Passing Patterns
(An Analysis of Passing Actions and Abilities of Players in Association Football)

Rohan George Philip

Abstract

This paper presents an analysis of passing actions and abilities of players in association football using Stats Perform's pass event data from one season. Two models, Logistic regression and a Multi Layered Perceptron, were used to predict pass probability. The paper then explores how passing probabilities varied across the football pitch. Pass quality ratings were estimated for each player in the dataset, and four passing styles were identified. The findings of this study offer valuable insights into the passing patterns of players and provide a framework for understanding the nuances of the game. This research contributes to the growing field of sports analytics and has implications for player recruitment, team tactics, and game strategy.

Keywords: *Sports Analytics; Machine Learning; Logistic Regression; Deep Neural Networks*

1. Introduction

Passing is a critical aspect of association football (Soccer), and its importance cannot be overstated. Passing accuracy, speed, and style are all essential components that contribute to the success of a team. Despite the significance of passing in the sport, there is a lack of comprehensive analyses on passing patterns and abilities of players in association football. This study aims to address this gap in the field of sports analytics by providing a detailed analysis of pass probabilities and quality ratings across one season of association football using Stats Perform's pass event data. The objective of this research is to identify passing patterns and styles of players that can help improve team tactics and strategy. To achieve this objective, two models, Logistic regression and a Multi Layered Perceptron, will be used to predict pass probability, and the way in which passing probabilities vary across the pitch will be observed. Pass quality ratings will be estimated for each player in the dataset, and four passing styles will be identified, with adjustments made for the predicted probabilities. This paper provides valuable insights into the passing patterns and abilities of players in association football, contributing to the growing field of sports analytics. The findings of this research have implications for player recruitment, team tactics, and game strategy, and provide a framework for understanding the nuances of the game.

2. Methods

The data used in this study was obtained from Stats Perform's pass event data from one season of association football. The dataset included the following variables: player_id, game_id, team_id, home_score, away_score, home_id, away_id, half, minute, second, outcome, x, y, endx, endy, hd_pass, cross, corner, throw, gk_throw, fk_taken, and chipped.

Player_id represents the ID for the player making the pass, game_id represents the ID for the game in which the pass occurred, and team_id represents the ID for the team of the player making the pass. Home_score and away_score represent the home and away team's score at full

time for the game, respectively, while home_id and away_id represent the home and away team's ID, respectively. The half variable indicates which half the pass occurred in, while the minute variable indicates which minute of the match the pass occurred in. The second variable represents the seconds part of the match clock, along with the minute variable above.

Outcome is a binary variable with 0 indicating an incomplete pass and 1 indicating a complete pass. The x and y variables represent the x and y coordinates of the pass in meters, respectively, with x going from 0-105 in the attacking direction and y going from 0-68 from right to left from the perspective of the attacking team. The endx and endy variables represent the end x and y coordinates of the pass in meters, respectively.

The hd_pass variable is a binary variable with 0 indicating that the pass was not a headed pass, while 1 indicates that the pass was a headed pass. The cross variable is a binary variable with 0 indicating that the pass was not a cross, while 1 indicates that the pass was a cross. The corner variable is a binary variable with 0 indicating that the pass was not a corner kick pass, while 1 indicates that the pass was a corner kick pass. The throw variable is a binary variable with 0 indicating that the pass was not a throw-in, while 1 indicates that the pass was a throw-in.

The gk_throw variable is a binary variable with 0 indicating that the pass was not thrown from the goalkeeper, while 1 indicates that the pass was thrown from the goalkeeper. The fk_taken variable is a binary variable with 0 indicating that the pass was not a free kick taken following a foul or stoppage of play, while 1 indicates that the pass was a free kick taken following a foul or stoppage of play. The chipped variable is a binary variable with 0 indicating that the pass was not in the air, while 1 indicates that the pass was in the air. Overall, the dataset included information on passing actions and passing abilities of players across one season of association football. The variables included in the dataset were used to predict pass probability, estimate pass quality ratings, and identify passing styles of players.

