Player Ranking by Pitch Control and Expected Threat

FC Twente Project Group 1

1st Vo Nhat Minh
Technical Computer Science
University of Twente
s2503042

2nd Phuoc Ho
Technical Computer Science
University of Twente
s2587130

3rd Duc Duc Tran
Technical Computer Science
University of Twente
s2482266

4th Vladimir Nikolov
Technical Computer Science
University of Twente
s2311747

***Abstract*—Analysis of players' performance has always been a great concern in the field of football analysis. Over the years, different approaches have been conducted to estimate the impact and performances of players during a match. These data can be used to get measure the actual values that a player can bring to a team and also help players themselves realize their weaknesses to improve through the training. This paper will discuss one approach that can be used to value the action and decision made by a player during a match and from that come up with a ranking metric for the overall performance of a team as well as the players, and this is the formula using the combination of Pitch Control and Expected Threat(xT) model.**

***Index Terms*—Football Analytics, Pitch Control, Player Ranking, Data Visualization, Expected Threat(xT)**

I. Introduction

A. Motivation

For any competitive sport, the athletes need to improve. They go through a rigorous training regime to strengthen their bodies and minds. However, victory in a sport like a football is not limited to individual contributions but also the overall strategy and teamwork. Given all the various exercises and games the players do, it is hard to identify strengths or weaknesses. Therefore, there is a growing interest in developing and perfecting models to analyze the play and team performance.

B. Context

The models use anonymised tracking data provided by Football Club Twente. This data includes information about the positions of players and the ball on the pitch, as well as the roles of those players and their teams. Inactive players, referees or other staff are excluded. Event-specific data is also included such as actions with information about outcomes and involved players. The data is listed in frames, with respective identification and timestamp. Due to the anonymity, a few assumptions are made. Mainly player, team and frame identifications remain the same within all games from the data set and are unique to what is represented. The game is "association football" referred to later as football. The relevant area or the area the players use is 105 by 68 meters referred to as the pitch.

C. Goal

The main intended result of this project is to find a metric for which player performance can be measured. By finding out the metrics, we focus on figuring out the way to understand the given football data, to get some key insights and metrics for evaluation players. This metric must be influenced by the individual skills of the player as well as the contributions to the team. In order to complete this objective the following sub-objectives will be necessary:

- Determining the the control region based on player's position using Pitch Control model
- Determining under what circumstances can a team or player become a threat to the opponent using Expected Threat model. For football, this happens when someone possesses the ball, so the objective is determining the likely hood of capturing the ball
- Analyze the impact of player's decision
- Find a way to combine all of the above into a singular ranking method. This is also the main objective.

Besides the functional goal for the analysis of this project, which includes metrics and approaches for analysis, we also focus on how to make the visualization clear and explainable, which helps understand the analyzing processes of this project. Hence, along with the technical approach, the analyzing and visualizing approach should be compatible with the processed data.

D. Summary of prerequisites

To achieve the goals of this project, an adaptation of the following prior models will be made:

- Pitch Control [1] - by William Spearman, 2017. This model the probability a team will control the ball at some location, given it was passed from another location. In particular, the model can show how well the teams control a given region. Offering accurate evaluation of the current game situation.
- Expected Threat (xT) [2] - by Sarah Rudd, 2011. This model uses Markov Chains [5] to calculate the probabilities of some actions based on the probabilities from a

series of past actions. For football shooting and moving the ball are the actions that are of interest. Based on past data it can be calculated how likely a player is to perform the actions with their respective probability of success, given some position on the pitch. This effectively can calculate the likelihood of succeeding, for example scoring a goal, thus this is a way to evaluate the threat. This model was later repurposed by Karun Singh [3], which is currently the preferred variant.

While pitch control evaluates current situations, it cannot predict the likeliness of outcomes. Expected Threat is more of a statistical analysis that does not use current events and does not account for opponent reactions. Therefore, to achieve the main project goal, a combination of both methods will be used to create a combined formula. The result should measure the threat of a player by combining the probability of possible actions occurring with the likeliness of situational development.

