Evaluating Sports Analytics Models: Challenges, Approaches, and Lessons Learned

Jesse Davis[1], Lotte Bransen[1,2], Laurens Devos[1], Wannes Meert[1], Pieter Robberechts[1], Jan Van Haaren[1,3] and Maaike Van Roy[1]

[1]Department of Computer Science, Leuven.AI, KU Leuven, Leuven, Belgium
[2]SciSports, The Netherlands
[3]Club Brugge, Belgium

Abstract

There has been an explosion of data collected about sports. Because such data is extremely rich and complicated, machine learning is increasingly being used to extract actionable insights from it. Typically, machine learning is used to build models and indicators that capture the skills, capabilities, and tendencies of athletes and teams. Such indicators and models are in turn used to inform decision-making at professional clubs. Unfortunately, how to evaluate the use of machine learning in the context of sports remains extremely challenging. On the one hand, it is necessary to evaluate the developed indicators themselves, where one is confronted by a lack of labels and small sample sizes. On the other hand, it is necessary to evaluate the models themselves, which is complicated by the noisy and non-stationary nature of sports data. In this paper, we highlight the inherent evaluation challenges in sports and discuss a variety of approaches for evaluating both indicators and models. In particular, we highlight how reasoning techniques, such as verification can be used to aid in the evaluation of learned models.

Keywords

sports analytics, challenges with evaluation, indicator evaluation, model evaluation, model verification, reliability

1. Introduction

Sports is becoming an increasingly data-driven field as there are now large amounts of data about both the physical states of athletes such as heart rate, GPS, and inertial measurement units (e.g., Catapult Sports) as well as technical performances in matches such as play-by-play (e.g., Stats Perform, StatsBomb) or optical tracking data (e.g., TRACAB, Second Spectrum, SkillCorner). The volume, complexity and richness of these data sources have made machine learning (ML) an increasingly important analysis tool. Consequently, ML is being used to inform decision-making in professional sports. On the one hand, it is used to extract actionable insights from the large volumes of data related to player performance, tactical approaches, and the physical status of players. On the other hand, it is used to partially automate tasks such as video analysis that are typically done manually.

EBeM'22: Workshop on AI Evaluation Beyond Metrics, July 25, 2022, Vienna, Austria
jesse.davis@kuleuven.be (J. Davis); lotte.bransen@kuleuven.be (L. Bransen); laurens.devos@kuleuven.be (L. Devos); wannes.meert@kuleuven.be (W. Meert); pieter.robberechts@kuleuven.be (P. Robberechts); jan.vanhaaren@kuleuven.be (J. Van Haaren); maaike.vanroy@kuleuven.be (M. Van Roy)
0000-0002-3748-9263 (J. Davis); 0000-0002-0612-7999 (L. Bransen); 0000-0002-1549-749X (L. Devos); 0000-0001-9560-3872 (W. Meert); 0000-0002-3734-0047 (P. Robberechts); 0000-0001-7737-5490 (J. Van Haaren); 0000-0001-8959-3575 (M. Van Roy)

At a high level, ML plays a role in team sports in three areas:

Player recruitment. Ultimately, recruitment involves (1) assessing a player's skills and capabilities on a technical, tactical, physical and mental level and how they will evolve, (2) projecting how the player will fit within the team, and (3) forecasting how their financial valuation will develop. (c.f., [1, 2, 3, 4])

Match preparation. Preparing for a match requires performing an extensive analysis of the opposing team to understand their tendencies and tactics. This is can be viewed as a SWOT analysis, which particularly focuses on the opportunities and threats. How can we punish the opponent? How can the opponent punish us? These findings are used by the coaching staff to prepare a game plan. Typically, such reports are prepared by analysts who spent many hours watching videos of upcoming opponents. The analysts must annotate footage and recognize reoccurring patterns, which is a very time-consuming task. Learned models can automatically identify patterns that are missed or not apparent to humans (e.g., subtle patterns in big data) [5], automate tasks (e.g., tagging of situations) [6, 7] that are done by human analysts, and give insights into players' skills.

Management of player's health and fitness. Building up and maintaining a player's fitness level is crucial for achieving good performances [8, 9]. However, training and matches place athletes' bodies under tremendous stress. It is crucial to monitor fitness, have a sense of

how much an athlete can do or, most importantly, when they need to rest and recover. Moreover, managing and preventing injuries is crucial for a team's stability and continuity which is linked to success.

One of the most common uses of ML for addressing the aforementioned tasks is developing novel indicators for quantifying performances. Typically, machine-learned models are trained on large historical databases of past matches. Afterwards, the indicator is derived from the model as the indicator cannot be used directly as a target for training because it is not in the data. One prominent example of such an indicator is expected goals (xG) [10], which is used in soccer and ice hockey to quantify the quality of the scoring opportunities that a team or player created. The underlying model is a binary classifier that predicts the outcome of a shot based on features such as the distance and angle to the goal, the assist type and the passage of play.[1] It is typically a more consistent measure of performance than actual goals, which are extremely important in these sports but also very rare. Even shots are relatively infrequent, and their conversion is subject to variance. The idea of xG is to separate the ability to get into good scoring positions from the inherent randomness (e.g., deflections) of converting them into goals.

Typically, an indicator should satisfy several properties. First, it should provide insights that are not currently available. For example, xG should tell you something beyond looking at goals scored. Second, the indicator should be based on domain knowledge and concepts from sports such that it is intuitive and easy for non ML experts to understand. Finally, the domain experts need to trust the indicator. This often boils down to being able to contextualize when the indicator is useful and ensuring some level of robustness in its value (i.e., it should not wildly fluctuate).

