**National Research University Higher School of Economics Faculty of Computer Science**

**Programme ‘Master of Data Science’**

**MASTER’S THESIS**

**Evaluating Performance in Women’s Soccer: A Machine Learning Approach**

**Student: Sedin Valerii Valerievich**

**Supervisor: Vasiliev Gleb Albertovich**

**Moscow, 2022**

# Abstract

Evaluating the performance of football players has gone a long way from judging by the result and counting goals scored through more detailed statistics like tackles made and passes completed to modern machine learning-based methods. Nowadays advanced models are utilized by many clubs alongside with expert opinions of scouts and coaches.

The disparity in popularity of men's and women's soccer leads to numerous consequences, among which there is a disparity in data available about men's and women's games. As a result, modern models are trained mostly or exclusively on men's data. However, there is no evidence that models trained this way are applicable to evaluate women. Moreover, several research studies state that there are differences in the way men and women behave and achieve their results on a football pitch.

This paper investigates the applicability of such models for women’s data. To achieve this purpose, we train a few models on men’s data, check their validity and efficiency and apply them to women’s data. Then models are trained on women’s data and their validity and efficiency are checked too.

This paper results in working models to assess women’s and men’s performance in football matches.

# Table of contents

[1 Introduction](#_heading=h.1fob9te) 5

[1.1 Metrics](#_heading=h.3znysh7) 5

[1.1.1 Shot-Based Metrics](#_heading=h.2et92p0) 5

[1.1.2 Possession-Based Metrics](#_heading=h.tyjcwt) 6

[1.2 Comparisons of women’s and men’s football](#_heading=h.3dy6vkm) 14

[1.3 Problem Formulation](#_heading=h.1t3h5sf) 15

[2 Data](#_heading=h.17dp8vu) 16

[3 Models Trained on Male Data](#_heading=h.3rdcrjn) 18

[3.1 Training the Models on Male Data](#_heading=h.26in1rg) 20

[3.2 Validating the Male Data Models](#_heading=h.lnxbz9) 21

[3.3 Applying Male Data Models to Test Data](#_heading=h.1ci93xb) 23

[3.4 Robustness Check](#_heading=h.3whwml4) 24

[3.5 Conducting Additional Experiments](#_heading=h.4ddeoix) 32

[4 Female Data Models](#_heading=h.35xuupr) 34

[4.1 Training and Validating the Models on the Female Data](#_heading=h.y3nxzrsmw76w) 34

[4.2 Testing the Female Data Scoring Model](#_heading=h.452snld) 35

[5 Discussion](#_heading=h.t8i7l8giea18) 36

[5.1 Further Use of the Model](#_heading=h.yj9j5zozn330) 36

[5.2 Theoretical results](#_heading=h.ilhz3gq2bz1f) 37

[6 Conclusion](#_heading=h.2k82xt6) 38

[References](#_heading=h.zdd80z) 39

# 1 Introduction

A result of a football match is often decided by few goals or no goals at all. Because of it, assessment of players’ performance with traditional statistics does not take into account a vast majority of events that happen on the football pitch. The other issue is assessing teams and players solely based on the score which might be a result of highly unlikely events.

For a long time assessing players was a duty of scouts and coaches, who watched specific players they were interested in, live and made conclusions based on their vision and opinions. The development of information technologies has led to the creation of additional tools which help clubs make their decisions and provide unbiased information about players.

## 1.1 Metrics

Modern advanced performance metrics can be divided into on-ball metrics and off-ball metrics. The latter address isolated aspects of football, such as space exploitation and positional advantage [11]. On-ball metrics rate actions that players do with the ball, such as passes, shots, tackles. There are shot-based and possession-based models.

### 1.1.1 Shot-Based Metrics

The most common advanced shot-based metric and one of the oldest is xG. The introduction of xG and different models built for calculating xG has transformed our understanding of football. xG is a shot-based metric. Most modern xG models take into account position on a pitch, defensive pressure on an attacker [16], shot type (for example, leg, header, other) [1]. xG models are created by applying machine learning methods to modern statistical data.

xG is an effective tool to assess the quality of finishing. A player who scores more than expected can be considered a good finisher. It is a proper tool to assess a team's performance too. A result of a game can be misleading as well as shots and shots on target count. xG shows how fair an outcome of a game is from a point of chances’ quality.

What xG is not good for is assessing other aspects of the game: passing, dribbling, defending, etc. It is built for measuring the quality of shots, so it does not assess passing performance, dribbling performance, tackling performance, etc. This leads to the invention of possession-based models. Here is how Nils Mackay, a data scientist in Opta/Stats Perform describes the shot-based models in his blog: "This [a shot-based model] is your regular xG model. Shooting from distance is less dangerous to the opposition than shooting from close range. Also, shooting from an angle is less dangerous than shooting from right in front of a goal. This is the basic concept behind most xG models..." [12].

### 1.1.2 Possession-Based Metrics

Possession-based metrics were introduced to rate players, whose primary work on the pitch is not scoring goals. Nils Mackay describes them the following way: "A possession-based model tries to estimate the probability that, given a certain event, the team with the ball will score within the same possession. Let's look at an example. Say Team A gets to take a corner kick. The model will try to estimate what fraction of corners ends in the goal within the same possession. This goal could be directly from the corner kick, from a direct header resulting from the corner kick, after a short corner and 35 more passes, etc. As long as all the events between the current event and the goal are from Team A, it counts." [12]

To complement that, xGChain and xGBuildup were proposed by Thom Lawrence from StatsBomb. The algorithm to calculate xGChain is the following:

“Find all the possessions each player is involved in.

Find all the shots within those possessions.

Sum their xG (you might take the highest xG per possession, or you might treat the shots as dependent events, whatever floats your boat).

Assign that sum to each player, however, involved they were." [10]

The author points out it is still highlighting mostly attacking players, but some defenders are observed at the top of his xGChain rating for season 2016/2017 too.

The idea behind xGBuildup builds upon xGChain. The algorithm is almost the same, but the last step is excluding an assist and a shot [10].

