ONLINEISSN (2958-793X) PRINTISSN (2960-1657)
Ethiopian Journal of Sport Science (EJSS)
Volume V.6, Issue I (2025)

Biomechanical and Body Composition Factors in Shot Put Performance: A Predictive Model Using Machine Learning

Daniel Getnet Tasew
Bahir Dar University, Ethiopia
danielgetnet706@gmail.com

Receivedin July.2025. Revised fromSep-Dec.2025,Accepted: Dec, 2025. Ethiopian Journal of Sport Science (EJSS), Volume VI, and Issue I, Published by Ethiopian Sport Academy 2025

Key words*:* *Biomechanical Features, Body Composition, Machine Learning, Shot Put Performance*

Abstract

This study examines the key biomechanical and body composition features influencing shot put performance, utilizing machine learning models to predict shot distances. Four models Random Forest, Gradient Boosting, Categorical Boosting, and extreme Gradient Boosting were employed to analyze a dataset of 42 elite athletes. Fifteen biomechanical features were assessed for importance using the Random Forest model. Through feature selection, release velocity, gender, shot path length, and body mass emerged as the four most influential predictors of shot put performance, while shot release height, technique, and angle of release were among the least influential factors. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). Of the models tested, Gradient Boosting showed the highest predictive accuracy, achieving an R^2 of 0.8248, an MAE of 0.4474, and an RMSE of 0.6500. Following hyper parameter tuning, the final model was evaluated on unseen data, demonstrating impressive predictive accuracy and further validating its robustness. These findings provide valuable insights into the relationship between biomechanical and body composition factors and shot put performance, offering practical applications for athletes and coaches seeking data-driven approaches to optimize performance. By utilizing the model developed in this study, athletes and coaches can use their own data to predict shot distance, enabling more targeted and effective training strategies.

.

Citedas:Daniel GetnetTasew*(2025):*Biomechanical and Body Composition Factors in Shot Put Performance: A Predictive Model Using Machine Learning*: Ethiopian Journal of Sport Science (EJSS) V.6 page 139-159*

1. Introduction

The shot put is a field event that requires executing a series of complex, high-speed movements within the confined space of the throwing circle (Lipovšek, Škof et al. 2011). An athlete's performance in shot put is shaped by biomechanical factors, including their body composition, physical capabilities, and motor skills, as well as the mechanics specific to their chosen throwing technique (Linthorne 2001, Terzis, Kyriazis et al. 2012). Key elements, such as the angle, velocity, and height of release, shot path length, and shoulder hip separation angle are crucial to achieving the longest possible throw distance (Linthorne 2001, Pavlović and Vrcić 2011). Given the complexity and nonlinear nature of these biomechanical interactions, traditional statistical models may struggle to accurately capture performance determinants. This has led to increased interest in advanced computational approaches such as machine learning (ML).

ML models offer advanced methods to capture complex and nonlinear relationships, leading to more accurate predictions of athletic performance (Meäyk and Unold 2011, Mürsel, Murat et al. 2022, Hore April. 2018). As a subset of artificial intelligence, ML enables computers to learn from data and improve their performance on tasks without explicit programming. By identifying patterns and relationships within datasets, ML models can effectively address various challenges, including making predictions, classifying data, and aiding in decision-making (Shalev-Shwartz and Ben-David 2014, Harleen and Vinita 2018).

Building on the success of ML in different sports, researchers have explored its applications in various performance prediction tasks, demonstrating its effectiveness across disciplines. For instance, Ofoghi et al. applied ML methods to develop predictive models for athlete performance in the track cycling championships (Ofoghi, Zeleznikow et al. 2010). Similarly, **Whiteside et al. (2017) employed** ML to evaluate hitting loads in tennis (Whiteside, Cant et al. 2017). **Maier et al. (2018) utilized ML techniques for** predicting biathlon shooting performance (Maier, Meister et al. 2018). In another study, **Musa et al. (2019)** proposed a classification model to predict the future success of young archers (Musa, Anwar et al. 2019). Moreover, ML has been employed to categorize the playing positions of elite junior Australian football players using technical skill indicators, alongside linear discriminant analysis (Woods, Veale et al. 2017).

A widely adopted method in this context is supervised learning, where models are trained on labeled datasets, enabling them to make predictions on new, unseen data (Kelleher and Tierney 2018). This approach enhances pattern recognition and supports informed decision making in various contexts. And also, this method is extensively applied across various domains, such as finance, healthcare, and sports, to provide meaningful insights and support data driven decision making (Shiliang 2013, Jordan and Mitchell 2015).

Despite the presence of field event athletes, including shot putters, in various clubs, there is a notable absence of well-structured training programs that address the

ONLINEISSN (2958-793X) PRINTISSN (2960-1657)
Ethiopian Journal of Sport Science (EJSS)
Volume V.6, Issue I (2025)

critical biomechanical and body composition factors essential for optimal performance. Although some athletes in our country possess the ideal somatotype for shot put, their performance remains significantly below that of professional athletes, hindering their ability to compete at continental and international levels, unlike their counterparts in running events. This disparity highlights a reliance on generalized training methods by coaches and athletes, often lacking scientific guidance on the specific areas that require emphasis for improvement. This gap underscores the need for a data-driven analysis to determine the key biomechanical and body composition predictors of shot put performance.

This study hypothesizes that ML models can effectively predict shot put performance by analyzing biomechanical and body composition variables, with specific factors such as release velocity, shot path length, and body mass significantly contributing to prediction accuracy. The study aims to identify the key biomechanical and body composition features that influence shot put performance, employing ML models to predict shot distances. Ultimately, the findings will contribute to the development of a model (or tool) designed to assist coaches, athletes, and sports organizations in optimizing training strategies and enhancing shot put performance through data-driven analysis.

