Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights

Joel Brooks
Massachusetts Institute of Technology
Cambridge, MA
brooksjd@mit.edu

Matthew Kerr
Massachusetts Institute of Technology
Cambridge, MA
mattkerr@alum.mit.edu

John Guttag
Massachusetts Institute of Technology
Cambridge, MA
guttag@mit.edu

ABSTRACT

Quantitative evaluation of the ability of soccer players to contribute to team offensive performance is typically based on goals scored, assists made, and shots taken. In this paper, we describe a novel player ranking system based entirely on the value of passes completed. This value is derived based on the relationship of pass locations in a possession and shot opportunities generated. This relationship is learned by applying a supervised machine learning model to pass locations in event data from the 2012-2013 La Liga season. Interestingly, though this metric is based entirely on passes, the derived player rankings are largely consistent with general perceptions of offensive ability, e.g., Messi and Ronaldo are near the top. Additionally, when used to rank midfielders, it separates the more offensively-minded players from others.

Keywords

machine learning; sports analytics; soccer analytics

1. INTRODUCTION

Although soccer is by far the world's most popular sport [19], published work in soccer analytics has yet to achieve the same level of sophistication as analytics being performed in other professional sports. Crude summary statistics such as goals, shots, and assists are still the most common way to compare player performance analytically. More work is emerging [23] that leverages the rich datasets available to make discoveries about soccer, but there has not been much focus on quantitative metrics for evaluating player performance.

In this paper, we describe a novel way of quantitatively measuring a player's *passing* performance using existing data. We chose to focus on passing because it is one of the more strategic elements of soccer. Currently, players are often considered good passers if they accumulate many assists. Assists identify when a player makes a pass that directly leads to a goal, but this measure alone is quite limited. For example, assists do not capture passes that would have been assists except for an errant shot, or an excellent save by the opposing team's goalkeeper. Opta [17] extends the idea of assists to include all passes that lead to shots (whether or not they lead to a goal) in their "key passes" metric, but both this metric and assists are only applicable to passes immediately preceding a shot. There may be players who are excellent passers that create many opportunities for their team, but rarely make the last pass before a shot or goal.

Instead, we want to be able to quantitatively measure the importance of *any* pass. We accomplish this by first training a classifier that uses information about the locations of a set of passes to identify when that group of passes results in a shot. Since we use a linear classifier, we can directly utilize the model weights to understand the relative importance of pass origins and destinations for generating shot opportunities. These weights allow us to compute an estimated value of any pass for creating shots. We can then rank players by the value of the passes they complete over the course of a season.

In this paper, we use data from the 2012-2013 La Liga season to:

1. Construct a model relating pass origins and destinations during a possession with the probability of a shot. This model accurately identifies whether a possession ends in a shot from the pass locations alone.

2. Show how the resulting weights offer insights into the offensive utility of passes.

3. Utilize this model to rank players by the frequency with which their passes are highly valued by the model.

The rest of the paper is organized as follows. In Section 2, we outline some related work on using machine learning for knowledge discovery in soccer and other sports. In Section 3, we describe our event-based dataset. In Section 4, we present our methodology for building a model for predicting shots at the end of possessions. In Section 5, we demonstrate that this model can be used to rank players in an objective manner by how often they complete passes our model rates as valuable. Finally, in Section 6, we summarize the overall contributions of this paper and discuss possible future work.

2. RELATED WORK

Much of the published research in sports analytics, especially research that utilizes spatiotemporal data, has focused on sports that are easily discretized, such as baseball,

KDD '16 August 13-17, 2016, San Francisco, CA, USA

ACM ISBN 978-1-4503-4232-2/16/08.

DOI: http://dx.doi.org/10.1145/2939672.2939695

American football and tennis [22]. These sports are easily broken up into individual events (e.g. at bats, plays or points) that have immediate outcomes, such as hits, yards gained or a point won. For example, Intille and Bobick used player tracking data to recognize different plays in American football [11]. It is more difficult to perform similar work for sports that are not as easily discretized, such as basketball and soccer, because the continuous nature of play makes the connections to outcomes less obvious. Previous work has used tracking data for basketball to classify offensive plays and the movement patterns of offensive players [12, 18]. Similar work in soccer has proved to be more difficult because there are not as clear distinctions between epochs in play in soccer as there are in basketball.

