Predicting the Receivers of Football Passes

Heng Li[1] and Zhiying Zhang[2]

[1] School of Computing, Queen's University, Kingston, Canada
hengli@cs.queensu.ca
[2] Microsoft, Bellevue, WA, USA
zhiyingz@microsoft.com

Abstract. Football (or association football) is a highly-collaborative team sport. Passing the ball to the right player is essential for winning a football game. Anticipating the receiver of a pass can help football players build better collaborations and help coaches make informed tactical decisions. In this work, we analyze a public dataset that contains 12,124 passes performed by professional football players. We extract five dimensions of features from the dataset and build a learning to rank model to predict the receiver of a pass. Our model's first, top-3 and top-5 guesses find the correct receiver of a pass with an accuracy of 50%, 84%, and 94%, respectively, when we exclude false passes, which outperforms three baseline models that we use to rank the candidate receivers of a pass. The features that capture the positions of the candidate receivers play the most important roles in explaining the receiver of a pass.

Keywords: Football pass prediction · Learning to rank · LambdaMART · Gradient boosting decision tree · LightGBM.

1 Introduction

In a football game, players pass the ball to their teammates in order to create good shooting opportunities or prevent the opposing team from getting the control of the ball. Accurately passing the ball to the right player is essential for winning a football game [1, 6].

Prior work [9, 6] studies how passing sequences lead to goals. Their findings have shaped the tactics of many football coaches. In this work, we build a learning to rank model [8] to anticipate the receiver of a football pass. We believe that football coaches and players can take our results into consideration when they make their tactics or make their passes/runs. For example, a player could learn from the important factors for explaining the receiver of a pass to improve his/her chance of receiving the ball. Anticipating receivers of passes can also help automated cameras always focus on the ball in a game.

This work analyzes a dataset which contains 12,124 passes performed by a Belgian football club in 14 games[3]. We want to answer the following research questions (RQs):

[3] https://github.com/JanVanHaaren/mlsa18-pass-prediction

Table 1. A summary of players' passing statistics.

	Back-field	Middle-field	Front-field	Overall
Passing accuracy	86%	83%	79%	83%
Median passing distance (m)	17	14	11	14
Passing forwards ratio	74%	61%	50%	62%

RQ1: How well can we model the receiver of a pass? We build a learning to rank model to predict the receiver of a football pass. An accurate model can help coaches and players make informed tactical decisions in a game.

RQ2: What are the important factors that explain the receiver of a pass? We analyze the model to find the most influential factors that explain the receiver of a pass. Understanding such influential factors can help coaches and players improve their tactics and passes/runs according to these factors.

Paper organization. The remainder of the paper is organized as follows. Section 2 explores the dataset that we use. Section 3 discusses our approaches for building and evaluating our prediction model. Section 4 presents the results for answering our research questions. Finally, Section 5 draws conclusions.

2 Data Exploration

Dataset overview. The dataset contains information about 12,124 football passes. For each pass, the dataset provides the information about the time of the pass since the start of the half, the coordinates of all the players on the pitch, the identifier of the player who passes the ball, and the identifier of the player who receives the ball.

Dealing with missing values. Our goal is to predict the receiver of a pass based on the information about the sender and other players on the pitch (i.e., the candidate receivers). Among the 12,124 passes, there is one pass that misses the coordinates of the sender, and one pass that misses the coordinates of the receiver. For another six passes, the senders and the receivers are the same players. We remove the above-mentioned eight data instances from our dataset. We end up with 12,116 valid passes in our dataset.

Overall, players' passing accuracy is 83%, and the passing accuracy decreases from the back field to the front field. Table 1 shows a summary of players' passing statistics. We define the passing accuracy as the ratio of the passes that reach a teammate. We divide the field into three equally sized areas along the long side of the field, namely back field, middle field and front field. We define a pass as a **back-field pass**, **middle-field pass**, or **front-field** pass when the sender is within the back field, the middle field and the front field, respectively.