1. Materials and methods

1.1. Study design

This study utilized a correlational design to examine the predictive relationship between biomechanical and body composition features on shot put performance.

1.2. Data set

Acquiring biomechanical data to predict athletic performance is inherently challenging, as it requires sophisticated video analysis tools and precise measurement techniques, which can be difficult for researchers to access. Capturing key biomechanical variables, such as release velocity, shot path length, and joint angles, demands high-speed motion capture systems and expert annotation, making data collection both time-intensive and resource-intensive.

For this study, data were sourced from the World Athletics reports of the 2017 and 2018 World Championships (Dinsdale, Thomas et al. 2017, Thomas, Dinsdale et al. 2018), which provided fifteen biomechanical and body composition information on 42 elite shot putters (22 males and 20 females). To handle missing values, the mean imputation method was applied, replacing missing data points with the average value of the respective feature. This approach helps maintain dataset integrity while minimizing bias, ensuring that machine learning models can effectively analyze performance predictors (Wijayasekara, Shyamala et al. 2022).

Table 1

Attribute and outcome variables used for the models

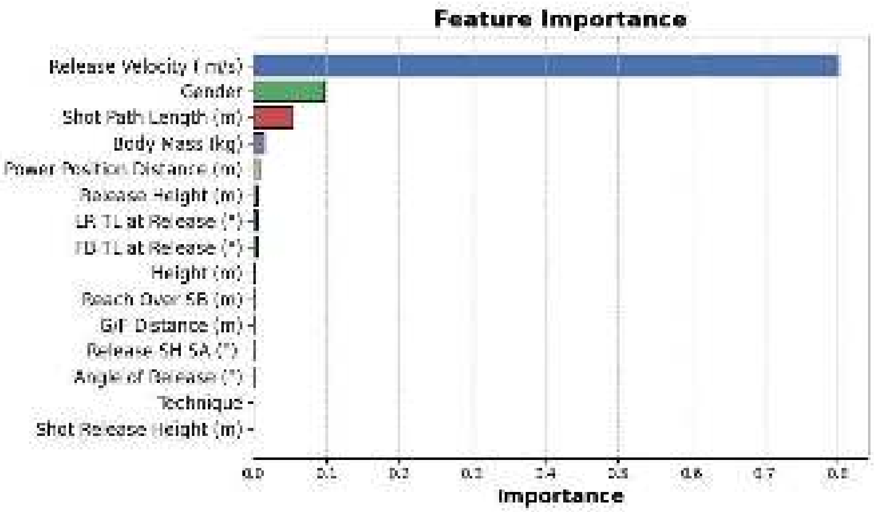
Attribute parameters			
	Unit		**Unit**
Gender		Forward Backward Trunk Lean (FB TL) at Release	(°)
Height	(m)	Left Right Trunk Lean (LR TL) at Release	(°)
Body Mass	(kg)	Shot Path Length	(m)
Technique		Shot Release Height	(m)
Release Velocity	(m/s)	Glide Flight (G/F) Distance	(m)
Angle of Release	(°)	Power Position Distance	(m)
Release Height	(m)	Shoulder Hip Separation Angle (SH SA) at Release	(°)
Reach Over Step Board (SB)	(m)		
Outcome parameter			
Performance (Distance)			(m)

1.1. Feature selection

A Random Forest (RF) model was used in this study to identify key biomechanical and body composition variables influencing shot put performance. Known for handling complex datasets and capturing nonlinear relationships, RF effectively determined the ten most influential predictors, which were then used in the final predictive analysis (Kursa and Rudnicki 2010).

Figure 1

Feature selection for predicting shot put performance using Random Forest



1.2. Machine learning models

In this study, Random Forest (RF), CatBoost (CB), Gradient Boosting (GB), and Extreme Gradient Boosting (XGB) ML models were utilized to predict and analyze the performance of shot putters based on biomechanical and body composition variables.

1.2.1. Random Forest

As Breiman (2001) states, RF is a widely used machine learning model known for its accuracy and versatility (Breiman 2001). As an aggregation learning method, it builds multiple decision trees, each predicting the average value of the target variable. By using a subset of variables for each split, RF reduces bias and variance errors, making it robust for handling large datasets with many features. Key characteristics include its ability to rank feature importance, with significant features appearing at the top of trees, improving interpretability and reducing generalization errors as more trees are added (Breiman 2001, Liaw and Wiener 2022).

Tuning key hyper-parameters in RF, such as **n_estimators**, which defines the number of trees, is critical to balancing performance and computational cost. While more trees typically improve accuracy by reducing variance, too many can increase computational demands without significant gains. Other parameters like **max_depth** control tree complexity, and **max_features** adjust the number of features considered for splitting, helping to prevent **over_fitting**. These features, along with **min_samples_split** and **bootstrap**, make RF a flexible and powerful model for predictive tasks (Fabian, Gaël et al. 2011).

1.2.2. Catagorical boosting regressor

Prokhorenkova et al. (2018) describe the CB as a machine learning model designed to predict continuous outcomes using GB decision trees (Prokhorenkova, Gusev et al. 2018). This newer algorithm is noted for its efficiency, precision,

and capability to manage categorical variables effortlessly, thereby reducing the need for extensive preprocessing. It starts by creating a series of weak decision trees, which are subsequently integrated to develop a strong predictive model. The CB model utilizes the following formula to predict continuous values:

$$F(x) = Fo(x) \sum_{m=1}^{M} \sum_{i=1}^{N} fm(xi)$$

Equation 1. CatBoostRegressor model formula

Where the output function $F(x)$ is the overall prediction function that CB aims to learn, $Fo(x)$ is the initial guess or the baseline prediction, the summation over with variable M represents the summation over the ensemble of trees and with variable N represents the summation over the training samples and the prediction of the mth tree for the ith training sample. Each tree in the ensemble contributes to the overall prediction by making its own prediction for each training sample (Dorogush, Ershov et al. 2018).

