Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v18n2p431

Statistical enhanced learning for modeling and prediction tennis matches at Grand Slam tournaments
By Buhamra, Groll

15 October 2025

Electronic Journal of Applied Statistical Analysis
Vol. 18, Issue 2, October 2025, 431-457
DOI: 10.1285/i20705948v18n2p431

Statistical enhanced learning for modeling and prediction tennis matches at Grand Slam tournaments

Nourah Buhamra*[a] and Andreas Groll†[a]

[a]*Department of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund*

15 October 2025

In this manuscript, we concentrate on a specific type of covariates, which we call statistically enhanced, for modeling tennis matches for men at Grand slam tournaments. Our goal is to assess whether these enhanced covariates have the potential to improve statistical learning approaches, in particular, with regard to the predictive performance. For this purpose, various proposed regression and machine learning model classes are compared with and without such features. To achieve this, we considered three slightly enhanced variables, namely elo rating along with two different player age variables. This concept has already been successfully applied in football, where additional team ability parameters, which were obtained from separate statistical models, were able to improve the predictive performance.

In addition, different interpretable machine learning (IML) tools are employed to gain insights into the factors influencing the outcomes of tennis matches predicted by complex machine learning models, such as the random forest. Specifically, partial dependence plots (PDP), individual conditional expectation (ICE) and accumulated local effect (ALE) plots are employed to provide better interpretability for the most promising ML model from this work. Furthermore, we conduct a comparison of different regression and machine learning approaches in terms of various predictive performance measures such as classification rate, predictive Bernoulli likelihood, and Brier score. This comparison is carried out on external test data using cross-validation, rolling window, and expanding window strategies.

keywords: Grand Slam tournaments, tennis matches, prediction, statistical enhance covariates, interpretable machine learning, expanding window.

* nourah.buhamra@tu-dortmund.de.
† groll@statistik.tu-dortmund.de.

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

1. Introduction

In recent years, various methodologies for statistical and machine learning based modeling of tennis matches and tournaments have emerged, and the existing methods for predicting the probability of winning matches in tennis have been expanded. Moreover, there is potential to calculate winning probabilities for an entire tournament when all individual matches can be predicted.

Recently, machine learning (ML) models have been employed to predict the outcomes of tennis matches. Somboonphokkaphan et al. (2009) introduced a method utilizing match statistics and environmental data to predict winners using a Multi-Layer Perceptron (MLP) with a back-propagation learning algorithm. MLP, a type of Artificial Neural Network (ANN), is effective for solving real-world classification problems and predicting outcomes, especially when handling large databases with incomplete or noisy data. Whiteside et al. (2017) proposed an automated stroke classification system to quantify hitting load in tennis, using machine learning techniques like a cubic kernel support vector machine. Wilkens (2021) expanded previous research by applying various ML techniques, including neural networks and random forests, with extensive datasets from professional men's and women's tennis singles matches. Despite these efforts, he found that the average prediction accuracy does not exceed 70%.

Sipko and Knottenbelt (2015) predicted match winners based on the probability of winning serve points, which subsequently indicates the overall probability of winning the match. They extracted 22 features from historical data, including abstract features like player fatigue and injury, and optimized models that outperformed Knottenbelt's Common-Opponent model using ML approaches such as artificial neural networks (ANNs). They suggest that ML-based techniques can innovate tennis betting, noting that ANNs generated a 4.35% return on investment, a 75% improvement in the betting market. Moreover, Gao and Kowalczyk (2021) developed a model that predicts tennis match outcomes with over 80% accuracy, surpassing predictions based on betting odds alone, and identifying serve strength as a crucial predictor. Their model used a random forest classifier, highlighting the importance of simple models even in the age of deep learning. Their comprehensive data set, compiled from ATP data from 2000 to 2016, includes a variety of features capturing physical, psychological, court-related, and match-related variables. Finally, a comprehensive overview of modeling and predicting tennis matches at Grand Slam tournaments by different regression approaches has been presented in Buhamra et al. (2024).

The main focus of this manuscript, however, is to analyze, whether so-called enhanced covariates have the potential to improve statistical and machine learning approaches, in particular, with regard to predictive performance. Generally, in recent years, there has been a growing interest in feature engineering. Effective feature engineering plays a crucial role in enhancing the performance of machine learning models by identifying and capturing relevant patterns and relationships within the data. This enables models to improve their predictive accuracy and extract meaningful insights from the data. For instance, Felice et al. (2023) introduce the concept of Statistically Enhanced Learning (SEL), a formalization framework for existing feature engineering and extraction tasks in ML. This approach has the potential to address challenges in ML tasks by optimizing feature selection and representation.

For example, in the context of modeling soccer, Ley et al. (2019) proposed a model to estimate flexible, time-varying team-specific ability parameters. The resulting estimates were then added to the set of (conventional) features in a random forest model, which turned out to be quite suc-

cessful for predicting the FIFA World Cup 2018 in Groll et al. (2019). Cefis and Carpita (2025b) introduce an enhanced expected goals (xG) model for football that combines tracking data, event data, and player performance indicators using logistic regression with sample-balancing techniques. Their model achieves both a significantly larger sensitivity and AUC compared to standard benchmarks, demonstrating improved goal prediction capability. For a recent discussion on the trade-off between predictive accuracy and model interpretability in xG modeling, see Cefis and Carpita (2025a).

Also, recent work has shown promising results in forecasting tennis outcomes using a variety of statistical and machine learning approaches. As an example, Candila and Palazzo (2020) employ neural networks for betting strategy optimization, while Del Corral and Prieto-Rodríguez (2010) assess the predictive value of ranking differences in Grand Slam matches. Meanwhile, Klaassen and Magnus (2003) take a more granular approach by modeling tennis at the point level to forecast match outcomes dynamically.

When using modern and complex ML models, another important aspect is interpretability of the fitted model. Hence, several studies have been conducted in the field of understanding and interpreting complex (black box-type) ML models. For example, Auret and Aldrich (2012) used variable importance measures, directly generated by the random forest models, and partial dependence plots, indicating that random forest models can reliably identify the influence of individual variables, even in the presence of high levels of additive noise.

Moreover, Goldstein et al. (2015) present both individual conditional expectation (ICE) plots and classical partial dependence plots (PDPs) on three different real data sets, namely depression clinical trial, white wine and diabetes classification in Pima Indians. They demonstrate how ICE plots can shed light on estimated models in ways PDPs cannot. Accordingly, ICE plots refine the PDP by graphing the functional relationship between the predicted response and the feature for individual observations. In particular, ICE plots highlight the variation in the fitted values across the range of a certain selected covariate, suggesting where and to what extent heterogeneities might exist.

More generally, Molnar et al. (2023) investigated the relationship between PDPs and permutation feature importance (PFI) methods in understanding the data generating process in machine learning models. They explored how these two techniques can provide complementary insights into the importance and effects of features on model predictions. Consequently, they formalize PDP and PFI as statistical estimators representing the ground truth estimands rooted from the data generating process. Their analysis reveals that PDP and PFI estimates can deviate from this ground truth not only due to statistical biases but also due to variations in learner behavior and errors in Monte Carlo approximations. To address these uncertainties in PDP and PFI estimation, they introduce the learner-PD and learner-PFI approaches, which involve model refits, and propose corrected variance and confidence interval estimators.

Unlike traditional black-box models, interpretable machine learning (IML) models aim to provide insights into the decision-making process, enabling users to make informed decisions and understand the implications of model outputs. Therefore, in this work, we focused on IML models, such as partial dependence plots (PDP; Friedman, 2001), individual conditional expectation (ICE; Goldstein et al., 2015) and accumulated local effect (ALE; Apley and Zhu, 2020) plots, to make our employed random forest model more interpretable. Additionally, we demonstrate how feature engineering techniques can be applied in the context of sports analytics to enhance

predictive modeling and gain insights from sports data. Specifically, we examine the use of covariates such as *Elo*, *Age.30* and *Age.int*, which all are not directly available, but are obtained from either a separate modeling approach (*Elo*) or via meaningful mathematical transformations (*Age.30*, *Age.int*), in order to enhance statistical models for tennis. For this application, various regression and machine learning approaches were considered, including linear regression, spline models, and random forest. These approaches were then compared based on various performance measures within an expanding window strategy for analyzing tennis data from Grand Slam tournaments. To conduct this analysis, a dataset was compiled using the R package `deuce` (Kovalchik, 2019). This data set included information on 6,586 matches from men's Grand Slam tournaments spanning the years 2011 to 2024. Several potential covariates were considered, including the players' age, ATP ranking and points, odds, elo rating, as well as two additional age variables. These additional age variables were designed to reflect the "optimal" age range of a tennis player, which is typically between 28 and 32 years (Weston, 2014).

The remainder of the article is structured as follows. Section 2 briefly introduces the data set and defines the objectives of this work. Then, in Section 3, different modeling approaches are introduced, including classical regression approaches and ML methods such as random forest. Besides, some interpretable machine learning techniques like partial dependence plots (PDP), individual conditional expectation (ICE) plots and accumulated local effect (ALE) plots are discussed, and various performance measures are defined. In Section 4, the modeling approaches are compared in terms of various performance measures, using an expanding window strategy. Additionally, interpretations for different model classes considering enhanced covariates and IML techniques are provided. Finally, Section 5 summarizes the main results and gives a final overview.