II. Expected Threat Model

Expected threat according to Sarah Rudd [2] defines a play in a match as sequences of consecutive on-the-ball actions where the same team possesses the ball until it results in 1 out of 2 final states: losing the ball or scoring a goal. The play sequence is then considered as time series data to put into a discrete-time Markov process to measure the goal likelihood given that play sequence. Therefore the basic idea behind xT is to divide the pitch into a grid, with each cell reflecting the probability (xT values) leading to a goal in the next N actions(play sequences). This approach allows us to value not only parts of the pitch from which scoring directly is more likely but also those from which an assist is most likely to happen. Actions that move the ball, such as passes and dribbles, can then be valued based solely on their start and end points, by taking the difference in xT between the start and end cell. Our implementation is based on Soccermatics's (position-based) xT model [4].

A. Data

The required data for Expected Threat is the Event Stream Data type which is the Tracking Data only for the player with the ball point of view, performing a certain action(passing, dribbling, etc). In our implementation, we use the Event Stream Data for 6 matches provided by FC Twente. From which, we extract 3 different types of events namely: moving ball events, shooting events and goal events. The moving ball events consider events that move the ball from 1 location to a different location of the field, and for that, we filter all "Pass" and "Take on(dribble)" event types from data "Fig. 2". Shooting events are the combination of all shoot attempts by players, so we filter all "Attempt saved", "Miss" and "Goal" events from the data "Fig. 1". Goal events are just "Goal" type events "Fig. 3". Those events are later converted as 3 heat map grid size 16x12 where the start locations of those events were plotted as below.

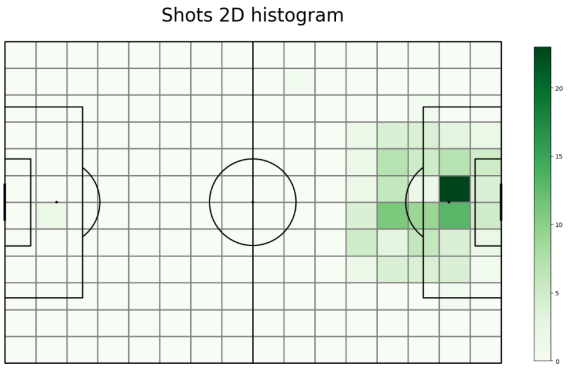

Fig. 1. Move ball events heat map.

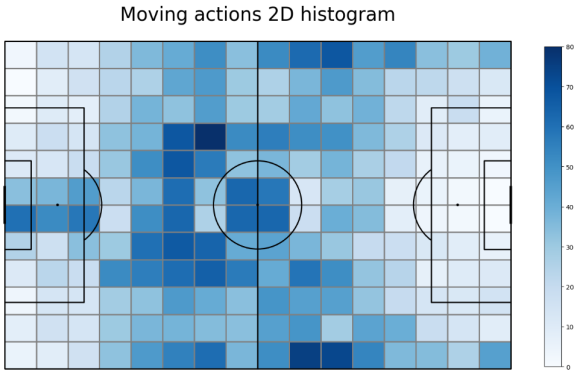

Fig. 2. Move ball events heat map.

B. Formula and Model

The Markov model view allows deriving these xT values from historical events data by iteratively solving the following equation defined by Karun Singh [3]:

$$xT_{x,y} = (s_{x,y} \times g_{x,y}) + (m_{x,y} \times \sum_{x'}^{12} \sum_{y'}^{16} T_{x,y \to x',y'} \times xT_{x',y'})$$

Where xT(x,y) is the likelihood of a goal occurring during the possession sequence starting at cell (x,y), s(x,y) is the

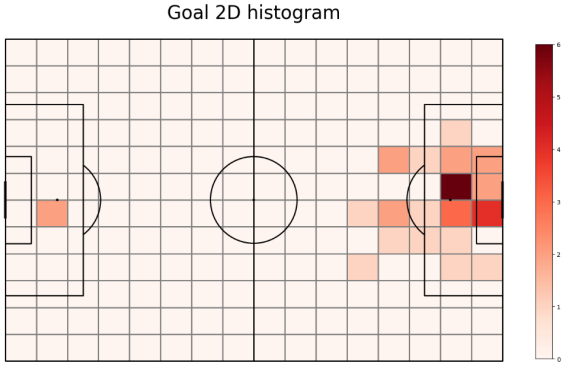

Fig. 3. Goal events heat map.

probability of a shot being taken - calculated from shooting events, g(x,y) is the probability of a goal being scored given that a shot is taken - calculated from goal events, m(x,y) is the probability of the ball being moved (pass or dribble) - calculated from moving ball events, and T(x,y)→(x',y') is a transition matrix which is the probability of successfully transitioning from the cell (x,y) to all other cells (x',y') on the pitch given that a move ball action is taken.