The median passing distance is 14 meters, and the passing distance decreases from the back field to the front field. Table 2 shows the

Table 2. Five-number summary of players' passing distance.

Min.	1st Qu.	Median	3rd Qu.	Max.
0	9	14	20	70

five-number summary of players' passing distance. While the maximum passing distance is 70 meters, 75% of the passes are within 20 meters. As shown in Table 1, the median passing distance for the back field, middle field, and front field is 17, 14, and 11 meters, respectively.

Players pass the ball forwards in 62% of the passes, and the ratio of forward passes decreases from the back field to the front field. In the back field, players pass the ball forwards in 74% of the passes, and the ratio decreases to 61% and 50% for middle-field passes and front-field passes, respectively.

Players present different passing characteristics in different areas of the field. Such differences suggest us to build and evaluate our model in different areas of the field separately.

3 Methodology for Predicting the Receivers of Football Passes

This section discusses our overall methodology, including our feature extraction process, modeling and evaluation approaches.

3.1 Feature extraction

From the dataset that we explain in Section 2, we extract five dimensions of features to explain the likelihood of passing the ball to a certain receiver. In total, we extract 54 features. A full list of our features is available at our public github repository[4]. We also share our extracted feature values online [5].

- **Sender position features.** This dimension of features capture the position of the sender on the field, such as the sender's distance to the other team's goal. We choose this dimension of features because players have different passing strategies at different positions, for example, players may pass the ball more conservatively in their own half.
- **Candidate receiver position features.** This dimension of features capture the position of a candidate receiver, such as the candidate receiver's distance to the sender. Senders always consider candidate receivers' positions when they decide to whom to pass the ball.

[4] https://github.com/henglicad/mlsa18-pass-prediction/blob/master/feature-list.md
[5] https://github.com/henglicad/mlsa18-pass-prediction/blob/master/features.tsv

- **Passing path features.** This dimension of features measure the quality of a passing path (i.e., the path from the sender to a candidate receiver), such as the passing angle. The quality of a passing path can predict the outcome (success/failure) of a pass.
- **Team position features.** This dimension of features capture the overall position of the team in control of the ball, such as the front line of the team. Team position might also impact the passing strategy, for example, a defensive team position might be more likely to pass the ball forwards.
- **Game state features.** This dimension of features capture the state of the whole game, such as the time when the sender passes the ball. **We do not use the time when the receiver receives the ball as a feature in our model, as it exposes information about the actual pass (e.g., pass duration).**

3.2 Modeling approach

We formulate the task of predicting the receiver of a football pass as a learning to rank problem [8]. For each pass, our learning to rank model outputs a ranked list of the candidate receivers. A good model should rank the correct receiver at the front of the ranked list. LambdaRank [2] is a general and widely-used learning to rank framework. LambdaRank relies on underlying regression models to provide ranking predictions. **LambdaMART** [2] is the boosting tree version of LambdaRank. It relies on a gradient boosting decision tree (GBDT) [5] to provide ranking predictions. There are quite a few effective implementations of LambdaMART, such as XGBoost and pGBRT, which usually achieve state-of-the-art performance in learning to rank tasks.

In this work, we use an efficient implementation of LambdaMART, **LightGBM** [7], which speeds up the training time of conventional LambdaMART implementations (e.g., XGBoost and pGBRT) by up to 20 times while achieving almost the same accuracy. We use an open source implementation of LightGBM that is contributed by Microsoft[6].

We use a 10-fold cross-validation to build and evaluate our model. The passes data is randomly partitioned into 10 subsets of roughly equal size. We build our model using nine subsets (i.e., the model building data) and evaluate the performance of our model on the held-out subset (i.e., the testing data). The process repeats 10 times until all subsets are used as testing data once.

In each fold, we further split the model building data into the training data and validation data. We train the model on the training data and use the validation data to tune the hyper-parameters of the model. We do a grid search to get the top three sets of hyper-parameter values according to the performance of the model on the validation data. Then, we build three models with these three set of hyper-parameters using the training data. We apply these three models on the testing data and get three sets of results. We then average the results for

[6] https://github.com/Microsoft/LightGBM

each receiver candidate and use the averaged results to rank the receiver candidates. We find that with such an ensemble modeling approach, the accuracy of our model improves up to 2%.