2. Data

Next, we shortly introduce our data set, which was compiled using the R package `deuce` (Kovalchik, 2019), and contains information on 6,586 matches from men's Grand Slam tournaments from 2011 to 2024. It is based on the data that was already used in Buhamra et al. (2024) and Buhamra et al. (2025), has been extended here and contains the following variables.

Victory: A dummy variable indicating whether the first-named player won the match (1: yes, 0: no), used as the response variable in all models. In addition, as we randomly assign which player is first- and which is second-named, the binary victory variable is almost evenly distributed, with wins accounting for approximately 50.08% and losses for 49.92% of the matches, indicating a balanced outcome distribution suitable for binary classification modeling.

Conventional covariates

The following three covariates are conventional covariates, which could be extracted more or less directly from public sources. They are all incorporated in our final data set in the form of differences, i.e. for each feature the value of the 2nd player is subtracted from the one of the 1st player.

Age: A metric predictor collecting the age of the players in years. Note that players' ages were not given directly and had to be deduced from the player's date of birth as well as the date of the respective match.

Rank: For each player, the rank at the start of the respective tournament was collected. The position in the ranking is based on the ATP ranking points.

Points The ATP world ranking points are awarded for each match won per tournament. Wins in later rounds of a tournament are valued higher than wins in the first rounds of a tournament. Points earned in a tournament expire after 52 weeks.

Note that principally, many more conventional covariates on the players could be collected, such as e.g. their height and handedness. Unfortunately, as those were not directly available to us (and scraping those would involve an enormous effort), we decided to keep the list of conventional covariates in this study rather short.

Enhanced covariates

Next, we introduce three covariates which we call enhanced, as they are not directly available, but are obtained from either a separate modeling approach (*Elo*) or via meaningful mathematical transformations (*Age.30, Age.int*), in order to enhance statistical models for tennis. Again, also these features are all incorporated as differences (value of 2nd player minus value of 1st player).

Elo: For each player, the overall Elo rating before a certain match is collected. The idea of the Elo rating system is that one can more accurately track and predict player performances over time, taking into account the relative strength of opponents and the outcomes of matches. Each player starts with an initial rating (often set around 1,500 points, though this can vary depending on the specific implementation). After each match, the players' ratings are updated based on the match outcome. If a higher-rated player wins, they gain fewer points than if a lower-rated player wins. The amount of points exchanged in a match depends on the difference in ratings between the two players. Hence, this covariate is considered "enhanced" as it involves complex calculations and provides a more delicate measure of player performance. Further details on the Elo rating in tennis can be found in Angelini et al. (2022) and Vaughan Williams et al. (2021)

Age.30: This variable is given by the absolute distance between the age of the players and reference age 30. This is based on the assumption that the standard *Age* variable introduced above does not contain enough information. For example, while a 25-year-old player typically has an age-advantage over a 20-year-old one, a 40-year-old player rather has a disadvantage over a 35-year-old one; and, in both cases, the age difference between the two players equals 5 years. Following Weston (2014), who argued that the optimal age of tennis players is between 28 and 32 years, we chose the mid-point as the reference age.

Age.int: This feature is based on the assumption that the optimal age of a tennis player lies within the interval $[28;32]$. Then, the smaller distance to the limits of this interval was derived, i.e. for players younger than 28 the distance to 28 was calculated and for players

older than 32 the distance to 32 was calculated. For players between 28 and 32 the distance was set to 0.

Note that another very informative enhanced feature, which is also typically easy to access, is based on bookmakers' odds. These are typically derived by sophisticated models, as the bookmakers' profit depends on them being adequate[1]. Moreover, further descriptive statistics such as mean, standard deviation (SD), median and quantiles for both conventional and enhanced covariates are provided in the appendix for clarity (see Table 3).

A detailed description of the variables included in the data set can be found in Section 2 of Buhamra et al. (2024) and Buhamra et al. (2025) .

Note at this point that the data set does not include matches in which one of the two players gave up or was unable to compete, e.g. due to injury, such that the other player won without actually playing the match. These matches do not contain any relevant information for the present analysis and, hence, in order not to distort the results, are excluded. Moreover, the data set does not contain any missing values.

Based on this data set, the best possible statistical enhance learning model for predicting tennis matches at Grand Slam tournaments is sought. Also, it will be examined whether a machine learning approach, namely a random forest, incorporating enhanced statistical covariates, is more powerful in predicting tennis matches compared to classical regression approaches that also incorporate the corresponding enhanced covariates. Then, for all proposed modeling approaches, including machine learning and classical regression methods, optimal models are determined based on certain performance measures in terms of an expanding window prediction approach. Here, our focus will be only on the expanding window strategy, which reveals a clear winner model, but also other technique such as leave-one-tournament-out cross-validation and a rolling window approach have been considered. The results for those approaches can be found in the appendix. Our main objectives are (i) to quantify how the predictive performance can be improved by incorporating enhanced variables, and (ii) to provide better interpretations for the random forest model using IML tools such as PDPs, ICE and ALE plots.

3. Methods

In the following, first the statistical and machine learning methods used in this work are briefly introduced in Section 3.1. We focus on both standard logistic regression, and generalized additive models based on P-splines. Moreover, the random forest as a representative from the field of machine learning is shortly presented. Then, in Section 3.2 several interpretable machine learning methods are explained, including partial dependence plots (PDP), individual conditional expectation (ICE) plots and accumulated local effect (ALE) plots. Finally, in Section 3.4 various performance measures are defined.

[1]However, due to the funding arrangements of one of the authors, who is supported by a ministry in a country where betting is strictly regulated, we have chosen not to further explore this feature in our analysis.

3.1. Statistical and machine learning methods

In this section, we introduce the classic logistic regression model for binary outcomes, followed by its extension to non-linear effects via spline-based approaches.

Logistic regression

Given observations $(y_i, x_{i1}, \ldots, x_{ip})$ for $i = 1, \ldots, n$ tennis matches,

$$\pi_i = P(y_i = 1 | x_{i1}, \ldots, x_{ip}) = E[y_i | x_{i1}, \ldots, x_{ip}]$$

is the (conditional) probability for $y_i = 1$, i.e. Player 1 winning the match, given covariate values $x_{i1}, \ldots, x_{ip}$.

We further specify a strictly monotonically increasing *response function* $h\colon \mathbb{R} \to [0,1]$,

$$\pi_i = h(\eta_i) = h(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}), \tag{3.1}$$

which relates the linear predictor η_i to π_i.

The logistic regression model, which uses the sigmoid function as response function, is the most famous candidate within the framework of Generalized Linear Models (GLMs).

The corresponding estimates $\hat{\beta}_0, \ldots, \hat{\beta}_p$ are obtained by numerical maximization of the underlying log-likelihood, e.g. by using Fisher scoring or the Newton-Raphson algorithms (see, e.g., Nelder and Wedderburn, 1972). This approach is implemented in the `glm` function from the `stats` package in R. For more details on GLMs, see Fahrmeir and Tutz (2001).

Spline-based approaches

In the classic logistic regression model introduced above, the covariates effects are strictly linear, see equation (3.1). However, in practice also non-linear influences might be relevant. In order to model these appropriately and flexibly, the GLM from above be extended to a so-called Generalized Additive Model (GAM; Wood, 2017). For this purpose, so-called *splines* can be used. In this work, we focus on *B-splines* (see, e.g., Eilers and Marx, 2021).

So instead of linear effects like the $\beta_j x_{ij}$ in equation (3.1), with B-splines a non-linear effect $f(x)$ of a metric predictor can be represented as

$$f(x) = \sum_{j=1}^{d} \gamma_j B_j(x),$$

where $B_j(x)$ are different B-spline basis functions, d denotes the number of basis functions used, and γ_j the corresponding spline coefficients. As an unpenalized estimation of a non-linear B-spline effect often overfits, typically the non-linear effect is smoothed by using *penalized B-splines*, i.e. *P-splines*.

Beside the problem of potential overfitting, the goodness-of-fit of the B-spline approach depends on the number of selected nodes. To avoid this problem, various penalization methods exist in the form of P-splines. Here, a penalized estimation criterion, which is extended by a penalty term, is used instead of the usual estimation criterion. For P-splines based on B-splines

see, e.g., Eilers and Marx (1996). For this approach, we rely on the `gam` function from the `mgcv` package (Wood, 2017) in R. Note that such P-spline based approaches have also been used in Buhamra et al. (2024) and Buhamra et al. (2025).

Random forest

In the following, a random forest will be used for comparison with the linear and non-linear regression approaches introduced above. This method is based on a (typically large) ensemble of trees, which were introduced by Breiman et al. (1984). However, as individual trees suffer from instability (Breiman, 1996b) an ensemble method called *bootstrapping and aggregating* (bagging; Breiman, 1996a) was introduced, which in general improves the predictive performance compared to a single regression tree and is rather easy to implement.