These desiderata illustrate that a key challenge in developing indicators is in how to evaluate them: none of the desiderata naturally align with the standard performance metrics used to evaluate learned models. This does not imply that standard evaluation metrics are not important. In particular, ensuring that probability estimates are well-calibrated is crucial in many sports analytics tasks. It is simply that one must both evaluate the indicator itself and the models used to compute the indicator's value. The goal of this paper is three-fold. First, we will highlight some of the challenges that arise when trying to evaluate work in the context of sports data. Second, we will discuss the various ways that indicator evaluation has been approached. Third, we will overview how learned models that the indicators rely upon have been evaluated. While we will briefly discuss some standard evaluation metrics, we will focus on a more speculative use of reasoning techniques for model evaluation. This paper focuses on the context of professional soccer, where we have substantial experience. However, we believe the lessons and insights are applicable to other team sports, or other domains than sports.

[1]For an interactive discussion of xG, see: https://dtai.cs.kuleuven.be/sports/blog/illustrating-the-interplay-between-features-and-models-in-xg

2. Common Sports Data and Analytics Tasks

This section serves as a short, high-level primer on the data collected from sports matches as well as typical styles of performance indicators and tactical analyses.

2.1. Data

While there are a variety of sources of data collected about sports, we will discuss three broad categories: physical data, play-by-play data and optical tracking data.

During training and matches, athletes often wear a GPS tracker with accelerometer technology (e.g., from Catapult Sports). These systems measure various physical parameters such as distance covered, number of high-speed sprints, and high-intensity accelerations. These parameters are often augmented with questionnaire data [11] to obtain subjective measurements about the difficulty of training such as the rating of perceived exertion (RPE) [12]. Such approaches are used to optimize an athlete's fitness level and ensure their availability and ability to compete.

Play-by-play or event stream data tracks actions that occur with the ball. Each such action is annotated with information such as the type of the action, the start and

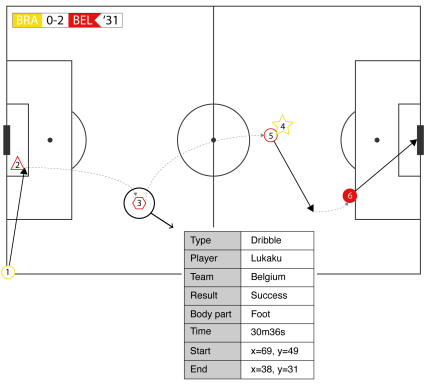

Figure 1: The sequence of actions leading up to Belgium's second goal during the 2018 World Cup quarter-final. Each on-the-ball action is annotated with a couple of attributes, as illustrated for Lukaku's dribble. *(Data source: StatsBomb)*

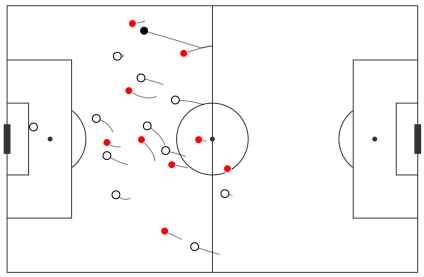

Figure 2: Illustration of a tracking data frame for the first goal of Liverpool against Bournemouth on Dec 7, 2019. The black lines represent each player's and the ball's trajectories during the previous 1.5 seconds. *(Data source: Last Row)*

end locations of the action, the result of the action (i.e., successful or not), the time at which the action was performed, the player who performed the action, and the team of which the acting player is a part of. Figure 1 illustrates six actions that are part of the game between Brazil and Belgium at the 2018 World Cup as they were recorded in the event stream data format. This data is collected for a variety of sports by vendors such as Stats Perform who typically employ human annotators to collect the data by watching broadcast video.

Optical tracking data reports the locations of all the players and the ball multiple times per second (typically between 10 and 25 Hz). This data is collected using a fixed installation in a team's stadium using high-resolution cameras. Such a setup is expensive and typically only used in top leagues. There is now also extensive work on tracking solutions based on broadcast video [13, 14]. Figure 2 shows a frame of tracking data.

2.2. Individual Performance Indicators

Performance indicators for individual players usually fall in one of two categories. The first type focuses on a single action such as a pass or shot. The second type takes a holistic approach by developing a unifying framework that can value a wide range of action types.

Single action. Single action indicators typically take the form of expected value-based statistics: they measure the expected chance that a typical player would successfully execute the considered action in a specific game context. For example, the aforementioned xG model in soccer assigns a probability to each shot that represents its chance of directly resulting in a goal. These models are learned using standard probabilistic classifiers such as logistic regression or tree ensembles from large historical datasets of shots. Each shot is described by the game context from when it was taken, and how this is represented is the key difference among existing models [10, 15, 16].

Such indicators exist for a variety of sports including American football (e.g., expected completion percentage for quarterbacks and expected yards after the catch for receivers),[2] basketball (e.g., expected field goal percentage [17]), and ice hockey (expected goals [18]).

All actions. Instead of building bespoke models for each action, these indicators use the same framework to aggregate a player's contributions over a set of action types. Regardless of sport, almost all approaches exploit the fact that each action a_i changes the game state from s_i to s_{i+1} (as illustrated in Figure 3). These approaches value the contribution of an action a_i as:

$$C(s_i, a_i) = V(s_{i+1}) - V(s_i), \tag{1}$$

where $V(.)$ is the value of a game state and s_{i+1} is the game state that results from executing action a_i in game state s_i.