In Expected Threat(xT) using the Markov model, the transition matrix is the core of the calculation. The transition matrix is a 4D array T[x][y][x'][y'] where (x,y) is the start cell, (x',y') is the target cell of action and we have T[x][y] is a 2D matrix showing the success probability of moving the ball from start cell x,y to all other cells on the fields. Based on past provided "moving ball events" data from 6 matches, we can group all passing or dribbling action starting from cell x,y and then calculate the probability of moving a ball from (x,y) to (x',y') is based on the number of actions from (x,y) to (x',y') divides by the total number of actions start from (x,y).

From the formula, we can see that to compute the xT value for a cell (x,y) it requires that we already know the xT value for all the other cells. Therefore, we need to initialize xT(x,y) = 0 for all (x,y), and evaluate the formula for several iterations until convergence (almost every xT values are not 0). During each iteration, the new xT for each zone is recomputed by using xT values from the previous iteration. For each iteration, we multiply with the transition matrix once and it is corresponding to simulate the prediction of the next actions.

C. Result

The result of xT that we computed after 5 iterations is a 12x16 grid showing the probability that given a ball in cell (x,y) attacking team (team control the ball) can score a goal in the next 5 actions sequence "Fig. 8". Note that the xT grid assumes the attacking direction is from left to right. Now with the given xT, we can value individual player actions by computing the difference in xT between the start and end locations of an event. In other words, we will say that an action that moves the ball from location (x,y) to location (x',y') has value xT(x',y')-xT(x,y).

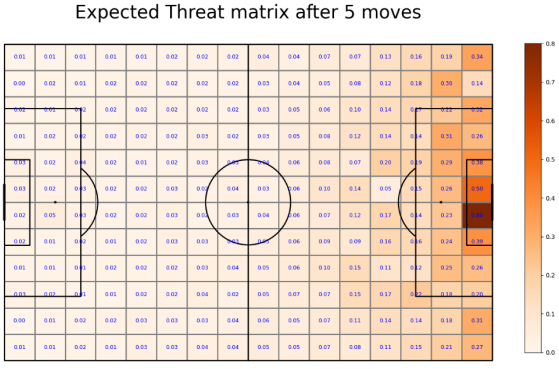

Fig. 4. Final xT heatmap

III. Pitch Control Model

The pitch control model was first introduced by William Spearman, defining pitch control as "the probability that a player can control the ball, assuming it moves directly to that location" [6]. Pitch control allows us to quantify the passing options of a player holding the ball directly.

A. Data

In order to calculate pitch control values, our model uses Tracking data of Players in both teams and the ball and Events data. We also transformed the tracking data (position in coordinates X and Y), originally the root of X and Y in the middle of the field to the root at the bottom left of the field for better visualization and calculation.

B. Formula and Model

Pitch Control values can be calculated for each player at a specific location, r, at time t by the following differential equation defined by William Spearman [1]:

$$\frac{dPPCF_j}{dT}(t, \overrightarrow{r}, T|s, \lambda_j) = (1 - \sum_k PPCF_k(t, \overrightarrow{r}, T|s, \lambda_j)) \times f_j(t, \overrightarrow{r}, T|s) \times \lambda_j \quad (1)$$

In this formula, $f_j(t, \overrightarrow{r}, T|s)$ represents the probability that player j at time t can reach location r within some time T, called "Time to intercept". We will further discuss Time to intercept in subsection (1).
We incorporate the time of flight of the ball by setting $PPCF_j(t, \overrightarrow{r}, T|s, \lambda_j) = 0$ where T is less than the time of flight of the ball at location r.
λ_j is the rate of control and represents the inverse of the mean time it would take a player to make a controlled touch to the ball. We further discuss it in subsection (2).