Key **hyperparameters** in CB Regressor that significantly influence its performance include**learning_rate**, which determines the size of the updates during training, and iterations, specifying the number of boosting rounds to be executed. The **depth** parameter controls the maximum depth of the trees, enabling the model to capture complex relationships while managing the risk of **over_fitting**. Additionally, parameters like **random_seed**enhance the reproducibility of results. The algorithm's inherent ability to manage categorical features and its robust performance in various predictive tasks make CB an essential tool in the machine learning toolkit (Fabian, Gaël et al. 2011).

1.2.3. Gradient boosting

Natekin, A. and A. Knoll (2013) explain that GB is a ML technique that uses multiple models to tackle both classification and regression tasks. It combines several weak learners, typically decision trees, to form a stronger predictive model (Natekin and Knoll 2013). In GB, each tree is constructed in a sequence, with each new tree aiming to correct the errors made by its predecessor. By refining the model through successive iterations, it can identify complex relationships between the input data and the target variable. This process of continuously improving the model is known as boosting, and it greatly enhances its predictive power (Friedman 2001). The prediction for a GB model can be expressed using the following formula:

$$F(x) = fo(x) + \sum_{m=1}^{M} f\mathrm{m}(\mathrm{x})$$

Equation 2. Gradient Boosting model formula

Where $F(x)$ is the final prediction for a given input x, fo (x) is the first weak model, which is, usually a simple constant or mean value, $fm(x)$ is the mth weak model, for m = 1, 2, . . . , M and M is the total number of weak models used in the gradient boost model (Friedman 2001).

In GB, the **learning_rate** controls how much each tree contributes, with smaller values requiring more trees but improving stability. The **n_estimators** defines the number of trees, while **max-depth** limits tree complexity to avoid over-fitting. **Random_state** ensures reproducibility, and **max_features** specifies the number of features to consider for splits. **Min_samples_split**and **min_samples_leaf** set the minimum samples needed for node splitting and leaf nodes. Subsample introduces randomness to reduce **over_fitting**, similar to bootstrap sampling (Fabian, Gaël et al. 2011).

1.2.4. Extreme gradient boosting

XGB is a powerful regression model and ensemble learning technique that leverages gradient boosting and decision trees to make accurate predictions (Chen and Guestrin 2016). While similar to other GB models, XGB includes performance enhancements such as improved computation speed and reduced over-fitting. By iteratively improving weak learners to form a strong predictor, the "boosting" approach in XGB strengthens model precision (Emadi, Bagherzadeh et al. 2020).

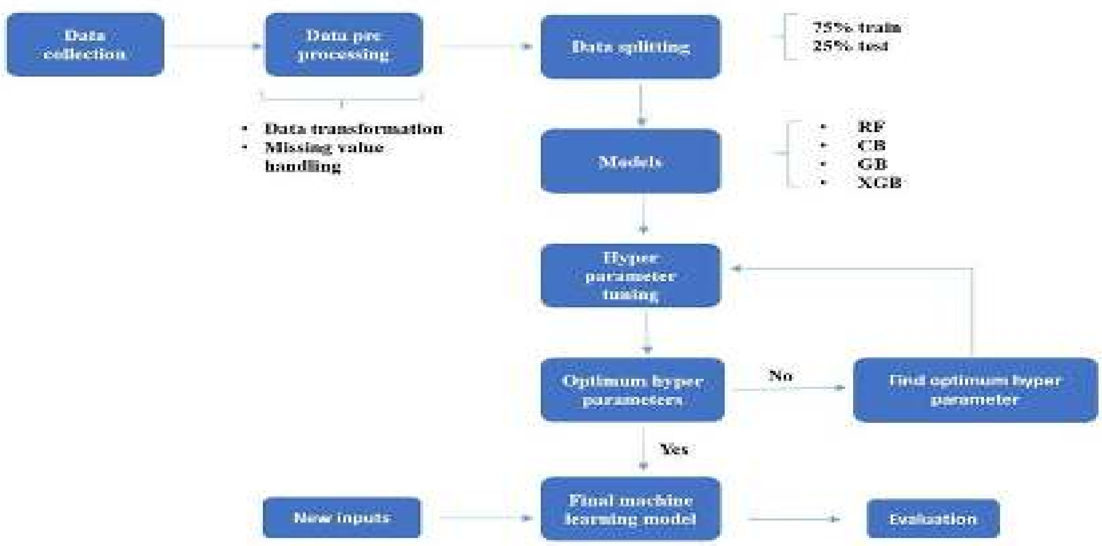
Key hyper-parameters in XGB significantly influence its training and performance. The **learning_rate** controls the step size during iterations, with a smaller rate often requiring more **n-estimators** for accuracy.

Max_depth determines tree complexity, while **random-state** ensures reproducibility. Parameters like **min_child_weight** prevent over-fitting by setting the minimum sum of instance weights in a child node. **Subsample** and **colsample_bytree** introduce randomness by selecting fractions of samples and features for training, respectively, and **gamma** provides regularization by setting a threshold for leaf node splitting. Careful tuning of these**hyperparameters**is crucial for optimizing performance and managing the bias-variance tradeoff (Fabian, Gaël et al. 2011).

1.3. Development of the model

The dataset was randomly divided into two parts: 75% for training and 25% for testing, minimizing potential bias. Models were built using the training data, focusing on evaluating their generalization ability and tuning hyperparameters for optimal performance. Once the best hyperparameters were identified, the model was tested on the unseen test set to assess its predictive performance. This process, as illustrated in **Figure 2**, ensured a fair evaluation of the model's accuracy and effectiveness.

