Theory and Practice of Science and Technology
2025, VOL. 6, NO. 10, 234-241
DOI: 10.47297/taposatWSP2633-456946.20250610

Maximizing ROI in Youth Tennis: A Multi-Stage Modeling Approach for Data-Driven Investment Decisions

Kangyue Zhu

The High School Affiliated Renmin University of China, Beijing 100086, China

ABSTRACT

This study develops an optimization decision model to enhance the development of young tennis players and maximize investment returns. A stage classification model using the Random Forest Classifier predicts athletes' development stages (S0 - S4) based on seven-dimensional features, including training time and competition level. The proposed two-stage optimization framework first adjusts controllable variables to help athletes progress to the next stage while estimating intermediate variables through regression models. The second stage employs a multi-objective optimization model that incorporates actual costs and aims to balance consistency with economic feasibility, under realistic constraints like weekly training limits. This method combines machine learning and optimization techniques to offer personalized, cost-effective development suggestions for tennis families with limited resources, demonstrating strong interpretability and operability.

KEYWORDS

Youth sports investment; Dynamic programming; Bi-level 0ptimizatio

1 Introduction

1.1 Problem Statement

Having 9 years of training and competition experience as a national second-level tennis player, I noticed how exhausting the sport can be to the young athletes, and their families. Most parents cannot afford their kids, and even the highest-paid players such as Novak Djokovic or Jannik Sinner have expressed how the expenses and a lack of initial tennis help are high. These are the stories that are comparable to what I have seen. These inspired me to utilize modeling tools to understand how smarter and efficient investment decisions can be made by families.

Tennis has been considered as a sport of the rich, but there are numerous wealthless families that venture into the sport. Extra training and rivalry costs render the maintenance of excellent players with a mundane historical preparation hard. This compelled me to learn how to achieve the optimum training outcomes by using minimum resources. My objective is to develop a mathematical model to assist young Chinese players to opt in training hours, match schedules and development strategies that achieve maximum results with minimal cost. Literature on optimizing investment into tennis is almost nonexistent, so I created my framework initially with the help of the Random Forest Classification and Gradient Boosting Regression methods. I would also like this work to provide the young athletes with greater opportunities and make tennis more accessible.

1.2 General Modeling Framework

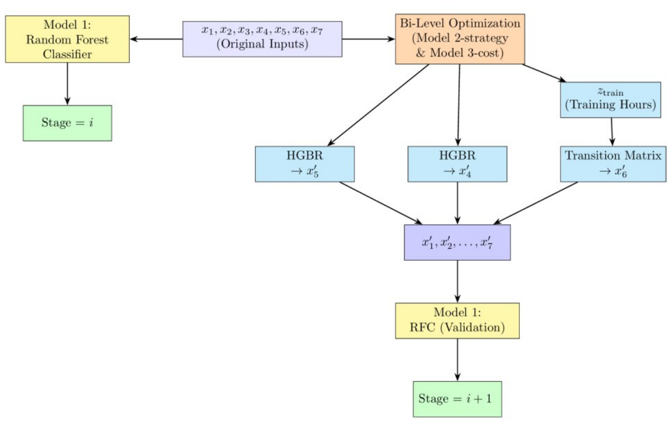

Figure 1 Bi-level optimization framework integrating regression models for stage advancement

2 Variables and Parameters

2.1 Explanatory Variables

The selection of variables for assessing tennis player development involves integrating theoretical insights from sports science with practical data collection challenges in youth sports. Key dimensions influencing progression include training volume and intensity, as skill acquisition necessitates sustained effort (Côté and Vierimaa 2014; Gulbin et al. 2013). Competition experience is vital for unique learning opportunities, while physical development is essential for performance, affected by training methods and maturation. Objective performance metrics provide standardized evaluations of tennis skills. Additionally, maintaining academic balance is critical, as long-term athlete development models stress the importance of academic success alongside athletic training. Note that we classified variable (x_3) (Tournament level) as an ordinal categorical variable, with the values 0–4 corresponding to different tiers of competition from the lowest amateur level (0) to the highest professional level (4). This classification is based on the hierarchical nature of tennis tournaments, where higher-level competitions typically feature stronger opponents, more rigorous match intensity, and greater pressure, all of which exert distinct impacts on a player's skill refinement and competitive adaptability. Meanwhile, variable (x_6) (Physical fitness), another ordinal categorical variable, is ranked from 1 (weak) to 4 (excellent) according to a standardized physical test battery for tennis players, covering indicators such as aerobic endurance, anaerobic power, agility, and strength. The binary classification of gender (x_7) is based on biological sex, a variable that may correlate with differences in physical characteristics, training adaptation, and competitive participation patterns in tennis. The continuous variables, including training time (x_1), win rate (x_4), and UTR score (x_5), are retained in their original quantitative form to preserve the precision of the data, as small variations in these metrics can reflect subtle changes in a player's training commitment and competitive performance. The discrete variable of annual match count (x_2) is treated as a count data type, as it represents the integer number of matches a player participates in per year, a key indicator of competitive exposure frequency.

