Machine learning based statistical inference in sports analytics

Robert Bajons[1] and Lucas Kook[1]

[1] Institute for Statistics and Mathematics, WU Vienna, Austria

E-mail for correspondence: robert.bajons@wu.ac.at

Abstract: Identifying which factors are predictive of an outcome (e.g., scoring a goal) in the presence of other features (e.g., position of the shooter) is a fundamental task in sports analytics. In practice, this task is commonly addressed using feature importance measures derived from machine learning algorithms. However, such algorithms typically come at the cost of limited interpretability and invalid statistical inference. Here, we achieve valid inference by using machine learning based nonparametric conditional independence tests to (i) determine strong shooters based on goals above expectation in soccer and (ii) determine the influence of statistically derived motion features on defensive coverage schemes in the NFL. We further relate these tests to a partially linear logistic regression model to facilitate interpretation.

Keywords: Sports analytics; conditional independence tests; machine learning.

1 Introduction

Sports analytics, fueled by the recent availability of high-resolution tracking data, has experienced a surge in the use of advanced statistical and machine learning (ML) models. A key focus of these applications is identifying the factors that influence a game, for instance, identifying top players, predictors of injuries or factors influencing the final score [Kovalchik, 2023]. Another popular approach in sports analytics has centered around statistically enhanced learning models [Felice et. al., 2025], which use advanced statistical techniques for feature extraction. These features are then leveraged to predict relevant outcomes, such as match results.

Commonly, the task of identifying influential factors is tackled by fitting machine learning models and analyzing traditional variable importance measures. However, with the gain in flexibility due to ML, uncertainty quantification, valid statistical inference, and interpretation become challenging. In this work, we consider the problem of testing whether a binary response Y is independent of features X given potentially high-dimensional conditioning variables Z. The tests

This paper was published as a part of the proceedings of the 39th International Workshop on Statistical Modelling (IWSM), Limerick, Ireland, 13–18 July 2025.

directly translate into inference on parameters in a partially linear logistic regression model (PLLM). This allows the identification of features which may aid in outcome prediction in an interpretable way, without making strong modeling assumptions and maintaining type I error control.
We present two applications of our framework: First, we provide a novel statistical view on goals above expectation (GAX), a statistic traditionally used to evaluate the shooting skills of soccer players. Secondly, we analyze a statistically enhanced model to predict defensive coverage schemes in the NFL. In particular, we focus on testing whether features derived from a hidden Markov model (HMM) help in predicting the defensive scheme for a given play.

2 Methodology

Let (Y, X, Z), where $Y \in \{0, 1\}$, $X \in \mathbb{R}^{d_X}$, $Z \in \mathbb{R}^{d_Z}$, be a random vector in which Y is governed by a PLLM,

$$Y := \mathbb{I}\left(X^\top \beta + g(Z) > \varepsilon\right) \quad \text{and} \quad X := f(Z) + \eta, \tag{1}$$

where $\beta \in \mathbb{R}^{d_X}$, $\varepsilon \sim \text{Logistic}(0, 1)$, $\mathbb{E}[\varepsilon|X, Z] = 0$, $\mathbb{E}[\eta|Z] = 0$, g, f are arbitrary measurable functions, and $\mathbb{I}(\cdot)$ denotes the indicator function. We are interested in testing whether Y is conditionally independent of X given Z (in short: $Y \perp\!\!\!\perp X \mid Z$), which in the PLLM corresponds to $H_0 : \beta = 0$. Under H_0, the PLLM makes no modeling assumptions about the relationship between Y, and Z. Without strong assumptions on g, parametric inference on β is invalid in this setting. We instead rely on the nonparametric Generalised Covariance Measure (GCM) test [Shah and Peters, 2020]. The GCM is defined as $\mathbb{E}[(Y - \mathbb{E}[Y|Z])(X - \mathbb{E}[X|Z])]$ and is zero under H_0. Given independent and identically distributed observations $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, the GCM can be estimated using arbitrary machine learning algorithms for $h(Z) := \mathbb{E}[Y|Z]$ and $f(Z) = \mathbb{E}[X|Z]$. The resulting test is valid under mild rate conditions on those regressions [Shah and Peters, 2020, Theorem 6], akin to conditions in debiased machine learning [Chernozhukov et al., 2016]. Additionally, the GCM allows for directional testing, i.e., alternatives of the form $H_1 : \beta > 0$. The finite sample procedure to conduct the GCM test is described in Kook and Lundborg [2024, Algorithm 1].

3 Computational details

All GCM tests were conducted using `comets` [Kook and Lundborg, 2024] and (5-fold) cross-validated `xgboost` models [Chen et al., 2024] for all regressions. The soccer data were obtained via the `R` package `StatsBombR` [Yam, 2025]. The NFL data were obtained from the NFL Big Data Bowl 2025 competition on Kaggle [Lopez et al., 2024]. Code for reproducing all results is available at `https://github.com/Rob2208/ml-sports-iwsm`.