Figure 2
Model development process



Adapted from Nigusie et al. 2024

1.4. Data analysis technique

Python software (version 3.11.0) was utilized to analyze the performance of ML models, with libraries like scikit-learn, pandas, and matplotlib used to evaluate metrics, visualize results, and fine-tune model parameters.

1.5. Evaluation of models

Performance metrics, also referred as error measures, are essential for evaluating models across various disciplines. These metrics offer a structured mathematical approach to determine how well the predicted outcomes match the actual results. Several metrics have been explored in the academic literature, with the most common being the coefficient of determination (**R^2**), Mean Absolute Error (**MAE**), and Root Mean Squared Error (**RMSE**). In this study, all these evaluation metrics were used.

2. Results

2.1. Performance of models

Four machine learning models RF, CB, GB, and XGB were evaluated for predicting shot put performance using biomechanical and body composition variables.

3.2.1. Random forest model

Figure 3 displays a scatter plot comparing predicted shot put distances from the **RF** model to actual measured values in the test dataset. The regression line in the plot highlights a moderate correlation, with the RF model achieving an R^2 of 0.7743. Optimized hyperparameters (**Table 2**) ensured reproducibility and mitigated over fitting.

Figure 3

Evaluating RF model performance on the test set

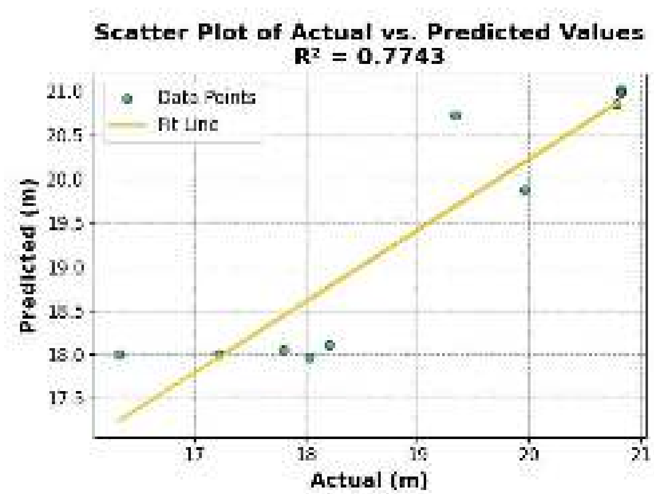


Table 2

Tuned hyperparameters, their ranges, step, and optimal values employed in the RF model

Hyperparametres	Range investigated	Step	Optimum value
n_estimator	[100, 500]	1	176
max_depth	[1, 10]	1	2
min_samples_split	[2, 10]	1	2
min_samples_leaf	[2, 10]	1	5
random_state	42		42

3.2.2. CatBoost model

The scatter plot in **Figure 4** compares actual shot put distances with predictions from the **CB** model, which achieved an **R^2** of **0.7586**, indicating a strong correlation. The optimized hyperparameters shown in **Table 3** enabled the model to effectively capture data patterns with good precision, ensuring consistent performance.

Figure 4

Evaluating **CB** model performance on the test set

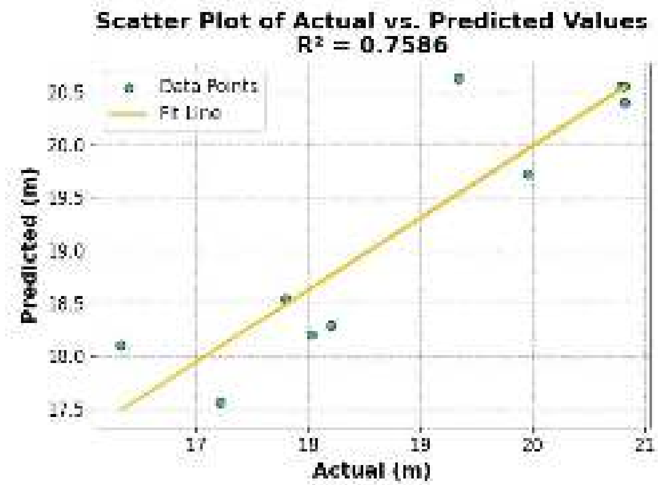


Table 3

Tuned hyperparameters, their ranges, step, and optimal values employed in the CB model

Hyperparametres	Range investigated	Step	Optimum value
iteration	[100, 500]	1	100
learning _rate	[0.01, 0.3]	0.01	0.28
Depth	[1, 10]	1	5
l2_leaf_reg	[1, 10]	1	8
random_state	42	_	42

3.2.3. Gradient boosting model

Figure 5 illustrates the scatter plot for GB predictions against actual values. GB demonstrated the strongest performance, with an **R^2** of **0.8248**. As highlighted in **Table 4**, GB effectively modeled complex relationships between biomechanical and body composition variables and shot put performance, benefiting from carefully tuned hyperparameters to achieve high accuracy and robustness.

Figure 5

Evaluating **GB** model performance on the test set**.**

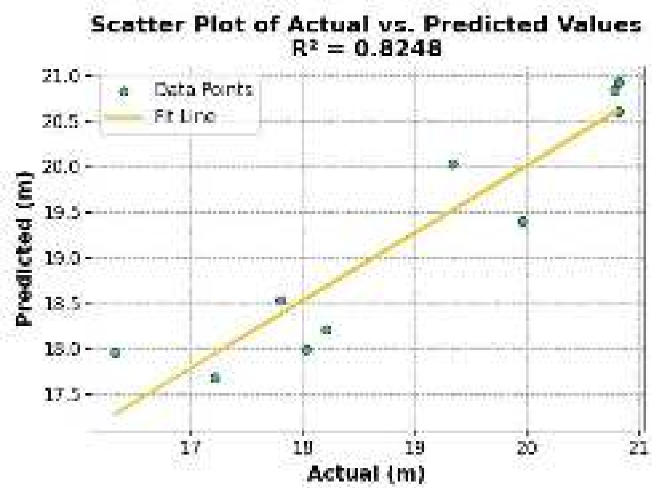


Table 4

Tuned hyperparameters, their ranges, step and optimal values employed in the GB model