The dataset is subjected to data cleaning procedures, where rows with NaN values were removed, and dimension values were clipped to a lower bound of 0. For pass probability modeling, 12 features including x, y, endx, endy, hd_pass, cross, corner, throw, gk, gk_throw, fk_taken and chipped were retained while the target variable was Outcome.

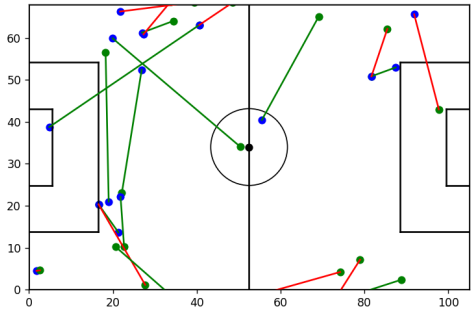
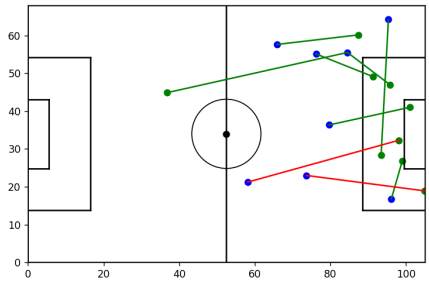


Figure 1. Sample of passes in the Stats Perform Dataset

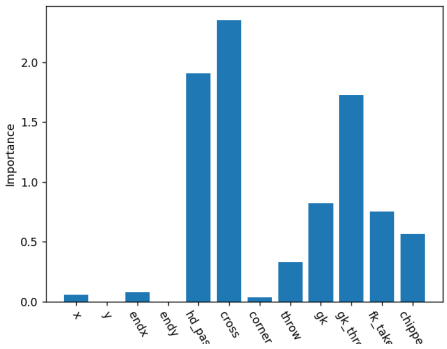
2.1. Pass Probability Prediction

A logistic regression model was implemented using Scikit-Learn's logistic regression API, with a maximum iteration count of 1000 and the 'liblinear' solver. The trained model achieved an accuracy of approximately 83.9% on the test set, and the pass success probability was extracted from the model. The performance of the model can be experienced using the interactive pitch (2a) in the Jupyter notebook. Left-click on the interactive plot to add a pass source, right click to add a destination. If the line produced by the model is green, the pass has a high probability of success. If the line is red, the pass has a high probability of failure. The coefficients of the logistic regression

model (2b) were analyzed to determine the influence of the predictor variables on the model. It was observed that hd_pass, cross, and gk_throw have the highest influence on the model, with these variables having a greater impact on the pass success probability. Surprisingly, the location variables, x, and y, were found to have a lower influence on the model.



(a) Sample of interactive plot with a LogReg model



(b) Feature Importance on Logistic Regression Model

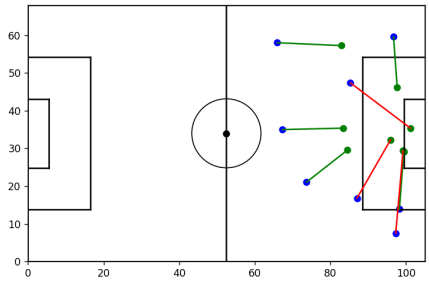
Figure 2. Logistic Regression model

An alternative to the Logistic Regression model is the Multi-Layered Perceptron model. The neural network contained three linear layers; one input layer, one hidden layer with 100 neurons, and one output layer with two output neurons. Each of the linear layers is interleaved with rectified linear unit non-linearities. The loss function used is the CrossEntropy loss.

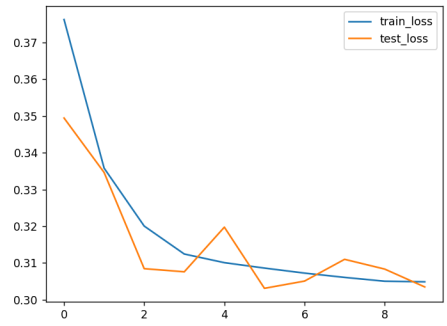
The optimizer used is Adam, with a learning rate of 1e-3. The momentum parameters are, 0.9 and 0.999 as is default in PyTorch. The model is trained for 10 epochs and achieves a test accuracy of 85%.