A *random forest* aggregates multiple decision trees to enhance prediction accuracy (Breiman, 2001). The key idea is that combining uncorrelated models (individual trees) reduces variance and improves predictions compared to using a single model. In this manuscript, classifications trees are considered in the *random forest* (where the most frequent class among the trees determines the outcome), as our target variable is binary. For this purpose, the `ranger` package in R by Wright et al. (2019) is used for fitting the random forest models. Also, the two hyperparameters `mtry` and`ntree` are quite important. For optimizing `mtry`, 10-fold cross-validation is used, and `ntree` is set to 400. Further details about these parameters can be found in Buhamra et al. (2025)

3.2. Interpretable machine learning

In the following, we discuss interpretable machine learning (IML) methods such as *partial dependence plots* (PDP; see Friedman, 2001), *individual conditional expectation* (ICE) plots (Goldstein et al., 2015), and *accumulated local effect* (ALE) plots (Apley and Zhu, 2020) in more detail. These methods aim to enhance the interpretability of complex, black box-type machine learning models. Particularly, they can be applied to such black box models to provide explanations for individual predictions or overall model behavior. For this purpose, the implementations from the R packages `pdp` by Greenwell (2017), `iml` by Molnar et al. (2018), and `ggplot2` by Wickham et al. (2016) are used (the latter one being generally applied for plotting). A nice overview of standard approaches for IML can be found in Molnar (2025).

Partial dependence plot (PDP)

Following Molnar (2025), the partial dependence plot (or PDP) illustrates the marginal effect of one or two features on the predicted outcome of a machine learning model (Friedman, 2001). It can reveal whether the relationship between a feature and the target is linear, monotonic, or more complex. The partial dependence function for regression is defined as follows

$$f_S(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}_C}[f(\mathbf{x}_S, \mathbf{x}_C)] = \int f(\mathbf{x}_S, \mathbf{x}_C)\, dP(\mathbf{x}_C)\,,$$

where $P(\cdot)$ is the distribution of the features in set C, $\mathbf{x}_S$ are the features for which the partial dependence function is plotted, while $\mathbf{x}_C$ represents the other features used in the machine learn-

ing model $f(\cdot)$, which are considered as random variables in this context. Typically, the set S contains only one or two features, namely those whose effect on the prediction one aims to understand. This implies that $\mathbf{x}_S$ denotes the features of interest and $\mathbf{x}_C$ represents the complementary features. Hence, one obtains a function that depends solely on the variables in S, while still accounting for interactions with the other variables.

The feature vectors $\mathbf{x}_S$ and $\mathbf{x}_C$ together form the complete feature space $\mathbf{x}$. Partial dependence operates by marginalizing the model's output over the distribution of the features $P(\cdot)$ in set C, allowing the function to reveal the relationship between the features in S and the predicted outcome. By doing so, we obtain a function depending solely on the features in S, while still accounting for interactions with other features.

In practice, the function f is unknown, so we estimate the partial dependence function by the fitted model $\hat{f}$. The empirical partial dependence function $\hat{f}_S$ is given by

$$\hat{f}_S(\mathbf{x}_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(\mathbf{x}_S, \mathbf{x}_{i,C}),$$

i.e. it is estimated by calculating averages on the training data. The partial function shows the average marginal effect on the prediction for given values of the features in S. Here, $\mathbf{x}_{i,C}$ denotes the actual values from the dataset for the features not of interest, and n represents the number of instances in the dataset. A key assumption of the PDP is that the features in C are uncorrelated with those in S. If this assumption does not hold, the calculated averages may include data points that are highly unlikely or even impossible. Further details can be found in Molnar (2025).

Individual conditional expectation (ICE)

The counterpart to a PDP (Friedman, 2001), which illustrates the average effect of a feature, is referred to as individual conditional expectation (ICE) plot and is applied for individual data instances. The approach was first introduced by Goldstein et al. (2015). ICE plots are used in machine learning to analyze the relationship between a feature and the predicted outcome for individual instances within a dataset. Unlike PDPs, which focus on the average effect of a feature, ICE plots offer insight into how changes in a specific feature or feature set impact the model's predictions for individual instances. Each line in an ICE plot represents the predicted outcome for a single instance as the feature value varies, revealing the variability and heterogeneity in feature effects across different instances. This helps identifying interactions between features, understanding complex model behaviors, and detecting outliers, thereby improving model interpretability and transparency.

3.3. Accumulated local effect (ALE) plots

To interpret the influence of individual predictor variables on the predicted outcome, we employ accumulated local effect (ALE) plots, as proposed by Apley and Zhu (2020). ALE plots offer a model-agnostic method for interpreting complex machine learning models by quantifying the local effect of input features on model predictions, while accounting for feature interactions and avoiding extrapolation beyond the observed data.

Unlike partial dependence plots (PDPs), which can produce biased estimates in the presence of correlated predictors, ALE plots estimate the local effect of a variable by computing finite differences in small intervals across the variable's domain. These local effects are then accumulated and centered, resulting in a global interpretation of how the variable affects the model prediction on average.

Let $\mathbf{x}_S$ denote the feature(s) of interest and $\mathbf{x}_C$ represent the complement set of features. The ALE function isolates the effect of $\mathbf{x}_S$ on the model output $f(\mathbf{x}_S, \mathbf{x}_C)$, while still capturing interactions with $\mathbf{x}_C$. The ALE function is computed as

$$\hat{f}_S^{ALE}(\mathbf{x}_S) = \int_{\mathbf{z}_{\min}}^{\mathbf{x}_S} \mathbb{E}_{\mathbf{x}_C}\left[\frac{\partial f(\mathbf{x}_S, x_C)}{\partial \mathbf{x}_S}\,\middle|\,\mathbf{x}_S = \mathbf{z}\right] d\mathbf{z}.$$

This integral is approximated using finite differences within quantile-based intervals of $\mathbf{x}_S$, averaged over the observed values of $\mathbf{x}_C$. The accumulated values are then centered so that the ALE function has a mean of zero, allowing interpretation in terms of relative effects. In our analysis, ALE plots are applied for the random forest model to assess and compare the relative importance of key predictors, such as *Points*, *Rank*, *Age.30*, and *Elo* on the probability of match victory.

3.4. Performance measures

In the following, performance measures are defined which we use to select the best model based on the predictive performance with regard to those measures (see also Buhamra et al., 2024, and Buhamra et al., 2025). Let $\tilde{y}_1, \ldots, \tilde{y}_n$ denote the true binary outcomes of a set of n matches, i.e., $\tilde{y}_i \in \{0,1\}, i = 1, \ldots, n$. Moreover, let $\hat{\pi}_{i1} =: \hat{\pi}_i$ denote the probability, predicted by a certain model, that player 1 wins match i. Then, the probability that player 2 wins the match is directly given by $\hat{\pi}_{i2} = 1 - \hat{\pi}_{i1}$.

Classification rate

The (mean) *classification rate* is given by the proportion of matches correctly predicted by a certain model, i.e.

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\tilde{y}_i = \hat{y}_i), \text{ where } \quad \hat{y}_i = \begin{cases} 1, & \hat{\pi}_i > 0.5 \\ 0, & \hat{\pi}_i \leq 0.5 \end{cases},$$

see, e.g., Schauberger and Groll (2018). Hence, large values indicate a good predictive performance.

Predictive Bernoulli likelihood

Following again Schauberger and Groll (2018), the (mean) *predictive Bernoulli likelihood* is based on the predicted probability $\hat{\pi}_i$ for the true outcome $\tilde{y}_i$. For n observations it is defined as

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_i^{\tilde{y}_i}(1-\hat{\pi}_i)^{1-\tilde{y}_i},$$

Once again, large values indicate good predictive performance.

Brier score

The *Brier score* (Brier, 1950) is based on the squared distances between the predicted probability $\hat{\pi}_i$ and the actual (binary) output $\tilde{y}_i$ from match i, and is defined as

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\pi}_i - \tilde{y}_i)^2 .$$

It is an error measure and, hence, low values indicate a good predictive performance.

4. Results

In the next section, the predictive power of the enhanced variables is presented for both regression and machine learning approaches. For this purpose, we use an expanding window (EW), a rolling window (RW) as well as a leave-one-tournament-out cross-validation (CV) approach. The best models are identified with respect to the previously defined performance measures. Additionally, we provide better interpretability of the machine learning approach, i.e. the random forest model, by using IML tools such as partial dependence plot (PDP), individual conditional expectation (ICE) and accumulated local effect (ALE) plots. All calculations and evaluations were performed using the statistical programming software R (R Core Team, 2024).

4.1. Enhanced variables predictive power

To investigate the predictive potential of so-called 'statistically enhanced covariates' in more detail, certain promising variables are considered, such as *Elo, Age.30* and *Age.int*, along with two conventional covariates (*Rank* or *Points*). All possible combinations of these covariates result in a total of 31 models for each of our proposed approaches, including linear effects, non-linear effects (splines), and the random forest. The results are given in Tables 1, with the best performers highlighted in bold. For the expanding window approach, results are presented here in detail, as this method clearly identifies the top-performing models among the 31 proposed. The results for the leave-one-tournament-out cross-validation and rolling window approaches are included in the appendix.

