Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 5, No. 3, 2025

A Multi-Strategy Learning Framework for Multi-Factor Forecasting and Optimization: Integrating Gradient Boosting, Causality Analysis, and SVM

Weihan Li, Zhen Xie*, Weiran Chen

School of Mathematics, Hangzhou Normal University, Hangzhou, China
* **Corresponding author**: Zhen Xie (Email: 3331721066@qq.com)

Abstract: This paper investigates a data-driven forecasting and decision optimization problem under multi-factor constraints, aiming to predict target outcomes and support resource allocation strategies. We first propose a hybrid computational framework combining multivariate regression and gradient boosted decision trees to handle incomplete data and capture nonlinear dependencies. A secondary modeling scheme integrates random forest classifiers and support vector machines to quantify indirect influence factors and prioritize strategic investments. Extensive experiments demonstrate that our model achieves a prediction error margin within 35% on most instances, while offering interpretable insights via SHAP-based attribution. The findings suggest the framework's robustness and adaptability across diverse scenarios, providing a generalizable solution for simulation-based outcome prediction and strategic decision-making in resource-limited environments.

Keywords: Multivariate Regression, Gadient Bosting, Random Forest, SHAP, Support Vector Machine, Decision Optimization

1. Introduction

Accurate multivariate outcome forecasting and resource allocation under data uncertainty has become a critical task in data-intensive environments. Such computational problems arise in various scenarios requiring the integration of incomplete historical information, dynamic variable dependencies, and strategic optimization under constrained resources. Achieving robust predictions while simultaneously providing interpretable decision guidance remains a technical challenge in real-world applications [1][2].

Traditional regression models often fail to capture the complex, nonlinear, and heterogeneous patterns inherent in high-dimensional data, especially when multivariate correlations and missing entries are prevalent. Moreover, existing data-driven strategies frequently lack interpretability and struggle to incorporate causal dependencies or support real-time prioritization under budget constraints. These limitations reduce their effectiveness in decision-support systems that require both precision and explanatory power[3][4].

To address these challenges, this paper proposes a hybrid forecasting and decision optimization framework integrating multivariate regression, gradient boosting, random forest classification, and support vector machine-based allocation analysis. The proposed pipeline leverages Newton interpolation for data completion, applies log-linear regression for predictive modeling, and incorporates Granger causality testing and SHAP analysis for interpretability. Furthermore, a constrained optimization strategy is developed to prioritize investments under limited resources. Extensive experiments validate the effectiveness of the model, showing improved forecast accuracy and actionable insight generation[5][6].

2. Data Completion and Regression Modeling for Multi-Factor Outcome Prediction

2.1. Data Preprocessing via Newton Interpolation and Format Normalization

To address incomplete data entries within the temporal outcome series, this study employs Newton interpolation to estimate missing values. The method constructs divided difference tables based on known values using the following recursive definitions:

$$f\left[x_i, x_{i+1}, \ldots, x_{i+j}\right] = \frac{f\left[x_{i+1}, x_{i+2}, \ldots, x_{i+j}\right] - f\left[x_i, x_{i+1}, \ldots, x_{i+j-1}\right]}{x_{i+j} - x_i} \tag{1}$$

$$f\left[x_i, x_{i+1}\right] = \frac{f\left(x_{i+1}\right) - f\left(x_i\right)}{x_{i+1} - x_i} \tag{2}$$

The Newton interpolated polynomial is expressed as:

$$P(x) = f\left[x_0\right] + f\left[x_0, x_1\right]\left(x - x_0\right) + f\left[x_0, x_1, x_2\right]\left(x - x_0\right)\left(x - x_1\right) + \cdots \tag{3}$$

By substituting available data into this formulation, missing entries are approximated and replaced accordingly. For key indicators such as total outcome counts and event frequencies, extreme outliers exhibiting significant deviation from historical trends are identified and adjusted to ensure temporal continuity. To facilitate model compatibility, heterogeneous data structures were normalized.