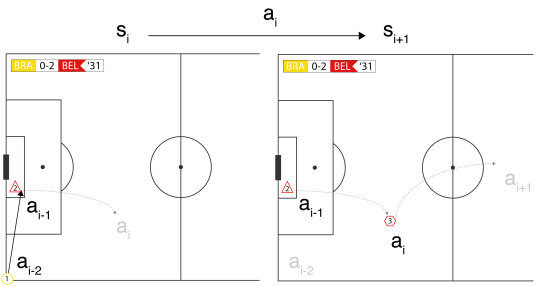

Figure 3: Lukaku's dribble (a_i) changes the game state from the pre-action state s_i to the post-action state s_{i+1}.

Approaches differ on how they value game states, with two dominant paradigms emerging: scoring-based and win-based. Scoring-based approaches take a narrower possession-based view. These approaches value a game state by estimating the probability that the team possessing the ball will score. In soccer, this may entail looking at the near-term probability of a goal in the next 10 actions or 10 seconds [2] or the long-term probability of scoring the next goal [19]. Win-based approaches look at valuing actions by assessing a team's chance of winning the match in each game state. That is, these approaches look at the difference in in-game win-probability between two consecutive game states [20, 21, 22, 23]. Such models have been developed for many sports, including basketball [24], American football [25], ice hockey [26, 3] and rugby [27].

[2] https://nextgenstats.nfl.com/glossary

2.3. Tactical Analyses

Tactics are short-term behaviors that are used to achieve a strategic objective such as winning or scoring a goal. At a high level, AI/ML is used for tactical analyses in two ways: to discover patterns and to evaluate the efficacy of a tactic.

Discovering patterns is a broad task that may range from simply trying to understand where on the field certain players tend to operate and who tends to pass to whom, to more complicated analyses that involve identifying sequences of reoccurring actions. Typically, techniques such as clustering, non-negative matrix factorization, and pattern mining are used to find such reoccurring behaviors [28, 29, 30].

Evaluating the efficacy of tactics is an equally broad task that can generally be split up into two parts: evaluating the efficacy of (1) a current and (2) a counterfactual tactic. Assessing the efficacy of currently employed tactics is typically done by focusing on a specific tactic (e.g., counterattack, pressing) and relating it to other success indicators (e.g., goals, wins) [31, 32]. In contrast, assessing the efficacy of counterfactual tactics is more challenging as it entails understanding what would happen *if* a team (or player) employed different tactics than those that were observed. This is extremely interesting and challenging from an AI/ML and evaluation perspective as it involves both (1) accurately modeling the current behavior of teams, and (2) reasoning in a counterfactual way about alternative behaviors. Such approaches have been developed in basketball and soccer to assess counterfactual shot [33, 34] and movement[3] [35] tactics.

3. Challenges with Evaluation

The nature of sports data and the tasks typically considered within sports analytics and science pose many challenges from an evaluation and analysis perspective.

Lack of ground truth. For many variables of interest, there are simply very few or even no labels, which arises when analyzing both match and physical data. When analyzing matches, a team's specific tactical plan is unknown to outside observers. One can make educated guesses on a high level, but often not for fine-grained decisions. Similarly, when trying to assign ratings to players' actions in a soccer match, there is no variable that directly records this. In fact, in this case, no such objective rating even exists.

Physical parameters can also be difficult to collect. For example, if one is interested in measuring fatigue[4] during a match or training session, some measures are invasive (e.g., blood lactate or creatine kinase). Similarly, in endurance sports such as distance running and cycling, monitoring athletes' aerobic fitness levels is important, which is often measured in terms of the maximal oxygen uptake (VO_{2max}) [36]. However, the test to measure this variable is extremely strenuous and disrupts training regimes, so it can only be measured sporadically.

Credit assignment. It is often unclear why an action succeeded or failed. For example, did a pass not reach a teammate because the passer mishit the ball or did their teammate simply make the wrong run? Similarly, for those actions that are observed, we are unsure why they arose. For example, does a player make a lot of tackles in a soccer match because they are covering for a teammate who is constantly out of position? Or is the player a weak defender that is being targeted by the opposing team?

Noisy features and/or labels. When monitoring the health status of players, teams often partially rely on questionnaires [11] and subjective measures like the rating of perceived exertion [12]. Players respond to such questionnaires in different ways, with some being more honest than others. There is a risk for deception (e.g., players want to play, and may downplay injuries). There are also well-known challenges when working with subjective data. Similarly, play-by-play data is often collected by human annotators, who make mistakes. Moreover, the definitions of events and actions can change over time.

Small sample sizes. There may only be limited data about teams and players. For example, a top flight soccer team plays between 34 and 38 league games in a season and will perform between 1500 and 3000 on-the-ball actions in a game.[5] Even top players do not appear every game and sit out matches strategically for rest.

Non-stationary data. The sample size issues are compounded by the fact that sports is a very non-stationary setting, meaning data that is more than one or two seasons old may not be relevant. On a team level, playing styles tend to vary over time due to changes in playing and management personnel. On a player level, skills evolve over time, often improving until a player reaches their peak, prior to an age-related decline. More generally, tactics evolve and change.

Counterfactuals. Many evaluation questions in sports involve reasoning about outcomes that were not observed. This is most notable in the case of defense, where defensive tactics are often aimed at preventing dangerous

[3] https://grantland.com/features/the-toronto-raptors-sportvu-cameras-nba-analytical-revolution/

[4] Note that there are different types of fatigue that could be monitored such as musculoskeletal or cardiovascular fatigue.