1) Time to intercept: We calculate f_j, taking into account the uncertainty that in real-life settings, players' unawareness, and errors by this formula, as a translation of the formula by William Spearman by Ramon Dop from the University of Amsterdam [7]:

$$f_j(t, \overrightarrow{r}, T|s) = \frac{1}{1 + e^{-\Pi \times \frac{\sigma * (T - ttc)}{\sqrt{3}}}}$$

Here, σ is the standard deviation (0.45) of the cumulative distribution function of the time it takes to control. ttc is the ideal time it take player j to arrive at location r with some assumptions. We calculate ttc by adopting a simple approximation for calculating arrival time. This approximation is a two-step process [8]:

1) There is an initial reaction time of 0.7s; during this time each player continues along their current trajectory.
2) After 0.7s, each player runs directly towards the target location at their maximum speed of 5m/s.

We illustrate it with the following figure (See Fig. 5.)
Hence, we can compute ttc using the formula [7]:

$$r_{p-new} = r_p + v_p * t_{react}$$

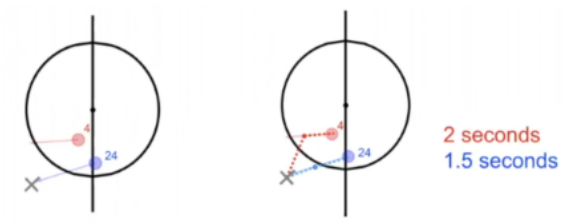

Fig. 5. Two steps ball interception, taken from [8].

$$ttc = d(r_b - r_{p-new})/v_p$$

Here, we assume the time to react t_{react} is 0.7 second, the maximum velocity and acceleration of the player are 5m/s and 7m/s/s respectively, and the ball moves linearly at a constant speed of 15m/s.

2) Control Force: Let us consider that a controlled touch is a stochastic process with a fixed rate. $\lambda_j \times \Delta t$ is then the probability that a player in the vicinity of the ball can control the ball in Δt. For our model, we used $\lambda = 4.3\ (1/s)$ suggesting each player typically takes 0.25s to control the ball.

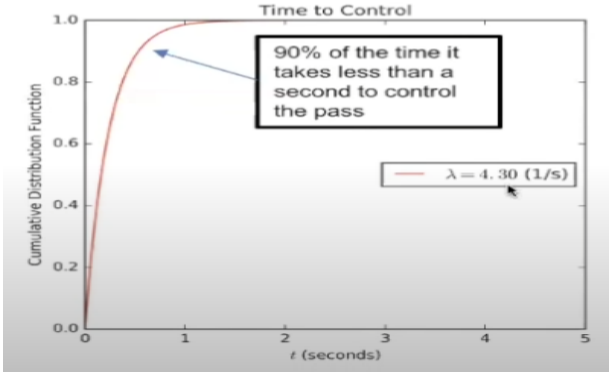

Fig. 6. Cumulative Distribution Function, taken from [8]

C. Result

We calculated pitch control values for W X H grids, 50 X 32 by default as we re-scaled grids together for faster calculation, for every event in the game. Based on the pitch control value, we are able to identify the control regions for each team per situation and from that, evaluate players' decisions when performing a moving ball action, see "IV. Player Action Valuing Metric".

Here, we calculate the Pitch control value for the 6th event of 1st match which is a 50 X 32 grid of values and visualise it in Fig. 7 and Fig. 8. Fig. 7 represents the event data of the 6th event and Fig. 8 represents the calculated PPCF value for the time frame which 6th event happens.