Though xGChain and especially xGBuildup allow to highlight roles of players who have large roles at earlier stages of attacks, they have their flaws. One of them is that they spread the value of xG equally among participants of an attack, whereas more modern research shows that different actions have different weights in creating a scoring chance. It also does not evaluate the defensive impact of players.

Therefore, Karun Singh introduced an Expected Threat (xT) framework in his blog [18]. The idea behind xT is to use the Markov model to determine if the change of the ball position in the current possession increases the probability of scoring within the next n actions. The pitch is divided into 192 zones (16x12), so the change of the ball position is recorded only when the ball is moved from one zone to another.

Tom Decroos et al. offer a framework to assess player performance holistically [5]. It is based on the idea that actions in football in most cases have one of the two following purposes: to score a goal or to avoid conceding a goal. Therefore, every action is converted to a format comfortable for use in machine learning tasks. It is called SPADL (soccer player action description language). The paper says, every action is described with the following parameters:

* StartTime – the action’s start time
* EndTime – the action’s end time
* StartLoc – the action’s start location (x, y)
* EndLoc – the action’s end location (x, y)
* Player – player who performed the action
* Team – the player’s team
* ActionType - the type of the action
* BodyPart – the player’s body part used for the action
* Result – the result of the action.

The 1.1.1 version of the Socceraction Python package converts match events to a slightly different format.

After a football match event stream is converted to the appropriate format, the authors developed the VAEP (Valuing Actions by Estimating Probabilities) framework. The CatBoost algorithm was chosen for showing the best empirical results. The authors suggest, this model is applied to every action in a match, so every action gets its rating in terms of increasing the probability of scoring and decreasing the probability of conceding a goal for a team, player who performs the action, plays for. To get the player rating, first, the VAEP value of every single action of a player is calculated: change in probability of scoring minus change in probability of conceding. All these values are summed, multiplied by 90, and divided by the number of minutes the footballer played in the game.

This paper also contributes a Python package to implement it [23]. It consists of a loader for popular event stream operators (Opta, WyScout, etc.), converter to SPADL format, implementation of the xT framework, implementation of the VAEP framework.

A comparison between xT and VAEP frameworks is described in a conference proceeding from 2019 [19]. The key points from this research are:

* xT ignores defensive actions like a tackle or an interception because as the result the ball does not change its zone
* xT does not punish actions that increase the risk of conceding a goal
* xT has greater interpretability because it values actions solely on the ball location
* xT is biased towards creative players who complete a lot of key passes and dribbles, while VAEP is biased towards strikers
* Both models are biased towards offensive players

The authors also observe four types of actions in specific contexts to get more insights about differences between xT and VAEP:

* VAEP assigns a more diverse range to a backward pass into its penalty box due to capturing both risk and reward of the pass, not only change in a threat towards the opposition goal
* VAEP rates the first ball progression of the counterattack significantly higher because VAEP considers not only location but the context of the action too.
* Small dribbles towards the opposition goal inside the opposition penalty box are rated higher and in a more diverse range by VAEP. The reason is that such dribbles are often very short, so the xT model does not record the ball location change, so the xT of such actions is often 0.
* A through ball near the opposition penalty box is rated higher by xT than by VAEP. Authors suggest this might happen because xT is better than VAEP at capturing positional advantage.

Both VAEP and xT papers give example player ratings within leagues of similar strength or only one league. The xT blog post describes players of the English Premier League. VAEP paper rates English and Spanish leagues players separately from the ones from "the smaller" Dutch, Belgian, and French leagues. It is due to one of the remaining challenges described in the VAEP paper: it is hard to compare players across the leagues of different strengths because in weaker leagues it is easier to perform valuable actions. Because of this, top players of the "weaker leagues'' get higher ratings than top players in more competitive leagues.

VAEP is utilized in a later paper by Tom Decroos et al. to analyze English Premier League players' performances. The results underline the model bias towards attacking players [6].

Lotte Bransen et al. Develop advanced passing measuring [4]. They set up a metric called Expected Contribution to the Outcome of the Match (ECOM). Passes get this attention because authors claim, passes constitute around 70% of on-ball action in a match. The approach is the following:

* Split a match into possession sequences
* Label sequences and their constituting passes using an expected-goals model
* Introduce a domain-specific distance function to measure similarity between passes
* Value each pass by computing the expected added reward of the pass using a k-nearest-neighbours search using their distance function
* Compute each player’s ECOM rating by aggregating their pass values and normalizing them for 90 minutes of play.

The function from the third step considers the following properties to calculate the distance between passes:

* Difference between lengths
* Differences between origins
* Differences between destinations
* Ball locations 5s, 10s, and 15s before a pass

The latter provides context for the model. This might be a considerable improvement compared to prior models, but it does not provide tools to assess defensive performance. The results of this paper display a significant bias towards midfielders.

Pieter Robberechts introduces VPEP (Valuing Pressure decisions by Estimation Probabilities) to measure pressing. This paper uses SPADL along with complex features, which distance and angle to goal for both action's start and end location and distance covered during the action in both the x and y directions. Between two consecutive actions elapsed time is computed and it is checked whether the ball has changed possession. Based on this, the authors make conclusions about the current speed of the game. Game context features are also considered by the model. They are the number of goals scored by each team and the goal difference. The last set of features used for this model is pressure features, which include the distance between the defender and the ball, the angle between the defender's goal, the ball, and the defender, and the time delay between the start of the action and the start of the pressure [17].

Stephan Wolfe et al. develop a different approach to assessing football players. They suggest adapting the Elo rating to measure performance [20]. However, it is computed only based on traditional statistics and, though it might give some useful insights about players, it is unclear how it can be used for comparing footballers who never face each other.

Pappalardo et al. Suggest their framework to assess performances called PlayerRank [13]. This framework consists of three phases.

At the learning phase, two major tasks are done: feature weighting and role detector training.