Similar to the logistic regression model, the MLP model can be interacted with through the interactive plot (3a). As before, if the line produced is green, the model predicts that the pass is likely to succeed and red if the pass is less likely to be successful.

Through the interaction, it is noticeable that pass location seems to have much more of an effect on success probability. For example, passes in and around the penalty box and particularly around the six yard zone are less likely to be successful than in the logistic regression model.



(a) Sample of interactive plot with an MLP model



(b) Train and Test Loss graph

Figure 3. Multi Layered Perceptron model

For each of the passes in the dataset, the probabilities are estimated using the MLP model. A softmax is performed on the output of the model and the resultant predicted values are added to the dataset. A heatmap of passing probabilities (4), that have been predicted based on the data available produces interesting results. Predictably, passes from around the centre circle (possibly kick-offs), and the goal kick zones have very high probabilities of success. The region in and around the penalty box sees a very low probability of passes being completed.

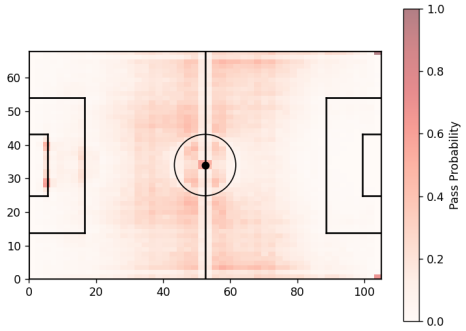
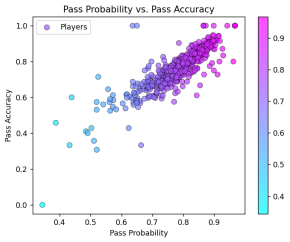


Figure 4. Heatmap of all the passes in the dataset and their associated probabilities of success

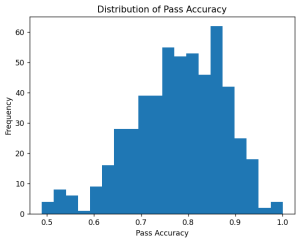
2.2. Player Passing Ability Prediction

We use our results from the previous section to help create an estimate of each player's passing ability. The pass probabilities and pass success rates for each player are generated using Pandas' group-by function to calculate the mean of the respective variables. This allows us to obtain an estimate of each player's passing ability, based on their performance in the dataset.

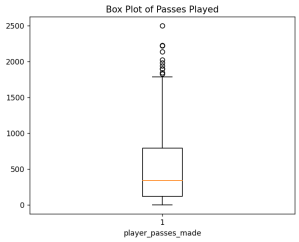
An observation of the scatter plot of Pass Probability vs Pass accuracy (5a) belies an extremely linear relationship. A Pearson's Correlation test produces a coefficient of 0.8588 with a p-value of 1e-209 which indicates a very high level of correlation. In order to produce an estimate of the pass accuracy, while correcting for the influence of the average pass probability of the player, a regression analysis was used.



(a) Probability vs Accuracy per player



(b) Distribution of Pass Accuracy



(c) Box Plot of passes made per player

Figure 5. Pass accuracy information

Pass probability was used as the independent variable and pass accuracy was used as the dependent variable. Once the regression line was fit, the residuals were calculated. These residuals were interpreted as the performance or lack thereof, that was being hidden by the average pass probability of a player. For example, a player who primarily passes in their own half is likely to have a higher passing accuracy than a player who primarily operates in more advanced zones, but is not necessarily a better passer. By adding the residuals to the passing accuracy scores, and then normalizing them, we end up with a relatively reliable estimate of a player's passing ability while correcting for the effect of the average pass probability of the player.

In order to identify the top passers, the first step taken was to generate a five-point summary (5c) of the dataset. We only consider players with over 121 passing actions, which is at the 25^{th} percentile. Once plotted (5b) , the accuracies follow a distribution that is fairly close to normal. The interactive plot of the top ten passing ratings amongst the players in the dataset (6) will allow users to hover over each bar to view information about each of the players.