Table 1 Definition of Variables Used for Athlete Classification

Variable Name	Symbol	Type	Description
Training time	$x_1 \in [0, 168]$	Continuous	Hours per week
Annual match count	$x_2 \in N^+$	Discrete	Matches per year
Tournament level	$x_3 \in \{0, 1, 2, 3, 4\}$	Ordinal categorical	Most frequently played type of Tournament
Win rate	$x_4 \in [0, 1]$	Continuous	Win percentage across all matches
UTR	$x_5 \in [0, 16]$	Continuous	UTR score; 0 if no UTR available
Physical fitness	$x_6 \in \{1, 2, 3, 4\}$	Ordinal categorical	Physical condition level from 4 (excellent) to 1 (weak)
Gender	$x_7 \in \{0, 1\}$	Binary categorical	Gender; 0 = Female, 1 = Male

Most Frequently Participated Tournament Level into four Subtype, Each having equivalent Level based on Tennis skill, Those four types are:

Table 2 Tournament Level Classification

Level	Tournament Category Description
0	Never participated in competition
1	Local or small-scale matches; point-based tournaments
2	Large city or national tournaments; below ITF J100
3	ITF J100 and above

In order to measure physical fitness, I design four levels of physical fitness level to classify a player's physical condition through a self-report system.

Table 3 Physical Fitness Level Classification

Level	Physical Fitness Description
4	Excellent: Able to handle high-intensity training, quick recovery, explosive power and endurance.
3	Good: Able to complete most training and matches, but occasionally fatigued or needs longer recovery.
2	Fair: Able to complete regular training, but often needs more rest due to fatigue.
1	Weak: Often fatigued during training and matches, slow recovery, limited physical support.

2.2 Youth Tennis Development Stage Variable

To evaluate the development of young tennis players, their growth is categorized into five stages (S0 - S4) from beginner to achievement realization. This classification relies on indicators like training time, competition frequency, and future development direction. These stage variables enhance classification models and support precise judgments in goal setting and path optimization, demonstrating theoretical validity and practical applicability in research and modeling.

Table 4 Developmental Stages of Tennis Players

Stage Variable	Stage Name	Stage Description
S0	Beginner Stage	Just started tennis; training time is very limited, movement is uncoordinated, and there is almost no experience in formal matches.
S1	Entry-Level Competition	Engages in regular weekly training, begins participating in small or local tournaments, occasionally wins; building experience and confidence.
S2	Progression Stage	Has developed a systematic training routine, frequently participates in city-level and higher tournaments; performance is steadily improving.
S3	Competitive/Professional	Training duration and intensity are close to professional levels; wins in high-level tournaments; ready to challenge stronger opponents professionally.
S4	Outcome Stage	Has earned scholarships or advancement opportunities through tennis achievements; approaching a high competitive level.

3 Model 1- random Forest Classification Model

This study classifies tennis players into stages S0–S4 using Random Forests, which build many randomized decision trees and combine their votes to improve accuracy and limit overfitting. Using scikit-learn, we apply this method to features like training time, match counts, win rate, and fitness to support reliable stage prediction.

3.1 Why RFC Fits this Study

Random Forest Classifier (RFC) is effective in categorizing junior tennis players into stages S0-S4 using various training and performance variables. It adeptly manages mixed data types without the need for scaling, making it suitable for datasets with continuous (training hours) and ordinal data (tournament level). RFC is particularly efficient for smaller datasets (100 to 200 players) and is robust against missing values, providing stable predictions. It captures the nonlinear aspects of tennis development, and its feature importance scores offer essential insights for coaches and families regarding training and investment planning.

3.2 Model Architecture and Implementation

Based on the theoretical background of the Random Forest method, we adopt the Random Forest Classifier (RFC) to help predict the developmental stage $y_i \in \{0, 1, 2, 3, 4\}$ of youth tennis players based on various performance and profile features.