Feature weighting is based on the idea that the relation between the team performance and the outcome of the match is strong enough to use it as the ground truth. The first phase of feature weighting is extracting the outcome of the match (victory or non-victory) and the vector of a team in the match. The team performance vector is obtained by summing the corresponding features over all players in that match. The features used in the paper cover most of the possible actions and action-outcome tuples, for example, "pass - simple pass - accurate", "pass - cross – not accurate", "foul - yellow card", "offside". The second phase of feature weighing is solving a classification task between the team performance and the outcome. It is done with a linear classifier. The authors suggest the Linear Support Vector Machine as an example. This classifier returns weights that are used in the rating phase.

Role detector training is conducted to determine a player's role on the pitch. In this research, the role is synonymous with the player's average position. It is done to rank players of different roles separately to avoid bias towards any positions. The case when a player participates in a significant number of episodes in different areas is solved by assigning him to more than one role and ranking him in different ratings.

The rating phase consists of two steps: individual performance extraction and player rating.

Individual performance extraction models a player's performance as an n-dimensional vector, where n is the number of features used. Each feature describes a specific aspect of a player's behavior, for example, the number of fouls committed, number of accurate passes. In the paper, an example with n = 76 is provided, but the framework is designed to work with any set of features.

Player rating is computed as the normalized scalar product of a player vector, extracted in the previous step, and weights from the learning phase. The original set of features does not include goals, because the weight of this feature would be significantly larger than any other. The researchers propose adjusted player rating to take goals into account which is computed as alpha \* norm\_goals + (1 - alpha) \* r, where norm\_goals is the number of goals scored in the match m normalized in the range [0; 1], r is non-adjusted player rating and alpha is the importance of goals in the new rating. To calculate a player’s rating over a series of matches authors suggest using the Exponential Weighted Smoothing Average (EWMA) to give more weight to the more recent matches.

At the ranking phase, players are ranked according to their roles using their ratings.

## 1.2 Comparisons of women’s and men’s football

Scientific comparisons with men’s football are relatively new. However, a large number of differences have already been observed.

Van Lange et al. learned that women are more willing to stop the game and let their teammates and opponents get medical care. According to the paper, sex is one of the factors along with age, stage of the game, and score [9].

Pedersen et al. studied anthropomorphic and physiological differences between male and female players. As a result, altering match length, goal size, ball size, and pressure were proposed to give female players conditions more equal to males with an amendment to average height and muscle mass [15].

Bradley et al. in the paper called "Gender differences in match performance characteristics of soccer players competing in the UEFA Champions League" studied match performance characteristics and learned that men cover greater distances during a game and cover greater distances at high speed [2]. In this research, it was also observed that female players show lower pass completion rates and lose the ball more often than male players.

Gomez et al. observed that female teams shoot more after team plays than male teams as males shoot after solo plays more than females [8]. This is particularly interesting because, with lower pass completion rates observed in the "Gender differences in match performance characteristics" paper, one would expect to use solo plays more than team plays.

Pappalardo et al. conducted research based on the latest men’s World Cup (2018) and women’s World Cup (2019). The authors studied shooting zones for men and women and showed that, with an amendment for physiological differences, women shoot more from close range and long-range, but middle range shots are used by men more [14]. This research also shows a different perspective on an average time of possession: this parameter is called "recovery time", which incites people to think about changes in possession not as losing the ball, but as recovering the ball. Pass completion rates in this paper are also observed to be lower for female teams than for male teams. The model presented in this research uses pass completion rate and recovery time to distinguish male and female football. This research also shows that female matches have more free kicks, duels, and other on-the-ball actions, but commit fewer fouls. PlayerRank metrics of female and male footballers are compared in this paper. It is observed that male players have a higher average PR (unpaired t-score = 9.01, p-value < 0.001), whereas standard deviations are similar (unpaired t-score = -0.4, p = 0.69).

## 1.3 Problem Formulation

A lot of differences between men’s and women’s soccer have been observed, but models are mostly trained and intended for men’s soccer. The question is if models trained on men’s data are suitable for rating women’s performance. This question is relevant for both shot-based and possession-based metrics. Pedersen et al. give a reason to assume the same shots by men and women might have different xGs [15]. Pappalardo et al. let us assume that similar actions by men and women might lead to different probabilities of scoring and conceding too [14]. In terms of modern metrics, VAEP scores could be different for men and women.

These facts raise two main questions:

* Are models trained primarily on male data suitable for evaluating women’s performance?
* If we do not find evidence that male data trained models are suitable, can train on female data produce an effective model?

To answer this question, we create a new model design, different from those used in the described papers. The purpose of doing that is to create a model with a clear indicator of its quality. In this paper, we train separate possession-based models for male soccer, test its efficiency, apply it to women's test dataset and check its efficiency on the new data. If proved inefficient, we go to the second question: train a model on women’s training data and test it on women’s data to check its efficiency after training on a new target.

# 2 Data

In this paper StatsBomb Open Data [24] will be used. StatsBomb is one of the leading football events operators. Its data can be converted to SPADL format [5]. It contains information about events of matches from 40 competitions. In this work, we will only use national club competitions, because they have more matches than international competitions. They also have a wider range of players of different skills, because international competitions are harder to qualify for and usually involve only elite players. This could add some unwanted bias to our model.

Filtered from unnecessary competitions data is divided into female and male competitions datasets. Each of them is split into training, validation, and test data, so after a common gender independent model is trained and examined, data can be used to conduct separate experiments. Each dataset is converted to SPADL [5]. For each action time, used to do it, is calculated. Then this time is categorized and each action is assigned a time interval (eg. from 1 to 2 seconds) instead of time spent. It is done so, because of set pieces. They take a lot more time than an average action. Output values depending on n of actions before scoring or conceding a goal are generated where n is in a range between 5 and 12. It is done to train various models and find out which one is the most accurate. Categorical values: type, result, body part, and time interval are encoded with one hot encoder.

As the result of applying the described filters the following competitions were left in the dataset:

* Spanish La Liga, seasons 2004/2005-2020/2021 (male);
* English Premier League, season 2004/2005 (male);
* English FA Women's Super League;
* USA National Women's Soccer League.

That gives a total of 362 female and 552 male matches.