Expanding window validation

We validated all proposed 31 models with respect to their predictive performance on new, unseen test data. The validation is performed using an expanding window forecasting approach, i.e., each time one of the remaining tournaments is used as the test data set in chronological order, and the training data set is constantly updated and enlarged. This scheme has already been previously used in Buhamra et al. (2025) and can be explained as follows:

1. First, all tournaments prior to 2024 are used as the training data set. Then, all models are fitted. Based on those, predictions are derived for the 2024 Australian Open matches, as this was the 1^{st} Grand Slam tournament in 2024.

2. Then, the training data is updated, by adding the matches of the Australian Open 2024. Then, the models are fitted again on the extended training data, and predictions are made for the French Open 2024, which is the 2^{nd} Grand Slam tournament in 2024.

3. Next, the matches of the French Open 2024 are added to the training data set and the models are fitted again. Based on those fits, Wimbledon 2024 is predicted.

4. Then, the Wimbledon 2024 matches will be added to the training data set, and again, the models are fitted and predictions are made for the final Grand Slam tournament in 2024, the US Open 2024.

Finally, the prediction results for all four tournaments are compared with the actual match outcomes, and the corresponding performance measures are calculated .

Table 1 presents the predictive performance of regression models with linear and non-linear effects, alongside random forest models, based on 31 possible feature combinations. We ensured that each model included at most one age-based variable.

Linear model: The classification rates varied between 0.438 and 0.731. The winning linear regression model, which included *Points*, *Rank*, *Elo*, and *Age*, achieved a classification rate value of 0.731 and turned out to be the best performing model among all models with respect to the other two performance measures, yielding a predictive likelihood value of 0.654, and the lowest Brier score of 0.168.

Splines model: For the non-linear models the winning model is identified based on the enhanced variable *Elo* and other conventional variables such as *Points*, *Rank* and *Age*. Specifically, the splines model including covariates *Points*, *Rank*, *Elo* and *Age* demonstrates a good performance with respect to the predictive likelihood and the Brier score performance measures. The corresponding values are 0.653 and 0.169, respectively. The model based on *Age.int* and *Points* only, achieved the overall best classification rate of 0.733.

Random forest: Similar to the case of the spline-based approaches, the winning model among the random forests can be identified based on the results for the predictive likelihood and Brier score measures. The results show that the random forest includes covariates *Points*, *Rank*, *Elo* and *Age.30* performs slightly better than the other models. It yielded a classification rate of 0.721, a predictive likelihood of 0.624, and a Brier score of 0.179.

Overall, if we considered both the proposed model types (i.e., linear regression, spline-based regression and random forests) and different forecasting startegies (such as EW, CV, and RW), certain models consistently prevail when these enhanced variables are present. The results of both leave-one-tournament-out CV approach and the rolling window approach for the regression models (with linear and non-linear effects), as well as for the random forest model, are presented in Tables 5 and 6, respectively, in the appendix. In general, across all the tables presented in the appendix, the results suggest a nearly identical trend: prediction accuracy improves when at least one of these enhanced variables (i.e., *Elo*, *Age.30*, *Age.int*) is included.

Finally, it has to be noted that focusing just on the mean predictive performance with respect to certain performance measure,s as done in Tables 1, 5 and 6, might be a too simple way of comparison, as it ignores the uncertainty of these values. Hence, we follow the *Model Comparison Set* approach proposed by Hansen et al. (2011), exemplarily for the expanding window approach from Table 1. We always compare the overall best performing model (highlighted by

the dark gray cell background color) to all other models, based on their individual match predictions via suitable statistical paired comparison tests (one-sided; $\alpha = 0.05$). More specifically, we use the McNemar test for the classification rate, and a standard t-test for the predictive likelihood and Brier score. Whenever a model is signifficantly outperformed by the winner model, the respective cell background is highlighted in slight gray color.

In more detail, we can see that e.g. all models which are only based on the age variables (see first three rows from Table 1) are significantly outperformed with respect to all performance measures by the respective winner model. Moreover, it turns out that the best linear model with respect to predictive likelihood and Brier score significantly outperform almost all other models, particularly the random forest models. Overall, we find that those models which the winner model does not significantly outperform, often include one or several of the statistically enhanced features (*Elo*, *Age.30*, *Age.int*).

4.2. Model interpretation

In this section, our interpretation is developed based on the refitting of the top-performing linear, spline, and random forest models introduced in Section 4.1. Consequently, we provide deeper insights for these models. Additionally, interpretable machine learning tools such as a partial dependence plot (PDP), an individual conditional expectation (ICE) and an accumulated local effect (ALE) plot are used to enhance our understanding of the random forest model.

4.2.1. Linear model

The linear model can be directly interpreted based on the p estimated regression coefficients $\hat{\beta}_j$ from Equation (3.1). Their values indicate how much the outcome variable is expected to change when the corresponding predictor variable changes by one unit, assuming all other covariates in the model are held constant. Hence, they allow direct interpretation of both strength and direction of the relationship between the predictors and the outcome variable.

Table 2 shows the coefficient estimates for the best candidate linear model based on the results from the previous Section 4.1. It incorporates *Points*, *Rank*, *Age* as conventional covariates, and *Elo* as enhanced covariate. Note that all predictors have been standardized (mean = 0, standard deviation = 1) prior to model fitting. The outcome variable is binary (win/loss), and the coefficients represent changes in the log-odds of winning per one standard deviation increase in the predictor.

For the *Rank* difference variable, the negative coefficient for the standardized variable indicates that a one standard deviation increase in the rank difference (i.e., the first-named player is relatively lower ranked or worse compared to the opponent) leads to a decrease in the linear predictor by approximately 0.087 units. Due to the logistic link function, this also indicates a decrease in the predicted probability of winning for the first-named player.

Analogously, for a one standard deviation increase in the *Elo* rating difference (i.e., the first-named player has a relatively larger *Elo* rating compared to his opponent), the linear predictor increases by approximately 1.282 units. Due to the logistic link function, this again translates to a larger predicted probability of winning for the first-named player. In practical terms, this

Table 1: Results of the expanding window approach for linear, spline, and random forest approaches; best results are highlighted in bold font; dark gray cell background: overall winning models; light gray cell backgrounds: models that are significantly outperformed by winning model.

Model	**Linear**			**Spline**			**Random Forest**		
	Cr	LLH	Bs	Cr	LLH	Bs	Cr	LLH	Bs
Age	0.442	0.500	0.251	0.452	0.500	0.250	0.612	0.550	0.234
Age.30	0.438	0.499	0.251	0.500	0.499	0.251	0.577	0.539	0.242
Age.int	0.438	0.499	0.251	0.502	0.499	0.251	0.583	0.537	0.243
Points	0.721	0.635	0.180	0.727	0.644	0.180	0.706	0.593	0.192
Elo	0.719	0.642	0.173	0.721	0.642	0.173	0.723	0.600	0.186
Rank	0.721	0.591	0.200	0.719	0.601	0.200	0.717	0.574	0.201
Elo, Age	0.723	0.647	0.170	0.729	0.647	0.170	0.725	0.600	0.186
Rank, Age	0.723	0.593	0.199	0.710	0.603	0.199	0.717	0.581	0.200
Points, Age	0.723	0.637	0.179	0.723	0.646	0.179	0.723	0.602	0.189
Elo, Age.30	0.717	0.644	0.172	0.713	0.644	0.172	**0.727**	0.598	0.187
Rank, Age.30	0.729	0.593	0.198	0.719	0.603	0.198	0.719	0.573	0.202
Points, Age.30	0.725	0.635	0.180	0.731	0.644	0.179	0.715	0.591	0.193
Elo, Age.int	0.715	0.644	0.172	0.710	0.644	0.172	0.725	0.600	0.186
Rank, Age.int	0.727	0.593	0.198	0.713	0.603	0.198	0.719	0.574	0.201
Points, Age.int	0.729	0.635	0.180	**0.733**	0.645	0.179	0.710	0.591	0.193
Points, Rank	0.721	0.642	0.177	0.723	0.647	0.178	0.712	0.606	0.186
Points, Elo	0.715	0.648	0.172	0.721	0.648	0.172	0.702	0.618	0.181
Rank, Elo	0.725	0.644	0.172	0.723	0.643	0.173	0.706	0.609	0.184
Points, Rank, Age	0.717	0.645	0.176	0.721	0.648	0.178	0.710	0.606	0.187
Points, Elo, Age	0.729	0.653	0.169	0.727	0.653	0.169	0.702	0.619	0.183
Elo, Rank, Age	0.725	0.649	0.170	0.727	0.648	0.170	0.710	0.615	0.182
Points, Rank, Age.30	0.727	0.644	0.176	0.727	0.648	0.178	0.717	0.604	0.187
Points, Rank, Age.int	0.729	0.644	0.176	0.725	0.648	0.178	0.710	0.604	0.187
Points, Elo, Age.30	0.717	0.650	0.170	0.715	0.650	0.170	0.715	0619	0.182
Elo, Rank, Age.30	0.721	0.646	0.171	0.723	0.645	0.171	0.717	0.609	0.184
Points, Elo, Age.int	0.715	0.651	0.170	0.715	0.650	0.171	0.710	0.616	0.182
Elo, Rank, Age.int	0.715	0.651	0.170	0.721	0.645	0.172	0.706	0.609	0.184
Points, Rank, Elo	0.719	0.650	0.171	0.723	0.648	0.171	0.712	0.621	0.181
Points, Rank, Elo, Age	**0.731**	**0.654**	**0.168**	0.727	**0.653**	**0.169**	0.719	0.622	0.181
Points, Rank, Elo, Age.30	0.725	0.652	0.170	0.721	0.650	0.170	0.721	**0.624**	**0.179**
Points, Rank, Elo, Age.int	0.723	0.652	0.169	0.717	0.650	0.170	0.714	0.621	0.180