3.2.1 Dataset Representation and Setup

We collected data from more than 100 young players, creating a dataset with n samples and seven predictors. Each player appears as one row in a feature matrix $X \in R^{n} \times ^{7}$, and the target vector y assigns a stage label from 0 to 4, covering levels S0 to S4. For missing feature values, we use class-wise median imputation so each empty entry x_{ij} is replaced with the median of its class, preserving class patterns.

3.2.2 Stratified Splitting

We divide the dataset into 80% training and 20% testing using stratified sampling:

$$X \rightarrow (Xtrain, Xtest), y \rightarrow (ytrain, ytest) \tag{1}$$

In the implementation of the Random Forest Classifier, 100 estimators are used to balance predictive performance and computational cost. The model consists of multiple decision trees, each trained on a bootstrap sample with random subsets of features. Prediction Function:

$$\hat{y} = majority\ voteh_m(x)_{(m=1)}^{M} \tag{2}$$

Each h_m is a decision tree, and the RFC predicts the majority class. It uses bagging, feature subsampling, and impurity-minimizing splits. Model performance on X_test is evaluated with Acc and macro F1, and variable importance is reported afterward.

$$Importance(x_j) = \frac{1}{M} \sum_{m=1}^{M} \sum_{t \in Tm} \Delta 1i(S_t, t) 1(feature\ used\ at\ t = x_j) \tag{3}$$

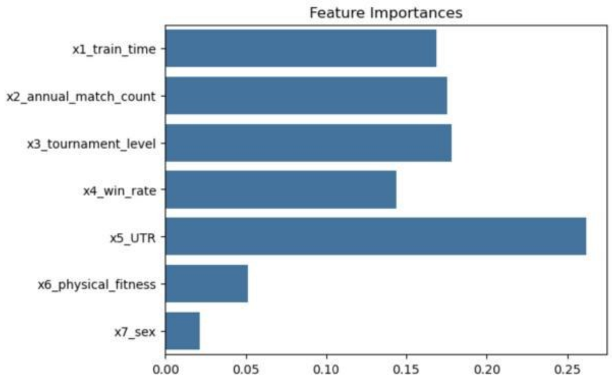

Figure 2 Feature Importances

```
Classification Report:

              precision    recall  f1-score   support

          S0       1.00      1.00      1.00         3
          S1       0.80      0.80      0.80         5
          S2       0.67      0.67      0.67         3
          S3       1.00      0.33      0.50         6
          S4       0.50      1.00      0.67         4

    accuracy                           0.71        21
   macro avg       0.79      0.76      0.73        21
weighted avg       0.81      0.71      0.70        21
```

Figure 3 Model Accuracy and F1 Score

The model shows UTR, match count, and tournament level as the strongest predictors of a player's stage. These factors reflect performance, exposure, and competitive level, allowing us to classify each young athlete's development using the full set of seven variables.

4 Model 2: Optimization for Minimal Adjustment Path to Ad- vancement

Our aim is to find the smallest adjustment to the player's feature vector x so the classifier produces a higher stage,

$$f(x') = y + 1 \tag{4}$$

We refine all seven inputs by controllability: training hours, match count, tournament level, and fitness are adjustable, while win rate, UTR, and gender are fixed. This structure supports focused, realistic development planning.

Table 5 Notation used for feature variables and model outputs

Symbol	Type	Description
$\mathrm{x} = (x_1, x_2, \ldots, x_7)$	Vector	Original input feature vector for one player
x_1	Controllable	Training hours per week
x_2	Controllable	Annual match count
x_3	Controllable	Tournament level (e.g., local, national, ITF)
x_4	Intermediate	Win rate (influenced by training and exposure, not directly adjustable)
x_5	Intermediate	UTR rating (intermediate outcome reflecting long-term development)
x_6	Intermediate	Physical fitness (can be improved through professional training)
x_7	Fixed	Gender
$y \in \{0, 1, 2, 3, 4\}$	Label	Predicted developmental stage output from Random Forest Classifier
x'	Vector	Adjusted input feature vector proposed by optimization model
$f(\cdot)$	Function	Trained Random Forest Classifier (RFC) model

4.1 Intermediate Variables Build Up

Our framework uses training effort and match count to shape win rate and UTR through gradient boosting models. These updated variables feed into a Random Forest Classifier to predict stage, and a bi-level optimizer then identifies the smallest changes needed for advancement.