After splitting them into training, validation, and test dataset, we get the following figures (Table 1).

| Dataset | Matches | Actions | Passes | Successful Passes |
| --- | --- | --- | --- | --- |
| Women training | 271 | 523142 | 215014 | 162594 |
| Men training | 414 | 914148 | 387044 | 325684 |
| Women validate | 91 | 171546 | 70213 | 52641 |
| Men validate | 138 | 305432 | 129326 | 108537 |
| Women test | 73 | 138288 | 56522 | 42320 |
| Men test | 111 | 246435 | 104155 | 87476 |

Table 1. Number of matches, actions, passes and successful passes in split datasets

The table should be read the following way: the training dataset for women consists of 271 matches. There are a total of 523142 SPADL actions in these matches. 387044 of them are passes. 325684 of them are successful passes.

# 3 Models Trained on Male Data

The idea behind the model is similar to the VAEP idea. In the case of scoring, if a goal occurs within n actions after the current action, we assume this action was good. However, instead of CatBoost, a neural network is used to rate the actions. The network consists of 6 sequential layers: linear nx -> 20 features, ReLU, linear 20 -> 10 features, ReLU, linear 10 -> 1, Sigmoid. nx is the number of features in the dataset after encoding. The model is implemented with PyTorch, based on the nn.Module class. All training, validating, and experimenting are conducted in Google Colab to utilize CUDA devices. The source code is available on GitHub: https://github.com/IjonTichy42/master\_thesis\_code.

For training evaluation and validation, a binary cross-entropy loss function is used. While validating, the model results are compared to the target with n actions before goal = 12.

For conceding goals there is a separate model, but the logic is the same. If soon after a specific action a goal is conceded, we suggest, this action was bad.

The important consequence of such architecture is that no action will have a rating equal to one. So, on the contrary to VAEP, goal scorer is not getting a rating boost equal to 1 – goal scoring probability after an assist. It should make the model less biased towards the strikers. On the other hand, great finishers will get very small rewards for overperforming and poor finishers will not be fined greatly for low effectiveness in front of goal. From the practical point of view, this issue can be addressed by comparing players’ goals count with their xG.

Another characteristic of this model is that the start location of action affects the action rating. This feature is not an accomplishment of a player from a strictly on-ball point of view. The ball was passed there by a different player than the one who received it. However, taking this into account allows the model to rate a player's ability to get the ball in a good position and to use the angle to the goal. For example, we expect a shot from the penalty spot to have a considerably higher rating than a shot from the sideline.

The output of rating actions of a particular game is a tensor of floats. Its length is equal to the length of actions in the game. Each of those numbers represents the model's assessment of action. These actions are tied to players and teams, so each player's performance can be assessed by calculating the mean of ratings of the player's action can be obtained by doing similar calculations, as well as the team rating. The latter will be used for model validation.

## 3.1 Training the Models on Male Data

First, we train models on male training data. As stated in the problem formulation section, models with the parameters obtained will be applied to both male and female data.

Initially, both scoring and conceding models are trained for 3500 epochs. The scoring model has a learning rate of 0.01 and conceding model of 0.0075. These rates were obtained empirically. The scoring learning curve goes a longer way down.

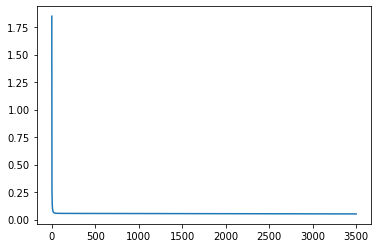


Figure 1. Learning curve for scoring. Male data. N = 5. Learning rate = 0.01

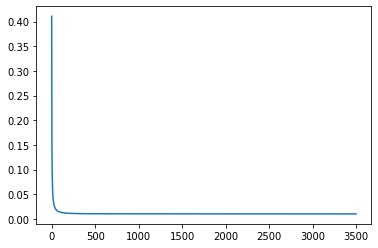


Figure 2. Learning curve for conceding. Male data. N = 5. Learning rate = 0.0075

Binary cross entropy criterion is used to evaluate loss while training.

## 3.2 Validating the Male Data Models

Validating the models, we assume that a team that makes more good actions and less bad actions. Scoring and conceding outputs are validated separately because the optimal n parameter can be different for scoring and conceding.

Validation of the scoring model goes the following way. We take the games from the validation dataset that ended with a victory of one of the teams. We assume the better team performed actions with higher ratings which led to the victory. Each trained model is tested this way to find out which one allows it to predict more outcomes. If two or more models show the same results, the one with higher prediction rates after excluding the goal-scoring actions is chosen, so the less biased towards finishers one is preferred. The algorithm is the following: all the actions that are goals (shots that succeeded in terms of SPADL) are detected and their respective ratings are excluded from the rating array. It is dictated by the fact that VAEP is biased towards the strikers and, taking into account similarities in model training, we suggest our model might have biases towards strikers too. Moreover, from the industry point of view, finishing efficiency is well assessed with the shot-based metrics, so developing the possession-based metric there is a reason to focus on actions other than scoring goals.

Validation of the conceding model goes a similar way. The only difference is that the output of it is an anti-rating, so the less the mean of these actions, the better the result we expect.

There are a total of 117 games in the male validation dataset that did not end up in a draw, therefore, suitable for this procedure.

The best result among the conceding models was shown by the one with n = 6. Based on it, we predict 80 out of 117 outcomes correctly. We apply a one-sided binomial test to evaluate the significance of this result. The null hypothesis is that by guessing we can obtain this result (parameter of binomial test p = 0.5). The alternative hypothesis is that our model adds to the probability to predict the correct outcome. The p-value we obtain is approximately 0.0004. Therefore, we reject the null hypothesis and assume our model can assess the badness of the actions.

The best result among the scoring models is 50 out of 117 outcomes. It was shown with n = 8. It is less than 50%, so the best model is useless or harmful for the purposes of this work.

As a result, we take an attempt to create a better model. We take the best one, tweak learning hyperparameters and train another model. The change is increasing the number of training epochs from 3500 to 10000. We use the n = 8 target as the one that allowed us to show the best results at the previous attempt.