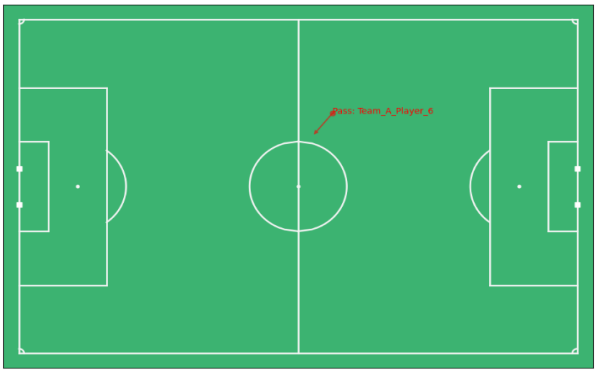

Fig. 7. Visualisation of Event data, pass of Team˙A˙Player˙6

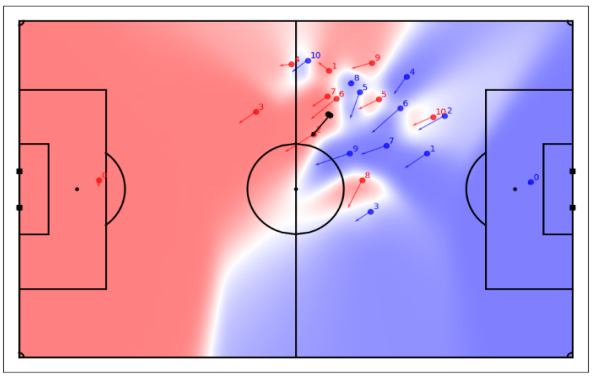

Fig. 8. Visualisation of calculated PPCF for event 6th. The red region is PPCF value for the left-sided team, blue region is PPCF value for the right-sided team. The deeper the colour is, the more control that team have over this region.

IV. Player Action Valuing Metric

When considering event stream data, a soccer match can be viewed as a sequence of actions. Each action is described by a number of properties such as its start location, its end location, its start time, and what type of action it was. The effect of an action is to move the game from state $S_{i-1} = a_1, ..., a_{i-1}$ to state $S_i = a_1, ..., a_{i-1}, a_i$. Since a game state is fully captured by a zone on the pitch, Expected Threat (xT) can value the actions that move the ball from the previous state to the next state based on the ball location and it is often referred to as "ball-progression models" (Yam, 2019) [5]. However, xT completely ignores the success probability of receiving the ball which can be influenced by current the positions and moving directions of teammates and opponents. Therefore, the pitch control model is an optimal solution to use in combination with xT to bring information about the current states of all players on the field. The final equation for Action Valuing is defined by Laurice Shaw [9] as follows:

$$ActionValue = xT_r \times PPCF_{r,t} - xT_{r'} \times PPCF_{r',t'}$$

Where r and r' are the start and end (x,y) location of an action event, t and t' are the start and end time frames of an action event. In our implementation, the action value is calculated

for only passing type action since there are not many take-on actions(dribble) and due to the division of the grid, most take-on actions do not bring much-added value.

With the ActionValue metric, all moving ball actions can now be quantified more precisely since some passings even though aiming at high-threat locations that teammates cannot control are just useless and, on the other hand, some passings with low-threat but increase the possession of the ball can be more valuable. The final performance ranking is then based on the $ActionValue$ of all moving ball actions that each player contributed to the team.

V. Result

For the result of our implementation, we choose the 1st match from FC Twente to illustrate how the model can be used to evaluate and analyze the performance of players as well as both teams during the match.

For the performance ranking, we generated $ActionValue$ calculation for all players, as seen in Table"Fig. **??**". From the result, we can see that among the top 5 players, the first 3 are from Team_A and it also reflects the final result that team A dominated the game with a final score of 4-1.

Name	Position	Team	xT_added
Team_A_Player_5	Midfielder	Team_A	1.09687
Team_A_Player_8	Striker	Team_A	0.83423
Team_A_Player_3	Defender	Team_A	0.79273
Team_B_Player_6	Midfielder	Team_B	0.75025
Team_B_Player_3	Defender	Team_B	0.63612

TABLE I
Top 5 players by xT_added.

"Fig. 9" represents bar charts of the average action value for each player in the same match. Even though Team_A_Player_5 generated the most threat value, his average action value is not so high. The reason is that Team_A_Player_5 is a $Midfielder$, so he can make more passings than other positions but his passing is not at top quality. On the other hand, the passings from Team_A_Player_16 and Team_B_Player_18 can pose a real danger to their opponents, they were also reflected as 2 of their passings during match 1 are key passings leading to goals.

Using the same match data, we plot in "Fig. 10" the average action value, but for the position, for both teams. The chart suggests that both teams did well in most of their passings in every position except $Goalkeeper$ although Team_A generally did a better job in passings than players from Team_B. In the plot, we also can immediately realize the huge negative value from $Goalkeeper$ of Team_B and it suggests his bad passings are a key factor leading to Team_B's lose.