[5] The number depends on what is annotated in the data (e.g., pressure events) and modeling choices such as whether a pass receival is treated as a separate action.

actions from arising such as wide-open three-point shots in the NBA or one vs. the goalie in soccer. Unfortunately, it is hard to know why certain actions were or were not taken. For example, it is difficult to estimate whether the goalie would have saved the shot if they had been positioned slightly differently. Similarly, evaluating tactics also involves counterfactual reasoning as a coach is often interested in knowing what would have happened if another policy had been followed, such as shooting more or less often from outside the penalty box in soccer.

Interventions. The data is observational and teams constantly make decisions that affect what is observed. This is particularly true for injury risk assessment and load management, where the staff will alter players' training regime if they are worried about the risk of injury. Managers also change tactics during the course of the game, depending on the score and the team's performance.

4. Evaluating an Indicator

A novel indicator should capture something about a player's (or team's) performance or capabilities. Evaluating a novel indicator's usefulness is difficult as it is unclear what it should be compared against. This problem is addressed in multiple different ways in the literature.

4.1. Correlation with Existing Success Indicators

In all sports, a variety of indicators exist that denote whether a player (or team) is considered or perceived to be good. Such indicators can be on either the individual or team level.

When evaluating individual players, there are a wealth of existing indicators that are commonly reported and used. First, there are indirect indicators such as a player's market value, salary, playing time, or draft position. Second, there are indicators derived from competition such as goals and assists in soccer (or ice hockey). It is therefore possible to design an evaluation by looking at the correlation between each indicator's value for every player [20, 26, 3]. Alternatively, it is possible to produce two rank-ordered lists of players (or teams): one based on an existing success indicator and another based on a newly designed indicator. Then the correlation between rankings can be computed.

Arguably, an evaluation that strives for high correlations with existing indicators misses the point: the goal is to design indicators that provide insights that current ones do not. If a new indicator simply yields the same ranking as looking at goals, then it does not provide any new information. Moreover, some existing success indicators capture information that is not related to performance. For example, salary can be tied to draft position and years of service. Similarly, a soccer player's market value or transfer fee also encompasses their commercial appeal. Even playing time is not necessarily merit-based.

Other work tries to associate performance and/or presence in the game with winning. This is appealing as the ultimate goal is to win a game.[6] For example, indicators can be based on correlating how often certain actions are performed with match outcomes, points scored, or score differentials [37, 38]. An alternative approach is to build a predictive model based on the indicators and see if it can be used to predict the outcomes of future matches [39].

4.2. The Messi Test

When evaluating indicators about player performance, one advantage is that there is typically consensus on who are among the very top players. While experts, pundits, and fans may debate the relative merits of Lionel Messi and Cristiano Ronaldo, there is little debate that they fall in the very top set of offensive players. An offensive metric where neither of those players scores well, is not likely to convince practitioners. In other words, if a metric blatantly contradicts most conventional wisdom, there is likely a problem with it. This is also called face validity [40]. Of course, some unexpected or more surprising names could appear towards the top of such a ranking, but one would be wary if all such names were surprising.

Unfortunately, this style of evaluation is most suited to analyzing offensive contributions. In general, their is more consensus on the offensive performances of individual players than their defensive performances, as good defense is a collective endeavor and more heavily reliant on tactics.

4.3. Make a Prediction

While backtesting indicators (and models) is clearly a key component of development, sports does offer the possibility for real-world predictions on unseen data. One can predict, and most importantly publish, the outcomes of matches or tournaments prior to their start. In fact, there have been several competitions designed around this principle [41] or people who have collected predictions online.[7]

This is even possible for player indicators, and is often done in the work on quantifying player performance [2, 3]. Decroos et al. [2] included lists of the top

[6] This is not always the case: Sometimes teams play for draws, rest players for strategic reasons, prioritize getting young players experience or try to lose to improve draft position.

[7] https://twitter.com/TonyElHabr/status/1414619621659971588

under-21 players in several of the major European soccer leagues for the 2017/2018 season. It is interesting to look back on the list, and see that there were both hits and misses. For example, the list had some players who were less heralded then such as Mason Mount and Mikel Oyarzabal, who are now key players. Similarly, it had several recognized prospects such as Kylian Mbappé, Trent Alexander-Arnold, and Frenkie de Jong who have ascended. Finally, there were misses like Jonjoe Kenny and David Neres. While one has to wait, it does give an immutable forecast that can be evaluated.

Because they do not allow for immediate results, such evaluations tend to be done infrequently. However, we believe this is something that should be done more often. It avoids the possibility of cherry-picking results and overfitting by forcing one to commit to a result with an unknown outcome. This may also encourage more critical thinking about the utility of the developed indicator. The caveat is that the predictions must be revisited and discussed in the future, which also implies that publication venues would be open to such submissions. Beyond the time delay, another drawback is that they involve sample sizes such as one match day, one tournament, or a short list of players.

4.4. Ask an Expert

Developed indicators and approaches can be validated by comparing them to an external source provided by domain experts. This goes beyond the Messi test as it requires both deeper domain expertise and a more extensive evaluation such as comparing tactical patterns discovered by an automated system to those found by a video analyst. Pappalardo et al. [38] compared a player ranking system they developed to rankings produced by scouts. Similarly, Dick et al. [42] asked soccer coaches to rate how available players were to receive a pass in game situations and compared this assessment to a novel indicator they developed.