Table 2: Estimated coefficients for the candidate linear model based on the results from Table 1 for standardized predictors, along with standard errors and 95% confidence intervals

Predictor	**Estimate**	**Std. Error**	**z value**	$\Pr(>\|z\|)$	**2.5%**	**97.5%**
Rank	-0.087	0.042	-2.093	0.036	-0.171	-0.007
Elo	1.282	0.064	20.006	0.000	1.157	1.408
Points	0.400	0.069	5.797	0.000	0.267	0.538
Age	-0.174	0.030	-5.725	0.000	-0.234	-0.115

means that larger *Elo* rating differences in favor of the first-named player are strongly associated with increased chances of winning the match, as reflected by the large positive coefficient.

The same positive relationship also holds for the predictor variable *Points*. For every one standard deviation increase in the points difference (i.e., when the first-named player has accumulated relatively more ranking points than his opponent), the model predicts an increase of approximately 0.400 units in the linear predictor. Consequently, the winning probability for the first-named player also increases. This positive coefficient reflects the intuitive implication that players with a larger difference in ranking points are more likely to win their matches.

For the variable *Age*, we observe a pattern similar to that of the predictor variable *Rank*, indicating a negative relationship. Specifically, a one standard deviation increase in the age difference is associated with a decrease of approximately 0.174 units in the linear predictor.

4.2.2. Splines

Graphical illustrations of spline effects are a useful way to understand and visualize potential non-linear effects of predictors on the response variable. The spline-based approaches from Section 3.1 are able to capture non-linear relationships and allow for a flexible, semi-parametric form of regression using smooth functions (splines) for predictors. Corresponding ploting functions provide interpretable visualizations. In this context, the `gam` function from the `mgcv` package (Wood, 2017) in R is used.

Figure 1 displays the effects for the covariates *Elo*, *Age*, *Points* and *Rank*, which actually appear to be linear in most of these covariates.

In particular, differences in *Elo* show a strong increasing effect. The effect of *Age* difference is relatively flat across most of the range, but non-linear, with only minor deviations at the extremes; however, these are accompanied by wide confidence intervals, indicating that *Age* differences do not play a significant role in determining outcomes. For the *Points* difference, the effect is also positive, though a bit less pronounced as for *Elo*. In contrast, the effect for the *Rank* differences is negative. Finally, note that all effects are statistically significant.

4.2.3. Random forest with interpretable machine learning

Next, the results for the random forest model with the best predictive performance from Section 4.1 are discussed. This model, which included *Points*, *Rank*, *Elo*, and *Age.30* as covariates,

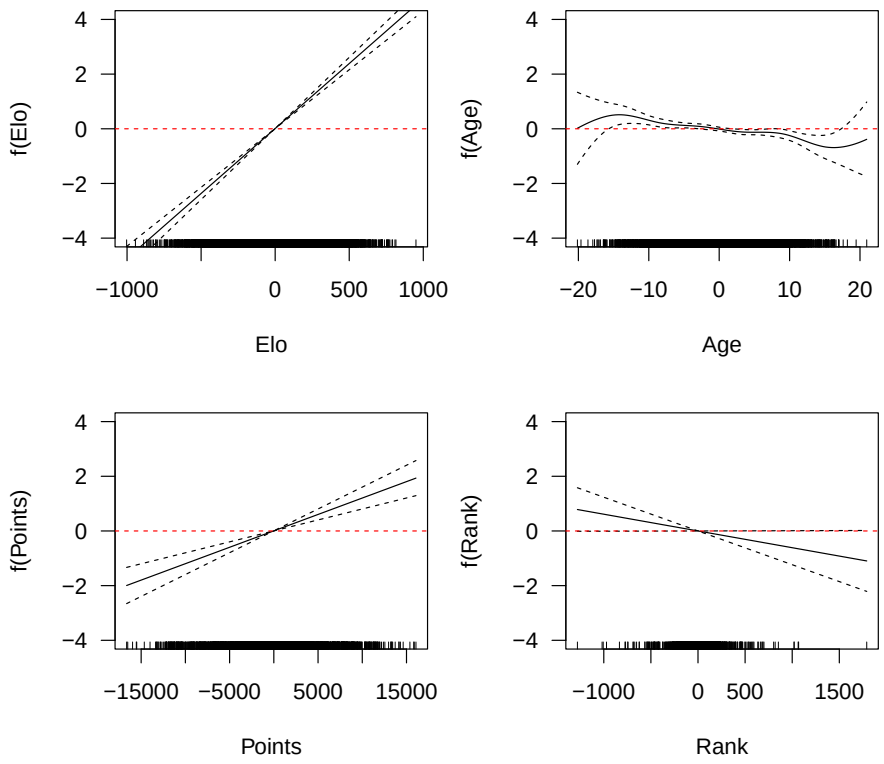

Figure 1: Spline effects for the covariates *Elo*, *Age*, *Points* and *Age*

is refitted to our entire data set. Then, both partial dependence plots (PDPs) and individual conditional expectation (ICE) plots are presented for better interpretability and visualization of the respective complex model. Furthermore, correlations and multicollinearity among *Elo*, *Rank*, *Points*, *Age*, *Age.int* and *Age.30* are assessed using the variance inflation factor (VIF; Kutner et al., 2004), see Table 4 in the appendix.

Since ICE plots—the individual equivalent of PDPs—are considered crucial in our case, the combined ICE plots and PDPs were used to provide a comprehensive visualization of the relationship between predictor variables and the predicted outcome.

In Figure 2, we generated a combined ICE plot and PDP. Each ICE line represents how the predicted winning probability changes with the predictor value for a single observation. The spread of the lines indicates variability in the effect of the predictor across different observations. The thick black line represents the PDP, showing the average effect of the respective predictor on the predicted outcome. This PDP line provides context to the ICE lines by highlighting the overall trend across all observations.

For instance, in the *Elo* panel, the PDP line increases steadily, showing that larger *Elo* differences (favoring the first-named player) are strongly associated with higher winning probabilities. The ICE lines closely follow this upward trend, confirming *Elo* as a dominant predictor.

In the *Rank* panel, the predicted probability of winning is first rather constant, then shows

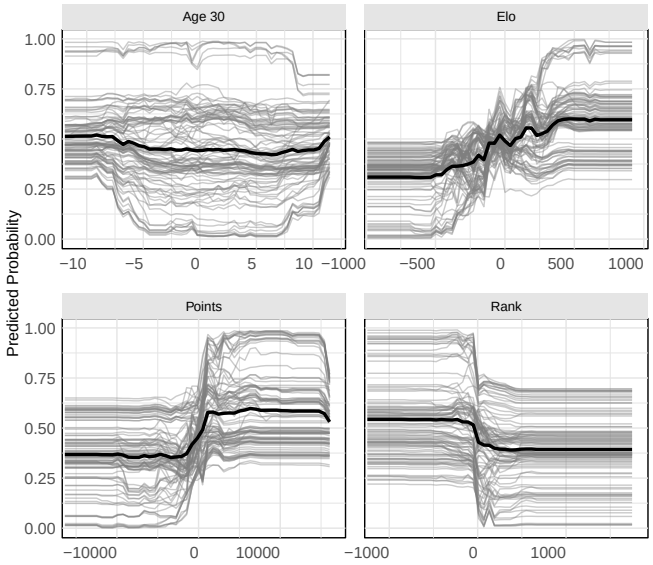

Figure 2: Combined PDP (thick black line) and ICE plots (grey lines) for the random forest model including covariates *Age.30*, *Elo*, *Points* and *Rank*

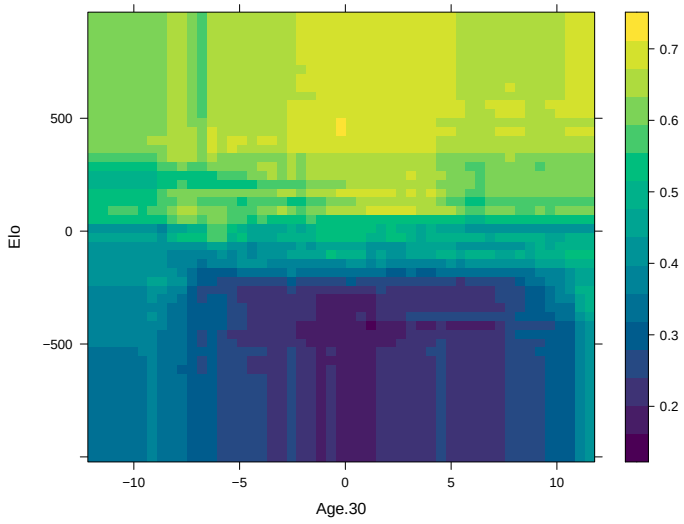

Figure 3: Heat map for the random forest model including *Elo* and *Age.30* as covariates

a steep decline as the *Rank* difference increases, followed by another nearly constant trend. The spread of the ICE lines indicates that individual responses vary considerably, but the general PDP trend suggests that larger positive *Rank* differences (i.e., the first-named player being ranked worse) reduce winning probability sharply for moderate differences, with only minor fluctuations before and thereafter.