Soccer analytics has focused on building probabilistic models to simulate game actions and predict the outcomes of matches or goals scored. Reep and Benjamin developed models for the success rates of different length passing sequences [21]. More recently, there has been work on predicting the outcomes of matches by using possession rates of different teams and other historical statistics to develop probabilistic models [7, 8]. Other work has identified a relationship between a goal being scored and the frequency of passes in the 5 minutes preceding that goal [20]. In contrast, our work focuses on predicting shots taken in possessions as opposed to the outcome of games or goals. Additionally, we focus on how these predictions can be useful for ranking players in a quantitative fashion, as opposed to the performance of the predictors themselves.

An increasing amount of spatiotemporal data for soccer is allowing analysts to study the underlying mechanics of the game in a manner that would not be possible with box score statistics alone. Bloomfield *et al.* used player tracking information to investigate the physical demands of soccer and the work rates of different players [3]. Gyarmati *et al.* leveraged ball-event data and passing sequences to cluster the playing styles of different teams [10]. Lucey *et al.* used ball-event data to infer the location of the ball throughout a game. Using this information, they constructed "entropy-maps" to characterize how different teams move the ball during a match [14]. In more recent work, the authors combine match statistics, event data and player tracking data to identify the teams playing in a given game with 70% accuracy [2].

Our feature representation is an extension of previous work we did for using pass locations to identify teams by their passing styles [13]. We also showed that these feature representations are promising for building quantitative player-rankings [4]. However, in this work, we extend these representations to have smarter ways of relating field locations, and a better breakdown of the understanding of offensive performance by different types of positional players.

Other recent work in sports analytics has focused specifically on quantitatively measuring specific player contributions. Macdonald [15] fits a ridge-regression model of a number of statistics of players in a hockey team to their expected goal-scoring rate. This allows a quantitative measure of an individual player's contribution to those expected goals while taking account of the contributions of the other players on the team. Cervone and others [5] fit a model relating actions and movements taken by the ball handler in basketball to the expected number of points scored over the course of a possession. They then demonstrate how this model can rank players based on the average value of every action they take over a course of a season. We similarly utilize predictive models for evaluating player performance, but our metric is focused specifically on the value of the location of a completed pass.

Finally, we draw upon work that trains interpretable discriminative models, and then utilizes those model weights for developing quantitative metrics for stratifying populations. This is commonly done in the medical literature for developing risk scores. Specifically, both the TIMI risk score [1] for risk of adverse cardiac events and the SAPS III score [16] for evaluation of patient status in intensive care units utilize logistic regression models to understand which patient variables have the strongest relationship with adverse outcomes. They then utilize their trained model weights for constructing simple metrics that are both easy to compute manually, and can accurately separate high risk patients from low risk patients. This is analogous to how we use a model trained on predicting shots for identifying important origin and destinations of passes, and then utilize these locations to rank players.

3. THE DATA

3.1 Data Overview

The data we used are hand-labeled annotations of each ball-event that took place during the course of a match (e.g. each pass, tackle, shot, etc). A ball-event is recorded any time a player makes a play on the ball, apart from dribbling. The dataset also includes additional information for each ball-event such as the location, the player involved, and the outcome.

We focused specifically on the *locations* of pass attempt origins and destinations because we hypothesized that pass location is a strong indicator of team strategy and personnel. We use data collected from the 2012-13 La Liga season. La Liga is the premier league in Spain and is comprised of 20 teams. On average, each team in La Liga attempted around 18,000 passes during the entire season, with Barcelona making the most passes by a wide margin at 30,283, and Levante attempting the smallest number of passes at 13,094. We present a complete list of teams ranked by their final standing in La Liga with the number of shots they took in the 2012-13 season in Table 1.

3.2 Pass Location Representation

Each pass event in the dataset has an origin and destination location. To aggregate pass origin or destination locations that are near each other, we discretized the field into 18 zones, as shown in Figure 2. This representation has previously been shown to identify critical zones on the field associated with offensive outcomes such as shots and goals [6]. Additionally, in previous work we showed that by discretizing the field into these 18 zones, we can build representations of pass locations that are highly indicative of team passing styles [4]. However, we did not want to represent each pass location as belonging to a single zone, since we do not believe the hard cutoffs between zones are representative of anything fundamental to the game of soccer.