Hyperparametres	Range investigated	Step	Optimum value
n_estimator	[100, 500]	1	103
learning _rate	[0.01, 0.3]	0.01	0.3
mean_sample_split	[2, 10]	1	10
Subsample	[0.1, 1.0]	0.1	0.6
random_state	42	_	42

3.2.4. Extreme gradient boosting

Figure 6 presents the scatter plot comparing actual and predicted distances for the XGB model. XGB performed well, achieving an $\mathbf{R^2}$ of **0.8088**, suggesting high predictive accuracy. Although slightly less precise than GB, XGB effectively captured underlying data relationships. **Table 5** lists the optimized hyperparameters for this model

Figure 6

Evaluating **XGB** model performance on the test set

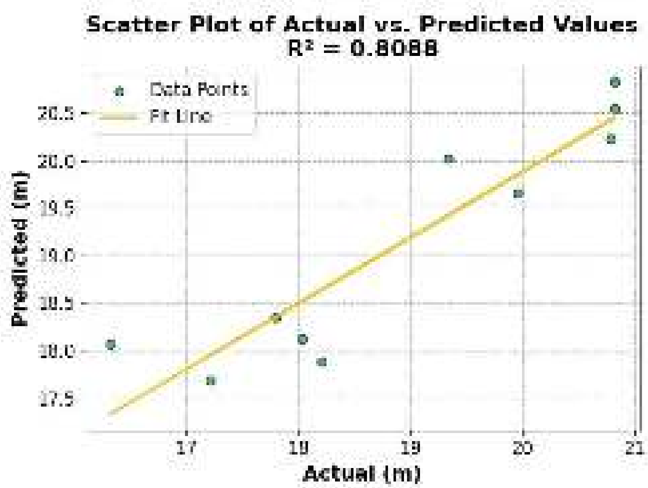


Table 5

Tuned hyperparameters, their ranges, step, and optimal values employed in the **XGB** model

Hyperparametres	Range investigated	Step	Optimum value
n_estimator	[100, 500]	1	101
learning _rate	[0.01, 0.3]	0.01	0.3
max_depth	[1, 10]	1	2
subsample	[0.5, 1]	0.1	0.89
random_state	42		42

2.2. Models evaluation metrics comparison

Performance metrics, including R^2, MAE, and RMSE, were used to assess each model. **Table 6** and **Figure 7** present the evaluation metrics for each model. Among them, the GB model stood out with the highest R^2, along with the lowest RMSE and MAE, confirming its position as the most precise and reliable model in this study. The XGB model closely followed in performance, while the RF model showed comparable results to XGB. In contrast, the CB model recorded the lowest R^2, along with higher RMSE and MAE values. This comparison underscores the GB model's superiority in terms of both predictive accuracy and stability.

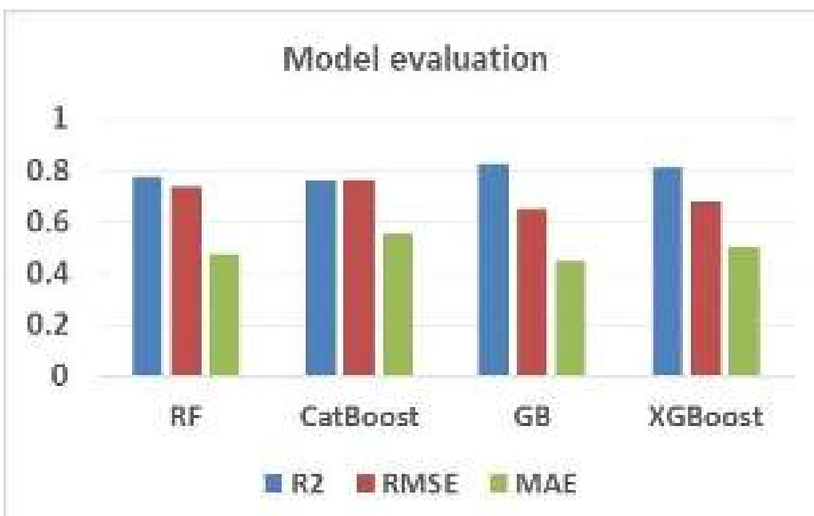
Table 6

Results of the evaluation metrics for each model utilized

Models	R^2	RMSE	MAE
RF	0.7743	0.7378	0.4748
CatBoost	0.7586	0.7632	0.5533
GB	**0.8248**	**0.6500**	**0.4474**
XGBoost	0.8088	0.6792	0.5012

Figure 7

Evaluation metrics for the four models utilized



2.3. Evaluation on unseen data

To further assess the generalization ability of the models, the GB model was tested on a separate dataset containing unseen shot put data. As indicated in **Table 7** below, the results confirmed the model's robustness, as it accurately predicted the shot put distances for new cases. For instance, the GB model predicted 19.26 m for a test case where the actual distance was 19.14 m, and 21.48 m for a case with a true value of 21.46 m. The small deviations between the predicted and actual values across different cases indicated that the GB model generalized well to unseen data, further supporting its applicability for predicting shot put performance based on biomechanical and body composition variables.

Table 7.