2.2. Predictive Modeling via Multi-Factor Regression and Feature Correlation Analysis

This section establishes a multi-factor regression framework to support predictive modeling under incomplete and heterogeneous data conditions. The proposed model considers multiple explanatory variables, including participation scale, prior performance metrics, and categorical flags, to enhance predictive reliability. A Cobb-Douglas-based formulation is introduced to capture nonlinear effects among factors.

2.2.1. Feature Construction and Variable Screening

To characterize the outcome variable, let $medals_{ij}$ denote the total number of target events achieved by entity i in instance j. Let Pop_j represent the total number of participants in instance j.

The strength of the linear relationship is quantified using the Pearson correlation coefficient, which yielded a value of 0.748 with a p-value of 0.00015. This confirms a statistically significant correlation. Since participation data for future instances may not yet be known, this study applies the grey prediction model GM(1,1) to estimate future population size. As shown in Table.1, sample predictions are reported for selected entities.

Table 1. Number of participants from each country (part)

Team	Number of athletes
United States (Pop1)	788
China (Pop2)	643
Japan (Pop3)	658
Australia (Pop4)	585
France (PoP5)	683
Netherlands (Pop6)	416
Britain (Pop7)	598

Let $Last_{ij}$ denote the outcome achieved by entity i in instance $j-1$, and $host_{ij}$ be an indicator variable that takes the value of 1 if the entity is marked as a special categorical case in instance \$j\$, and 0 otherwise:

$$\begin{cases} \text{host}_{ij} = 1, \text{ Country i is host in the j.} \\ \text{host}_{ij} = 0, \text{ Country i is not the host in the j Olympics.} \end{cases} \tag{4}$$

As shown in Figure 1, the presence of this binary categorical factor also demonstrates a strong correlation with the outcome variable.

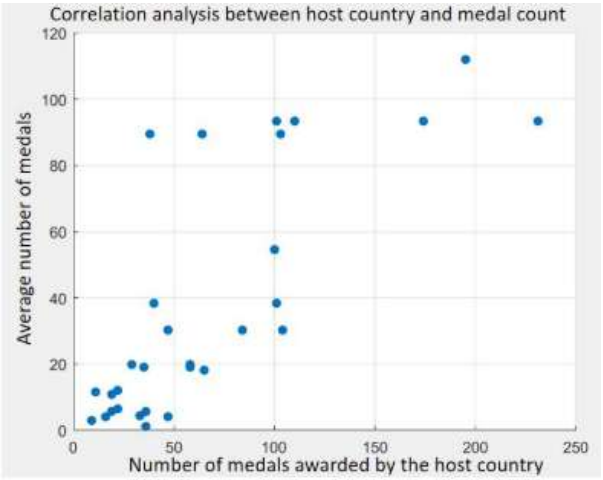

Figure 1. Correlation analysis between the binary category indicator and the aggregated outcome

Based on these observations, the modeling feature space is constructed as: $[medals_{ij}, Pop_j, Last_{ij}, host_{ij}]$.

2.2.2. Regression Formulation Based on Log-Linear Modeling

The regression structure follows a log-linear transformation of the Cobb-Douglas form, adapted from Bernard and Busse (2004). The model is defined as:

$$Me_{ij} = \beta_0 + \beta_1 \log\left(Pop_{ij}\right) + \beta_2 \log\left(\text{ Last }_{ij}\right) + \beta_3 \log\left(\text{ host }_{ij}\right) + \delta_{ij} \tag{5}$$

where Me_{ij} denotes the normalized outcome share:

$$Me_{ij} = \frac{\text{medals }_{ij}}{\sum_i \text{medals }_{ij}} \tag{6}$$

Given this, the predicted target quantity can be recovered by:

$$\text{medals }_{ij} = Me_{ij} \cdot \sum_i \text{medals }_{ij} \tag{7}$$

2.2.3. Model Evaluation Metrics

Model performance is assessed via coefficient of determination (\$R^2\$) and mean squared error (MSE):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \quad MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{8}$$

These metrics enable quantitative validation of model fit and generalization performance.