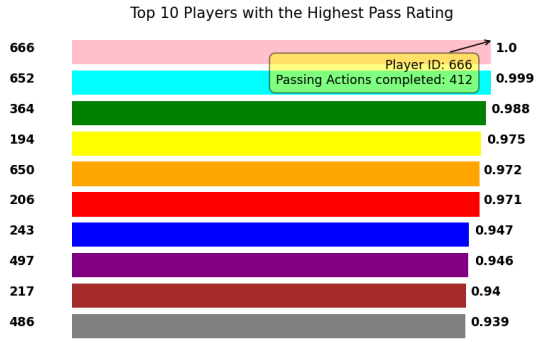


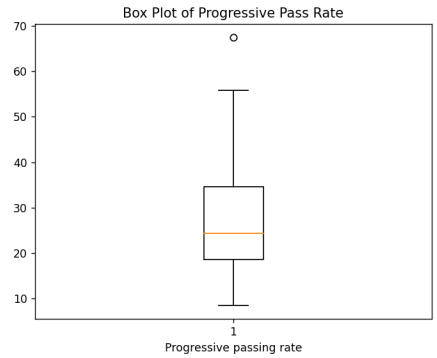
Figure 6. Top 10 Passers in the dataset

2.3. Observing Passing Styles of Players

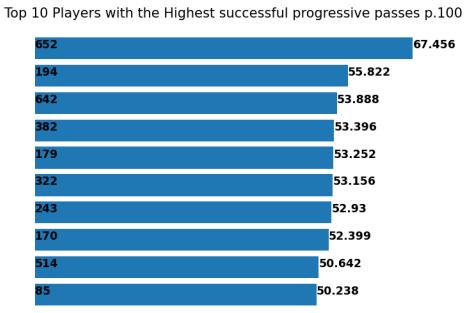
In order to identify the passing styles of players, I have divided them into four categories: progressive passers, final third entrants, crossers, and long-ball specialists. While I expect there to be overlap between these groups, this categorization will help us to assign specific traits to each player. This final stage of our analysis aims to provide a deeper understanding of each player's passing abilities and how they contribute to their team's overall performance.

2.3.1. Progressive Passers

For the progressive passers, the first step is to calculate the distance of each pass, in order to make sure that we only focus on passes that meet certain thresholds. Using a box plot (7a), I chose this threshold to be 9.5 meters which corresponds to the 25^{th} percentile. The other constraints include rows in which endx is greater than x, gk, and gk_throw = 0 and outcome = 1. I



(a) Box Plot of Progressive Passing rates



(b) Top 10 Progressive Passers in the dataset

Figure 7. Progressive Passers

then identified the players with the highest estimated progressive passes completed per 100 pass

attempts. Using the formula below

$$\frac{\textbf{Count of progressive passes for each player}}{\textbf{Count of passes attempted for each player}} * \textbf{player pass probability} * 100$$

Similar to the above bar plots, this one (7b) can be interacted with by hovering over each bar.

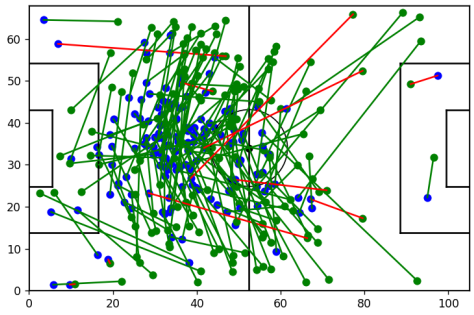
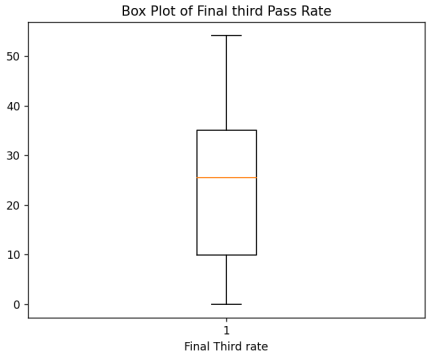