With $ActionValue$ we can see that it can show many key insights about the performance of players and both teams. It can be used to identify the player with bad performance, also the player that can bring more impact on the game if he is allowed to have more passings. The model can also be used to evaluate every action of players and help them to realize their weaknesses to improve for the next match.

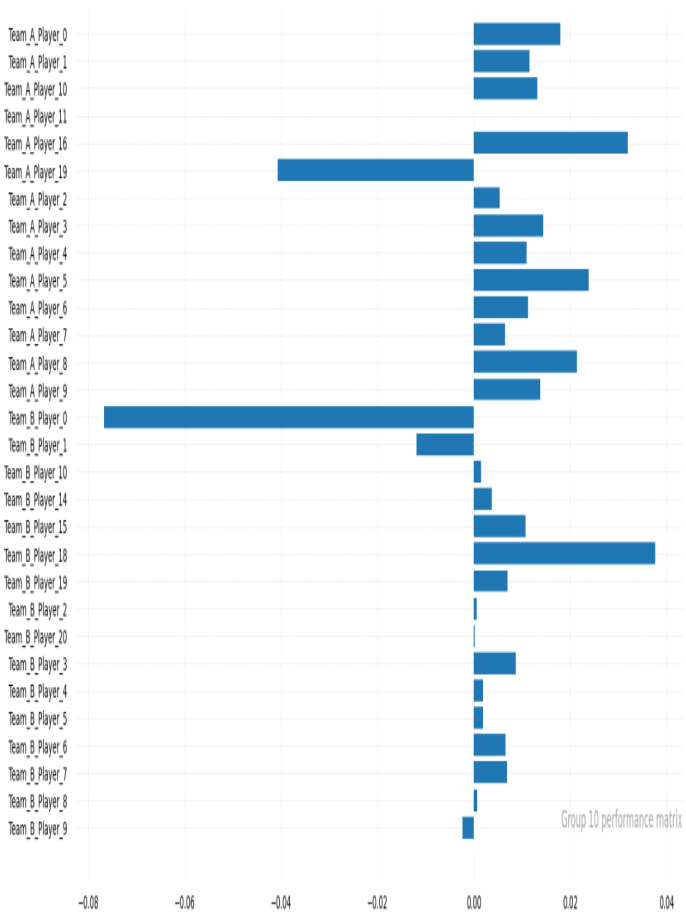

Fig. 9. Average action value per player from match 1

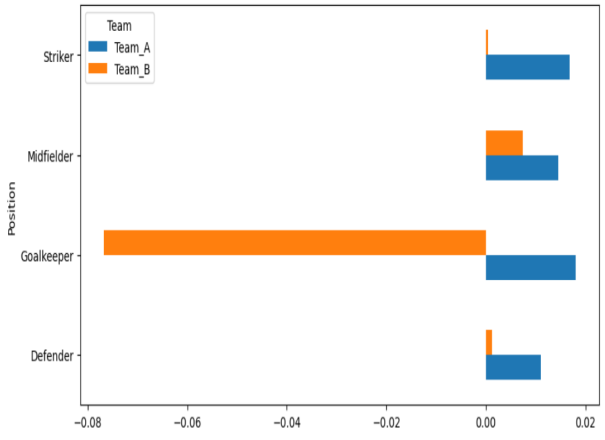

Fig. 10. Average action value per position from both teams in match 1

VI. Conclusion

The project aims to come up with a different metric to value football players' performance not only by their goals and assists but also by the impact of the game on every decision they made. While there are many other metrics for analysing player performance, the combination of Expected Threat and Pitch Control can provide more contextual information about player performance from their previous games. While the Expected Threat model is a good way to analyze and highlight the attacking region and especially the attacking pattern which can lead to goals from both teams, the Pitch Control model can bring information about how good a team formation was and how many important regions each team captured. From the data from 6 matches of FC Twente, we have shown how to train these models and how they can be applied in combination as a metric to evaluate, rank and analyze players' performance during the match. Although in the project we mostly evaluate on-ball actions like passing type, Pitch Control and Expected Threat model can also be used for off-ball actions analysis, for example, by calculating the value of pitch control captured by attacking and defending teams or evaluating added ActionValue for an off the ball run. In future work, we hope to improve the metric so that it becomes more complete to evaluate all types of player's actions.