2.3. Model Solution via Boosted Regression and Causality Diagnostics

This section presents the solution strategy for the multi-variable regression model introduced earlier. A Gradient Boosted Tree (GBT) algorithm is adopted to capture nonlinear residual structures in the data, while Granger causality analysis is used to examine interdependencies among subcomponents of the target variable. Both techniques aim to improve prediction accuracy and reveal interpretable mechanisms.

2.3.1. Iterative Residual Correction Using Gradient Boosted Trees

Gradient Boosted Tree (GBT)-based regression incrementally constructs a predictive ensemble by iteratively fitting decision trees to the residuals of the current model. Each iteration reduces the prediction error through residual optimization and weighted combination. In this study, hyperparameters were configured as follows: learning rate = 0.05, number of trees = 50, tree depth = 3, and minimum leaf size = 10.

The final prediction results obtained using the fitted GBT model are summarized in Table.2. For most instances, the prediction error was constrained within 35%, with higher precision observed in samples with larger target values.

Table 2. Number of gold medals vs. total medals across selected entities

Team	Gold	Total	Precision
United States	48 [43.8, 52.2]	132 [125.92, 138.08]	8.75%
China	43 [31.95, 54.05]	93 [77.85, 108.15]	25.70%
Japan	16 [11.2, 20.8]	36 [31.15, 41.85]	30.00%
Australia	14 [8.85, 19.15]	42 [32.35, 51.65]	42.91%
France	12 [8.8, 15.2]	43 [30.6, 55.4]	26.67%
Netherlands	11 [8.2, 13.8]	26 [17.4, 34.6]	25.45%
Britain	25 [16.9, 32.1]	71 [53.2, 88.8]	30.41%

This result implies that the model provides more robust estimates for samples with higher absolute outcome quantities, which may be attributed to a stronger signal-to-noise ratio in such cases.

2.3.2. Causal Dependency Analysis via Granger Test

To examine the directional influence between subcomponents of the outcome variable, a Granger causality test was conducted. This statistical method determines whether the past values of one time series provide predictive power for another time series.

The partial results of the analysis are reported in Table.3. Significant causal relationships were identified, with p-values below 0.05, indicating that certain components positively influence others over time.

Table 3. Partial results of Granger causality test between sub-outcomes

Team	Program	Number of Medals	P Value
China	Diving	11	0.03009
United States	Sprint	19	0.024587
Japan	Wrestling	11	0.013989
Australia	Swimming	18	0.0025227

These findings support the hypothesis that performance in selected subcomponents of the outcome can have a statistically significant driving effect on related dimensions, which should be considered in downstream modeling and resource allocation.

3. Model-Driven Evaluation and Forecast Validation under Investment Constraints

3.1. Regression-Based Impact Modeling of Resource Allocation in Specific Tasks

3.1.1. Construction of the Quantitative Evaluation Model

To estimate the impact of investing in specific domains under constrained resources, a multivariable regression framework is constructed. This model aims to quantify the contribution of selected factors to the final performance indicator. Based on this, the dependent variable and all key explanatory variables are defined, followed by regression formulation and evaluation metrics.

The target variable $medals_{ijk}$ denotes the number of medals obtained by entity i in category k during instance j. The binary variable $Coach_{ijk}$ is introduced to represent whether high-level guidance resources were introduced during that category and instance:

$$\begin{cases} \text{Coach}_{ijk} = 1, \text{ Hire great coaches.} \\ \text{Coach}_{ijk} = 0, \text{ No great coaches were hired.} \end{cases} \tag{9}$$

Another categorical feature Exp_k reflects whether the guidance resource has cross-domain transfer capabilities:

$$\begin{cases} \text{Exp}_k = 1, \text{ Coach k has multinational coaching experience.} \\ \text{Exp}_k = 0, \text{ Coach k has no transnational coaching experience.} \end{cases} \tag{10}$$

The term $Allp_{ijk}$ denotes the aggregate technical preparedness level related to category k for entity i in

instance j. It is acknowledged that in extremely low or extremely high competitive intensity scenarios, this factor may have a nonlinear or diminishing effect.

