Mastering the midfield transition pass using Expected Possession Value

Milan Klaasman
milanklaasman@gmail.com

1. Introduction

This analysis will focus on investigating the midfield transition pass. A transition pass is defined by the first pass within 5 seconds after capturing the ball in open-play. Transitions are valuable game situations and have become a central part of tactics. Hobbs et al. (2018) found that if a team counters immediately, i.e. creating a chance less than 15 seconds after regaining the ball, the likelihood of scoring is 12.4%. However, in case a team aims to maintain possession the likelihood of scoring is only 8.0% [1]. In the aforementioned situations, the likelihood of a team obtaining a shot is 35.1% while countering vs 10.6% in case of maintaining possession. Additionally, counter-attacks involve passes with a higher reward value, while they are on average not riskier than passes made during other transitions [1]. Thus, stimulating good passing decisions may lead to a higher reward, whilst not necessarily increasing the risk of possession loss. Hughes et al. (2019) [2] revealed that successful teams, top 4 of 2014-2015 Champions league season, created more scoring opportunities starting from the defence, defensive midfield and offensive midfield but not the offensive zone. In their research Hughes et al. (2019) conclude that the immediate player actions after winning the turnover are critical to the outcome of the transition. These findings support the importance of midfield transition passes and the added value of the analysis described in this paper.

After establishing the importance of midfield transition passes, the question of how to value these passes arises. Commonly used passing metrics like pass completion rate and packing [3], have the limitation that the risk of a pass is not considered. For this reason, a risk-reward model for passes is crucial to evaluate transition passes. Others, such as Power et al. (2017) have analysed the risk and reward of passes using tracking data [4]. Spearman et al. (2017) took another approach to model pass probability, by using physics-based models, which are built on the concepts of interception and control time [5]. On their turn, Fernandez et al. (2019) took a deep learning approach to quantify the Expected Possession Value (EPV) [6].

While these researches are of great scientific importance, they have relatively little practical implications on their own. For this reason, the goal of this analysis is to create rules for high-value midfield transition passes, e.g. the chances of a high-value transition pass are smaller when the closest teammate is further than 10 meters from the passing player. This will result in a recommendation that the closest player should stay within 10 meters of the passing player during a transition moment. To do this, game-state features and corresponding values should be created, that describe rules that are correlated to optimal midfield transition passes. Using the previous example, the closest teammate would be the "feature" and the value is a "maximum distance of 10 meters". Finding ideal values for these features, that are linked with high-value optimal passes, allows to:

- Improve the chance of creating scoring opportunities and win matches.
- Improve player's pre-orientation skills to find these features.
- Improve player's decision making, to increase the chance of a high-value pass.

To do this, an EPV model is implemented that is based on the approach of Spearman et al. (2017) since this model is relatively easy to implement while still holding a lot of value. With this model, it is possible to assign value to passes, after which the optimal pass can be calculated, by taking the maximum EPV value of all positions on the pitch. To relate these optimal passes with game-state characteristics, multiple features will be created based on tracking and events data, to summarize the current game state. To create clear, concise and easy to implement insights that are directly interpretable for practitioners, two models that meet these requirements will be:

- Linear Regression, to highlight the most important features
- Decision Tree Classifier, which enables creating rules for a high-value transition pass

For this reason, a highly interpretable model, like a Decision Tree, is favourable over other models that show higher accuracy like ensemble methods as Random Forest and Gradient Boosting Machines. Since a Decision Tree will generate conditions on which the tree is split. Using these conditions rules can be formed for high-value transition passes. In the next section, the modelling process and techniques are discussed.

2. Modelling process and techniques

In this section, the dataset is described, followed by an explanation of the implemented pass evaluation model, called Expected Possession Value (EPV), which is a combination between a Pitch Value model and a Pitch Control model, based on the approach coined by Spearman et al (2017) [5]. In the analysis section, the process of feature engineering is described, which explains the game-state of the transition pass. Lastly, the used Machine Learning models to link these game-state features with successful and optimal passes for the analysis of this approach are discussed.

2.1 Dataset

The given dataset consists of 52 matches of U-17 World Cup data. A pass is labelled as a midfield transition pass if it meets the following requirements:

- Pass or cross
- Within 5 seconds after a change in possession (in open-play)
- Within 3 events after a change in possession
- First pass after the transition

The midfield is defined by the middle third of the pitch. In the dataset, there are 2265 midfield transition passes found that meet the requirements.

2.2. Building a pass evaluation model

In this section, the building blocks for the used pass evaluation model are discussed. As coined in the introduction, to accurately assign value to passes a risk-reward model is crucial. Using the PV framework, the reward of a pass is established. By using a PC model the risk of retaining a ball is modelled. The pass evaluation model, EPV, is equal to the product of the PV model and the PC model. In the following sections, will elaborate on these steps.