Model prediction with unseen data

Gender	Height (m)	Body Mass (kg)	Release Velocity (m/s)	Release Height (m)	Reach Over SB (m)	FB TL at Release (°)	LR TL at Release (°)	Shot Path Length (m)	Power Position Distance (m)	Actual Observation distance (m)	GB Model Prediction distance (m)
F	1.75	110	12.95	2.11	0.18	6	-20	2.42	1.15	19.14	19.26
F	1.78	102	12.63	2.1	0.02	-6	-9	2.51	0.72	18.18	18.51
M	1.88	115	13.68	2.2	0.22	-13	6	2.84	0.27	21.46	21.48
M	1.91	134	13.49	2.24	-0.04	-14	-12	2.9	0.76	21.09	20.84
M	1.99	122	13.43	2.22	0.1	-1	1	2.76	1.26	20.8	20.45

3. Discussion

3.1. Feature Importance

Throwing distance in shot put is strongly influenced by release velocity, with numerous studies confirming its critical role (Linthorne 2001, Rodríguez, Riera et al. 2002, Gutiérrez-Davila, Rojas et al. 2009, Ohyama, Fujii et al. 2009). Their findings align with the results in this study, where release velocity emerges as the top contributor to performance, confirming its importance in maximizing throwing distance. The ability to optimize release velocity allows athletes to harness greater kinetic energy, translating into enhanced throwing power and distance, underscoring its pivotal role in the biomechanics of shot put (Athletics 2010). As such, coaches and athletes should prioritize improving this variable to achieve peak performance.However, the relationship between other factors such as release angle, height, and horizontal release distance must also be carefully considered, as these elements collectively contribute to optimizing performance.

Changes in release angle, for instance, can have a subtle yet significant impact on official shot put distances. While release angle is a critical biomechanical factor in optimizing performance, its influence on throwing distance is generally less pronounced compared to the effect of release velocity (Schofield, Cronin et al. 2019). Studies has shown that the optimal release angle for maximizing shot put distance typically ranges between 35° and 40°, depending on factors such as athlete height, release velocity, and environmental conditions (Zatsiorsky 1997, Linthorne 2001). However, deviations from this optimal range, even by a few degrees, tend to have a relatively minor impact on overall performance compared to variations in release velocity. This is because the relationship between release angle and distance follows a parabolic trajectory, where small deviations near the optimal angle result in only marginal changes in distance (Hay 1993).

In addition to release angle, the height at which the shot is released also contributes to overall performance, albeit to a lesser extent. Release height is influenced by the athlete's anthropometric characteristics, such as body height and arm length, as well as their technique during the final phase of the throw (Bartlett 2007). While a higher release height can theoretically reduce air resistance and slightly improve the trajectory of the shot, its effect on the overall distance is often considered secondary to the influence of release velocity and shot path length (Linthorne 2001).

Studies have demonstrated that increasing release height by 10 cm, for example, may only result in a distance gain of approximately 1-2%, which is relatively insignificant compared to the gains achievable through improvements in release velocity (Gutierrez-Davila, Rojas et al. 2013). Furthermore, the shot path length defined as the horizontal distance covered by the shot during the throwing motion plays a more substantial role in determining the final distance, as it directly influences the momentum and energy transfer from the athlete to the shot(Mackala, Fostiak et al. 2015).

In fact, the shot path length has been identified as a significant predictor of performance in biomechanical analyses, reflecting its critical contribution to the model's ability to predict outcomes. The length of the shot path not only affects the dynamics of the throw but also influences the timing and coordination of the athlete's movements, which are essential for maximizing performance (Schofield, Cronin et al. 2019). Therefore, while release height and angle are important considerations, optimizing shot path length and release velocity should remain the primary focus for athletes and coaches aiming to enhance shot put performance.

Body mass is a also observed as an important feature in predicting shot put performance due to its significant impact on an athlete's ability to generate force and power during the throw. In the context of this study, Body mass emerged as one of the most important variables in the predictive model. The relationship between body mass and shot put performance can be attributed to the fact that an optimal body mass enhances both strength and stability, which are fundamental for executing an effective shot put. (De Rose E and L. 1978, Kyriazis, Terzis et al. 2010). Athletes with an appropriate body mass are better equipped to generate higher levels of force, which directly influences the release velocity of the shot. Furthermore, Body mass plays a vital role in maintaining proper body mechanics throughout the throw, ensuring that the athlete can maximize their efficiency and technique.

Feature importance analysis of this study highlights that certain variables, such as Technique, Shot ReleaseHeight, Shoulder hip separation angle at release, and Release height, contribute minimally to predicting shot put performance compared to more dominant factors like release velocity and shot path length. Among these, Technique shows surprisingly low importance, which may be attributed to the high level of technical consistency among elite athletes. While technique is crucial for executing the throw efficiently, its variability is often reduced at the elite level, making it less predictive of performance differences (Bartlett 2007, Schofield, Cronin et al. 2019). This suggests that once athletes reach a certain proficiency, further refinements in technique may yield diminishing returns in terms of distance improvement.

Similarly, Shot release heightand release heightexhibit low importance scores, reinforcing the idea that anthropometric factors play a secondary role in shot put performance. Although a higher release height can theoretically provide a slight advantage by reducing air resistance and optimizing the shot's trajectory, its impact on overall distance is minimal compared to the influence of release velocity (Gutierrez-Davila, Rojas et al. 2013).

3.2. Performance of models

The results of this study provide valuable insights into the predictive power of different machine learning models applied to shot put performance analysis. Among the four models tested, **GB** emerged as the most accurate and effective model for predicting shot put distances. The high **R^2** value of **0.8248** and the low **RMSE of 0.65** suggest that GB is highly capable of identifying the complex relationships between biomechanical and body composition variables and shot put performance in this data set. These findings align with previous studies that have highlighted the effectiveness of boosting algorithms, particularly Gradient Boosting, in tasks requiring the modeling of intricate, non-linear relationships (Friedman 2001). GB's ability to iteratively improve upon weak learners by minimizing errors in a sequential manner makes it an ideal choice for capturing complex patterns in high-dimensional data, such as the biomechanical factors influencing shot put performance (Chen and Guestrin 2016).