The new model gives 68 correct predictions out of 117 matches. A similar to described before a one-sided binomial test is used. It gives a p-value of approximately 0.0478. This allows us to reject the null hypothesis and assume, this model rates actions that lead to the victory, higher than others.

Blending these two models via subtracting conceding ratings from scoring allows us to predict 81 out of 117 outcomes. So, we assume it is useful too.

## 3.3 Applying Male Data Models to Test Data

First, these models are applied to the male test dataset. Model efficiency is evaluated the same as on the validation step. For the scoring model one-sided test gives us a p-value equal to approximately 0.0198 (58 out of 95). Conceding model's result is approximately 0.0002 (65 out of 95). Their blend's result is approximately 0.0001 (66 out of 95). This proves, our models can rate actions in any football game, not only in the ones from the validation dataset.

This is proof of the quality we discussed in the introduction.

Given the results we have achieved, the hypotheses for the female test dataset application of these experiments are different.

The null hypothesis for the scoring model: the scoring model trained on the men's football matches and effective for assessing men's actions cannot be applied to assess the women's performances. The alternative hypothesis is that this scoring model is suitable for assessing women's performances. Similar null and alternative hypotheses are for conceding and general models.

The application to the female test dataset gives us the following p-values: 0.1791, 0.5522, and 0.2559 for scoring, conceding, and blend respectively. No result allows us to reject the null hypothesis.

Assessing players' performance takes place at this step. Each player is rated by calculating the mean of her or his action in every match. Then the mean of all ratings of a particular player is assigned to her or him as the final rating of a player based on her or his actions for the chosen matches. None of the models have proven to be effective on the female dataset. However, women's ratings are calculated to compare them to the ratings generated by the women's models we train later.

## 3.4 Robustness Check

To gain deeper insights on how the model works we assess the performances of players in a particular match and compare the ratings we obtain with the ratings of the Whoscored site. The algorithm of Whoscored is not revealed [21], however, there is evidence that it uses more traditional metrics: in a Q&A its co-founder has stated that even a match result gives bonuses to player ratings [25]. In the Whoscored algorithm, every player starts with a rating of 6.0. Then, with every action, the footballer's rating is updated. So, a 6.0-7.0 rating is a mediocre game, > 7.0 is a good game, < 6.0 is a poor performance. In this work, it is used because Whoscored ratings are a conventional way to check a footballer's performance [21].

It is not considered to be ground truth, but a different approach that helps us highlight the limitations and advantages of our model.

The match to be assessed is the one with match id 16182. It is a Barcelona versus Real Valladolid game. It is picked for two reasons: it is the first game in the men’s test dataset and it took place in 2019, so we can be sure relatively modern algorithms of Whoscored were used to describe it. We have no evidence if Whoscored updates older matches’ ratings after updating their algorithms, so if it were a 2004 game, we would have to look deeper into our dataset to find a more up-to-date match.

Here is a brief overview of this match: it was decided by one goal, scored from a penalty by Messi in the 43rd minute of the game. Barcelona took 20 shots, 8 of which were on target. Real Valladolid took 10 shots, none of which were on target [22].

From the tournament point of view, Barcelona was one of the leaders, eventually winning the title that season, meanwhile Rayo Vallecano was one of the worst teams, ending up 16th out of 20. Victory is rewarded by 3 points, draw brings each team a point, and loss goes unrewarded. Taking this into account, we assume that a victory was the only result suitable for Barcelona, meanwhile Rayo Vallecano would have probably been content with a draw.

Here are the ratings we have obtained:

| Player name | Player rating |
| --- | --- |
| Arturo Erasmo Vidal Pardo | 0.015643 |
| Carles Aleña Castillo | 0.011586 |
| Gerard Piqué Bernabéu | 0.029896 |
| Ivan Rakitić | 0.017384 |
| Jordi Alba Ramos | 0.004301 |
| Kevin-Prince Boateng | 0.009705 |
| Lionel Andrés Messi Cuccittini | 0.013816 |
| Luis Alberto Suárez Díaz | 0.011725 |
| Marc-André ter Stegen | 0.027018 |
| Ousmane Dembélé | 0.006037 |
| Philippe Coutinho Correia | 0.009425 |
| Sergi Roberto Carnicer | 0.020285 |
| Sergio Busquets i Burgos | 0.018270 |
| Thomas Vermaelen | 0.009397 |

Table 2. Barcelona scoring ratings

| Player name | Player rating |
| --- | --- |
| Antonio Jesús Regal Angulo | 0.005870 |
| Anuar Mohamed Tuhami | 0.010834 |
| Daniele Verde | 0.018870 |
| Enes Ünal | 0.012161 |
| Fernando Calero Villa | 0.005870 |
| Francisco José Olivas Alba | 0.011889 |
| Joaquín Fernández Moreno | 0.005598 |
| Jordi Masip López | 0.004501 |
| José Ignacio Martínez García | 0.019662 |
| Laureano Antonio Villa Suárez | 0.004807 |
| Miguel Alfonso Herrero Javaloyas | 0.013617 |
| Sergio Gontán Gallardo | 0.006539 |
| Sergio Guardiola Navarro | 0.021387 |
| Stiven Ricardo Plaza Castillo | 0.012188 |

Table 3. Real Valladolid scoring ratings

| Player name | Player rating |
| --- | --- |
| Arturo Erasmo Vidal Pardo | 0.000506 |
| Carles Aleña Castillo | 0.000332 |
| Gerard Piqué Bernabéu | 0.007244 |
| Ivan Rakitić | 0.003201 |
| Jordi Alba Ramos | 0.000065 |
| Kevin-Prince Boateng | 0.001368 |
| Lionel Andrés Messi Cuccittini | 0.000167 |
| Luis Alberto Suárez Díaz | 0.000852 |
| Marc-André ter Stegen | 0.006026 |
| Ousmane Dembélé | 0.000020 |
| Philippe Coutinho Correia | 0.000021 |
| Sergi Roberto Carnicer | 0.004899 |
| Sergio Busquets i Burgos | 0.001410 |
| Thomas Vermaelen | 0.000417 |