2.2.1 Pitch value model

The PV model makes use of a transition matrix of different states. To do this, the pitch is divided into *N* evenly sized partitions, each representing a state. Where possessions are thought of as sequences of 'states', with the possible next states dependent only on the current state, this is commonly referred to as a Markov process. By iterating over many following events for each state, an equilibrium can be found in terms of what the probability of scoring a goal or losing ball possession is for each given state. Resulting in each partition of the pitch possessing a probability value of scoring, i.e. a possession value. Generating a PV framework requires a sizable dataset. Most commonly, at least 1 season worth of data is used. Since the dataset is considered too small to generate an accurate PV model, a pre-calculated PV model is used [8] (See Figure A1).

2.2.2 Pitch Control model

The PC model is inspired on the work of Spearman et al (2017) [5]. They implemented multiple physics-based models to calculate the ball trajectory, time to intercept and time to control. To make this model more robust against non-viable passes. The PC model makes it possible to validate whether a pass is possible, taking into account the time it takes for a player to reach a position on the pitch. To achieve this, the pitch is divided into sections that are 'controlled' by either the attacking or the defending team, i.e. the probability that a team will retain the ball if the ball is passed to that position (See Figure A2). The main benefit of combining this model with the PV model is that a passing option can have a positive expected added value even when the ball is passed back towards a team's own goal, e.g. a position where the probability of retaining possession is greater.

2.2.3 Combining the models

By now both the PV and the PC model have been explained. In this section, it is described how to combine both models to calculate the Expected Possession value (EPV) of passes to a specific destination on the pitch. Let R_B be the end position of a pass A towards B. The expected value at end position R_B can be calculated by multiplying the probabilities of the Pitch Control model and the Pitch Value model *at* R_B:

Eq 1: $$EPV_B = Pitch\ Control\ (R_B)\ x\ Pitch\ Value\ (R_B)$$

The same holds for the current expected value at R_A. Therefore, the expected added value (*AV*) of the passing option R_A to R_B is derived using:

Eq 2: $$AV_{R_a \rightarrow R_b} = EPV_B - EPV_A$$

Figure A3 shows the EPV value of a pass. Above the requirements of midfield transition passes are established and the importance of building a risk-reward pass evaluation model that both models the risk, using PC, and the reward of a pass, using PV. This approach enables to approximate the optimal pass, by calculating the AV for each partition of the pitch for a particular situation. The optimal pass is equal to the potential pass with the highest AV. Enabling to analyse the interaction between optimal transition passes and game-state features, which will be discussed in the next section.

2.3 Analysis

In the analysis, the aim is to link game-state features to optimal passes to generate valuable insights into how to utilize the potential benefits of transition passes, as described in the introduction. To do this, game-state features are generated and machine learning approaches are utilized to link these to optimal passes. The dataset already consisted of 4 features that are used for the analysis. A total of 31 additional features are created (See Table A1 for all features). These features can be divided into two categories:

- Events-based features (9)
- Tracking-based features (22)

As a target value, the added value of the optimal pass (*max_EEPV_added*) will be used. The goal here is to generate easy to interpret results that can be used in practice, e.g. if the closest teammate is more than 10 meters away the chances of a high-value transition pass is smaller. For this reason, two models are implemented, Linear Regression (LR) allows linking the created features to the continuous target variable. Secondly, a Decision Tree Classifier (DT) that is used to find important values for the features that correlate with a higher probability of an optimal pass. This to create rules for a high-value transition passes, that can be used to mimic the decision-making process a player undergoes during a transition pass. Before implementing these models, multiple data preparation and data cleaning steps are performed:

- Training- and test-set split
- Outlier detection (5 rows were deleted from the dataset, based on incorrect data)
- Skewness correction of the target value (See Figure A4)
- k-Nearest Neighbours-imputation

In the sake of brevity, the above steps are not discussed in-depth. For feature selection, a feedforward selection model is applied (See Figure A5). This method starts with no variables in the model and then adds variables to the model one by one until any variable included in the model can add any significant contribution to the outcome of the model. Instead of implementing such a function, due to time constraints, it was chosen to select the number of features using a maximum threshold of features for the feedforward selection based on the figures (See Figure A5). This approach ensures that a minimum amount of features are selected which enhances interpretability.

To create the DT, it is necessary to change the continuous target variable *max_EEPV_added*, into multiple classes. The Jenks algorithm is applied to cluster the target variable [8]. This algorithm allows for finding natural breaks in a 1D-array. Three classes are selected, aiming to represent a low-value pass, a medium-value pass and a high-value pass. To validate the models, a benchmark prediction; a prediction using all features, and a prediction using the selected features are created. The average taken over all train targets is used to create a benchmark validation for the LR. The DT is benchmarked using an ignorant prediction of always classifying the most common class, which resulted. In the next section, the results of this approach will be discussed.