For the *Points* panel, we observe the exact opposite trend compared to the *Rank* differences, just even a bit more pronounced, again with the individual ICE lines varying considerably.

Also in the *Age.30* panel, the ICE lines spread rather widely, reflecting heterogeneous effects. The PDP indicates a slightly decreasing trend as *Age.30* differences increase, with a slightly increasing trend for large, positive differences. However, many ICE lines deviate from this average, implying that for some individual players the effect is rather different.

Overall, this combined plot effectively conveys both the average influence of each predictor (PDP) and its varying impact on individual observations (ICE). It allows for precise interpretation of how predictors influence the outcome across different levels and individuals.

Finally, in Figure 3, a heat map plot for the partial dependency of *Elo* and *Age.30* is presented. The plot examines how the linear predictor (and, hence, the wining probability) changes jointly with *Age.30* and *Elo*. If the color changes are not uniform across the heat map, this suggests a complex interaction between *Elo* and *Age.30*. For example, larger *Elo* differences have a stronger impact on the predictor for certain age ranges than for others. (e.g., players with an age difference to their opponent closer to 0 benefit more from a larger Elo difference than players that are much younger or older than their opponent). Generally, a region that transitions from blue to yellow as *Elo* differences increase indicates that larger *Elo* scores are associated with larger winning probabilities, especially when *Age.30* differences are within a certain range.

Overall, this heatmap illustrates the joint effect of *Elo* and *Age.30* differences on the predicted probability of winning a tennis match. *Elo* differences exert the strongest influence, i.e., players with larger *Elo* ratings relative to their opponents have a substantially increased probability of winning, while those with lower *Elo* differences face reduced chances of success. In contrast, *Age.30* differences show only a moderate impact. Players who are considerably older or younger the their opponent exhibit only slight shifts in the predicted winning probability. Importantly, the interaction pattern indicates that *Elo* remains the dominant predictor, and even substantial age differences cannot offset the advantage provided by a larger *Elo* rating.

To better understand the contribution of individual features to the model's predictions, we visualize feature importance using an accumulated local effect (ALE) plot (see Figure 4). This plot shows how selected features influence the model's output on average, while avoiding the extrapolation issues often encountered in partial dependence plots (PDPs). Specifically, we calculate the range of ALE values for each feature, and use the ALE range in a bar chart to illustrate their relative importance.

The ALE plot provides insight into the localized influence of features on the model's predictions by evaluating the range of ALE values across observed feature differences. In our case, the *Elo* difference exhibits the largest ALE range, indicating that it has the most substantial local effect on the model output and is thus the most influential feature. This suggests that differences in players' *Elo* ratings substantially affect the predicted outcome, with the model being particularly sensitive to variations in this variable across matches. The *Points* difference and the *Age.30* difference show moderate ALE ranges, meaning they also contribute to localized prediction variability, though to a minor extent. In contrast, the *Rank* difference demonstrates the smallest ALE range, implying that it contributes the least to localized changes in the model's predictions and therefore holds the lowest relative importance among the considered variables.

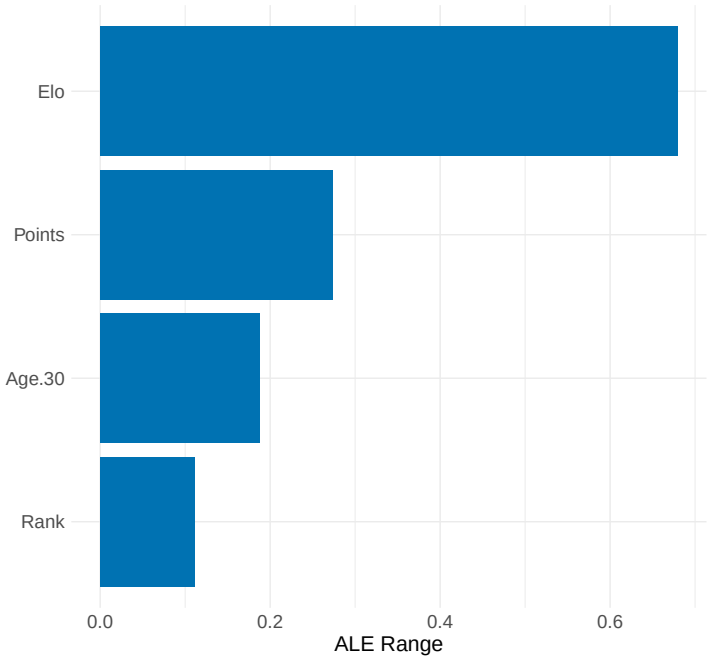

Figure 4: Variable importance via ALE-based range values, including *Elo*, *Points*, *Age.30* and *Rank* as covariates

5. Summary and overview

In this work, we compared different model approaches for modeling tennis matches in Grand Slam tournaments focusing on two main aspects: statistically enhanced covariates and interpretable machine learning tools. First, we demonstrated how these enhanced covariates can be applied in the context of sports analytics to improve predictive modeling performance and gain insights from sports data. Then, to better understand the interpretation of complex ML models, exemplarily for a random forest, we presented partial dependence plots (PDPs) to visualize the average partial relationship between the predicted response and one or more features, along with individual conditional expectation (ICE) plots, a tool for visualizing the model estimated by any supervised learning algorithm, and accumulated local effect (ALE) plots which provides insight into the localized influence of features on the model's prediction.

Moreover, note that we extended the data set provided and analyzed in Buhamra et al. (2024, 2025) to include the years 2023 and 2024, which were compiled using the R package `deuce` (Kovalchik, 2019).

It contains information on 6,586 matches in 55 men's Grand Slam tournaments from the years 2011-2024. It also includes covariate information on the age difference of both players (*Age*), the difference in their ranking positions (*Rank*) and ranking points (*Points*), in Elo numbers (*Elo*), as well as the two additional age-based variables, *Age.30* and *Age.int*, which were constructed such that they take into account that the optimal age of a tennis player is between 28 and 32 years.

Different regression models, which were already considered in Buhamra et al. (2024, 2025),

were compared with machine learning approaches, in particular a random forest model, for modeling and predicting tennis matches. Since there are only two possible outcomes in tennis (win or loss), all models were based on a binary outcome and thus focused on modeling the probability of the first-named player winning.

The different modeling approaches were compared with respect to their prediction performance on unseen matches via an expanding window strategy. The following regression and ML approaches were included in this comparison:

- Logistic regression with linear effects: all possible combinations of the three enhanced covariates along with the conventional variables *Point* and *Rank* were considered. This resulted in 31 models.
- Logistic regression with non-linear effects (splines): Again, the same 31 combinations of enhanced and conventional covariates were considered.
- A random forest model as a machine learning approach: Also here, the same combinations of enhanced and conventional covariates were considered, resulting in 31 models.

Via the expanding window approach, the models were compared in terms of classification rate, predictive Bernoulli likelihood and Brier score. Since each approach resulted in 31 different models, the model with the best predictive performance measures was selected in each case. Overall, the values vary between the different approaches and over proposed performance measures. The spline-based regression model based on *Points* and *Age.int* achieved the best classification rate among all other models with a value of 0.733. In contrast, the linear regression model including covariates *Points*, *Rank*, *Elo* and *Age* yielded the best predictive performance in terms of predictive likelihood and Brier score compared to all other model approaches. Generally, one could say that models consistently perform better when at least one of the enhanced variables is included.

Additionally, we investigated a CV-type approach and a rolling window approach. The rolling window approached was based on a (varying) training dataset always containing 12 tournaments that were used to predict the outcome of the next tournament. The results for these two approaches are provided in the appendix. Principally, results are varying among the three different training-test-subdivision approaches, but generally led to similar results as the expanding window strategy.

To gain a comprehensive understanding and interpretation of each approach, we analyze the coefficients of the linear regression model. Additionally, we employ spline graphs to visually represent and interpret the relationship between predictor variables and the response variable within the context of Generalized Additive Models (GAMs). These graphs can capture non-linear relationships, which is particularly advantageous for complex datasets where linear models may be insufficient. Furthermore, we introduce interpretable machine learning (IML) tools such as partial dependence plots (PDP) and individual conditional expectation (ICE) plots. These tools help in comprehending and interpreting the predictions made by complex (black box-type) machine learning models, in the present case a random forest model.

By examining PDP plots, we obtained insights into how each predictor variable affects the model's predictions. For instance, the PDP for *Points* exhibits a general upward trend, indicating

that earning more *Points* has a positive effect on the predicted outcome. Similarly, the PDP for *Age.30* suggests that being closer to the optimal age of 30 years is only slightly associated with larger winning probabilities.