Ideally, such an expert-based evaluation considers aspects beyond model accuracy. Ultimately, an indicator should provide "value" to the workflow of practitioners. Hence, it is relevant to measure how much time it saves an analyst in his workflow, whether an indicator can provide relevant new insights and whether the expert can correctly interpret the model's output. This type of evaluation checks whether indicators fulfill the needs of users (i.e., usefulness and usability) and also arises in human-computer interaction [43].

However, this type of evaluation can be difficult as not all researchers have access to domain experts, particularly when it comes to high-level sports. Moreover, teams want to maintain a competitive advantage, so one may not be able to publish such an evaluation.

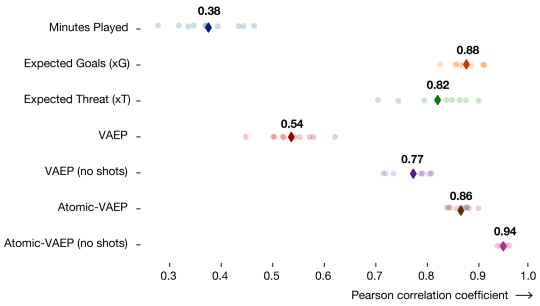

Figure 4: Pearson correlation between player performance indicators for ten pairs of successive seasons in the English Premier League (2009/10 – 2019/20). The diamond shape indicates the mean correlation. The simple "minutes played" indicator is the least reliable, while the Atomic-VAEP [8] indicator is more reliable than its VAEP [2] predecessor and xT [45]. As shots are infrequent and have a variable outcome, omitting them increases an indicator's reliability. The xT indicator does not value shots. Only players that played at least 900 minutes (the equivalent of ten games) in each of the successive seasons are included.

4.5. Reliability

Indicators are typically developed to measure a skill or capability such as shooting ability in basketball or offensive contributions. While these skills can and do change over a longer timeframe (multiple seasons), they typically are consistent within a season or even across two consecutive seasons. Therefore, an indicator should take on similar values in such a time frame.

One approach [39, 44] to measure an indicator's reliability is to split the data set into two, and then compute the correlation between the indicators computed on each dataset. An example of such an evaluation is shown in Figure 4. Methodologically, one consideration is how to partition the available data. Typically, one is concerned with respecting chronological orderings in temporal data. However, in this setting, such a division is likely suboptimal. First, games missed by injury will be clustered and players likely perform differently right when they come back. Second, the difficulty of a team's schedule is not uniformly spread over a season. Third, if the time horizon is long enough, there will be aging effects.

Franks et al. [46] propose a metric to capture an indicator's stability. It tries to assess how much an indicator's value depends on context (e.g., a team's tactical system, quality of teammates) and changes in skill (e.g., improvement through practice). It does so by looking at the variance of indicators using a within-season bootstrap procedure.

[8] https://dtai.cs.kuleuven.be/sports/blog/introducing-atomic-spadl-a-new-way-to-represent-event-stream-data/

Another approach [29] is to look at consecutive seasons and pose the evaluation as a nearest neighbors problem. That is, based on the indicators computed from one season of data for a specific player, find a rank-ordered list of the most similar players in the subsequent (or preceding) season. The robustness of the indicator is then related to the position of the chosen player in the ranking.

5. Evaluating a Model

Evaluating the models used to produce the indicator involves two key aspects. First, it is important to ensure that the model will behave as expected on unseen data. This is particularly important for sports since the data can have errors or noise (e.g., incorrect annotations, sensor failures, errors in tracking data) and rare or unexpected events. Hence, one wants to reason about the model. Second, there are standard evaluation metrics that are important to use to ensure, e.g., that probability estimates are accurate.

5.1. Reasoning about Learned Models

Verification is a powerful alternative to traditional aggregated metrics to evaluate and inspect a learned model. Verification attempts to reason about how a learned model will behave [47, 48, 49, 50]. Given a desired target value (i.e., prediction), and possibly some constraints on the values that the features can take on, a verification algorithm either generates one or more instances that satisfy the constraints, or it proves that no such instance exists. This is similar to satisfiability checking. In practice, verification allows users to query a model, i.e., reason about the model's possible outputs and examine what the model has learned from the data. It can be used to investigate how a model behaves in certain sub-areas of the input space. Examples of verification questions are:

- Is a model robust to small changes to the inputs? For example, does a small change in the time of the game and the position of the ball significantly change the probability that a shot will result in a goal? This relates to adversarial examples (c.f. image recognition).
- Related to the previous question, but with a different interpretation: given a specific example of interest, can one or more attributes be (slightly) changed so that the indicator is maximized? This is often called a *counterfactual* explanation, e.g., *if* the goalie would have been positioned closer to the near post, how would that have affected the estimated probability of the shot resulting in a goal? We want to emphasize that, this is not a causal counterfactual (because the considered models are not causal models).
- Does the model behave as expected in scenarios where we have strong intuitions based on domain knowledge? For example, one can analyze what values the model can predict for shots that are taken from a very tight angle or very far away from the goal. One can then check whether the predictions for the generated game situations are realistic.

Typical aggregated test metrics do not reveal the answers to these questions. Nevertheless, the answers can be very valuable because they provide insights into the model and can reveal problems with the model or the data.

We have used verification to evaluate soccer models in two novel ways. First, we show how it is possible to debug the training data and pinpoint labeling errors (or inconsistencies). Second, we identify scenarios where the model produces unexpected and undesired predictions. These are shortcomings in the model itself. We use Veritas [51] to analyze two previously mentioned soccer analytics models: xG and the VAEP holistic action-value model.