3. Results

In this section, the results are shown corresponding the goal to find game-state features that are associated with optimal transition passes. Achieving this by showcasing feature importances and the validation scores of the models. To investigate the link between features and the target variable, two feature importance charts from both the Linear Regressor (LR) and Decision Tree (DT) are created (See Figure A8). The LR shows that the average speed of all players is the most important feature. In the case of the Decision Tree, the percentage of pitch control towards the goal shows the highest importance. Additionally, Start X position of the pass and pass length seem important features. In Table A2, the validation of the models is presented. Both models score best for the approach using all features. with LR having a mean-squared error of 0.02 against a benchmark of 0.03 and DT having an accuracy of 0.617 against 0.443 respectively. Figure A9 shows the normalized confusion matrices of the DTs. It can be concluded that the models proposed perform better than the benchmarks, however, an accuracy of 0.617 for a DT with 3 classes can be considered lacking predictiveness. Additionally, it is remarkable that the selected feature models perform worse than the models with all features, where the feedforward selection plots (See Figure A5) would suggest otherwise.

In the previous paragraph, the performance of the models is explained. To extract general rules a DT is designed to showcase the conditions of the splits in the tree. These conditions will be used to create the rules. To do this, the best performing DT (i.e. using all features) is used with a maximum depth of 4 (See Figure A10)

To create the rules, the rule-target is focused on the high-value and medium value pass classes, i.e. good passes. The most important split is based on *pitch control percentage towards the goal* (PC%), followed by the *pass length* and *start position* (*Start X*). These 3 variables are chosen as a showcase to set up rules for optimal midfield transition passes. From the figure, it can be concluded that a higher PC% leads to a higher chance of good passes. In terms of pass length, the left branch shows a pass length split on 32.96m and in the right branch a split on 22.42m. With longer passes being better in terms of risk-reward. Lastly, the closer the transition moment (*Start X*) happens towards goal the higher the added value. Both the findings of PC% and starting position, seem logical since the EPV model is a combination of pitch value and pitch control.

Based on these findings multiple recommendation rules for optimal transition passes can be formed:

- The passing player should be aware of pitch control, space of teammates far up the pitch and the transition location (the closer to the goal means more value).
- The player must intend to pass forwards after a transition and avoid short passes of less than 20m as the first transition pass. By doing this the chances of a high-value transition pass are increased.
- Attacking players should make runs towards the goal, creating as much space (in front of the ball) for themselves and teammates while staying within an ideal passing distance of more than 20m from the passing player. This can be done by finding depth in play.
- Moreover, attacking players should aim to win duels high up the pitch, to create as much value as possible.

In practice, with these rules, it is possible to teach teams and players how to increase the probability of a high-value midfield transition passes. To do this, the recommendation is that practitioners should create awareness for these key game-state features (PC%, pass length and Start X), furthermore, it is possible to stimulate players to adapt their looking behaviour to recognize these identifiers more quickly, which will result in better decision making.

5. Conclusion

Using modelling techniques this analysis has created general rules for high-value transition passes. By creating a risk-reward model, Expected Possession Value, to find 2265 optimal passes in a dataset of 52 matches. These passes were linked to created 35 tracking and events-based game-state features. Using Linear Regression and a Decision Tree Classifier, important game-state features were linked to high-value optimal passes. Finding that pitch control towards the goal, the pass length and the start position are the most important features. Using the Decision Tree, multiple rules are created that allow to:

- Improve the chance of creating scoring opportunities and win matches.
- Improve player's pre-orientation skills to find these features.
- Improve player's decision making, to increase the chance of a high-value pass.

6. Future Work

Since the assessment process is only 3 weeks long, there were multiple concessions on the design choices:

- Data quality, while the timestamp of the tracking and events data are closely aligned, this does not hold for the actual positional data. Many providers deal with this problem, originating from the human annotation. Using syncing methods, it is possible and recommendable to align the event positions with the tracking data, for the tracking data is most commonly more accurate. The same would hold for automatic events detection, which will create more accurate data points and thus better analysis and added value for practitioners.
- The EPV model can have many additions to make the model more robust and even more tailored towards transition passes. An example would be using the probability that a pass will be successful, conditioned on the pass distance.
- In this approach, there are a limited amount of features created, with more investigation more suitable features for transition moments can be created. Additionally, limited data makes it hard for models to perform well.
- The performance of the Machine Learning models can be improved. Choices for interpretability, like model choice, where a Decision Tree is very interpretable, however, it is an unstable predictor. Since leaving out one feature may lead to very different results.