In addition, a heat map visualization illustrating the joint effect of the two covariates *Elo* and *Age.30* on the predicted outcome is presented. By examining the color gradient and regions in the heat map, we can infer how changes in these two covariates influence the outcome, highlighting any non-linear interactions and dependencies between those.

Finally, to evaluate feature importance, we use an ALE-based importance plot. This method is model-agnostic and provides a clear measure of how much each feature influences the model's predictions. ALE calculates the average local effect of a feature across its value range. A larger ALE range means the feature has a stronger effect on predictions (such as *Elo* in our analysis). The bar plot shows this visually: features with larger bars have more influence, while those with shorter bars affect the model less. Unlike other methods such as permutation importance or the Gini index, ALE takes into account feature interactions and correlations. This makes it a more reliable and interpretable alternative to methods like PDPs, especially when features are related.

In future research, additional IML method such as local interpretable model-agnostic explanations (LIME; Ribeiro et al., 2016) can also be used. Furthermore, one could investigate more complex machine learning models, such as deep learning approaches (Bishop, 1995; LeCun et al., 2015). Additionally, similar to soccer (Groll et al., 2019), one could also focus on tournament outcomes. For example, the probability of a certain player winning the tournament could be determined. This approach takes advantage of the fact that the tournament bracket is fully drawn before the start, allowing us to predict the earliest round in which two players could meet. Unlike in soccer, where group stage outcomes influence subsequent matches, this setup simplifies predictions. However, using only the match-specific betting odds for the first round presents a challenge. Models that exclude odds as covariates might be preferable, despite potentially lower prediction performance due to the significant influence of odds. Alternatively, models could be developed that use pre-tournament odds for each player to win the entire tournament, rather than odds for individual matches. Moreover, additional statistically enhanced covariates could be produced. For example, similar to the historic match abilities for soccer teams developed by Ley et al. (2019), such abilities could also be developed for tennis players. We are currently working on such an approach and plan to include those ability parameters into our models in the future (first results in this direction can be found in Bartmann et al., 2025). Finally, as one of the anonymous reviewers pointed us to the `welo` R package (Candila, 2023), which enables direct access to ATP, WTA, and Grand Slam match data for both male and female players, we plan to analyze in future research to what extend our results can be carried over to professional women's tennis.

References

Angelini, G., Candila, V., and De Angelis, L. (2022). Weighted Elo rating for tennis match predictions. *European Journal of Operational Research*, 297(1):120–132.

Apley, D. W. and Zhu, J. (2020). Visualizing the effects of predictor variables in black box

supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):1059–1086.

Auret, L. and Aldrich, C. (2012). Interpretation of nonlinear relationships between process variables by use of random forests. *Minerals Engineering*, 35:27–42.

Bartmann, H., Groll, A., and Michels, R. (2025). Estimation of match abilities for tennis players via a maximum likelihood approach. In Boccuzzo, G., Bovo, E., Manisera, M., and Salmaso, L., editors, *Innovation & Society - Statistics and Data Science for Evaluation and Quality*, pages 480–487. Cleup.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

Breiman, L. (1996a). Bagging predictors. *Machine learning*, 24:123–140.

Breiman, L. (1996b). Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24(6):2350–2383.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and regression trees. wadsworth int. *Group*, 37(15):237–251.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Buhamra, N., Groll, A., and Brunner, S. (2024). Modeling and prediction of tennis matches at grand slam tournaments. *Journal of Sports Analytics*, 10(1):17–33.

Buhamra, N., Groll, A., and Gerharz, A. (2025). Comparing modern machine learning approaches and different forecasting strategies for modeling tennis matches at grand slam tournaments. *Journal of Sports Analytics*. to appear.

Candila, V. (2023). welo: an r package for weighted and standard elo rates. *Italian Journal of Applied Statistics*, 35(1):1–18.

Candila, V. and Palazzo, L. (2020). Neural networks and betting strategies for tennis. *Risks*, 8(3):68.

Cefis, M. and Carpita, M. (2025a). Accuracy and explainability of statistical and machine learning xG models in football. *Statistics*, 59(2):426–445.

Cefis, M. and Carpita, M. (2025b). A new xG model for football analytics. *Journal of the Operational Research Society*, 76(1):1–13.

Craney, T. A. and Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality engineering*, 14(3):391–403.

Del Corral, J. and Prieto-Rodríguez, J. (2010). Are differences in ranks good predictors for grand slam tennis matches? *International Journal of Forecasting*, 26(3):551–563.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:89–121.

Eilers, P. H. and Marx, B. D. (2021). *Practical smoothing: The joys of P-splines*. Cambridge University Press.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York, 2nd edition.

Felice, F., Ley, C., Groll, A., and Bordas, S. (2023). Statistically enhanced learning: a feature engineering framework to boost (any) learning algorithms. *arXiv preprint arXiv:2306.17006*.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29:337–407.

Gao, Z. and Kowalczyk, A. (2021). Random forest model identifies serve strength as a key predictor of tennis match outcome. *Journal of Sports Analytics*, 7(4):255–262.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65.

Greenwell, B. M. (2017). pdp: An R package for constructing partial dependence plots. *The R Journal*, 9(1):421–436.

Groll, A., Ley, C., Schauberger, G., and Van Eetvelde, H. (2019). A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports*, 15:271–287.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Klaassen, F. J. and Magnus, J. R. (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research*, 148(2):257–267.

Kovalchik, S. (2019). deuce: resources for analysis of professional tennis data. *R package version 1.4*.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2004). *Applied Linear Statistical Models*. McGraw-Hill/Irwin, 5th edition.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Ley, C., Wiele, T. V. d., and Eetvelde, H. V. (2019). Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches. *Statistical Modelling*, 19(1):55–73.

Molnar, C. (2025). *Interpretable Machine Learning*. 3 edition.

Molnar, C., Casalicchio, G., and Bischl, B. (2018). iml: An R package for Interpretable Machine Learning. *Journal of Open Source Software*, 3(26):786.

Molnar, C., Freiesleben, T., König, G., Herbinger, J., Reisinger, T., Casalicchio, G., Wright, M. N., and Bischl, B. (2023). Relating the partial dependence plot and permutation feature importance to the data generating process. In *World Conference on Explainable Artificial Intelligence*, pages 456–479. Springer.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, A 135:370–384.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, 41(5):673–690.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international*

conference on knowledge discovery and data mining, pages 1135–1144.

Schauberger, G. and Groll, A. (2018). Predicting matches in international football tournaments with random forests. *Statistical Modelling*, 18(5-6):460–482.

Sipko, M. and Knottenbelt, W. (2015). Machine learning for the prediction of professional tennis matches. *MEng computing-final year project, Imperial College London*, 2.

Somboonphokkaphan, A., Phimoltares, S., and Lursinsap, C. (2009). Tennis winner prediction based on time-series history with neural modeling. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, pages 18–20. Citeseer.

Vaughan Williams, L., Liu, C., Dixon, L., and Gerrard, H. (2021). How well do Elo-based ratings predict professional tennis matches? *Journal of Quantitative Analysis in Sports*, 17(2):91–105.

Weston, D. (2014). Using age statistics to gain a tennis betting edge. `http://www.pinnacle.com/en/betting-articles/Tennis/atp-players-tipping-point/LMPJF7BY7BKR2EY`.

Whiteside, D., Cant, O., Connolly, M., and Reid, M. (2017). Monitoring hitting load in tennis using inertial sensors and machine learning. *International journal of sports physiology and performance*, 12(9):1212–1217.

Wickham, H., Chang, W., and Wickham, M. H. (2016). Package 'ggplot2'. *Create elegant data visualisations using the grammar of graphics. Version*, 2(1):1–189.

Wilkens, S. (2021). Sports prediction and betting models in the machine learning age: The case of tennis. *Journal of Sports Analytics*, 7(2):99–117.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, London, 2nd edition.

Wright, M. N., Wager, S., Probst, P., and Wright, M. M. N. (2019). Package 'ranger'. *Version 0.11*, 2.

A. Appendix

The descriptive statistics in Table 3 are provided for the covariate differences of both competing players. As additionally, it was randomly assigned which player is first- and which second-named, all preprocessed features now have means close to being centered and exhibit roughly symmetric distributions. The *Rank* variable has a reasonable interquartile range from -48 to 46, while the *Points* variable shows wide variability, with a large standard deviation of approximately 3,283. Also the *Elo* variable has a rather large spread (SD $= 269.792$), indicating substantial variation in player skill levels. While also the *Age.30* variable has a mean very close to zero, it is worth mentioning that the median for *Age.int* is exactly 0.000, as due to its definition many observations have a zero value: the variable is a feature engineered to capture age effects around the optimal age interval of 28–32 years.