Table 4. Barcelona conceding ratings

| Player name | Player rating |
| --- | --- |
| Antonio Jesús Regal Angulo | 0.000096 |
| Anuar Mohamed Tuhami | 0.000898 |
| Daniele Verde | 0.005623 |
| Enes Ünal | 0.002497 |
| Fernando Calero Villa | 0.001273 |
| Francisco José Olivas Alba | 0.000067 |
| Joaquín Fernández Moreno | 0.000122 |
| Jordi Masip López | 0.000006 |
| José Ignacio Martínez García | 0.007172 |
| Laureano Antonio Villa Suárez | 0.000015 |
| Miguel Alfonso Herrero Javaloyas | 0.001064 |
| Sergio Gontán Gallardo | 0.000153 |
| Sergio Guardiola Navarro | 0.012939 |
| Stiven Ricardo Plaza Castillo | 0.000704 |

Table 5. Real Valladolid conceding ratings

The general ratings are presented with the Whoscored ratings.

| Name | Model Rating | Whoscored Rating |
| --- | --- | --- |
| Arturo Erasmo Vidal Pardo | 0.005774 | 7.4 |
| Carles Aleña Castillo | 0.009936 | 7.0 |
| Gerard Piqué Bernabéu | 0.013247 | 8.0 |
| Ivan Rakitić | 0.014182 | 6.1 |
| Jordi Alba Ramos | 0.004235 | 7.1 |
| Kevin-Prince Boateng | 0.008337 | 6.4 |
| Lionel Andrés Messi Cuccittini | 0.013648 | 8.0 |
| Luis Alberto Suárez Díaz | 0.010873 | 6.4 |
| Marc-André ter Stegen | 0.020992 | 6.4 |
| Ousmane Dembélé | 0.006017 | 7.1 |
| Philippe Coutinho Correia | 0.009404 | 6.5 |
| Sergi Roberto Carnicer | 0.015386 | 7.5 |
| Sergio Busquets i Burgos | 0.016861 | 7.5 |
| Thomas Vermaelen | 0.008980 | 7.1 |

Table 6. Barcelona general and Whoscored ratings

| Name | Model Rating | Whoscored Rating |
| --- | --- | --- |
| Antonio Jesús Regal Angulo | 0.005774 | 6.3 |
| Anuar Mohamed Tuhami | 0.009936 | 7.4 |
| Daniele Verde | 0.013247 | 5.9 |
| Enes Ünal | 0.017440 | 6.0 |
| Fernando Calero Villa | 0.010888 | 6.0 |
| Francisco José Olivas Alba | 0.011822 | 6.4 |
| Joaquín Fernández Moreno | 0.005476 | 6.9 |
| Jordi Masip López | 0.004495 | 8.8 |
| José Ignacio Martínez García | 0.012490 | 6.2 |
| Laureano Antonio Villa Suárez | 0.004792 | 6.0 |
| Miguel Alfonso Herrero Javaloyas | 0.012554 | 6.5 |
| Sergio Gontán Gallardo | 0.006387 | 6.1 |
| Sergio Guardiola Navarro | 0.008448 | 6.8 |

Table 7. Real Valladolid general and Whoscored ratings

The mean of all test dataset ratings is 0,0114.

Our model's ratings and Whoscored ratings have a few key discrepancies. Judging by Whoscored, Jordi Masip López, Real Valladolid goalkeeper had an outstanding game, whereas Marc-André ter Stegen, Barcelona goalkeeper, was run of the mill. We suggest the ratings for them were calculated this way because of Masip's shot-stopping performance (8 shots on target, only 1 goal) and ter Stegen's low goalkeeping workload (0 shots on target). Our model, however, suggests that Masip had a very modest game, while ter Stegen was the best player on the pitch and one of the best in scoring input specifically. Based on this, we can say that our models do not reward a good shot-stopping performance generously, but rate the keepers from the same point of view as it does with other players. It is a common feature of possession-based models. Shot-stopping performance is described by shot-based metrics [7].

Based on the Whoscored algorithm, we assume Messi was the best field player in that game. Our model rates his performance high too, but there are some players with better ratings. To interpret that fact, we investigate the number of actions performed by each player in the game.

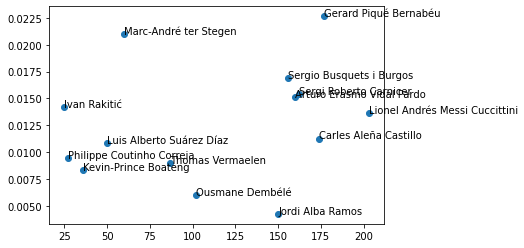


Figure 3. Ratings and action count for Barcelona players in the match against Real Valladolid

The X-axis is actions count. It represents the involvement of a player in the game. The Y-axis is ratings. It represents the effectiveness of a player.

It turns out, Messi was the most involved footballer that night with 203 actions, while the second-placed Pique performed only 177 actions. It allows us to suggest that Messi’s actions, despite being not the best from our model’s perspective, played an important role in keeping possessions alive. The simple way to get a more complete picture is to use our model’s ratings along with the actions value counts by players.

## 3.5 Conducting Additional Experiments

Passing is an essential part of a game of football and the most frequent on-ball action. In the dataset we use, passing occurs 962274 times, more than any other action. We investigate if good passing aligns with an outcome of a game. We use the same models as for the general assessment. Everything except passes is removed from the games of the test dataset.

First, we test our models on the male dataset. The null hypothesis for the scoring model is that the result is not significantly affected by the average scoring threat of passes alone. The alternative hypothesis is that the advantage in the scoring threat of passes alone has a sufficient impact on an outcome. 62 correct predictions out of 95 results into the p-value of a one-sided T-test equal to approximate 0.0019. Thus, we reject the null hypothesis.

The null hypothesis for the conceding model is similar: the result is not significantly affected by the average conceding threat of passes alone. The alternative hypothesis is that the disadvantage in the conceding threat of passes alone has a sufficient impact on an outcome. The conceding model predicts 61 out of 95. The p-value is 0.0037, so the null hypothesis is rejected.