6.1 Additional research

In this section, two additional examples of analysis are showcased, with confidence it can be stated analysis of optimal transition passes allows for more similar endeavours. The difference between the actual pass and the optimal pass is used to create a ranking system for players who performed best during the U-17 World Cup on the midfield transition pass. The top-performing players are shown in Table A3. With Italian Atalanta player Matteo Ruggeri performing best. By analysing pass sonars using the *EEPV_added* (See Figure A7), it can be concluded from the data that a lot of unsuccessful passes are long passes in the direction of the goal. This could also be linked with clearances in the dataset. However, plotting these pass sonars or additional field heatmaps may hold valuable information.

7. Literature

1. Hobbs, J., Power, P., Sha, L., & Lucey, P. (2018, February). Quantifying the value of transitions in soccer via spatiotemporal trajectory clustering. In *MIT Sloan Sports Analytics Conference*.
2. Hughes, M., & Lovell, T. (2019). Transition to attack in elite soccer.
3. Steiner, S., Rauh, S., Rumo, M., Emery, N., Sonderegger, K., & Seiler, R. (2017). Packing in football: a differential ecological perspective on passes.
4. Power, P., Ruiz, H., Wei, X., & Lucey, P. (2017, August). Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1605-1613).
5. Spearman, W., Basye, A., Dick, G., Hotovy, R., & Pop, P. (2017). Physics-based modeling of pass probabilities in soccer. In *Proceeding of the 11th MIT Sloan Sports Analytics Conference*.
6. Fernández, J., Bornn, L., & Cervone, D. (2019, March). Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer. In *13th Annual MIT Sloan Sports Analytics Conference*.
7. Friends of Tracking (2020) https://github.com/Friends-of-Tracking-Data-FoTD/LaurieOnTracking
8. Rey, S. J., Stephens, P., & Laura, J. (2017). An evaluation of sampling and full enumeration strategies for Fisher Jenks classification in big data settings. *Transactions in GIS*, *21*(4), 796-810.

8. Appendix

8.1 Figures

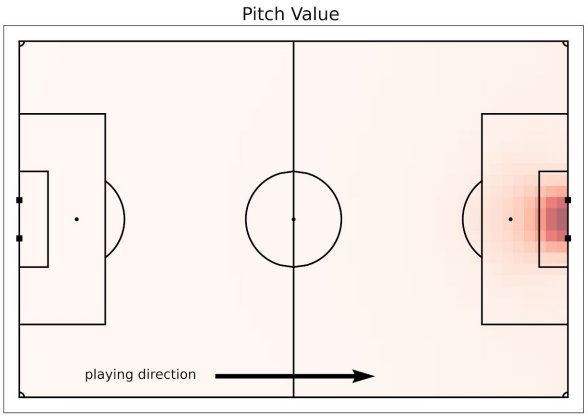


Figure A1. The Pitch Value model, which shows the probability of scoring a goal in the following events after possession in that partition of the pitch. As can be seen, values are higher (darker blue), closer to the goal. With a playing direction of left to right.

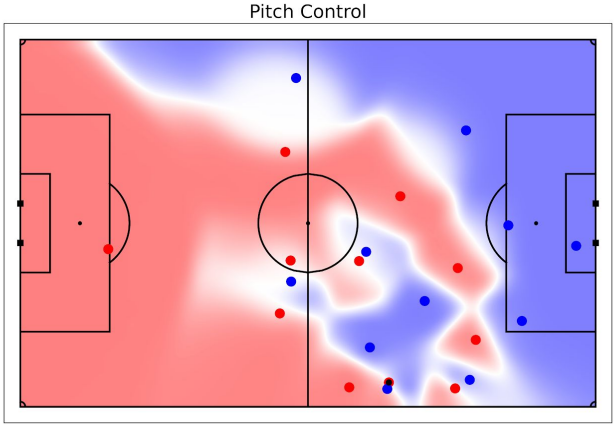


Figure A2. This figure shows an example of the Pitch control model given a passing situation. The red and blue colours represent the pitch control by the red and blue team respectively. White areas represent places where both teams can capture the ball if it is played to that position. (This Figure created with using the dataset)

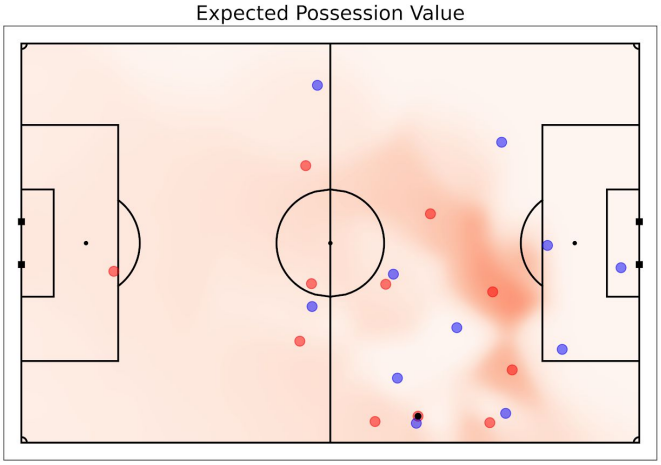


Figure A3. This figure shows an example of the EPV model in a given situation. With a darker red location representing a higher possession value for the attacking team. (This Figure created using the dataset)