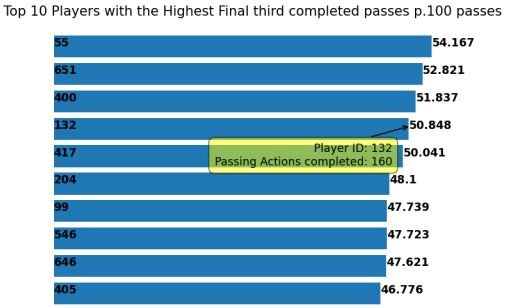
Figure 8. Passmap of the Top Progressive passer

2.3.2. Final Third Entrants

The next style of passing that I focused on was final third entrants. For the purposes of this project, I described these passes as being those that either took place or culminated in the final third of the football pitch.



(a) Box Plot of Final Third Passing rates



(b) Top 10 Final Third Passers in the dataset

Figure 9. Final Third Passers

Similar to the progressive passes formula, the formula used to estimate the number of successful final third passes per 100 passes attempted is as below

$$\frac{\textbf{Count of final third passes for each player}}{\textbf{Count of passes attempted for each player}} * \textbf{player pass probability} * 100$$

As has been the case thus far, the bar charts (9b) are interactive.

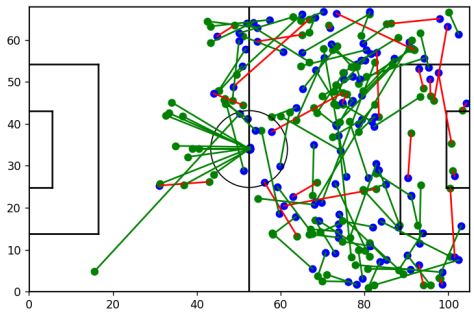
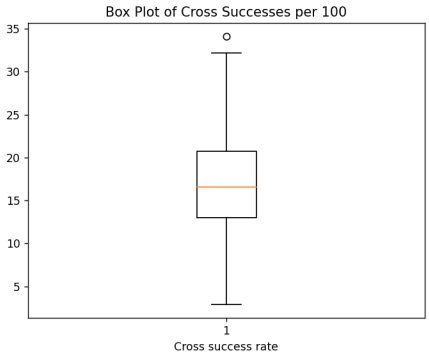


Figure 10. Passmap of the Top Final third passer

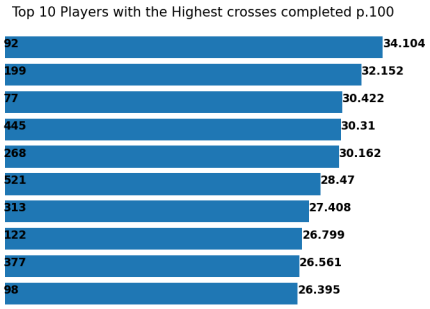
2.3.3. Crossers

I have defined crossers as being individuals with a high success rate from crosses attempted, as opposed to any type of pass attempted. This sets up a slightly different approach to the metric.

In order to ensure that the players I put forward as prolific crossers are not misrepresented, I restrict my search to players who have attempted over 25 crosses in a season, which puts them in the 75^{th} percentile (11a).



(a) Box Plot of crosses per 100 attempted



(b) Top 10 Crossers in the dataset

Figure 11. Crossers

Similar to the progressive passes formula, the formula used to estimate the number of successful crosses per 100 crosses attempted is as below

$$\frac{\textbf{Count of successful crosses for each player}}{\textbf{Count of crosses attempted for each player}} * \textbf{player pass probability} * 100$$

As has been the case thus far, the bar charts (11b) are interactive.