4 Evaluating shooting skills of players in soccer

Expected Goals (xG) are the output of a statistical model assigning a probability of success to shots using shot-specific features and are one of the most popular

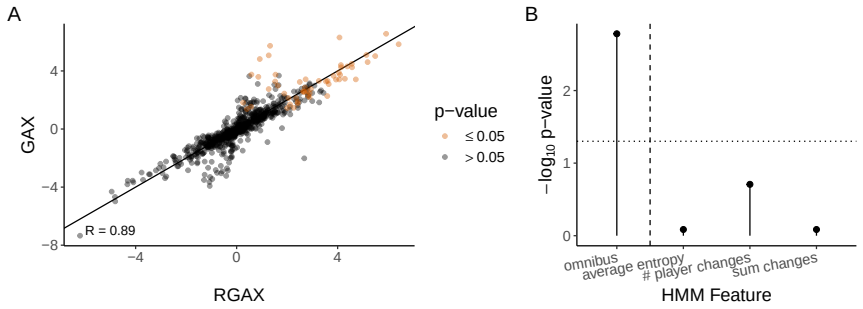

FIGURE 1. A: Comparison between GAX and RGAX based on an XGB model for both regressions on the soccer data. The colors indicate whether GCM-based p-values meet the 5% significance level without multiple testing corrections. $R = 0.89$ denotes the correlation between RGAX and GAX. B: HMM feature significance based on the GCM with an XGB model for the regressions using a Holm correction for multiple testing. The horizontal dotted line represents $-\log_{10}(0.05)$ and the vertical line separates the omnibus and feature-specific tests.

metrics in modern soccer analytics [Anzer and Bauer, 2021]. Popular xG models are based on flexible machine learning algorithms, such as extreme gradient boosting machines, that account for non-linear and interaction effects of the shot-specific features. As a measure of a shot's success, it is commonly used to evaluate the shooting skills of players by considering goals above expectation (GAX), i.e., the difference between actual (Y_i) and estimated expected goals ($\hat{h}(Z_i)$) for the ith shot,

$$\text{GAX} := \sum_{i=1}^{n} (Y_i - \hat{h}(Z_i))X_i,$$

where Z_i denotes shot-specific features and X_i is a binary indicator for whether the player was the shooter ($X_i = 1$) or not ($X_i = 0$). In contrast, we propose using residualized GAX (RGAX),

$$\text{RGAX} := \sum_{i=1}^{n} (Y_i - \hat{h}(Z_i))(X_i - \hat{f}(Z_i)),$$

where $\hat{f}(Z)$ is an estimate of $f(Z) = \mathbb{E}[X|Z]$. This additional regression accounts for whether a player would take the shot under the circumstances described by Z_i. RGAX coincide with the sample version of the GCM and serve as a basis for valid inference on which player significantly impacts the outcome of a shot.

We apply our framework to the 2015/16 season of the top five European leagues. Figure 1A shows a comparison of the proposed RGAX against traditional GAX. GAX and RGAX are highly correlated ($R = 0.89$), showing strong agreement between the two player evaluation metrics. Importantly, in contrast to GAX, the proposed RGAX yields a p-value for the hypothesis whether a player significantly impacts the outcome of a shot at the 5% level (without multiple testing adjustment). Taken together, RGAX are a viable alternative to and statistical improvement over the established GAX for player evaluation.

5 Defensive coverage prediction in the NFL

Classifying whether a team is playing man or zone coverage is one of the most critical strategic decisions of the defense in American football. Using NFL tracking data, we derive a model to predict this binary coverage outcome Y (0: zone, 1: man) for a given play. In addition to using naïve motion features Z, we are interested in identifying whether a set of three statistically enhanced features $X \in \mathbb{R}^3$ (based on a HMM model for player-based latent guarding assignment) helps in predicting defensive schemes. First, we test the null hypothesis $Y \perp\!\!\!\perp X \mid Z$, corresponding to the omnibus test $H_0 : \beta = 0$, where $\beta \in \mathbb{R}^3$, in the PLLM in (1). In a second step, we test HMM feature significance via $H_0^j : Y \perp\!\!\!\perp X_j \mid (Z, X_{-j})$, $i \in \{1, 2, 3\}$, where the index $-j$ denotes removing feature j from X. Figure 1B shows that the omnibus test is highly significant, indicating that the inclusion of HMM features help in predicting defensive schemes. Yet, possibly due to high correlation among the HMM features, the individual H_0^j, $j \in \{1, 2, 3\}$, are not rejected at conventional significance levels. Thus, it is not clear which of the three features is most informative.

6 Conclusion

While ML is gaining in popularity in sports analytics, it is important to maintain valid statistical inference. Here, we leverage machine-learning based inference in the context of nonparametric conditional independence testing and double machine learning in a PLLM to (i) infer player shooting skills, and (ii) test whether additional tracking features improve coverage prediction.

Acknowledgments: The authors would like to thank Rouven Michels and Jan-Ole Koslik for help in deriving HMM features for the NFL coverage model.

References

Anzer, G., and Bauer, P. (2021). A Goal Scoring Probability Model for Shots Based on Synchronized Positional and Event Data in Football (Soccer). *Frontiers in Sports and Active Living* **3**, 53.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J. (2024). `xgboost`: Extreme Gradient Boosting. `R` package version 1.7.7.1. `https://CRAN.R-project.org/package=xgboost`.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, **21**(1):C1 – C68.

Felice, F., Ley, C., Bordas, S. P. A., Groll, A. (2025). Boosting any learning algorithm with Statistically Enhanced Learning. *Scientific Reports*, **15**, 1605.

Kook, L., and Lundborg, A.R. (2024). Algorithm-Agnostic Significance Testing in Supervised Learning With Multimodal Data. *Briefings in Bioinformatics*, **25**(6).

Kovalchik, S. A. (2023). Player Tracking Data in Sports. *Annual Review of Statistics and Its Application*, **10**, 677 – 697.

Lopez, M., Bliss, T., Blake, A., Mooney, P., and Howard, A. (2024) NFL Big Data Bowl 2025. Kaggle. `https://kaggle.com/competitions/nfl-big-data-bowl-2025`.

Shah, R.D., and Peters, J. (2020). The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. *The Annals of Statistics*, **48**(3):1514 – 1538.

Yam, D. (2025). `StatsBombR`: Cleans and pulls StatsBomb data from the API. R package version 0.1.0. `https://github.com/statsbomb/StatsBombR`.