VII. Critical Reflection

In the previous sections, we have discussed how we are using Pitch Control and Expected Threat in evaluating and ranking players, also, the result that we have obtained from this analysis brought up some insights and knowledge about how we can utilize the football data for specific data analysis. In this section, we will focus on discussing the problems we encountered, what we have learned and which techniques we have used based on the knowledge from the course Data Science and AI at the University of Twente to apply for the analysis of this project.

A. Preprocessing phase(Data Quality)

Using what we learn in the Data Quality part of the module, we have conducted multiple analyses on the football data to detect noises and errors in the dataset. For example, we have observed a data quality problem when executing Pitch Control Analysis, one noise data we have found is the incompatibility between the Event Stream Data and Tracking Data frame and the correct location of events, that is the start position of the ball does not match the start position of the player where the ball has been kicked. This is an example of an inaccurate data quality problem, which can reduce the accuracy of the prediction for the analysis.

In the preprocessing phase of analyzing football data, we have combined many preprocessing techniques that we learned, including filtering the columns, removing invalid values (nan value), grouping values when related, or even merging different datasets to remove noises. By applying these techniques, we can improve the data quality we will use for further analysis and make the data more analyzable and accurate. The benefits of applying these preprocessing and analysis are that we can get and use the data easily from one single dataset, resulting in better training performance for our models. However, it still encountered some issues, especially when manipulating the data without changing the original data.

B. Analyzing phase(Knowledge Representation & Reasoning and Explainable AI)

Although Knowledge Representation & Reasoning is not being used in actual implementation, the knowledge from the lecture and tutorial about time series is really useful to help us formulate the domain problem. The part about pre-condition and effects of events mentioned in KRR lecture, helps us understand the dataset better to immediately start the work since football data is a sequence of action events, and the upcoming action is related to the previous one. Also, the practice of formulating the domain problem helps us easier to reason about the dataset.

For the Explainable AI part of the course, we find this is the core application of our time series analysis problem for understanding, plotting, and predicting the data based on historical data. Our approach using for evaluating player actions and ranking their performance. Since we use two models in combination to measure players' actions, plotting them should be understandable and intuitive, so customers can understand why a decision was made. Therefore, the common Explainable AI technique we use for different analyses of this project is plotting heatmap. From the result, we can see the output of Pitch Control is represented as a heatmap so we can clearly recognize the control region of each team on the field. The result of Expected Threat is visualized as a grid where each cell shows the probability of attacking danger with darker color indicating the higher probability value that helps customers realize which cells are more important for an attacking option. We also use different types of plots for visualization to get a better understanding of the data. The final result is plotted in both table formats and bar plots so the customer can immediately understand the collected and calculated statistics and make a comparison about the performance of each player.

C. Building model phase(Machine Learning)

Football data is one example of time series data, which has been changed continuously and unpredictable over a period, in this case, are frames or events of the entire match. However, football data is a special sub-type of time series data having a specific way of handling it and there are not many built-in models to apply for training the football data directly. The knowledge from Machine Learning about Time Series does have us get better intuition and understanding of the data, but the used model for our implementation is quite different from what we learned in the lecture. To actually handle football data and implement models for analyzing purposes, we have to refer to various sources and papers about football analysis and also different mathematic models like Markov model.

VIII. Ethics Perspective

In previous sections, we have shown what techniques we have applied for the analysis process of this project. However, we have not covered the transparency of the data and the ethical perspective of this process. In this section, we will discuss different ethical aspects which are the value of collected data, the consequences and results of this analysis when ethics are involved, and the framework we use to ensure an ethical perspective throughout this process.

A. Ethics in data

As mentioned in the previous section, we encountered different data quality problems during the execution of our analysis, and the involvement of ethics is one aspect of the data quality issues. One of the issues that affected this analyzing process is the confidentiality of the user data. The data we used, provided by the FC Twente research team, is collected from different football matches so these data vary. Since this data is for the analysis of player evaluation, we should avoid misuse of sensitive and biased data, such as human skin colour, religion, and player position. This analytic data does not include attributes related to religion and racism, only player code, team, and role have been provided and are unique for each player. However, the biased conclusion may come from the player's position. In detail, in our analysis, the goalkeeper's contribution to the game is the least, compared to the midfielder or striker, and it even reaches a negative value. This biased is because the technique we used for the evaluation of players is based on their moving ball impact action. However, as a goalkeeper, most of their passings should be safe and under the control of teammates instead of a risky one, and from the xT grid, we can see that most action value grids are on the opposite side. Therefore, the evaluation of goalkeepers could be biased compared to players from different positions.