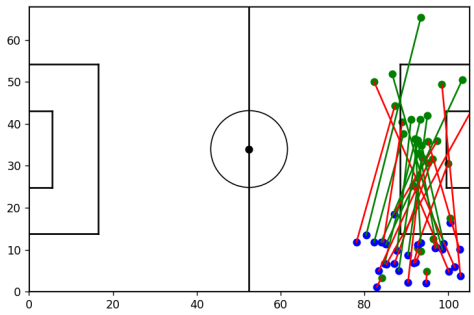
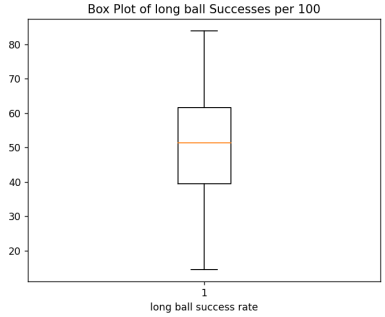


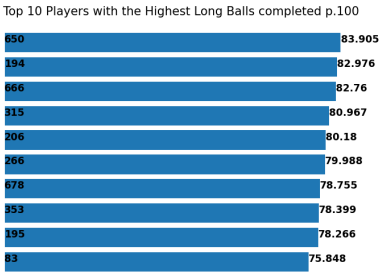
Figure 12. Passmap of the Top Crosser

2.3.4. Long Ball Specialists

I have defined long-ball specialists as players who have a high success rate on the long balls they attempt. I have defined a long ball to be a pass that has a distance of at least 23 meters. I also restrict my search to players who have attempted at least 75 long balls (13a).



(a) Box Plot of successful Long Balls per 100



(b) Top 10 Long Passers in the dataset

Figure 13. Long Ball Specialists

$$\frac{\textbf{Count of successful long balls for each player}}{\textbf{Count of long balls attempted for each player}} * \textbf{player pass probability} * 100$$

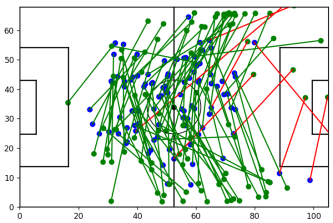


Figure 14. Passmap of the Top Long Ball Player

3. Future Work

While the pass event data used in this project is largely expressive, it fails to tell the whole story of the pass and therefore of the match. I believe that a few more points of information would go a long way toward making the probabilities of pass success more certain. For example, one such data point would be the count of players at a one-meter radius around each point on the trajectory of the pass. This will provide the model with information about how likely the pass is to get intercepted along its path. Another such feature would be the angle of the pass. This would require information about the orientation of the player. For example, a pass attempted by a prone player is far likelier to be misplaced than one from an individual who is upright. This information can be extracted by finding low-dimensional representations of pose data. Yet another set of potentially telling features would be information about the contact with the ball (ie) the point of contact with the ball, the trajectory of flight, the force behind the pass, or a measure of the degree of the Magnus effect on the football.

All of these features would allow for a more robust and specific measure of expected pass accuracy for each passing event. This in turn can be used further down the pipeline to estimate player passing ability and also player passing styles. Similar to [1] Spearman et al's "Physics-based modeling of Pass Probabilities in Soccer", with the right data, it is also possible to estimate the value of a pass from downstream events.

These and more would be potential extensions of this project.

4. Conclusion

In conclusion, this paper has analyzed the passing actions and abilities of players in association football using Stats Perform's pass event data from one season. The findings suggest that there are significant variations in passing probabilities across the football pitch, and players have different passing styles. The use of two models, Logistic regression and a Multi Layered Perceptron, allowed for the prediction of pass probability, and the estimation of pass quality ratings for each player provided valuable insights into the nuances of the game. This research contributes to the growing field of sports analytics and has important implications for player recruitment, team tactics, and game strategy. Overall, this study provides a foundation for future research in this area and highlights the importance of analyzing pass event data in understanding the game of football.

5. Acknowledgement

I would like to extend my gratitude to Dr.Patrick Lucey, and Stats Perform as a whole for allowing me access to their data and for affording me the opportunity to work on what was a very exciting project.

References

[1] W. Spearman, A. Basye, G. Dick, R. Hotovy, and P. Pop, "Physics-based modeling of pass probabilities in soccer," in *Proceedings of the 2017 Winter Simulation Conference*, IEEE, Dec. 2017, pp. 2462–2473.