswas, et al. A comparison between different classifiers for tennis match result prediction [J]. Malaysian Journal of Computer Science, 2019, 32(2): 97-111.

[5] Ishwarya, Nithya. Relative Analysis and Performance of Machine Learning Approaches in Sports; proceedings of the 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), F 2-4 Dec. 2021, 2021 [C].

[6] Lalwani, Saraiya, Singh, et al. Machine learning in sports: A case study on using Explainable models for predicting outcomes of volleyball matches [J]. arXiv preprint arXiv:220609258, 2022.

[7] Moustakidis, Plakias, Kokkotis, et al. Predicting football team performance with explainable ai: Leveraging shap to identify key team-level performance metrics [J]. Future Internet, 2023, 15(5): 174.

[8] Muslim, Dasril. Company bankruptcy prediction framework based on the most influential features using XGBoost and stacking ensemble learning [J]. International Journal of Electrical and Computer Engineering (IJECE), 2021, 11(6): 5549-57.

[9] Polk, Yang, Hu, Zhao. TenniVis: Visualization for Tennis Match Analysis [J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 2339-48.

[10] van Meurs, Buszard, Kovalchik, et al. Interpersonal coordination in tennis: assessing the positional advantage index with Australian Open Hawkeye data [J]. International Journal of Performance Analysis in Sport, 2020, 21(1): 22-32.