Instead, we extended our previous representation of each pass location (either a location of a pass origin or a destination) to be a continuous-valued vector of length 18, where each element in the vector represents the closeness of a

Rank	Team	Shots
1	Barcelona	528
2	Real Madrid	710
3	Atletico Madrid	513
4	Real Sociedad	557
5	Valencia	567
6	Malaga	443
7	Betis	456
8	Rayo Vallecano	567
9	Sevilla	553
10	Getafe	473
11	Levante	402
12	Atletico Bilbao	482
13	Espanyol	413
14	Valladolid	393
15	Granada	470
16	Osasuna	514
17	Celta de Vigo	460
18	Mallorca	485
19	Deportiva La Coruna	527
20	Zaragoza	461

Table 1: **Teams in the La Liga 2012-13 season.** Teams are ranked by their final standing in the league at the season end.

given pass location to the corresponding zone. We then compute the vector representation for a pass location l as $\mathbf{r}^l = [r_1^l \dots r_{18}^l]$, where each element is

$$r_i^l = \frac{c_i}{max(d(l, z_i), 1)} \quad (1)$$

where $d(l, z_i)$ is the Euclidean distance between the coordinates of l and the center of zone i (z_i), and c_i is an indicator variable that takes the value 1 if zone i is one of the N closest zones (by Euclidean distance to the center of the zone) to l, and 0 otherwise. We utilize the $max(d(l, z_i), 1)$ to prevent passes located very near the center of a zone from leading to extremely large values. Finally, we normalize $\mathbf{r}^l$ by its L1-norm so that the $\sum_{i=1}^{18} r_i^l = 1$.

The resulting vector $\mathbf{r}^l$ has N non-zero values. This vector describes how close a pass location is to the N^{th} closest zones. For our task, $N = 2$ seemed to provide the best results. Intuitively, this results in a location representation that emphasizes the zone in which the pass occurred, but accounts for cases where a pass location is near the border of two zones. This representation allows aggregation or averaging of multiple passes in a sequence of passes without restricting any one pass in that sequence to have an origin or destination in just a single zone.

4. METHODS

4.1 Feature Extraction

To extract the features that we used to build our predictive models, we first segment each game into a discrete sequence of observations. We segment the game at the level of *possessions*. A possession in soccer is defined as a period of time when a single team retains the ball without an interruption in play or loss of the ball to the opposing team. Possessions contain a sequence of passes between players on the same team. We only considered possessions that had a minimum number of passes in order to remove epochs of play where a team only had the possession of the ball for a short period of time. We set this minimum to be 3 complete passes since we found a minimum of 4 or greater led to a steep drop in overall performance.

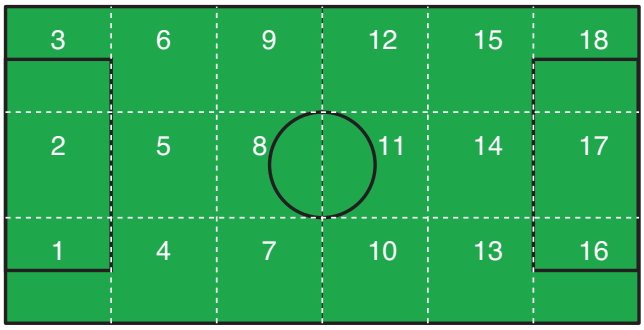

Figure 2: **The playing area split up into 18 zones.** The left side of the field (zones 1-3) is the defensive side of the field, and the right side (zones 16-18) is the offensive area.

We extracted features from each possession to construct the feature vectors for our learning algorithms. We constructed the feature vector for a single pass with origin and destination locations (l_o, l_d) in a possession by:

1. Computing $\mathbf{r}^{l_o}$ and $\mathbf{r}^{l_d}$, the representations of the pass origin and destination, respectively, using the method described in Section 3.2.
2. Constructing the matrix $\mathbf{R}^{l_{od}} = \mathbf{r}^{l_o} \otimes \mathbf{r}^{l_d}$, the outer product of the origin and destination representations. These origin-destination-pair features provide information about which pairs of origins and destinations are more likely to lead to shots (e.g. knowing that the ball was passed into the middle from the corner as opposed to just knowing that the ball reached the middle).
3. Constructing the feature vector as the concatenation of $\mathbf{r}^{l_o}$, $\mathbf{r}^{l_d}$, and a flattened $\mathbf{R}^{l_{od}}$.