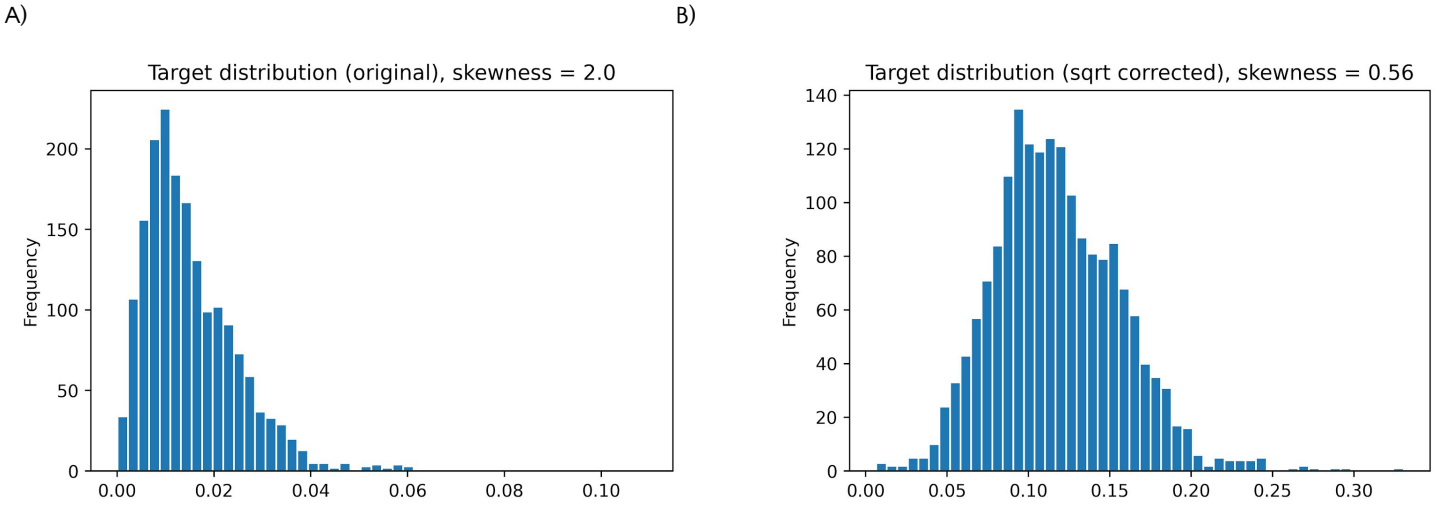


Fig. A4 This Figure shows the distribution of the target values before correction (A) and after square root correction (B). The skewness is diminished from -2 (A) to 0.56 (B). Since only a slightly minimum value occurs in the dataset, there was a probability for skewness. By adjusting for skewness, the model performance is improved.