A VIF of 1 indicates no multicollinearity, meaning the variable is completely independent of the others in the model. When VIF $\in [1;5]$, this suggests moderate correlation, which is generally acceptable and not considered problematic. However, if VIF>5, this indicates a potential

Table 3: Descriptive statistics of conventional and enhanced covariates

Variable	**Mean**	**SD**	**Q1**	**Median**	**Q3**
Age	-0.065	5.861	-4.020	-0.020	3.869
Rank	-0.141	112.189	-48.000	1.000	46.000
Points	-27.423	3,283.117	-975.750	-4.000	954.500
Elo	-2.162	269.792	-184.500	-4.208	181.781
Age.30	-0.009	3.840	-2.519	-0.012	2.500
Age.int	-0.005	3.401	-2.100	0.000	2.087

Table 4: Variance inflation factor (VIF) results and interpretation for checking correlations and multicollinearity between *Elo, Rank, Points* and *Age* base features.

Variable	**VIF**	**Interpretation**
Elo	2.32	No concern, below 5.
Rank	1.41	Very low, no multicollinearity.
Points	1.97	Low, multicollinearity is not an issue.
Age	1.65	Also low, no concern.
Age.int	33.29	Very large, strong multicollinearity with other variables.
Age.30	32.63	Very large again, strong multicollinearity.

multicollinearity issue, and further investigation is warranted. A VIF > 10 signals serious multicollinearity, and in such cases, it is advisable to consider removing, combining, or transforming the involved variables to improve model stability and interpretability (see, e.g., Kutner et al., 2004; Craney and Surles, 2002; O'Brien, 2007).

Here, just for the two age-based variables very large VIF values are recorded, which is because they are rather similar and, hence, exhibit strong pairwise correlation. But as those two variables were never jointly incorporated in our models, this is not an issue.

Table 5: Results of the leave-one-tournament-out CV approach for linear, spline, and random forest models; best results are highlighted in bold font.

Specific Model	**Linear**			**Spline**			**Random Forest**		
	Cr	LLH	Bs	Cr	LLH	Bs	Cr	LLH	Bs
Age	0.490	0.499	0.250	0.489	0.499	0.250	0.641	0.563	0.222
Age.30	0.496	0.499	0.250	0.501	0.500	0.250	0.634	0.559	0.223
Age.int	0.497	0.499	0.250	0.503	0.500	0.250	0.629	0.560	0.224
Points	0.724	0.621	0.188	0.723	0.632	0.183	0.722	0.608	0.188
Elo	0.744	0.655	0.172	0.744	0.655	0.172	0.744	0.624	0.178
Rank	0.724	0.588	0.202	0.723	0.599	0.199	0.720	0.607	0.193
Elo, Age	0.724	0.630	0.183	**0.748**	0.657	**0.171**	**0.745**	0.623	0.178
Rank, Age	0.724	0.630	0.183	0.719	0.599	0.199	0.723	0.611	0.192
Points, Age	0.724	0.630	0.183	0.727	0.633	0.183	0.724	0.607	0.188
Elo, Age.30	0.743	0.655	0.172	0.743	0.655	0.172	0.741	0.622	0.179
Rank, Age.30	0.723	0.588	0.201	0.721	0.599	0.199	0.719	0.608	0.193
Points, Age.30	0.727	0.621	0.187	0.723	0.632	0.183	0.723	0.606	0.188
Elo, Age.int	0.743	0.656	0.172	0.743	0.656	0.172	0.744	0.623	0.179
Rank, Age.int	0.722	0.589	0.202	0.720	0.611	0.199	0.717	0.607	0.193
Points, Age.int	0.726	0.621	0.187	0.725	0.632	0.183	0.721	0.606	0.188
Points, Rank	0.724	0.630	0.183	0.724	0.636	0.182	0.724	0.624	0.183
Points, Elo	0.742	0.657	0.171	0.744	0.656	0.172	0.741	0.639	**0.175**
Rank, Elo	**0.745**	0.656	0.172	0.745	0.655	0.172	0.744	0.637	0.176
Points, Rank, Age	0.744	0.657	0.171	0.725	0.637	0.182	0.724	0.626	0.183
Points, Elo, Age	0.744	0.657	0.171	0.725	0.637	0.182	**0.745**	0.639	**0.175**
Elo, Rank, Age	0.744	0.657	0.171	0.725	0.637	0.182	0.744	0.636	0.176
Points, Rank, Elo	0.744	0.657	0.171	0.745	0.656	**0.171**	0.742	0.643	**0.175**
Points, Rank, Age.30	0.727	0.630	0.183	0.725	0.636	0.182	0.722	0.622	0.184
Points, Rank, Age.int	0.727	0.631	0.183	0.725	0.637	0.182	0.723	0.623	0.184
Points, Elo, Age.30	0.743	0.657	0.171	0.742	0.656	0.172	0.742	0.638	**0.175**
Elo, Rank, Age.30	0.743	0.656	0.172	0.744	0.656	0.172	0.742	0.635	0.176
Points, Elo, Age.int	0.742	0.656	0.172	0.742	0.656	0.172	0.739	0.637	0.176
Elo, Rank, Age.int	0.743	0.656	0.172	0.743	0.656	0.172	0.743	0.635	0.177
Points, Rank, Elo, Age	0.744	**0.659**	**0.170**	0.744	**0.658**	**0.171**	0.742	**0.644**	**0.175**
Points, Rank, Elo, Age.30	0.744	0.658	0.171	0.743	0.656	**0.171**	0.740	0.642	**0.175**
Points, Rank, Elo, Age.int	0.744	0.658	0.171	0.743	0.656	**0.171**	0.739	0.642	**0.175**

Table 6: Results of the rolling window approach for linear, splines, and random forest approaches; best results are highlighted in bold font.

Specific Model	**Linear**			**Spline**			**Random Forest**		
	Cr	LLH	Bs	Cr	LLH	Bs	Cr	LLH	Bs
Age	0.524	0.501	0.249	0.521	0.503	0.249	0.601	0.539	0.237
Age.30	0.539	0.505	0.248	0.536	0.505	0.248	0.595	0.536	0.240
Age.int	0.541	0.505	0.248	0.531	0.504	0.248	0.598	0.534	0.241
Points	0.718	0.617	0.189	0.719	0.631	0.185	0.708	0.607	0.196
Elo	**0.740**	0.650	0.174	0.738	0.651	0.174	0.723	0.626	0.185
Rank	0.718	0.592	0.201	0.717	0.605	0.198	0.697	0.601	0.203
Elo, Age	0.718	0.629	0.185	**0.742**	0.652	**0.173**	**0.733**	0.628	0.184
Rank, Age	0.718	0.629	0.185	0.715	0.606	0.198	0.699	0.601	0.200
Points, Age	0.718	0.629	0.185	0.720	0.632	0.185	0.709	0.612	0.195
Elo, Age.30	0.739	0.650	0.174	0.738	0.651	0.174	0.731	0.627	0.185
Rank, Age.30	0.712	0.593	0.200	0.716	0.606	0.198	0.704	0.602	0.200
Points, Age.30	0.719	0.618	0.188	0.719	0.631	0.185	0.709	0.609	0.195
Elo, Age.int	0.738	0.650	0.174	0.739	0.650	0.174	0.729	0.627	0.185
Rank, Age.int	0.713	0.593	0.200	0.713	0.606	0.198	0.700	0.603	0.200
Points, Age.int	0.718	0.618	0.189	0.719	0.631	0.185	0.704	0.612	0.196
Points, Rank	0.718	0.629	0.185	0.719	0.635	0.184	0.708	0.616	0.190
Points, Elo	0.739	0.652	0.173	0.740	0.651	0.174	**0.733**	0.628	0.181
Rank, Elo	0.739	0.651	0.174	0.739	0.651	0.174	**0.733**	0.628	0.181
Points, Rank, Age	0.722	0.629	0.185	0.720	0.635	0.184	0.709	0.617	0.190
Points, Elo, Age	0.738	0.653	**0.173**	0.719	0.635	0.184	0.732	0.631	**0.180**
Elo, Rank, Age	0.738	0.653	**0.173**	0.719	0.635	0.184	0.731	0.629	0.182
Points, Rank, Age.30	0.718	0.629	0.185	0.719	0.635	0.184	0.708	0.615	0.189
Points, Rank, Age.int	0.719	0.629	0.185	0.719	0.635	0.184	0.708	0.614	0.191
Points, Elo, Age.30	0.739	0.652	**0.173**	0.737	0.651	0.174	0.725	0.629	0.182
Elo, Rank, Age.30	0.739	0.651	0.174	0.740	0.651	0.174	0.729	0.627	0.183
Points, Elo, Age.int	0.739	0.651	0.174	0.739	0.651	0.174	0.725	0.627	0.182
Elo, Rank, Age.int	0.738	0.651	0.174	0.738	0.651	0.174	0.729	0.627	0.183
Points, Rank, Elo	0.738	0.653	**0.173**	0.738	0.652	0.174	0.728	**0.632**	0.181
Points, Rank, Elo, Age	0.738	**0.654**	**0.173**	0.738	**0.653**	**0.173**	0.727	**0.632**	**0.180**
Points, Rank, Elo, Age.30	0.739	0.653	**0.173**	0.737	0.652	0.174	0.729	**0.632**	0.181
Points, Rank, Elo, Age.int	0.738	0.653	**0.173**	0.737	0.652	0.174	0.724	0.631	0.181