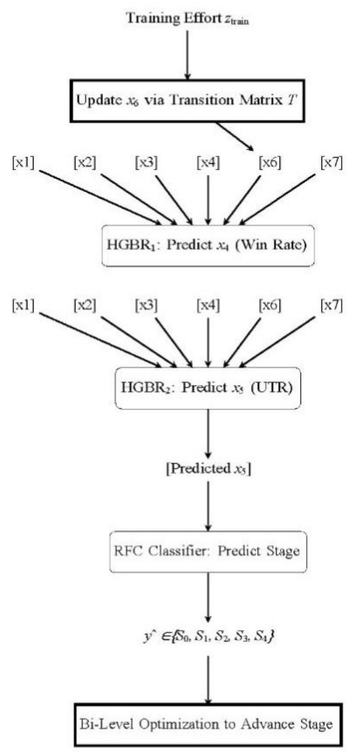

Figure 4 Two-Stage Prediction and Optimization Pipeline

4.1.1 Physical Fitness Level Modeling

In our framework, the fitness variable x6 ∈ {1, 2, 3, 4} reflects a player's physical condition and readiness, which shapes their predicted developmental stage. Although fitness is partly determined by natural ability, it improves through consistent training. In our optimization model, x6 influences the stage output f(x) from the Random Forest Classifier but cannot be set directly, since it changes with training investment. To capture this, we use a fitness advancement matrix linking weekly training hours to fitness progression, maintenance, or limited regression.

Fitness Advancement Matrix $T \in R^{4\times4}$Let:

The matrix defines how fitness progresses. Rows show the current level and columns show the target level. Each Tij gives the minimum weekly training hours needed to move from level i to j, while a dash marks an impossible transition. Given z_train, the model selects the highest level whose requirement is met. If training is adequate, the athlete moves up; if it only meets maintenance, they stay put; if it falls short, they drop one level but never below 1. For example, 16 hours lifts level 2 to 4, while 3 hours moves level 3 down to 2. This supports realistic progression modeling.

4.1.2 Win Rate and UTR – The HGBR Approach

In our two-stage optimization framework, win rate (x4) and UTR (x5) act as key performance indicators that guide training adjustments. Their nonlinear links to controllable variables are modeled with Histogram based Gradient Boosting Regression, which offers strong accuracy for small datasets:

Win Rate Estimator:

UTR Estimator:

$$\widehat{x}_4 = g(x_1, x_2, x_3, x_6, x_7) \tag{5}$$

$$\widehat{x}_5 = h(x_1, x_2, x_3, x_4, x_6, x_7) \tag{6}$$

Mathematical Framework Gradient Boosting is an additive model constructed iteratively:

$$\widehat{y} = F_M(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{7}$$

Each weak learner $h_m(x)$ is trained to minimize the loss with respect to the residual from the previous step:

When using squared loss:

$$r_i(m) = \frac{-\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \tag{8}$$

$$r_i(m) = yi - Fm - 1(xi) \tag{9}$$

Histogram-based Optimization To speed up split-point search in decision trees, HGBR discretizes con- tinuous features into B bins, drastically reducing the number of candidate splits per feature. This improves computational complexity from:

$$O\left(nlogn\right)\ to\ O(BlogB) \tag{10}$$

We implemented both estimators using HistGradientBoostingRegressor from scikit-learn, evaluated via three metrics:

Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \tag{11}$$

Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i) \tag{12}$$

R-squared score:

$$R^2 = 1 - \sum\frac{(y_i - \widehat{y}_i)^2}{(y_i - \bar{y})^2} \tag{13}$$

Model MSE MAE R^2

Table 6 Model Results for Intermediate Variable Estimation

Win Rate (x_4)	1.989	1.216	0.875
UTR (x_5)	0.019	0.116	0.669

Feature Importance Analysis To understand the contribution of each input feature, we employed Permu- tation Importance. The results are shown in the following pie chart:

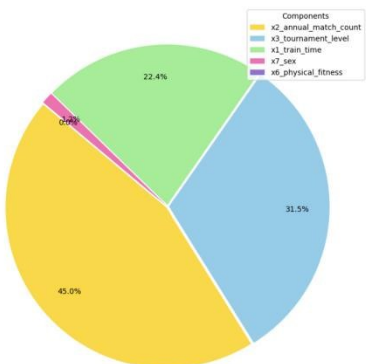

Figure 5 x_4 Feature Importance

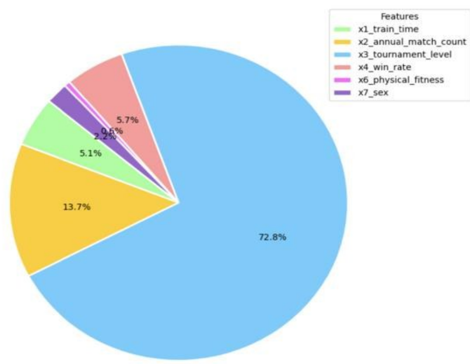

Figure 6 x_5 Feature Importance

4.2 Minimal Adjustment Strategy Optimization

To find the smallest change needed for a player to advance one stage, we optimize only the controllable variables x1, x2, x3, and x6. Win rate and UTR are updated through HGBR models, forming a bi-level system. The goal is to keep adjustments minimal while satisfying the constraint f(x′) =y+1. In practice, moving from S2 to S3 may require raising weekly training to 15 hours, increasing matches to 40, entering level 3 events, and improving fitness, with temporary dips in performance indicators expected.

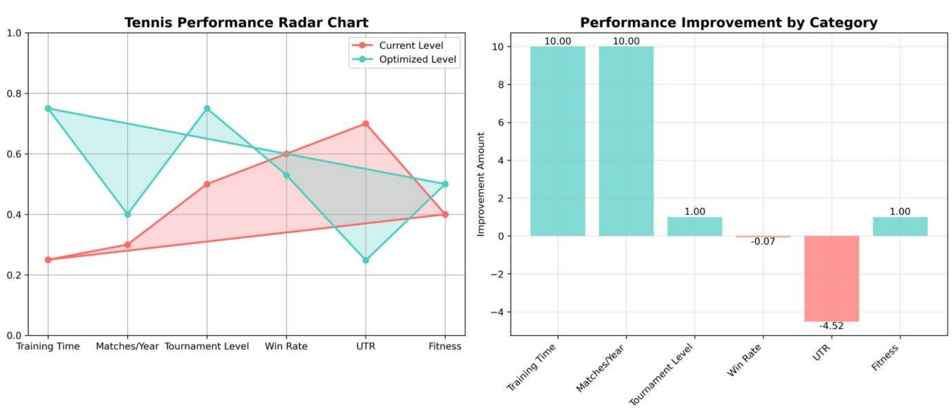

Figure 7 Tennis Performance Optimization. Radar chart (left panel) compares current and optimized perfor- mance across six tennis metrics. Bar chart (right panel) shows recommended improvements.

5 Model 3: Dynamic Optimization Model

5.1 Financial Cost By Category

We control four variables: training hours, annual match count, tournament level, and physical fitness, which rises through dedicated conditioning. Win rate and UTR cannot be set directly; they emerge from training and competition. Training cost varies by stage, with beginners paying lower hourly rates thanks to increased coaching accessibility in China's youth tennis market.

$$C_{training}(x_1|S_0)\in[200,300]RMB/hour \tag{14}$$

S1 Early Development: Players begin structured training and participate in small-scale competitions.

$$C_{training}(x_1|S_1)\in[300,500]RMB/hour \tag{15}$$

S2 Intermediate Competitive Stage: Players receive systematic training and frequently compete in regional tournaments,and the training frequency increases significantly.

$$C_{training}(x_1|S_2)\in[400,800]RMB/hour \tag{16}$$

S3 Pre-professional Stage: Athletes train at near-professional intensity and often compete in high-level events, Training cost varies more widely videly at this stage

$$C_{training}(x_1|S_3)\in[700,1500]RMB/hourThe \tag{17}$$

$C_{\text{training}}(x_1 \mid S_3) \in [700, 1500]$ RMB/hour The total weekly training investment is then modeled as:

$$Ctrainingtotal = x1\ Ctraining(x1|Si)RMB \tag{18}$$

Cost of Tounament Participation (x_2 and x_3)

Tournament cost is modeled as Ctournament(x3) multiplied by x2, with x3 defining four cost tiers. Level 0 has no fees. Level 1 local events cost about 200 RMB. Level 2 national events require travel and lodging, totaling roughly 12,800–13, 800 RMB. Level 3 ITF events, involving longer international trips, range from 22,000–24,000 RMB. This tiered structure supports better planning. Physical fitness cost is tied to weekly training hours, priced at 400 RMB per hour for the season.