The origin and destination representations have one value per zone, so each accounts for 18 features. The flattened origin-destination-pair representation is of length 324. As a result, each pass was converted into a feature vector of length 360. To get a feature vector for a possession, we average the feature vectors across all completed passes in that possession. This feature vector represents a summary of the locations of pass origins and destinations of a possession.

Each feature vector is labeled with $y \in \{1, -1\}$ according to how the possession ended. Possessions that ended in a shot taken by the offensive team were assigned a label of 1, and all others were assigned a label of -1.

4.2 Experimental Design and Testing

Upon converting each possession in a game to a feature vector, we then used these feature vectors to train models to relate passing strategy in a possession to shots taken. We first split the data into a training and holdout set. We split the data by using the first 80% of games chronologically as the training set, and setting aside the final 20% as the holdout set. We did this split to simulate the scenario of applying our models to newly generated data.

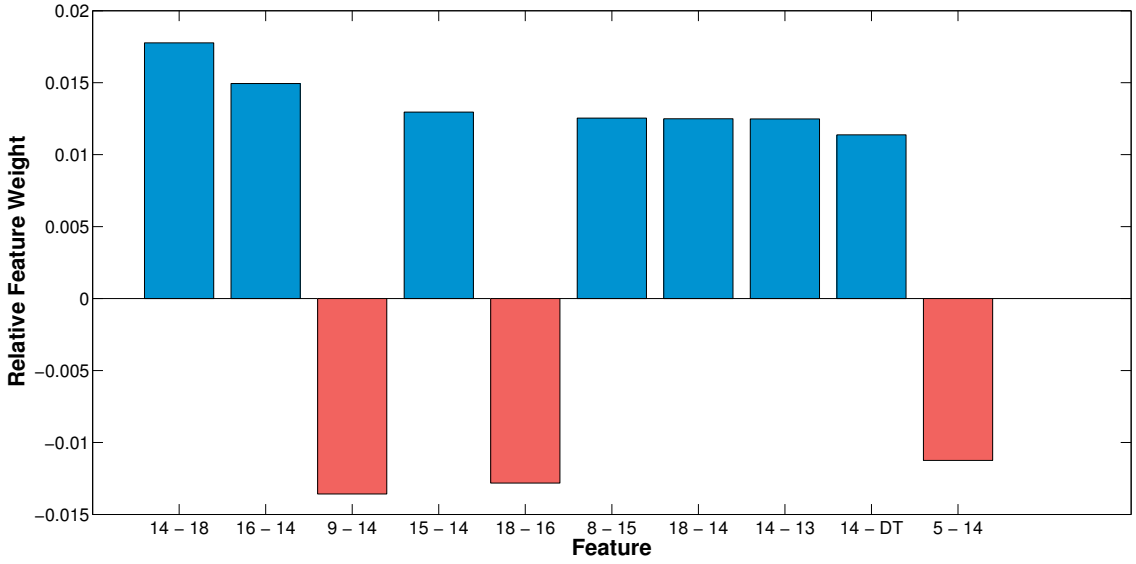

Figure 1: **Top 10 most influential features for predicting when a possession will end in a shot**. All features are normalized by the total sum of absolute weights. Each number corresponds to a zone. "DT" designates a pass destination feature, and features labeled with two zones, "$Zone_1$ - $Zone_2$," designate an origin-destination feature from $Zone_1$ to $Zone_2$.

Using the training set, we trained an L2-regularized Support Vector Machine (SVM) model using the LIBLINEAR package [9]. We used LIBLINEAR's asymmetric cost parameter option in order to account for the extreme class imbalance between positive and negative examples in the training set. We utilized 5-fold cross validation to find the optimal class-specific cost parameters on the training set. The folds were constructed at the *game* level so possessions in a single game were not split across multiple folds. We chose the cost parameters that had the maximum average Area Under the Receiver Operating Characteristic Curve (AUROC) on the five test folds, and used those parameters to train the final model. The final model was then tested on the holdout set.