The null hypothesis for the blend: the result is not significantly affected by computing the general quality of teams' passing. Alternative hypothesis: the general quality of passing significantly impacts the outcome of the game. The blend predicts 69 out of 95 winners correctly and the one-sided T-test p-value is less than 0.0001, which provides us with evidence to reject the null hypothesis.

The same set of experiments was applied to the female test dataset. The null hypothesis for the scoring model is the following: the model trained on men's actions, which effectively assesses men's passing input into the score, is not suitable for assessing women's passing. The alternative is that we can effectively assess women's actions with this model. The model shows the passing advantage of the team that won the game in 33 cases out of 58. The p-value of the one-sided binomial test is approximately 0.1791. Thus, we do not reject the null hypothesis. The results for the rest of the models are worse than this one, so we assume they are not suitable for women's football too.

# 4 Female Data Models

# 4.1 Training and Validating the Models on the Female Data

The results we obtained by applying our models trained on men's data to women's competitions do not prove we can use the same rating system to distinguish good actions from bad ones in women's soccer. Therefore, we train and validate our models the same way as on the men's data.

The best results for scoring were shown by the model trained for 10000 epochs with a learning rate equal to 0.01 of the n=11 target. We apply a one-sided binomial test; null-hypothesis: simply by guessing we could obtain this result; alternative: we need additional knowledge to predict this many outcomes. 48 out of 75 outcomes were predicted correctly, so the p-value for the test is 0.0101. Therefore, we reject the null hypothesis.

For the conceding model, however, we could not find a set of parameters that would return ratings based on which we can consistently predict the outcome. For 3500 epochs experiments, the best results were predicted by the models with small n (5 to 7). However, tweaking the learning rate and the number of epochs did not improve the results considerably. The blend of models failed to provide us with credible results either.

## 4.2 Testing the Female Data Scoring Model

Similar tests on the female test data provided us with 37 correct predictions based on all actions' ratings out of 58 matches that did not end up in a draw. The hypotheses for one-sided binomial tests for all actions are: null is the probability to predict this many outcomes is 0.5, therefore our model is useless; the alternative is this probability is higher than 0.5. 0.024 p-value allows us to reject the null hypothesis, so we assume our model is effective.

Rating passing performances have the same efficiency, so we assume that our model is effective for rating passing, as well as that passing is a key parameter for evaluating women’s performances in a football match.

These results prove that the machine learning approach is appropriate for evaluating performance in women’s soccer.

The failure with the conceding model raises a lot of questions. Our suggested possible interpretations are:

* The amount of data used for training female conceding models was not sufficient;
* The structure of women’s games is different from men’s ones.

The first suggestion implies that collecting and publishing data about women’s soccer should be continued, so more powerful models can be trained.

The second suggestion implies that women’s poor decisions on the football pitch have a weaker connection with conceding goals. This raises the question, how modern performance assessment models should fine female players for their mistakes.

# 5 Discussion

# 5.1 Further Use of the Models

We have created models that have proven their quality on the test datasets. This means they can be utilized in the industry. Particularly the results shown by the male data models allow us to suggest a more flexible approach than simply using a blend of scoring and conceding models along with the involvement measurements.

The approach we suggest potentially allows us to conduct a deeper analysis. Separate ratings for scoring and conceding provide us insights into how threatening a player has been towards the opponent and his own goal. For example, in the match we analyzed in section 3.4 Sergi Roberto has a lower general rating than Sergio Busquets. However, a look at separate ratings tells us he was better in creating threats to the opposition, but at the same time took riskier decisions for his team. Based on the team style of play, the context of the match, and the position of a player it might be a desirable trait as well as undesirable. The context of the game is another thing that can be taken into account. The match we analyze can be logically split into two logical segments:

1. Before Barcelona scored on the 43rd minute;
2. After Barcelona scored.

In the first segment Barcelona were probably more interested in scoring and more willing to take risky decisions, whereas after scoring, risks were unnecessary. On the other hand, Rayo Vallecano in the first segment was probably looking to keep their goal safe and had to take risky decisions in the second segment, even though they could lead to more danger for their own goal. Our approach allows us to split a single game into segments and rate actions of each segment separately based on the team’s target in the current segment.

This model is harder to interpret. Whoscored has a clear scale of ratings. VAEP and xT models offer a range of probabilities, which means that there eventually is a scale from 0 to 1 for every action too. Meanwhile, our model offers fractional ratings with no real-life units. The suggested solution for this is scaling the action ratings from 0 to 1 and calculating the coefficient, making the numbers look more representable. However, it still would not map our ratings to any units. Moreover, the best action ever might have not been met in our datasets, meaning that there still would be a probability to obtain a >1 rating for a particular action.

Apart from comparisons with conventional ratings, there is one feature highlighted by this analysis. The conceding ratings are lower, so our general model might be too rewarding for risky decisions.

## 5.2 Theoretical results

Based on these results, we suggest that possession-based models should be trained separately for men and women to produce more accurate results for women. On the other hand, this implies that adding female data to a training dataset for a model, intended to evaluate men’s performance, would probably decrease its quality too.

The suggested interpretation is that women play soccer somehow differently. So, the strategies of their play probably should alter from men's soccer strategies, even though passing performance has significant input in outcomes for both genders. We do not know what factors lead to this situation. Hopefully, the growth of women's football popularity will create more equal training and working conditions, and that will narrow down the list of possible causes.

The results of this work lead to some questions that require further discussion and research.

The difference shown by the possession-based models along with the facts investigated in the "Scaling Demands of Soccer According to Anthropometric and Physiological Sex Differences: A Fairer Comparison of Men's and Women's Soccer" paper leads to the hypothesis that shot-based models trained separately too. However, in the recent paper "Women's Football Analyzed: Interpretable Expected Goals Models for Women." Bransen et al. have investigated this question and found out that xG models are applicable cross-gender despite some differences [3].