First, we analyzed an xG model in order to identify "what are the optimal locations to shoot from outside the penalty box?". We used Veritas to generate 200 examples of shots from outside the penalty box that would have the highest probability of resulting in a goal, which are shown as a heatmap in Figure 5. The cluster in front of the goal is expected as it corresponds to the areas most advantageous to shoot from. The locations near the corners of the pitch are unexpected. We looked at the shots from the 5 meter square area touching the corner and counted 11 shots and 8 goals, yielding an extremely high 72% conversion rate. This reveals an unexpected labeling behavior by the human annotators. Given the distance to the goal and the tight angle, one would expect a much lower conversion rate. A plausible explanation is that annotators are only labeling actions as a shot in the rare situations where the action results in a goal or

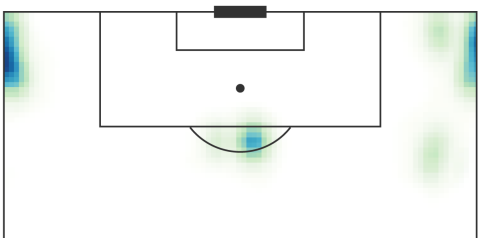

Figure 5: A heatmap showing where Veritas generates instances of shots from outside the penalty box with the highest xG values.

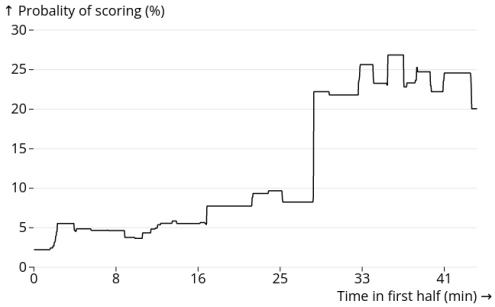

Figure 6: For specific action sequences, the time remaining in the game has a large variable effect on the probability of scoring in the VAEP action model. This variability is unexpected and reveals a robustness issue with the model.

a save. Otherwise, the actions are labeled as a pass or a cross.

Second, we analyzed VAEP [2], a holistic-action model for soccer. The models underlying this indicator look at a short sequence of consecutive game actions and predict the probability of a goal in the next 10 actions. Unlike xG models, all possible actions (passes, dribbles, tackles, …) are considered, not just shots. For the data in an unseen test set, the model produces well-calibrated probability estimates in aggregate. However, we looked for specific scenarios where the model performs badly and found several instances that are technically possible, but very unlikely. More interestingly, Veritas generated instances where all the values of all features were fixed except for the time in the match, and found that the probability of scoring varied dramatically according to match time. Figure 6 shows this variability for one such instance. The probability gradually increases over time, which is not necessarily unexpected as scoring rates tend to slightly increase as a match progresses. However, about 27 minutes into the first half the probability of scoring dramatically spikes. Clearly, this behavior is undesirable: we would not expect such large variations. This suggests that time should probably be handled differently in the model, e.g., by discretizing it to make it less fine-grained.

Such an evaluation is still challenging. One has to know what to look for, which typically requires significant domain expertise or access to a domain expert. Moreover, the process is exploratory: there is a huge space of scenarios to consider and the questions have to be iteratively refined.

5.2. Standard Metrics

Many novel indicators involve using a learned model that makes probabilistic predictions, making calibration the standard choice for a classical evaluation. Calibration can be evaluated in a number of different ways such as using reliability diagrams [52], the Brier score [53], logarithmic loss, and the multi-class expected calibration error (ECE) [54]. It is less clear when one of these metrics may be more appropriate than another. Here, it may be worth considering if the probabilities will be summed (e.g., for computing player ratings) or multiplied (e.g., modeling decision making) [55]. It is important to remember that these metrics depend on the class distribution, and hence their values need to be interpreted in this context. This is important as scoring rates can vary by competition (e.g., men's leagues vs. women's leagues) [56].

6. Discussion

Evaluating learning systems in the context of sports is an extremely tricky endeavor that largely relies on expertise gained through experience. On the one hand, the outputs of learned models are often combined in order to construct novel indicators of performance, and the validity of these indicators needs to be assessed. Here, we would like to caution against looking at correlations to other success metrics as we believe that a high correlation to an existing indicator fails the central goal: gaining new insights. We also believe that the reliability and stability of indicators is important, and should be more widely studied. Still, what remains the best approach for evaluating a specific problem is often not clear, and the field would benefit from a broader discussion of best practices.

On the other hand, it is also necessary to evaluate the models used to construct the underlying systems and indicators. Here, we believe that evaluating models by *reasoning about their behavior* is crucial: this changes the focus from a purely data-based evaluation perspective to one that considers the effect of the data on the model. The ability to have insight into a model's behavior also facilitates interactions with domain experts. Critically reflecting on what situations a model will work well in and which situations it may struggle in, helps build trust and set appropriate expectations.

Still, using reasoning is not a magic solution. When a reasoner identifies unexpected behaviors, there are at least two possible causes. One cause is errors in the training data which are picked up by the model and warp the decision boundary in unexpected ways (e.g., Figure 5). Some errors can be found by inspecting the data, but given the nature of the data, it can be challenging to know where to look. The other cause is peculiarities with the model itself, the learning algorithm that constructed the model, or the biases resulting from the model representation (e.g., Figure 6). Traditional evaluation metrics are completely oblivious to these issues. They can only be discovered by reasoning about the model. Unfortunately, it remains difficult to correct a model that has picked up

on an unwanted pattern. For example, the time's effect on the probability of scoring can only be resolved via representing the feature in a different way, relearning the model, and reassessing its performance. Alas, this is an iterative guess-and-check approach. We believe that reasoning approaches to evaluation are only in their infancy and need to be further explored.