5. RESULTS

5.1 Classification Results

Our model that predicts whether a possession will end with a shot has an AUROC of 0.79 and an F-score of 0.31. We plot the ROC curve in Figure 3. We used this model to investigate the relationship between our features and shots by looking at the relative importance of each zone. We show the feature weights for the top 10 features in Figure 1. The weights presented in the figure are normalized by the total sum of the absolute value of all the weights to reflect a feature's relative importance. We can see that no single feature dominates and that the top features have a mixture of positive and negative weights. This provides further evidence that our model captures a trade-off between where passes are likely to lead to shot opportunities in the future and is not based solely on simple rules such as "shots happen when there are passes near the opposing team's goal." In fact, zone 17, the area directly in front of the opposing team's goal, does not appear in the top 10 features.

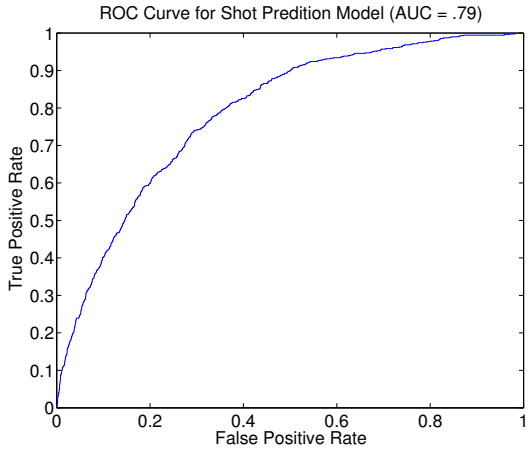

Figure 3: **ROC Curve for the Shot Prediction Model**

Looking at the relative feature importance provides insight into which pass origins and destinations generally lead to shot opportunities. Many of these top features involve zone 14, the "critical zone" centered in the field in front of the penalty box. Possessing the ball in this zone has been strongly identified to be associated with positive offensive outcomes [6]. This is supported by our model, where 6 of the top 10 features are positively associated features that involve zone 14. However, not all of our top features involving zone 14 have positive weights. Passes to zone 14 from either zone 5 or zone 9 both have negative weights. Thus long completed passes to zone 14 are not positively associated with shot opportunities by our model. So while getting the ball into zone 14 can often lead to shot opportunities, it depends a great deal on how you get it there.