The failure of all goal conceding models for women is an indicator that there is probably a better way to evaluate failure and risk in women's soccer and in soccer in general. Not only goals are not a frequent event in football, but also on-ball actions that directly lead to conceding are either not frequent or have too few things in common. One possible suggestion might be using a change of possession as an indicator of failure. However, it is obvious that players in some positions would be a lot more prone to lose possession than others. One way to remove bias is to rate the position where the possession was lost based on the average possession loss position or on the goal conceding probability after losing possession in a specific position.

# 6 Conclusion

The research we have conducted answers the questions formulated in the introduction in the following way.

Although we cannot state that models trained primarily on male data are suitable for evaluating women's performance, we found no evidence in favor of their applicability

Training on the female data has created an efficient model to evaluate female soccer players’ performances.

The additional outcome of this work is a set of models trained to evaluate a player's performance in football.

# References

[1] Anzer, Gabriel & Bauer, Pascal. (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. 10.3389/fspor.2021.624475.

[2] Bradley, Paul & A, Dellal & Mohr, Magni & Castellano, Julen & Wilkie, Anna. (2013). Gender differences in match performance characteristics of soccer players competing in the UEFA Champions League. Human movement science. 33. 10.1016/j.humov.2013.07.024.

[3] Bransen, Lotte, and Davis, Jesse. “Women’s Football Analyzed: Interpretable Expected Goals Models for Women.” AI for Sports Analytics (AISA) Workshop at IJCAI 2021, 2021, pp. AI for Sports Analytics (AISA) Workshop at IJCAI 2021; 2021.

[4] Bransen, Lotte, Van Haaren, Jan and van de Velden, Michel. "Measuring soccer players’ contributions to chance creation by valuing their passes" Journal of Quantitative Analysis in Sports, vol. 15, no. 2, 2019, pp. 97-116. https://doi.org/10.1515/jqas-2018-0020

[5] Decroos T., Bransen L., Van Haaren J., Davis J. Actions Speak Louder Than Goals: Valuing Player Actions in Soccer. KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining July 2019 Pages 1851–1861 https://doi.org/10.1145/3292500.3330758

[6] Decroos T., Bransen L., Van Haaren J., Davis J. VAEP: An Objective Approach to Valuing On-the-Ball Actions in Soccer. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. Sister Conferences Best Papers. Pages 4696-4700. https://doi.org/10.24963/ijcai.2020/648

[7] Gelade, Garry. "Evaluating the ability of goalkeepers in English Premier League football" Journal of Quantitative Analysis in Sports, vol. 10, no. 2, 2014, pp. 279-286. https://doi.org/10.1515/jqas-2014-0004

[8] Gómez López, Maite & Álvaro, Jordi & Barriopedro, Maria. (2008). Behaviour patterns of finishing plays in female and male soccer.

[9] Lange, Paul & Manesi, Zoi & Meershoek, Robert & Yuan, Mingliang & Dong, Mengchen & Van Doesum, Niels. (2018). Do male and female soccer players differ in helping? A study on prosocial behavior among young players. PLOS ONE. 13. e0209168. 10.1371/journal.pone.0209168.

[10] Lawrence T. Introducing xGChain and xGBuildup. URL: https://statsbomb.com/2018/08/introducing-xgchain-and-xgbuildup/

[11] Llana S., Madrero P., Fernández J. The right place at the right time: Advanced off-ball metrics for exploiting an opponent’s spatial weaknesses in soccer // MIT Sloan Sports Analytics Conference – 2020

[12] Mackay N. What is a possession-based model (and why does it matter)? URL: https://mackayanalytics.nl/2016/11/11/what-is-a-possession-based-model-and-why-does-it-matter/

[13] Pappalardo, Luca & Cintia, Paolo & Ferragina, Paolo & Massucco, Emanuele & Pedreschi, Dino & Giannotti, Fosca. (2019). PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach. ACM Transactions on Intelligent Systems and Technology. 10. 1-27. 10.1145/3343172.

[14] Pappalardo, Luca & Rossi, Alessio & Pontillo, Giuseppe & Natilli, Michela & Cintia, Paolo. (2021). Explaining the difference between men's and women's football.

[15] Pedersen Arve Vorland, Aksdal Ingvild Merete, Stalsberg Ragna. "Scaling Demands of Soccer According to Anthropometric and Physiological Sex Differences: A Fairer Comparison of Men’s and Women’s Soccer" // Frontiers in Psychology, 10, 2019, 762 p. URL=https://www.frontiersin.org/article/10.3389/fpsyg.2019.00762 DOI=10.3389/fpsyg.2019.00762

[16] Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. Journal of Human Sport and Exercise. 12. 10.14198/jhse.2017.12.Proc2.05.

[17] Robberechts, Pieter. "Valuing the Art of Pressing" StatsBomb; StatsBomb Innovation In Football Conference, Date: 2019/10/11 - 2019/10/11, Location: London, UK. Proceedings of the StatsBomb Innovation In Football Conference; 2019; pp. 1 – 11

[18] Singh K. Introducing Expected Threat (xT). URL: https://karun.in/blog/expected-threat.html

[19] Van Roy M., Robberechts P., Decroos T., Davis J. Valuing On-the-Ball Actions in Soccer: A Critical Comparison of xT and VAEP. AI in Team Sports Organising Committee; The AAAI-20 Workshop on Artifical Intelligence in Team Sports, Date: 2020/02/08 - 2020/02/08, Location: New York, USA Proceedings of the AAAI-20 Workshop on Artifical Intelligence in Team Sports; 2020; pp. 1 – 8

[20] Wolf, Stephan, Schmitt, Maximilian, and Schuller, Björn. ‘A Football Player Rating System’. 1 Jan. 2020 : 243 – 257.

[21] https://1xbet.whoscored.com/Explanations

[22] https://1xbet.whoscored.com/Matches/1316753/Live/Spain-LaLiga-2018-2019-Barcelona-Real-Valladolid

[23] https://github.com/ML-KULeuven/socceraction

[24] https://github.com/statsbomb/open-data

[25] https://medium.com/@dannypage/q-a-about-whoscored-ratings-160eebbbbccf