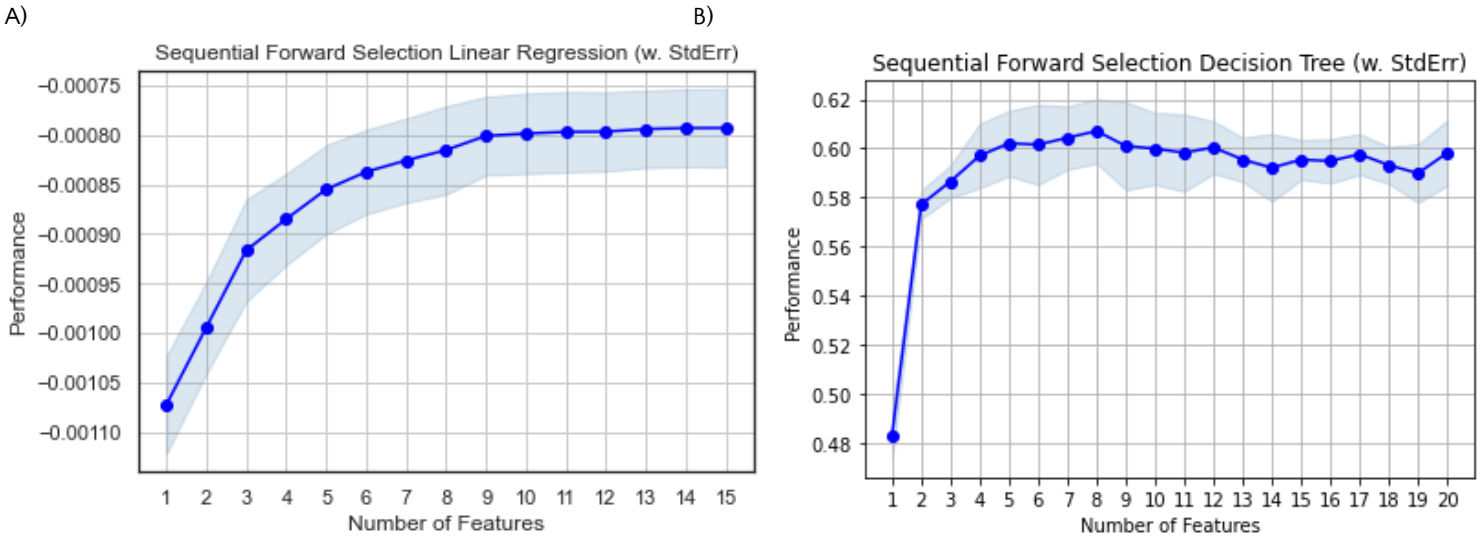


Figure A5 This Figure shows the sequential feedforward selection process for both the Linear Regression and the Decision Tree Classifier (B). For both models, the figures show that after approximately 10 features the model does not improve its prediction scores.

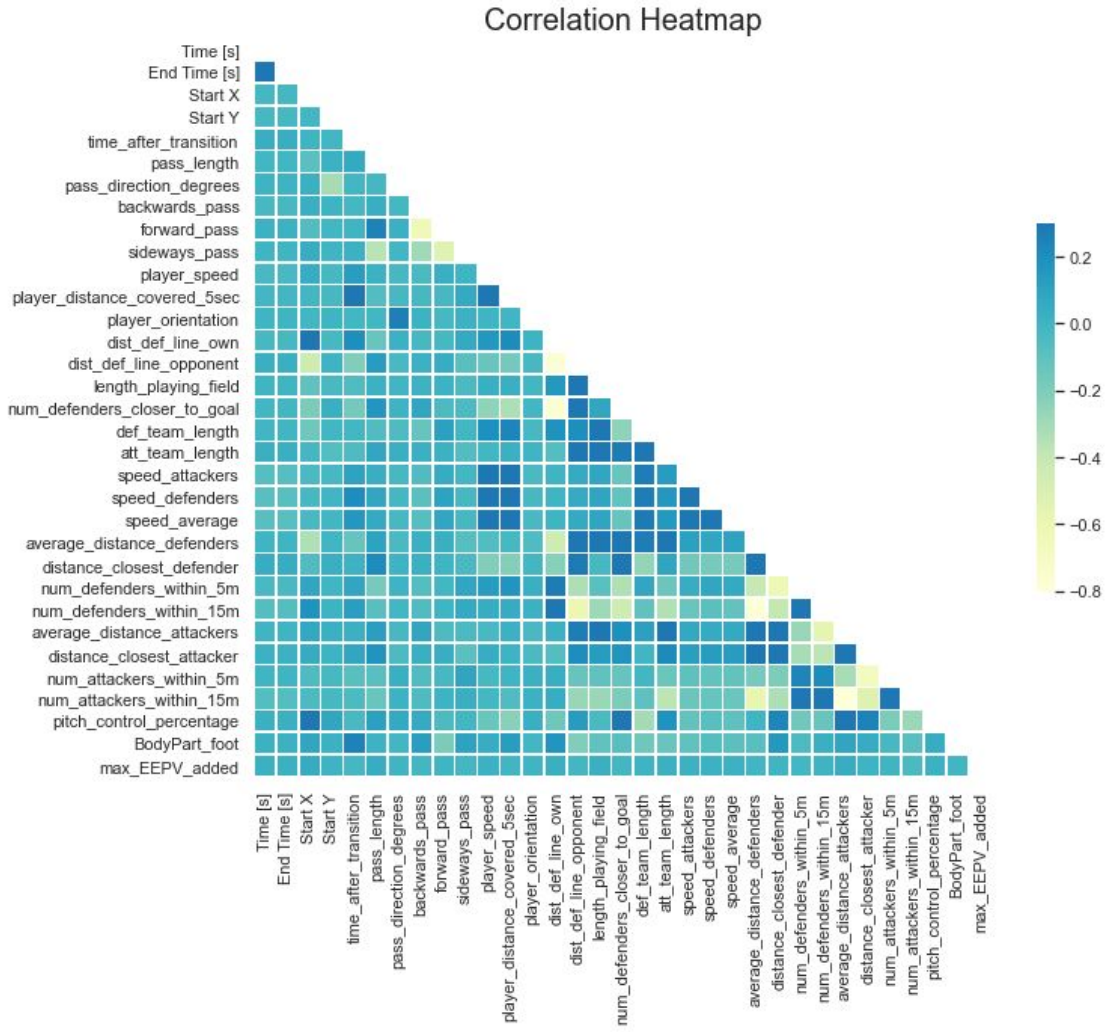
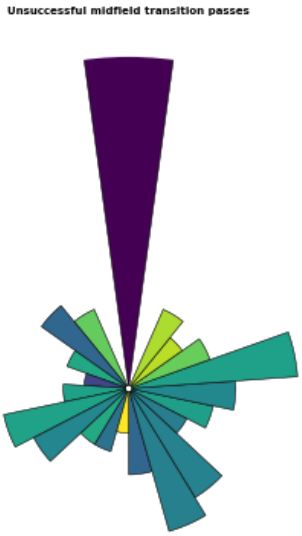


Figure A6 This Figure shows the correlation heatmap of all variables with the target value *max_EEPV_added*. It is clear from this figure that there is no high correlation between a feature and the target variable. The features show a high positive or negative correlation among each other.