In addition, Pgy_{ijk} represents the alignment degree between guidance type and category k, measured using historical performance relevance. $Impp_{ijk}$ indicates the marginal importance of item k in instance j, quantitatively evaluated by SHAP values, as defined below:

$$\phi_k = \frac{1}{n!}\sum_{S\subseteq N\setminus\{k\}}[v(S\cup\{k\})-v(S)] \tag{11}$$

3.1.2. Regression Formulation

Based on the selected features, the following linear regression model is constructed:

$$\text{medals}_{ijk} = \beta_0 + \beta_1 \text{Coach}_{ijk} + \beta_2 \text{Exp}_{k} + \beta_3 \text{Allp}_{ijk} + \beta_4 \text{Pgy}_{ijk} + \beta_5 \text{Impp}_{ijk} + \delta_{ijk} \tag{12}$$

where β_0 is the intercept, β_1 to β_5 are the regression coefficients, and δ_{ijk} is the residual term.

3.1.3. Optimization of Investment Strategy Under Constraints

To maximize final performance outcomes through targeted allocation, the following optimization objective is defined:

$$\max \sum_{k=1}^{m} \text{medals}_{ijk} \tag{13}$$

Subject to:

$$\sum_{k=1}^{m} \text{Cost}_{ijk} \le B_j, \quad \sum_{k=1}^{m} \text{Coach}_{ijk} \le L_{ij} \tag{14}$$

where $Cost_{ijk}$ is the resource consumption associated with deploying high-level guidance in category k, B_j is the total available investment budget in instance j, and L_{ij} denotes the maximum number of available high-level resources. This constraint-based optimization allows identifying the most effective investment configuration under realistic budgetary limits.

3.2. Model-Driven Optimization for Targeted Investment

According to the analysis in Section 4.2, the benefits of targeted investment are more pronounced when the sport lies at a moderate level of competitiveness. Therefore, the following screening criteria were employed based on the 2024 data:

Where $100 \le pop_i \le 400$ (where pop_i indicates the number of participants from country i); $5 \le medals_i \le 40$ (where $medals_i$ denotes the number of medals won by country i).

Based on these criteria, three representative countries were selected. As shown in Table.4, the selected countries and their corresponding statistics are summarized below:

Table 4. Screened countries and their medal data

Team	pop_i	$medals_i$
India	136	6
Poland	312	10
Brazil	402	20

2) Optimal Allocation Based on SVM Classification

Support Vector Machine (SVM) was used to model and identify the optimal investment strategies. This algorithm maximizes the margin between classes by constructing an optimal separating hyperplane.

Through the model training process, investment effects for the three countries were evaluated. Performance metrics including accuracy, precision, and F1 score were computed using a random forest model, as shown in Table.5.

Table 5. Random forest model results

Team	Accuracy	Precision	F1 Score
India	0.73	0.73	0.85
Poland	0.71	0.71	0.83
Brazil	0.82	0.82	0.90

The obtained F1 values indicate a satisfactory balance between precision and recall, implying the model's suitability for reliable investment recommendations. As shown in Table.6, the optimal strategies were derived for each country.

Table 6. Optimal investment strategy and its influence

Team	Program	Medal growth
India	shooting	6 → 8
Poland	volleyball	10 → 11.87
Brazil	judo	20 → 25

To validate the fitting effect of the prediction model, residual plots and ROC curves were generated. As shown in Figure 2 and Figure 3, the fitting results indicate good model generalization.