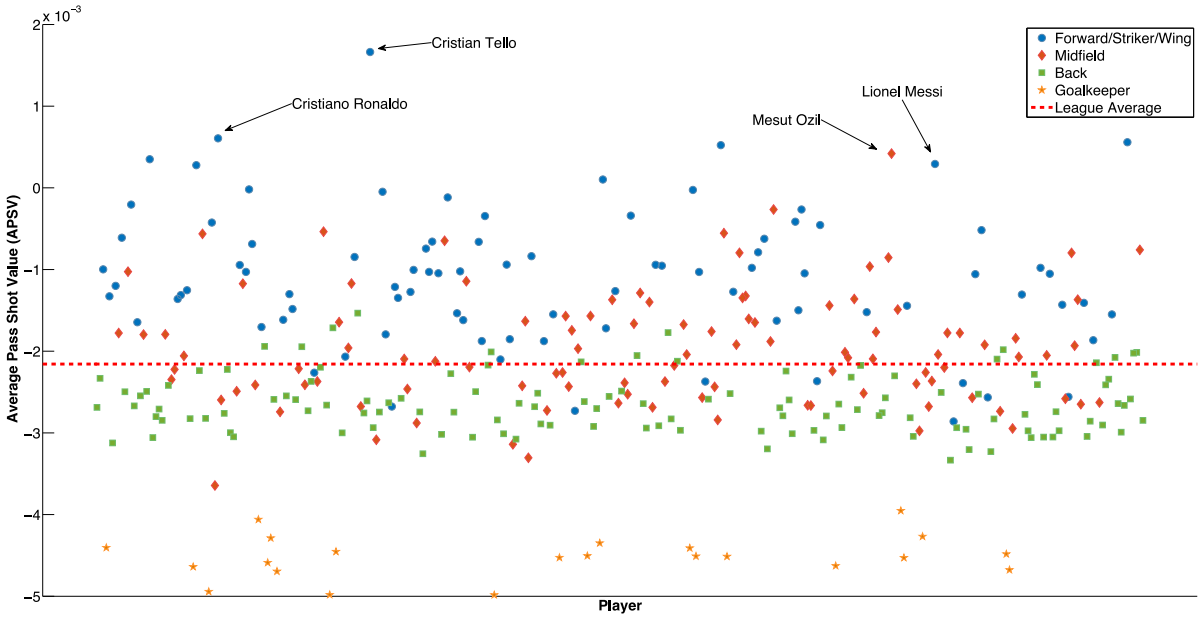

Figure 4: **APSV of passes for all players with more than 200 passes in the 2012-13 La Liga season.** Each item in the plot represents the APSV of a single player. Players have an associated color and shape chosen by their position. The APSV across the entire collection of players is represented by the dashed red line.

Many of the top features also involve the corner areas. The top two features overall are both passes between the corner areas and zone 14. This suggests that many passes that lead to shot opportunities later in the possession involve getting the ball into the corner of the field. However, there is also a strongly negatively associated feature weight with a ball sent from one corner to another (zone 18 to zone 16). This suggests that crosses across the field are harmful for generating shots if they result in the ball being played to the other corner.

5.2 Player Rankings by Shot Prediction Models

In the previous section, we described how we trained a model relating a possession to the outcome of the possession ending in a shot. The resulting model is a vector $\mathbf{w} = [\mathbf{w}^o, \mathbf{w}^d, \mathbf{w}^{od}]$, where $\mathbf{w}^o, \mathbf{w}^d, \mathbf{w}^{od}$ are the vectors of feature weights corresponding to the different zones for pass origins, destinations, and origin-destination pairs, respectively. These weights provide a conceptual map of the field that suggests which pass locations are most likely to lead to a shot opportunity later in the possession. We can use this map to rate a given pass by its association with shot opportunities in our model.

We took every completed pass in the La Liga 2012-13, and using our model computed an estimate of its relative importance for generating a shot. This importance, called Pass Shot Value (PSV), is computed for a pass with an origin in zone i and a destination in zone j as:

$$PSV(i,j) = w_i^o + w_j^d + w_{ij}^{od} \quad (2)$$

Thus, the PSV for a pass is the sum of the feature weights for its corresponding origin, destination and origin-destination pair. For example, a pass from zone 3 to zone 4 would have a PSV of the sum of the model weight for an origin in zone 3, the weight for the destination in zone 4, and the weight of the pair of having an origin of 3 and a destination of 4. We then computed the Average Pass Shot Value (APSV) for all players in La Liga who had over 200 completed passes in the 2012-13 season. We found that 200 passes filtered out everyone besides the top 15-20 players per team by number of completed passes, leaving the main contributors over that season. We again excluded passes that did not occur in a possession of 3 or more completed passes. Note that we only use the closest zone to the origin and destination for computing PSV. This is because we assume the predictive model has accounted for relationships between zones that are near each other.

We plotted the APSV for these players in Figure 4, which shows how our model would rank each player by their average tendency to complete passes that lead to a shot. Unsurprisingly, the APSV metric is biased towards offensive oriented players. In fact, the separation of the different positions by APSV indicates that APSV alone would be a fairly strong predictor of position. The goalkeepers are completely separate from the field position players. As such, APSV is most useful when comparing players within a position.

We ranked the top ten players by position category in Table 2. We also listed the top players by goals and assists in Table 3. Note that APSV is almost always negative. Most possessions do not end in a shot, and thus, most of the model's features are negatively associated with a shot opportunity being generated. Therefore, players make passes with a negative model value the vast majority of the time. In spite of this, some offensive players and one midfielder (Mesut Ozil) have a positive APSV. This suggests that generally their passes were rated by the model to be positively associated with shots at the end of a possession.

Rank	Player
1	Cristian Tello
2	Cristiano Ronaldo
3	Sergio Garcia
4	Karim Benzema
5	Gonzalo Higuain
6	Lionel Messi
7	Jonathan Viera
8	Angel Di Maria
9	Nolito
10	Jorge Molina

(a) **Offense**

Rank	Player
1	Mesut Ozil
2	Diego Buonanotte
3	Kaka
4	Emiliano Armenteros
5	Miguel de las Cuevas
6	Julio Baptista
7	Arda Turan
8	Jose Barkero
9	Cesc Fabregas
10	Ever Banega

(b) **Midfield**

Rank	Player
1	Marcelo Vieira
2	Eliseu
3	Dani Alves
4	Filipe Luis
5	Aly Cissokho
6	Diego Colotto
7	Martin Demichelis
8	Nacho Monreal
9	Sergio Sangez
10	Lolo

(c) **Defense**

Table 2: **Top 10 players in La Liga 2012-13 by APSV**. We separate the positions into three categories: Offense, Midfield, and Defense.