A)



B)

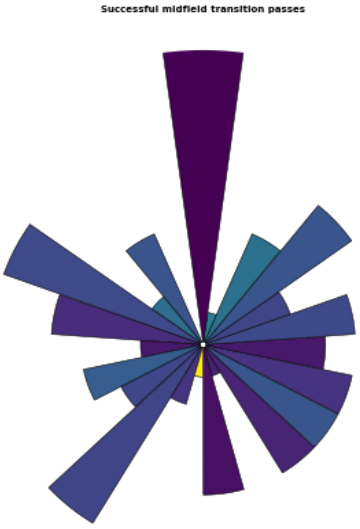
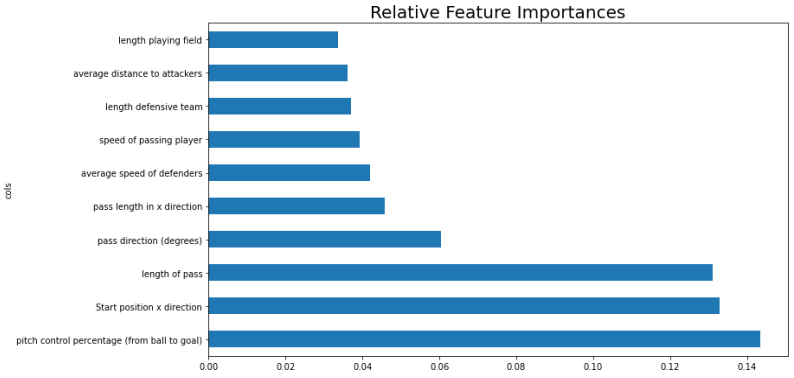


Figure A7. show two pass-sonars, with a vertical upwards direction representing the direction towards the goal. Figure A shows unsuccessful transition passes in the dataset and Figure B shows successful midfield transition passes. The length of the sonars represents the average pass length in that direction. The colour represents the number of passes in that direction. With yellow representing a few passes and purple many passes.

A)



B)

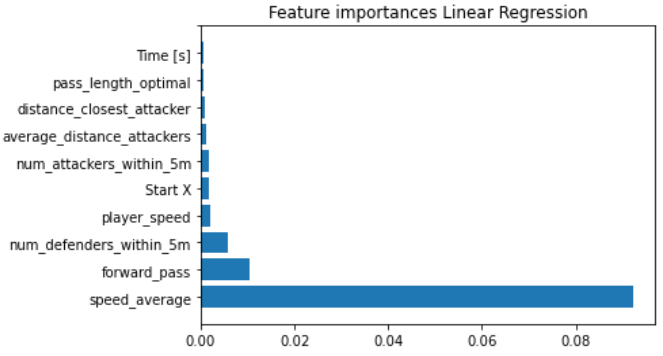
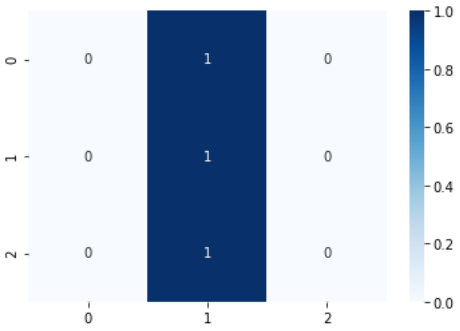
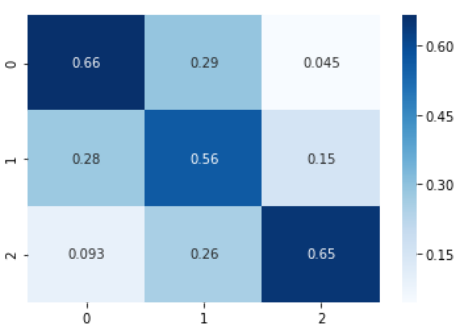


Figure A8. Shows the feature importances of the Decision Tree Classifier (A) and the Linear Regression (B). Both pitch control percentage towards the goal and the average speed score relatively high

A)



B)



C)

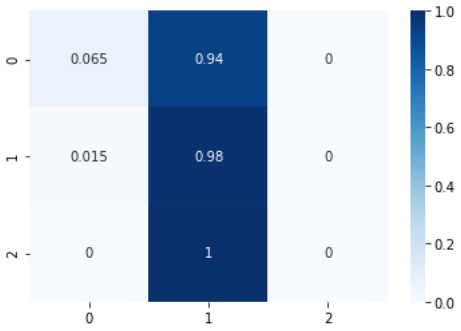


Figure A9 Shows the normalize confusion matrices of the performance of the three different Decision Tree models, namely the benchmark model (A), the model using all feature (B) and the selected features model. The selected features model performs surprisingly bad. This might be due to the manually chosen maximum of features.