3.3 Baseline models

In order to evaluate the performance of our LightGBM ranking model, we compare it with several baseline models. As discussed in Section 2, 75% of the passes are within 20 meters (i.e., short passes), and 62% of the passes are forward passes. Therefore, we derive baseline models that tend to select the nearest teammates and the teammates in the forward direction as the receiver.

- **The RandomGuess model** selects the receiver of a pass by a random guess. It randomly ranks the candidate receivers.
- **The NearestPass model** selects the nearest teammate of the sender as the top candidate receiver. It ranks the candidate receivers by their distance to the sender, from the teammates of the sender to the opponents, and then from the closest to the furthest.
- **The NearestForwardPass model** selects the nearest teammate of the sender that is in the forward direction (relative to the sender) as the top candidate receiver. It ranks the candidate receivers by their relative position to the sender, from the teammates of the sender to the opponents, then from the players in the forward direction to the players in the backward direction, and finally from the closest to the furthest.

3.4 Evaluation approaches

We use **top-N accuracy** and **mean reciprocal rank (MRR)** to measure the performance of our model. Top-N accuracy measures the accuracy of the model's top-N recommendations, i.e., the probability that the correct receiver of a pass appears in the top-N receiver candidates that are predicted by the model. For example, top-1 accuracy measures the probability that the correct receiver of a pass is the first player in the predicted list of receiver candidates.

Reciprocal rank is the inverse of the rank of the correct receiver of a pass in an ranked list of candidate receivers predicted by the model. MRR [3] is the average of the reciprocal ranks over a sample of passes P:

$$\mathrm{MRR} = \frac{1}{|P|} \sum_{p=1}^{|P|} \frac{1}{\mathrm{rank}_p} \tag{1}$$

where rank_p is the rank of the correct receiver for the pth pass. The reciprocal value of MRR corresponds to the harmonic mean of the ranks. MRR ranges from 0 to 1, the larger the better. While top-N accuracy captures how likely the correct receiver appears in the top-N predicted receivers, MRR captures the average rank of the correct receiver in the predicted list of receiver candidates.

As discussed in Section 3.2, we use a 10-fold cross-validation to build and evaluate our model. Therefore, we use a mean top-N accuracy and MRR across the 10 folds in Section 4.

Table 3. The accuracy of our model for predicting the receiver of a pass (excluding false passes).

	Back-field	Middle-field	Front-field	Overall
Top-1 accuracy	53%	46%	55%	50%
Top-3 accuracy	84%	81%	91%	84%
Top-5 accuracy	93%	93%	97%	94%
MRR	0.70	0.66	0.73	0.68

3.5 Feature importance

In order to understand the importance of the features in our model, we use the feature importance scores that are automatically provided by a trained LightGBM model. Gradient boosting decision trees (e.g., LightGBM) provide a straightforward way to retrieve the importance scores of each feature [4].

After the boosting decision trees are constructed, for each decision tree, the importance of a feature is calculated by the amount that the feature improves the performance measure at its split point (i.e., split gains). The importance of each feature is then accumulated across all of the decisions trees in the model.

4 Results

This section discusses the answers to our research questions.

4.1 RQ1: How well can we model the receiver of a pass?

Our model can predict the receiver of a pass with a top-1, top-3 and top-5 accuracy of 50%, 84%, and 94%, respectively, when we exclude false passes (i.e., passes to the other team). Table 3 shows the performance of our model when we exclude false passes. The "Back-field", "Middle-field", "Front-field" and "Overall" columns show the performance of our model for back-field passes, middle-field passes, front-field passes and all passes, respectively. A top-3 accuracy of 84% for all passes means that the actual receiver of a pass has a 84% chance to appear in our top-3 predicted candidates. The MRR value for all passes is 0.68, which means on average, the correct receiver is ranked 1.5th (i.e., 1/0.68) out of 10 or less receiver candidates (i.e., all teammates of the sender).