While this paper discussed evaluation in the context of sports, we do feel that some of the challenges and insights are relevant for other application domains where machine learning is applied. For example, evaluation challenges also arise in prognostics, especially when it is impossible to directly collect data about a target such as time until failure. In both domains, we do not want to let the athlete nor machine be damaged beyond repair. Also, we perform multiple actions to avoid failure, making it difficult to attribute value to individual actions or identify root causes. Another example is how to deal with subjective ratings provided by users, which often occurs when monitoring players' fitness and was also a key issue in the Netflix challenge. Finally, in terms of approaches to evaluation, there is also more emphasis within ML in general on trying to ensure the robustness of learned models by checking, for example, how susceptible they are to adversarial attacks.

Acknowledgments

This work was supported by iBOF/21/075, Research Foundation-Flanders (EOS No. 30992574, 1SB1320N to LD) and the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program.

References

[1] L. Bransen, P. Robberechts, J. Van Haaren, J. Davis, Choke or shine? quantifying soccer players' abilities to perform under mental pressure, in: MIT Sloan Sports Analytics Conference, 2019.

[2] T. Decroos, L. Bransen, J. Van Haaren, J. Davis, Actions Speak Louder Than Goals: Valuing Player Actions in Soccer, in: Proc. of 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 1851–1861.

[3] G. Liu, O. Schulte, Deep reinforcement learning in ice hockey for context-aware player evaluation, in: Proc. of 27th Int. Joint Conference on Artificial Intelligence, 2018, pp. 3442–3448.

[4] A. Franks, A. Miller, L. Bornn, K. Goldsberry, Counterpoints: Advanced defensive metrics for NBA basketball, in: MIT Sloan Sports Analytics Conference, 2015.

[5] L. Shaw, S. Gopaladesikan, Routine inspection: A playbook for corner kicks, in: MIT Sloan Sports Analytics Conference, 2021.

[6] P. Bauer, G. Anzer, Data-driven detection of counterpressing in professional football, Data Mining and Knowledge Discovery 35 (2021) 2009—-2049.

[7] A. Miller, L. Bornn, Possession sketches: Mapping NBA strategies, in: MIT Sloan Sports Analytics Conference, 2017.

[8] S. L. Halson, Monitoring training load to understand fatigue in athletes, Sports Med 44 (2014) 139–147.

[9] P. C. Bourdon, M. Cardinale, A. Murray, P. Gastin, M. Kellmann, M. C. Varley, T. J. Gabbett, A. J. Coutts, D. J. Burgess, W. Gregson, N. T. Cable, Monitoring athlete training loads: consensus statement, Int J Sports Physiol Perform 12 (2017) 161–170.

[10] S. Green, Assessing the performance of Premier League goalscorers, 2012. URL: https://www.statsperform.com/resource/assessing-the-performance-of-premier-league-goalscorers/.

[11] M. Buchheit, Y. Cholley, P. Lambert, Psychometric and physiological responses to a preseason competitive camp in the heat with a 6-hour time difference in elite soccer players, Int J Sports Physiol Perform 11 (2016) 176–181.

[12] G. Borg, Psychophysical bases of perceived exertion, Med sci sports exer 14 (1982) 377–381.

[13] N. Johnson, Extracting player tracking data from video using non-stationary cameras and a combination of computer vision techniques, in: MIT Sloan Sports Analytics Conference, 2020.

[14] A. Arbués Sangüesa, A journey of computer vision in sports: from tracking to orientation-base metrics, Ph.D. thesis, Universitat Pompeu Fabra, 2021.

[15] P. Lucey, A. Bialkowski, M. Monfort, P. Carr, I. Matthews, Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data, in: MIT Sloan Sports Analytics Conference, 2015.

[16] P. Robberechts, J. Davis, How data availability affects the ability to learn good xg models, in: Workshop on Machine Learning and Data Mining for Sports Analytics, 2020, pp. 17–27.

[17] V. Sarlis, C. Tjortjis, Sports analytics — evaluation of basketball players and team performance, Information Systems 93 (2020) 101562.

[18] B. Macdonald, An expected goals model for evaluating nhl teams and players, in: MIT Sloan Sports Analytics Conference, 2012.

[19] J. Fernández, L. Bornn, D. Cervone, A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions, Machine Learning 110 (2021) 1389–1427.