The **XGB** model also demonstrated strong predictive accuracy, with an **R^2** of **0.8088**. Although slightly lower than GB, the performance of XGB remains commendable, especially considering its ability to handle large datasets efficiently and its robustness to overfitting. While its predictive power in this dataset was slightly lower than that of GB, XGB's strong generalization ability and computational efficiency make it a suitable alternative for predicting shot put performance in real-world applications, particularly when dealing with large-scale datasets (Ke, Meng et al. 2017).

RF also performed reasonably well, with an **R^2** value of **0.7743**, but its results were not as strong as those of GB and XGB. RF is known for its versatility and ability to handle large datasets, and while it provided moderate performance in this study, it was not able to capture the underlying patterns in the data as effectively as the boosting algorithms. The moderate performance of RF can likely be attributed to the fact that it relies on a series of decision trees built independently, which might not be as effective in identifying complex interactions between variables as the boosting techniques, which build models sequentially (Breiman 2001).

CB, despite its strengths in handling categorical data, demonstrated the lowest predictive accuracy among the four models, with an **R^2** of **0.7586**. However, CB's performance was still respectable, and its ability to handle categorical variables without requiring extensive preprocessing made it a valuable model to consider for applications involving categorical data (Prokhorenkova, Gusev et al. 2018). Nonetheless, CB still provided valuable insights into the relationships between body composition and biomechanical features and their effects on shot put performance.

The performance of all models was significantly influenced by the careful optimization of hyperparameters. Critical hyperparameters, such as the **number of trees (n_estimators)**, **tree depth (max_depth)**, and **learning rate**, were systematically fine-tuned to achieve optimal predictions. These parameters were selected because they are among the most commonly used and recommended for optimization in tree-based models, as they directly influence model complexity, training efficiency, and generalization performance(Chen and Guestrin 2016, Ke, Meng et al. 2017). For example, the lower **max_depth** values used in Random Forest (RF) and XGB likely mitigated overfitting, enabling these models to generalize more effectively to unseen data. Conversely, the higher depth values employed in GB allowed it to capture more intricate interactions within the data, contributing to its superior performance. This underscores the importance of hyperparameter tuning in striking a balance between model complexity and predictive accuracy, ensuring that each algorithm performs at its best.

The evaluation of the models on unseen data further demonstrated the robustness of the Gradient Boosting (GB) model. The small deviations between predicted and actual values for the test cases confirmed that GB was able to generalize effectively to new datasets. This strong generalization capability highlights machine learning models reliability in real-world applications, where models must perform well on data they have not encountered during training.

4. Conclusions

Analysis of feature in this study revealed that release velocity, gender, shot path length, and body mass were the most influential predictors of shot put distance. These key variables contributed significantly to the model's predictive accuracy, underscoring their critical role in shot put performance. Recognizing these top features provides a clearer understanding of the physical attributes and mechanics most closely linked to shot put performance.

The strong performance of the developed GB model on unseen data underscores its effectiveness as a predictive tool, particularly when incorporating key biomechanical and body composition variables. These findings not only advance predictive analytics but also provide valuable insights for athletes and coaches, directing training efforts toward the most impactful performance factors. By utilizing the model's code provided in this study, along with specific biomechanical and body composition data, athletes and coaches can predict the performance of shot put athletes using their own data, facilitating data-driven training and performance optimization.

5. Code (predictive tool) availability

The data and code utilized in this study, as well as predictive tools for making predictions with your own data, are available on my GitHub ripository.; https://github.com/Daniel-Getnet/Shot-Put-performance

6. Statement of declaration

The author declare that there is no conflict of interest with respect to the research, authorship, and publication of this article. No financial or personal relationships have influenced the research presented in this paper.

References

Athletics, W. (2010). Biomechanical Analysis of the Shot Put at the 2009 IAAF World Championships in Athletics. New Studies in Athletics. **3:** 12-15.

Bartlett, R. (2007). Introduction to Sports Biomechanics: Analysing Human Movement Patterns. . London, Routledge.

Breiman, L. (2001). "Random Forests." Machine Learning**45**: 5-32.

Chen, T. and C. Guestrin (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, ACM.

Chen, T. and C. Guestrin (2016). XGBoost: A Scalable Tree Boosting System. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA.

De Rose E and B. L. (1978). "Distance assessment in shot put based on lean body mass. ." Nutrition, dietetics and sport: 233-240.

Dinsdale, A., A. Thomas and A. Biss (2017). "Biomechanical report for the shot put men's event at the IAAF World Championship, London 2017."

Dorogush, A. V., V. Ershov and A. Gulin (2018). "CatBoost: Gradient Boosting with Categorical Features Supported. ." Data Mining and Knowledge Discovery **33**(2): 576-588.

Emadi, M., A. Bagherzadeh, H. Fathi and A. Arabameri (2020). "Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran " Remote Sensing**12**(14).

Fabian, P., V. Gaël, G. Alexandre, M. Vincent, T. Bertrand, G. Olivier, B. Mathieu, M. Andreas, N. Joel, L. Gilles, P. Peter, W. Ron, D. Vincent, V. Jake, P. Alexandre, C. David, B. Matthieu, P. Matthieu and D. Édouard (2011). "Scikit-learn Machine learning in Python." The Journal of machine learning research**12**: 2825-2830.

Friedman, J. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." Annals of Statistics**29**(5): 1189-1232.

Gutierrez-Davila, M., F. J. Rojas and M. A Ortega (2013). "Response timing in the shot put: A comparison of elite and sub-elite athletes." Journal of Sports Sciences**31**(6): 631-640.

Gutiérrez-Davila, M., J. Rojas, J. Campos, J. Gámez and A. Encarnación (2009). "Biomechanical analysis of the shot put at the 12th IAAF World Indoor Championships." New Studies in Athletics**24**(3): 45-61.