5.2 Real-World Costs Minimization

We integrate stage prediction, targeted adjustment, and cost analysis into one cost-aware model. It balances minimal changes to a player's original plan with affordable financial investment. The final objective combines an L2 term to keep adjustments smooth and an L1 term to control expensive shifts:

$$\min_{\mathbf{x}'} \quad \lambda_1 \sum_{i=1}^{4} w_i (x_i' - x_i)^2 + \lambda_2 \sum_{i=1}^{4} c_i |x_i' - x_i| \tag{19}$$

$$\text{subject to:} \quad f(x_1', x_2', x_3', x_6') = y + 1 \tag{20}$$

$$x_1' \leq 42 \tag{21}$$

$$x_6' = g(z_6) \tag{22}$$

$$x_6' \in \{1, 2, 3, 4\} \tag{23}$$

The objective combines two penalties. The L2 term, weighted by w_i and λ_1, keeps adjusted variables close to their originals to prevent unrealistic shifts. The L1 term, scaled by real-world costs c_i and λ_2, encourages inexpensive, sparse adjustments. Together, they balance stability with cost-efficient training changes.

5.2.1 Stage Advancement Constrain

$$f(x_1, x_2, x_3, x_6) = y + 1 \tag{24}$$

This constraint enforces that the adjusted variable configuration, once passed through the trained classifier $f(\cdot)$, results in a developmental stage output that is strictly one level higher than the current stage y.

5.2.2 Weekly Training Constrain

$$x_1' \leq 42 \tag{25}$$

This constraint sets a maximum weekly training time to prevent overtraining, fatigue, and potential injuries. The limit of 42 hours/week is derived from a schedule of 6 training days per week and 7 hours per day.

5.2.3 Discretized Variable Handling Constrain

To handle the discrete fitness level variable $x'_6 \in \{1, 2, 3, 4\}$, we introduce an auxiliary continuous variable $z_{\text{train}} \in \mathbb{R}_{\geq 0}$, representing weekly physical training hours, so that:

$$x_6' = g(z_6) \tag{26}$$

where $g(\cdot)$ follows the fitness advancement matrix $T \in \mathbb{R}^{4\times 4}$ The optimization is conducted over $z_t rain$, so it will be subject to the induced constraint:

$$x_6' \in \{1, 2, 3, 4\} \tag{27}$$

This ensures compatibility with the classifier and training-cost model.

6 Conclusion and Results

This study introduces a multi-stage modeling framework that incorporates machine learning classification, bi-level optimization, and cost-aware decision making to enhance Return on Investment in youth tennis for underprivileged families. A Random Forest Classification model achieved 71% accuracy in categorizing players across developmental stages S0-S4, identifying UTR score, annual match count, and tournament level as key factors. The framework aids families in resource allocation, supports coaches in creating cost-effective programs, and equips policymakers with tools for targeted scholarships. Future research is recommended to include longitudinal validation and dataset expansion. This work illustrates the potential of advanced analytics in fostering equitable access to competitive tennis development.

References

[1] Cˆot´e, Jean and Matthew Vierimaa (2014). "The developmental model of sport participation: 15 years after its first conceptualization". *Science & sports* 29, S63–S69.

[2] Federation, International Tennis (Feb. 2025). *ITF World Tennis Tour Juniors Entry Fees FAQ*. Online doc- ument, ITF. Accessed: 2025-09-14. URL: https://www.itftennis.com/media/14141/entry- fee- faq- wttj.pdf.

[3] Friedman, Jerome H (2001). "Greedy function approximation: A gradient boosting machine". *Annals of statistics*, pp. 1189–1232.

[4] Geurts, Pierre, Damien Ernst, and Louis Wehenkel (2014). "Extremely randomized trees". *arXiv preprint arXiv:1407.7502*. URL: https://arxiv.org/abs/1407.7502.

[5] Gulbin, Jason P et al. (2013). "An integrated framework for the optimisation of sport and athlete development: A practitioner approach". *Journal of sports sciences* 31.12, pp. 1319–1331.

[6] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. 2nd. Springer.

[7] Ke, Guolin et al. (2017). "LightGBM: A highly efficient gradient boosting decision tree". *Advances in Neural Information Processing Systems*. Vol. 30.

[8] Tennis123 (2025). *Tennis Tournament Listings and Registration Platform*. Accessed: 2025-09-14. URL: https: //www.tennis123.net/match.

[9] Tikhonov, A. N. and V. Y. Arsenin (1977). *Solutions of ill-posed problems*. Washington, DC: Winston and Sons.

[10] Vicente, L. N. et al. (1996). "Mathematical Programming with Equilibrium Constraints". *Macmillan Ency-clopedia of Operations Research and Management Science*.

[11] Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the elastic net". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320.