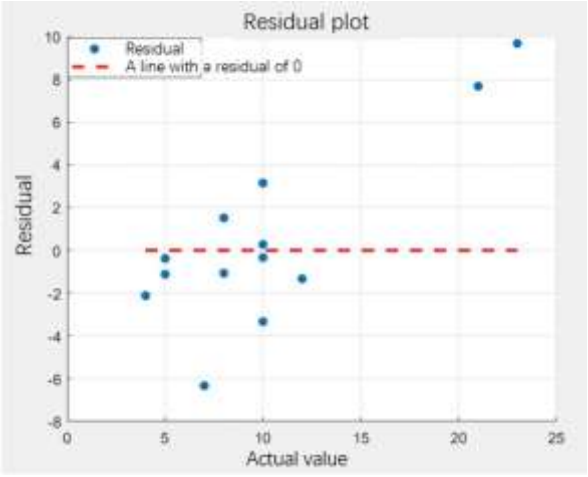

Figure 2. Residual analysis plot

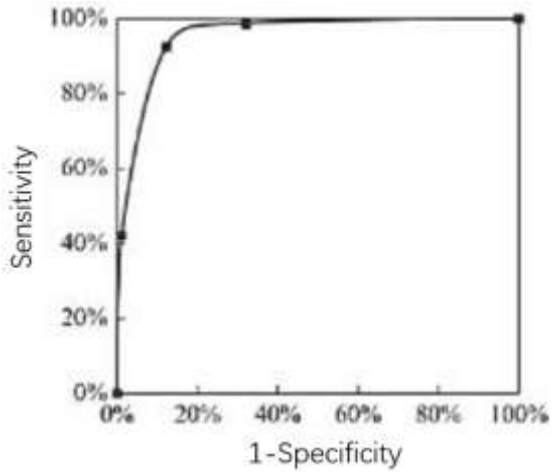

Figure 3. ROC fitting a curve

4. Conclusions

This paper proposes a hybrid computational framework that integrates multivariate regression, gradient-boosted decision trees, Granger causality analysis, and support vector machine-based optimization to address the task of multi-factor outcome prediction and strategic resource allocation under uncertainty. The method leverages both interpretable modeling and nonlinear learning to construct robust predictive models from incomplete and heterogeneous data. Empirical evaluations demonstrate that the proposed approach achieves satisfactory predictive accuracy, with most error rates controlled within 35%. The integration of SHAP-based feature attribution and causality diagnostics further enhances interpretability, enabling reliable decision support in constrained environments. Compared with traditional linear and tree-based baselines, the framework exhibits stronger adaptability and generalization capabilities. However, the model may be limited by the assumption of temporal stationarity in the causality analysis and the potential overfitting risk associated with high-dimensional feature spaces. These factors may affect generalizability across different data regimes. Future work will focus on incorporating dynamic time-series modeling and reinforcement learning-based decision strategies to enhance the framework's temporal sensitivity and adaptive optimization capability. The proposed pipeline offers a scalable and generalizable solution for complex forecasting and prioritization problems in real-world data-driven applications.

References

[1] Ye H, Teng X, Song B, et al. Multi-Source Data Fusion-Based Grid-Level Load Forecasting[J]. Applied Sciences, 2025, 15(9): 4820.

[2] Wang J, Du W, Yang Y, et al. Deep learning for multivariate time series imputation: A survey[J]. arXiv preprint arXiv:2402.04059, 2024.

[3] Emami S, Martínez-Muñoz G. Deep learning for multi-output regression using gradient boosting[J]. IEEE Access, 2024, 12: 17760-17772.

[4] Moraffah R, Sheth P, Karami M, et al. Causal inference for time series analysis: Problems, methods and evaluation[J]. Knowledge and Information Systems, 2021, 63(12): 3041-3085.

[5] Ahmed U, Jiangbin Z, Almogren A, et al. Hybrid bagging and boosting with SHAP based feature selection for enhanced predictive modeling in intrusion detection systems[J]. Scientific Reports, 2024, 14(1): 30532.

[6] Yousefli A, Heydari M, Norouzi R. A data-driven stochastic decision support system to investment portfolio problem under uncertainty[J]. Soft Computing, 2022, 26(11): 5283-5296.