Harleen, K. and K. Vinita (2018). "Predictive Modelling and Analytics for Diabetes using a Machine Learning Approach." Applied Computing and Informatics **18**(1): 90-100.

Hay, J. (1993). The biomechanics of sports techniques. Prentice Hall, Englewood Cliffs.

Hore, S. B., T. (April. 2018). A machine learning based approach towards building a Sustainability Model for NBA players. 2nd Int. Conf. Inventive Commun. Comput. Technol. (ICICCT).

Jordan, M. and M. Mitchell (2015). " Machine learning: trends, perspectives, and prospects." science **349**(6245): 255–260.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Ma and T.-Y. Liu (2017). " LightGBM: A highly efficient gradient boosting decision tree." Advances in Neural Information Processing Systems**30**: 3146-3154.

Kelleher, J. and B. Tierney (2018). A Comprehensive Introduction to Concepts, Techniques, Tools, and Applications. , Cambridge, MA: MIT Press.

Kursa, M. and W. Rudnicki (2010). "Feature selection with the Boruta package." Journal of Statistical Software **36**(11): 1-13.

Kyriazis, T., G. Terzis, G. Karampatsos, S. Kavouras and G. Georgiadis (2010). "Body composition and performance in shot put athletes at preseason and at competition. ." International Journal of Sports Physiology and Performance**5**(4): 417-424.

Liaw, A. and M. Wiener (2022). "Classification and Regression by randomForest." R News**2**(3): 18-22.

Linthorne, N. (2001). "Optimum release angle in the shot put " Journal of Sports Sciences **19**(5): 359-372

Lipovšek, S., B. Škof, S. Štuhec and M. Čoh (2011). "Biomechanical factors of competitive success with the rotational shot put technique. ." New Studies in Athletics**1**(2): 101-109.

Mackala, K., M. Fostiak and K. Kowalski (2015). "Journal of Human Kinetics: Name of the journal." **45**(1): 135-148.

Maier, T., D. Meister, S. Trösch and J. Wehrlin (2018). "Predicting biathlon shooting performance using machine learning." Journal of Sports Sciences **36**(14): 1620-1627.

Meäyk, E. and O. Unold (2011). "Machine learning approach to model sport training." Comput. Hum. Behav**27**: 1499-1506.

Mürsel, O., U. Murat, B. Işık and Ç. Murat (2022). "Using Machine Learning Algorithms for Jumping Distance Prediction of Male Long Jumpers." Journal of Intelligent Systems: Theory and Applications.**5**: 145-152.

Musa, R. M., P. Anwar, Z. Taha, S. Chang, A. Fakhri and M. Abdullah (2019). "A machine learning approach of predicting high potential archers by means of physical fitness indicators." PLOS ONE **14**(1): 1-12.

Natekin, A. and A. Knoll (2013). "Gradient boosting machines, a tutorial." Frontiers in neurorobotics**7**: 21.

Ofoghi, B., J. Zeleznikow, C. MacMahon and D. Dwyer (2010). A Machine Learning Approach to Predicting Winning Patterns in Track Cycling Omnium. International Conference on Artificial Intelligence (AI) / Held as Part of World Computer Congress Brisbane, Australia.

Ohyama, K., H. Fujii, M. Murakami, T. Endo, H. Takesako, K. Gomi and K. Tauchi (2009). " A biomechanical analysis of the men's shot put at the 2007 World Championships in Athletic." Bulletin of Studies in Athletics of JAAF**5**: 132-142.

Pavlović, R. and M. Vrcić (2011). "Influence of anthropometric and kinematic parameters on result successfulness of shot put finalists " Journal of Physical Education Research **6**(1): 29-37.

Prokhorenkova, L., A. Gusev, A. Vorobev, A. Dorogush and V. Ershov (2018). "CatBoost: Unbiased boosting with categorical features." Advances in Neural Information Processing Systems **31**: 1-10.

Rodríguez, G., M. Riera, L. López Elvira, T. Bermejo and A. Jódar (2002). "Biomechanics of the Shot Put: A Study of the Individual Technical Model." Archives of Sports Medicine **19**(89): 187-195.

Schofield, M., J. Cronin, P. Macadam and K. Hébert-Losier (2019). "Rotational shot put: A phase analysis of current kinematic knowledge." Sports Biomechanics **21**(3): 278-296.

Schofield, M., J. Cronin and P. J. McNair (2019). "The role of biomechanics in optimizing shot put performance." Journal: Sports Biomechanics **18**(2): 123-135.

Shalev-Shwartz, S. and S. Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. . United state of America, Cambridge University Press.

Shiliang, S. (2013). "A survey of multi view machine learning." Neural Comput. Applic. **23**: 7-8.

Terzis, G., T. Kyriazis, G. Karampatsos and G. Georgiadis (2012). "Muscle strength, body composition, and performance of an elite shot-putter." International Journal of Sports Physiology and Performance **7**(3): 394-396.

Thomas, A., A. Dinsdale, A. Bissas and S. Merlino (2018). "Biomechanical report for the World Indoor Championships 2018: Shot put men.".

Whiteside, D., O. Cant, M. Connolly and M. Reid (2017). "Monitoring Hitting Load in Tennis Using Inertial Sensors and Machine Learning." International Journal of Sports Physiology and Performance **12**(9): 1212-1217.

Wijayasekara, G., K. Shyamala and F. Zawawi (2022). "Missing Value Imputation Techniques in Machine Learning: A Comprehensive Review." Data Science and Engineering **7**(3): 201-228.

Woods, C., J. Veale, J. Fransen, S. Robertson and N. Collier (2017). "Classification of playing position in elite junior Australian football using technical skill indicators." Journal of Sports Science **36**(1).

Zatsiorsky, V. (1997). Kinematics of Human Motion. Champaign, Human Kinetics.