Rank	Player	Goals
1	Lionel Messi	46
2	Cristiano Ronaldo	34
3	Radamel Falcao	28
4	Alvaro Negredo	25
4	Roberto Soldado	24
6	Ruben Castro	18
6	Piti	18
8	Gonzalo Higuain	16
9	Carlos Vela	14
9	Helder Postiga	14
9	Artiz Aduriz	14

(a) **Top Goal Scorers**

Rank	Player	Assists
1	Andres Iniesta	16
2	Mesut Ozil	13
3	Lionel Messi	12
4	Karim Benzema	11
4	Cesc Fabregas	11
6	Cristiano Ronaldo	10
6	Ivan Rakitic	10
8	Ibai Gomez	9
8	Carlos Vela	9
8	Koke	9
8	Alexis Sanchez	9

(b) **Top Players by Assists**

Table 3: **Top 10 players in La Liga 2012-13 by Goals and Assists**

We grouped forwards, strikers, and wingers into an "offense" category. The top players by APSV in the offense category include renowned players such as Ronaldo and Messi. They were the top two scorers in La Liga that season and they both finished the season in the top 10 for assists as well. In fact, we find that a significant ($\rho = 0.27, p < 0.05$) correlation between APSV and goals scored for the season for players in the "offense" category. Notice, however, that neither goals or assists are used in compute APSV, suggesting that these players are not only effective scorers but also effective passers. Others in this category were considered to be strong offensive players and appear in the list of top players by goals and assists as well.

Even within the midfield and defense categories, we see players that are identified to play an "offensive" style. Midfielders in the top 10 by APSV like Ozil and Kaka play most often in the "attacking midfielder" position. Marcelo Viera and Eliseu play as backs, but are known for being capable at playing in the wing position as well. Others like Dani Alves are pure backs but are known for contributing on the offensive end.

6. CONCLUSION

In this paper we presented a novel method of utilizing soccer event data to understand the relationship between pass location and shot opportunities. We showed that the locations of the origins and destinations of passes in a possession relate strongly to whether that possession will end in a shot. Using supervised machine learning techniques, we built a model for predicting whether a possession will end in a shot. The model had an AUROC of 0.79.

We used the features of this model to create a map to understand the relative importance for generating shot opportunities of passing from one location to another. We then used this map to build a data-driven ranking of players by weighing a pass by its relative importance for generating a shot later in the possession. When we ranked all players in La Liga 2012-13 with more than 200 passes with this metric, we see some of the elite attacking players at the top. This ranking also correlates well with standard offensive box score metrics such as goals and assists, even though neither were directly used in its computation. We believe this warrants further investigation into its utility as a player comparison tool. Furthermore, we have outlined a framework for constructing data-driven player metrics. For example, using actions taken by players on defense for predicting *defensive* outcomes may be used to help rank players by defensive ability.

We believe that our results show that appropriate analyses of pass event data in soccer can provide sometimes non-obvious insights. However, soccer is a complicated sport with constantly changing game situations. Incorporating temporal information (e.g. duration of possession, time elapsed between passes) in any analyses would provide more situation-specific insights. Also, utilizing player-tracking data as a source dataset would better allow investigation into the strate-

gic aspects of the game that are not directly involved with the ball. Expanding our features to include sequential information could give a more detailed understanding of how passing strategy relates to outcomes. Lastly, if a team had a large collection of event data from their own games, they could build team-specific models that could provide a better analysis into which strategies are most promising in their system. Further investigation will better reveal how useful this type of analysis can be for gaining a deeper understanding of the world's most popular game.

Finally, we would like to extend this work to other sports, such as basketball. From a technical perspective, this methodology for creating data-driven player metrics could be applied to any sports datasets with events and location information. Just as in soccer, passing performance is primarily measured in basketball by assists. By using the methods from this paper, we could rank players by the value of all passes they complete, and not those just before a shot.