The evaluation approaches we have used for this project may also encounter a problem of fairness of the data and context of analysis. For example, we have a conflict in the teamwork and team division in a match since, in the current result, we have concluded that strikers have higher scores (strikers provide more threat and contribute more to all the situations of the game), while teamwork is not the attribute which using to evaluation player's ranking. The result is not entirely true and is an example of a biased conclusion on player position.

B. Ethics in analyzing (Ethics Framework)

To adapt to ethical principles and perspectives, we have used the FAIR framework for applying ethics for the entire process of this project. FAIR stands for Findability, Accessibility, Interoperability, and Reusability. These requirements of the data that we need to acquire help ensure ethics in the analyzing process.

The data is findability. We have stored the data in a structured way so that it can easily be found and used. The data has been stored in different datasets, depending on the content of the information. For example, in our football data, we have event data, which stores the players' actions and position throughout a match; ball data stores the information about the ball location in all the events of the game; and players data which specify to different players.

The data is also accessible for the development of this project; however, the user will not have access to this data. We have hidden the data so that it cannot be seen and accessed from outside, but it can still be easily explored and analyzed by developers. We have implemented it so that different information can be accessed by using different modes, such as ball data, event data, or even data of all players of a specific game team. With this approach, we can control and choose the specific data for proper analysis.

Besides, these data also ensure to be interoperable, which means it is used in different approaches and for several analyses. The information in the data should not cause biases in any circumstances. In detail, in this project, both the Pitch control and expected Threat approach required the events data, however the way we processed those data are different, so we have generated multiple datasets for each specific calculation and utilization. However, the correctness of the original event data remains the same and the mappings of these data are also the same for the different datasets.

They are also reusable since we can apply the same data for different approaches, i.e., Pitch Control and Expected Threat. With many techniques and criteria, we can still use the same data, which means we can discover different information from this data, also further strategies can be used for player ranking using this data.

From an ethical perspective, the football data we have used for analysis processes is FAIR, consistent, fairly, and equally used for different purposes.

References

[1] W. Spearman, "Beyond expected goals," in Proceedings of the 12th MIT sloan sports analytics conference., 2018.

[2] S. Rudd, "A Framework for Tactical Analysis and Individual Offensive Production Assessment in Soccer Using Markov Chains," in New England Symposium on Statistics in Sports, 2011.

[3] K. Singh, "Introducing Expected Threat(xT)," 2019. Retrieved November 3, 2022. Available: https://karun.in/blog/expected-threat.html

[4] "Calculating xT (position-based) — Soccermatics documentation". Retrieved November 3, 2022. Available: https://soccermatics.readthedocs.io/en/latest/gallery/lesson4/plot_ExpectedThreat.html

[5] Yam, D. 2019. Attacking contributions: Markov models for football. Retrieved November 3, 2022. Available: https://statsbomb.com/articles/soccer/attacking-contributions-markov-models-for-football/

[6] Spearman, William. (2016). "Quantifying Pitch Control". 10.13140/RG.2.2.22551.93603. Available: https://www.researchgate.net/publication/334849056_Quantifying_Pitch_Control

[7] Ramon Dop, "Pitch control metrics to improve the predictions of moments leading to goals in football", 2022. Available: https://staff.fnwi.uva.nl/a.visser/education/masterProjects/Thesis_Ramon_Dop_MsDS.pdf

[8] Friend of Tracking, "Advanced football analytics: building and applying a pitch control model in python.", 2020. Available on Youtube: https://www.youtube.com/watch?v=5X1cSehLg6s&ab_channel=FriendsofTracking

[9] Friends of Tracking. (2020, June 1). Beyond pitch control: valuing player actions and passing options. [Video]. YouTube:https://www.youtube.com/watch?v=KXSLKwADXKI&t=1459s&ab_channel=FriendsofTracking