Our model can predict the receiver of a pass with a top-1, top-3 and top-5 accuracy of 41%, 70%, and 81%, respectively, when we consider all passes. Table 5 shows the performance of our model when we consider all passes (including false passes). The performance of our model decreases when we consider false passes (i.e., passes to the other team). False passes are very difficult to predict because it is not the sender player's intention to pass the ball to the other team. The MRR value for all passes is 0.58, which means the correct receiver is averagely ranked 1.7th (i.e., 1/0.58) out of all 21 or less candidate receivers (i.e., all players excluding the sender).

Table 4. Comparing the accuracy of our model with baseline models (excluding false passes).

	LightGBM	RandomGuess	NearestPass	NearestForwardPass
Top-1 accuracy	**50%**	10%	33%	27%
Top-3 accuracy	**84%**	30%	70%	54%
Top-5 accuracy	**94%**	50%	86%	71%
MRR	**0.68**	0.29	0.55	0.47

Table 5. The accuracy of our model for predicting the receiver of a pass (considering all passes including passes to the other team).

	Back-field	Middle-field	Front-field	Overall
Top-1 accuracy	45%	38%	43%	41%
Top-3 accuracy	72%	68%	72%	70%
Top-5 accuracy	82%	80%	83%	81%
MRR	0.61	0.56	0.60	0.58

Table 6. Comparing the accuracy of our model with baseline models (considering all passes including passes to the other team).

	LightGBM	RandomGuess	NearestPass	NearestForwardPass
Top-1 accuracy	**41%**	5%	27%	23%
Top-3 accuracy	**70%**	14%	58%	45%
Top-5 accuracy	**81%**	24%	71%	59%
MRR	**0.58**	0.17	0.47	0.40

Our model performs better for back-field and front-field passes, while performing worse for middle-field passes. Table 3 and Table 5 also shows the performance of our model for back-field, middle-field and front-field passes, separately. Surprisingly, the performance of our model is the worst for middle-field passes. A player in the middle area may have more passing options, thereby increasing the difficulty to predict the right receivers.

Our model perform better than the baseline models. Table 4 and Table 6 compare the performance of our LightGBM model with the RandomGuess, NearestPass, and NearestForwardPass models which are described in Section 3. Our LightGBM model consistently show much better performance than the three baseline models in terms of the top-N accuracy and MRR. The NearestPass model, which tends to pass the ball to the nearest teammates, achieve a better performance than the NearestForwardPass, which tends to pass the ball to the nearest teammates in the forward direction relative to the sender. Both of the NearestPass and NearestForwardPass baseline models achieve a much better performance than randomly guessing the receiver of a pass.

Table 7. The combined importance of each feature dimension.

Feature dimension	Combined feature importance
Candidate receiver position	6723
Team position	3493
Sender position	2688
Passing path	1753
Game state	343

Our model can predict the receiver of a pass with a top-1, top-3 and top-5 accuracy of 50%, 84%, and 94%, respectively, when we exclude false passes, outperforming three baseline models. Our model performs better when the sender of a pass is in the back or front area of the field.

4.2 RQ2: what are the important factors that explain the receiver of a pass?

The features that capture the candidate receivers' positions play the most important roles in explaining the receiver of a pass. Table 7 shows the combined importance of each feature dimension in our model. The combined importance is a sum of the importance scores of all the individual features in a dimension. The features from the dimension of candidate receiver position, which capture a candidate receiver's position on the pitch and his/her position relative to the teammates and opponents, have the biggest combined importance score in our model. The team position features, which captures the overall position of the team in control of the ball, are the second important dimension in explaining the receiver of a pass. The third important feature dimension (i.e., sender position) captures the sender's position on the pitch and his/her position relative to the teammates and opponents. The passing path features, which captures the characteristics of a passing path (i.e., the path from the sender to a candidate receiver), also play a significant role in explaining the receiver of a pass.