7. REFERENCES

[1] E. M. Antman, M. Cohen, P. J. Bernink, C. H. McCabe, T. Horacek, G. Papuchis, B. Mautner, R. Corbalan, D. Radley, and E. Braunwald. The TIMI risk score for unstable angina/non–st elevation mi: a method for prognostication and therapeutic decision making. *Jama*, 284(7):835–842, 2000.

[2] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews. Identifying team style in soccer using formations learned from spatiotemporal tracking data. In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*, pages 9–14. IEEE, 2014.

[3] J. Bloomfield, G. K. Jonsson, R. Polman, K. Houlahan, and P. O'Donoghue. Temporal pattern analysis and its applicability in soccer. In L. Anoilli, S. Duncan Jr., M. Magnusson, and G. Riva, editors, *The Hidden Structure of Interaction: From Neurons to Culture Patterns*. IOS Press, Amsterdam, The Netherlands, 2005.

[4] J. Brooks, M. Kerr, and J. Guttag. Using machine learning to draw inferences from pass location data in soccer. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, forthcoming 2016.

[5] D. Cervone, A. D'Amour, L. Bornn, and K. Goldsberry. Pointwise: Predicting points and valuing decisions in real time with nba optical tracking data. MIT Sloan Sports Analytics Conference, 2014.

[6] A. Coghlan. The secret of zone 14. *New Scientist*, 1999.

[7] C. Collet. The possession game? A comparative analysis of ball retention and team success in European and international football, 2007-2010. *Journal of Sports Sciences*, 31(2):123–136, 2013.

[8] A. C. Constantinou, N. E. Fenton, and M. Neil. pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems*, 36:322–339, Dec. 2012.

[9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[10] L. Gyarmati, H. Kwak, and P. Rodriguez. Searching for a unique style in soccer. *arXiv:1409.0308 [physics]*, Sept. 2014. arXiv: 1409.0308.

[11] S. S. Intille and A. F. Bobick. A framework for recognizing multi-agent action from visual evidence. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, AAAI '99/IAAI '99, pages 518–525, Menlo Park, CA, USA, 1999. American Association for Artificial Intelligence.

[12] M. Kates. Player motion analysis: Automatically classifying nba plays. Master's thesis, Massachusetts Institute Of Technology, September 2014.

[13] M. Kerr. Applying machine learning to event data in soccer. Master's thesis, Massachusetts Institute Of Technology, June 2015.

[14] P. Lucey, A. Bialkowski, P. Carr, E. Foote, and I. Matthews. Characterizing multi-agent team behavior from partial team tracings: Evidence from the english premier league. In *AAAI*, 2012.

[15] B. Macdonald. An expected goals model for evaluating nhl teams and players. MIT Sloan Sports Analytics Conference, 2012.

[16] R. P. Moreno, P. G. Metnitz, E. Almeida, B. Jordan, P. Bauer, R. A. Campos, G. Iapichino, D. Edbrooke, M. Capuzzo, J.-R. Le Gall, et al. SAPS 3 - From evaluation of the patient to evaluation of the intensive care unit. part 2: Development of a prognostic model for hospital mortality at icu admission. *Intensive care medicine*, 31(10):1345–1355, 2005.

[17] OptaPro. http://www.optasportspro.com/.

[18] M. Perše, M. Kristan, S. Kovačič, G. Vučkovič, and J. Perš. A trajectory-based analysis of coordinated team activity in a basketball game. *Computer Vision and Image Understanding*, 113(5):612–621, 2009.

[19] B. R. And the silver goes to... *The Economist*, Sept. 2011.

[20] A. Redwood-Brown. Passing patterns before and after goal scoring in FA premier league soccer. *International Journal of Performance Analysis in Sport*, 8(3):172–182, Nov. 2008.

[21] C. Reep and B. Benjamin. Skill and chance in association football. *Journal of the Royal Statistical Society. Series A (General)*, 131(4):581–585, Jan. 1968.

[22] SlamTracker. http://www.wimbledon.com/en_GB/slamtracker/.

[23] Soccer Analytics — Presented by Prozone | MIT Sloan Sports Analytics Conference. http://www.sloansportsconference.com/?p=9740.