[20] S. Pettigrew, Assessing the offensive productivity

of NHL players using in-game win probabilities, in: MIT Sloan Sports Analytics Conference, 2015.
[21] B. Burke, WPA explained, 2010. URL: http://archive.advancedfootballanalytics.com/2010/01/win-probability-added-wpa-explained.html.
[22] P. Robberechts, J. Van Haaren, J. Davis, A bayesian approach to in-game win probability in soccer, in: Proc. of 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2021, pp. 3512–3521.
[23] M. Bouey, NBA win probability added, 2013. URL: https://www.inpredictable.com/2013/06/nba-win-probability-added.html.
[24] D. Cervone, A. D'Amour, L. Bornn, K. Goldsberry, POINTWISE: Predicting Points and Valuing Decisions in Real Time with NBA Optical Tracking Data, in: MIT Sloan Sports Analytics Conference, 2014.
[25] D. Romer, Do firms maximize? Evidence from professional football, Journal of Political Economy 114 (2006) 340–365.
[26] K. Routley, O. Schulte, A Markov game model for valuing player actions in ice hockey, in: Proc. 31st Conference on Uncertainty in Artificial Intelligence, 2015, pp. 782–791.
[27] T. Kempton, N. Kennedy, A. J. Coutts, The expected value of possession in professional rugby league match-play, Journal of sports sciences 34 (2016) 645–650.
[28] Q. Wang, H. Zhu, W. Hu, Z. Shen, Y. Yao, Discerning tactical patterns for professional soccer teams: An enhanced topic model with applications, in: Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 2197–2206.
[29] T. Decroos, J. Davis, Player vectors: Characterizing soccer players' playing style from match event streams, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2019, pp. 569–584.
[30] J. Bekkers, S. S. Dabadghao, Flow motifs in soccer: What can passing behavior tell us?, Journal of Systems Architecture 5 (2019) 299–311.
[31] J. Fernandez-Navarro, L. Fradua, A. Zubillaga, A. P. McRobert, Evaluating the effectiveness of styles of play in elite soccer, International Journal of Sports Science & Coaching 14 (2019) 514–527.
[32] S. Merckx, P. Robberechts, Y. Euvrard, J. Davis, Measuring the effectiveness of pressing in soccer, in: Workshop on Machine Learning and Data Mining for Sports Analytics, 2021.
[33] N. Sandholtz, L. Bornn, Markov decision processes with dynamic transition probabilities: An analysis of shooting strategies in basketball, Annals of App Stat 14 (2020) 1122–1145.
[34] M. Van Roy, P. Robberechts, W.-C. Yang, L. De Raedt, J. Davis, Leaving goals on the pitch: Evaluating decision making in soccer, in: MIT Sloan Sports Analytics Conference, 2021.
[35] H. M. Le, Y. Yue, P. Carr, P. Lucey, Coordinated multi-agent imitation learning, in: Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 1995–2003.
[36] M. J. Joyner, Modeling: optimal marathon performance on the basis of physiological factors, Journal of Applied Physiology 70 (1991) 683–687.
[37] I. McHale, P. Scarf, D. Folker, On the development of a soccer player performance rating system for the english premier league, Interfaces 42 (2012) 339–351.
[38] L. Pappalardo, P. Cintia, P. Ferragina, E. Massucco, D. Pedreschi, F. Giannotti, Playerank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach, ACM Trans. Intell. Syst. Technol. 10 (2019) 59:1–59:27.
[39] L. M. Hvattum, Offensive and defensive plus–minus player ratings for soccer, Applied Sciences 10 (2020).
[40] A. Z. Jacobs, H. Wallach, Measurement and fairness, in: Proc. of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, p. 375–385.
[41] W. Dubitzky, P. Lopes, J. Davis, D. Berrar, The open international soccer database for machine learning, Machine learning 108 (2019) 9–28.
[42] U. Dick, D. Link, U. Brefeld, Who can receive the pass? A computational model for quantifying availability in soccer, Data Mining and Knowledge Discovery (2022).
[43] W. Xu, Toward human-centered AI: A perspective from human-computer interaction, Interactions 26 (2019) 42–46.
[44] M. Van Roy, P. Robberechts, T. Decroos, J. Davis, Valuing on-the-ball actions in soccer: A critical comparison of xT and VAEP, in: 2020 AAAI Workshop on AI in Team Sports, 2020.
[45] K. Singh, Introducing expected threat, 2019. URL: https://karun.in/blog/expected-threat.html.
[46] A. M. Franks, A. D'Amour, D. Cervone, L. Bornn, Meta-analytics: Tools for understanding the statistical properties of sports metrics, Journal of Quantitative Analysis in Sports 12 (2016) 151–165.
[47] M. Kwiatkowska, G. Norman, D. Parker, PRISM 4.0: Verification of probabilistic real-time systems, in: Proc. 23rd Int. Conf. on Computer Aided Verification, 2011, pp. 585–591.
[48] S. Russell, D. Dewey, M. Tegmark, Research priorities for robust and beneficial artificial intelligence, AI Magazine 36 (2015) 105–114.
[49] A. Kantchelian, J. D. Tygar, A. Joseph, Evasion and hardening of tree ensemble classifiers, in: Proc. of the 33rd International Conference on Machine

Learning, 2016, pp. 2387–2396.
[50] G. Katz, C. Barrett, D. L. Dill, K. Julian, M. J. Kochenderfer, Reluplex: An efficient smt solver for verifying deep neural networks, in: Computer Aided Verification, 2017, pp. 97–117.
[51] L. Devos, W. Meert, J. Davis, Versatile verification of tree ensembles, in: Proc. of the 38th International Conference on Machine Learning, 2021, pp. 2654–2664.
[52] A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, in: Proc. of the 22nd Int. Conf. on Machine learning, 2005, p. 625–632.
[53] G. W. Brier, Verification of forecasts expressed in terms of probability, Monthly weather review 78 (1950) 1–3.
[54] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: Proc. of the 34th Int. Conf. on Machine Learning, 2017, pp. 1321–1330.
[55] T. Decroos, J. Davis, Interpretable prediction of goals in soccer, in: AAAI 2020 Workshop on AI in Team Sports, 2020.
[56] L. Pappalardo, A. Rossi, M. Natilli, P. Cintia, Explaining the difference between men's and women's football, PLoS ONE 16 (2021).