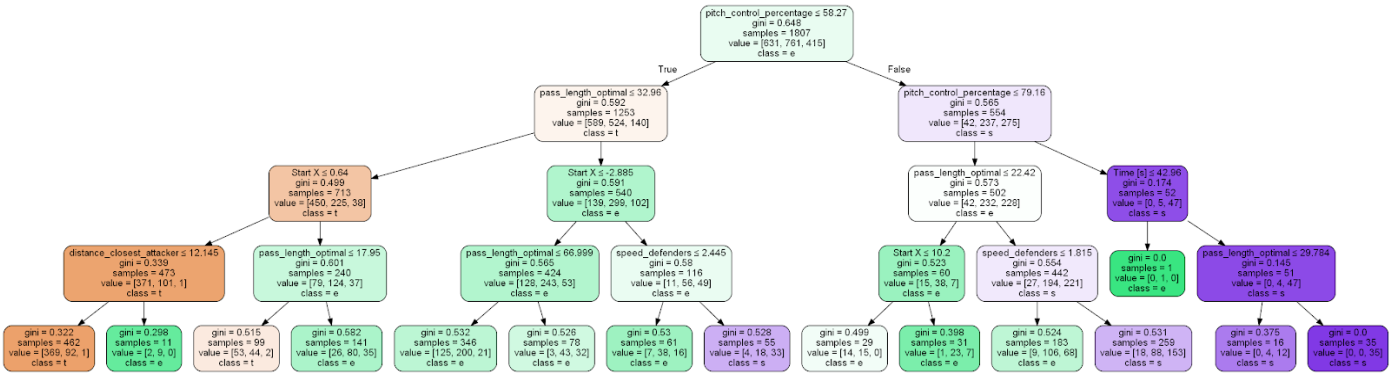


Figure A10. Shows the decision tree using all created features in the data frame. The maximum depth of the tree is set to 4. The most important features are *pitch_control_percentage*, *pass_length_optimal* and *Start_X*.

8.2 Tables

Feature Table
Time-based
- Start time of the event - End time of the event
Events-based:
Pass attributes: - X Start position - Y Start positon - X End position - Y End positon - Pass length (actual pass) - Pass length (optimal pass) - Pass direction (the degree in which the ball is passed) - Backwards pass - Forwards pass - Sideways pass
Tracking-based:
Passing players attributes: - Speed - Acceleration - Trajectory (distance covered) (last 5 seconds) - Players orientation - Average distance teammates
The number of players: - Defenders within a radius of 15 meters - Attackers within a radius of 15 meters - Defenders within a radius of 5 meters - Attackers within a radius of 5 meters - Defenders in front of the ball - Defenders between ball and goal - Distance to the closest defending player - Distance to the closest attacking player
Teams attributes: Movement - Average speed all players (Models static vs fluid game state) - Average speed attacking team - Average speed defending team - Average acceleration all players (Models static vs fluid game state) Distances: - Distance to defender’s defending line

- Distance to attacker's defending line - Team length attacking team - Team length defending team - Team width attacking team - Team width defending team Pitch Control - Attacking team's Pitch control percentage between ball an goal.

Table A1. This table shows an overview of game-state features that will be created based on both event and tracking data.

	Baseline (mean prediction)	Before Feature Selection	After Feature Selection
Linear Regression	MAE: 0.0300 MSE: 0.0013 RMSE: 0.0360 R-Squared: -0.0003	MAE: 0.0201 MSE: 0.0007 RMSE: 0.0261 R-Squared: 0.5074	MAE: 0.0205 MSE: 0.0007 RMSE: 0.0264 R-Squared: 0.4962
Decision Tree Classifier	accuracy: 0.443 f1_score: 0.272	Accuracy: 0.617 f1_score: 0.616	Accuracy: 0.458 F1_score: 0.313

Table A2. This Table shows the validation scores of both the Linear Regression model and the Decision Tree Classifier.

player_name	Number of passes	Mean difference with optimal
Matteo Ruggeri	21	0.008794
Lorenzo Pirola	25	0.009325
Jesus Gomez	22	0.012115
Chrislain Matsima	21	0.013908
Ian Maatsen	22	0.013676

Table A2. Shows the players with the lowest difference with the optimal pass in the given situations. Only players with more than 20 transitions passes are selected.