The most important features capture the candidate receivers' positions relative to the sender and the opponents. Table 8 lists the top ten features that are most important in our model and their respective importance scores. A full list of our features' important scores is available online[4]. Among the top 10 important features, there are eight features from the dimension of candidate receiver position. All of the top six features are from the dimension of candidate receiver position, among which three features capture a candidate receiver's relative position to the sender, and the other three features capture a candidate receiver's relative position to the components. The other two features from the dimension of candidate receiver position capture a candidate receiver's position on the pitch and his/her relative position to the teammates, respectively. Among the top 10 important features, there are also one from the sender

Table 8. The ten most important features and their importance scores.

Dimension	Feature	Importance	Description
Receiver position[1]	receiver_closest_opponent_dist	715	The distance between a candidate receiver and his/her closest opponent
Receiver position	norm_receiver_sender_x_diff	660	Normalized x-axis difference between a candidate receiver and the sender
Receiver position	abs_y_diff	653	The absolute value of the y-axis difference between the sender and a candidate receiver
Receiver position	distance	616	A candidate receiver's distance to the sender
Receiver position	receiver_closest_opponent_to_sender_dist	602	The distance between the sender and a candidate receiver's closest opponent
Receiver position	receiver_closest_3_opponents_dist	558	The average distance between a candidate receiver and his/her three closest opponents
Receiver position	receiver_to_center_distance	521	The distance between a candidate receiver and the center of the pitch
Receiver position	receiver_closest_3_teammates_dist	508	The average distance between a candidate receiver and his/her three closest teammates
Sender position	sender_closest_opponent_dist	498	The distance between the sender and his/her closest opponent
Passing path	min_pass_angle	467	Pass angles are the angles between the line from the sender to a candidate receiver (i.e., the pass line) and the lines from the sender to the opponents along the pass line. The min_pass_angle is the minimum pass angle for a pass line

[1] "Receiver position" is short for "candidate receiver position".

position dimension (i.e., the sender_closest_opponent_dist feature), and one from the passing path dimension (i.e., the min_pass_angle feature).

The features that capture the positions of the candidate receivers, in particular, relative to the sender and the opponents, play the most important roles in explaining the receiver of a pass.

5 Conclusions

This work proposes a novel approach to predict the receivers of football passes. We analyze a dataset containing 12,124 passes from 14 real-world football games and discuss players' passing characteristics. We find that players present different passing characteristics in different areas of the field. We then extract 54 features along five dimensions and build a LightGBM model to predict the receiver of a pass. Our model achieves a top-1, top-3, and top-5 accuracy of 50%, 84%, and 94%, respectively, when we exclude false passes. Our model outperforms three baseline models that we use to rank the candidate receivers of a pass. We find that the features that capture the positions of the candidate receivers play the most important roles in explaining the receiver of a pass. We believe that our approaches and findings can help football practitioners better understand the factors that impact the receiver of a pass and make informed tactical decisions.

References

1. Ali, A.: Measuring soccer skill performance: a review. Scandinavian Journal of Medicine & Science in Sports **21**(2), 170–183 (2011)
2. Burges, C.J.: From ranknet to lambdarank to lambdamart: An overview. Learning **11**(23-581), 81 (2010)
3. Craswell, N.: Mean Reciprocal Rank, pp. 1703–1703. Springer US, Boston, MA (2009)
4. Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning, vol. 1. Springer series in statistics (2001)
5. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
6. Hughes, M., Franks, I.: Analysis of passing sequences, shots and goals in soccer. Journal of Sports Sciences **23**(5), 509–514 (2005)
7. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 3146–3154 (2017)
8. Liu, T.Y., et al.: Learning to rank for information retrieval. Foundations and Trends® in Information Retrieval **3**(3), 225–331 (2009)
9. Reep, C., Benjamin, B.: Skill and chance in association football. Journal of the Royal Statistical Society. Series A (General) **131**(4), 581–585 (1968)