Rating Soccer Defenders

Jason van der Merwe
jasonvdm@stanford.edu

Jack Craddock
jwcrad@stanford.edu

Bridge Eimon
beimon@stanford.edu

MCS100 Stanford University

December 4, 2014

Abstract

The project is motivated by a deficiency in the status quo of soccer statistics. Statistical measures such as goals, assists, passing percentage are effective and intuitive measures of offensive performance, but what stat column reveals the efficacy of a soccer defender? As it stands today, there is no single statistic widely accepted as a comprehensive measure of a defenders performance. Our desire was to discover which quantitative factors best represent the performance and value of a defender. In this paper we apply various regression and machine learning techniques to discover insights and determine the best indicators of a defenders performance, ultimately using bayesian ridge regression.

1. Data Retrieval and Dependent Variables

We found a data set online for the 2011/2012 Premier League season. This data, published by a sports statistics gathering company, Opta, included match-by-match statistics for every player in the Premier League. The stats recorded included goals, assists, touches, successful long passes, unsuccessful duels and more. We wanted to look at just defenders, so we wrote a script that filtered the data to represent only defenders that started for their team. In soccer, defenders are rarely substituted so starters remained on the pitch. This gave us a total of 80 defenders. We looked solely at the statistics that represented pure defender ability, for example clearances, and not goals. While a defender who can score goals is impressive, the primary role of a defender is to play defense and prevent goals from being scored against his team. In this way, we could identify the defensive traits that were most important.

Our last step was to decide upon a dependent variable (the y label in machine learning). We needed a ground truth for each player. We first turned to player salary with the belief that team managers hold considerable wisdom and intuition and that they (and the market) have a respectable idea of how valuable and efficient defenders are. Salary caps are not an issue in the Premier League like they are in many American sports, so we assumed little to no intervention in the market. Additionally, regression would account for the fact that some teams have much larger budgets than others. We found player salaries on various websites and news articles on the web.

We wanted some other dependent values to compare to, so we also gathered the player rankings from FIFA 13, the video game. While these ratings are based on EA Sports own model, they are generally regarded as very accurate (and are updated to reflect the players performance in real life).

The last dependent variable we wanted to test was goals conceded. If an attackers responsibility is to score goals, then a defenders is to prevent goals. Goals conceded was a good measure for this. The data set we used had enough data for us to calculate the goals conceded by each defender.

2. Statistical Models

Linear Regression: We began by using linear regression to model the data set of our defenders. By creating a best fit line, we were able to construct a baseline model. This initial model, however, performed very poorly as the dimensionality of our dataset was too large and our sample size was too small. Because of this, we quickly moved on to our next iteration.

Principal Component Analysis: The next step we tried was to apply a Principal Component Analysis algorithm on our feature matrix. Principal Component Analysis is a dimensionality reduction technique, which we hoped would allow our model to perform better. While Principal Component Analysis certainly provided a boost to our performance, the learned weight vector no longer corresponded to our features, and so we had no way of knowing which of our original features were most important in ranking defenders. We therefore moved to our next model.

Support Vector Regression: Our second attempt was to use a Support Vector Regression model. After the Linear Regression failed, we had hoped that by using Support Vector Regression, we would eliminate outliers and therefore be able to deal with a more uniform model, which would mitigate many of the effects of our high dimensional feature space. Upon implementing our Support Vector Regression, however, we found that only by using a nonlinear kernel could we achieve passable performance. Unfortunately, when using a nonlinear kernel, the weight vector no longer corresponds to the original feature space, so we weren't able to identify the most important statistical factors in defender efficacy and so we had to discard the Support Vector Regression.

Bayesian Ridge Regression: Our final model was a Bayesian Ridge Regression model. Upon examining the weights of our earlier models, we realized that the weights had huge variability. In order to solve this problem, we used ridge regression to constrain the weights. Furthermore, we realized that many of the features in our dataset were unimportant, and only a few should really matter. Thus, we settled upon Bayesian Ridge Regression which imposes a Gaussian prior over the weight vector. This caused many of the weights to be forced into having little to no effect, effectively reducing the dimensionality of our feature space while at the same time giving us a weight vector perfectly corresponding to our features and also with easily highlighted values for the most important features.

3. Results

Predictions

We have results from three different runs of our algorithm corresponding with using Bayesian Ridge Regression using Salary, FIFA Score, and Goals Conceded respectively as our dependent variable.

Player Salary Dependent Variable: Using Salary as our dependent variable was by far the least successful. As seen in figure 1a in the appendix, the predictions do decently for average defender salary, but poorly when the defender is paid very little or a great deal of money. This is because in the premier league, there is no salary cap, so teams with lots of money can pay their players much more than lower level teams without much money. This in turn causes huge variation in player salary, which regression tends to overestimate low values while underestimating high values.

FIFA Score Dependent Variable: As seen in figure 1b, when using FIFA Score as the dependent variable, our model performs much better. In figure 1b, the scale has decreased, meaning that most predictions are now within 2 or 3 scores away from their true value. Additionally, the predicted values mirror the trends of the true values, rising and falling in synchronicity, even if not to the same exact values. This is evidence that the calculated FIFA Score is a good estimate of defenders ability, which makes it a very good model.

Goals Conceded Dependent Variable: Figure 1c shows the prediction results against the true values of goals scored against players. For the most part, using goals conceded as a dependent variable works out quite well as our predictions are even better than when using FIFA Score as the dependent variable. The glaring exception is in the middle of the chart where 8 players are significantly underestimated. These players belong to the Blackburn Rovers and the Bolton Wanderers, two teams who were moved up a league in 2011, and were relegated down to the lower league again at the end of the season. This is evidence of the vast skill difference between the worst teams in the premier league and the best teams in the premier league, and is the reason for the discrepancy in predicted vs. actual value for goals conceded.

Factor Analysis

Ultimately, our goal was to determine which factors had the largest impact in a players value and efficacy. Our learning algorithm learned by updating a weight vector that corresponded to each of the features we included in the data. After training our model, the weights with the highest positive values, and the weights with the greatest negative values, were the weights that had the largest impact. The visualization of these weights can be found in figures 2a, 2b and 2c in the appendix. The green circles represent positive weight factors, factors that increased a players efficacy rating. Grey circles represent negative factors that detract from a players value. The radius of the circle is proportional to the value of the feature coefficient.

Player Salary Dependent Variable (figure 2a): The first observation here is that recoveries and touches in the final third are the most positive factors in a good defender. This makes sense as recoveries reflect the desire and hustle of a player, as well as his ball skill, ability and decision making. Touches in the final third mean that defenders are holding possession so well that their team is spending a lot of time in their opponents third. On the other side, unsuccessful passes are an obvious detractor from player value. The two interesting insights in this visualization are that unsuccessful long balls are a positive factor while clearances are a negative factor. To explain this, we can look at the definition for each. A clearance is any ball quickly kicked out of player by a defender, to avoid any danger. These usually result in an inbounds throw by the other team - ie, a loss of possession, which is bad. An unsuccessful long ball is an alternative to the clearance: instead of simply kicking the ball out of play, a good defender tries to turn a dangerous situation into a possible fast break. If you watch the best defenses play, you will see defenders such as Vincent Kompany and John Terry do exactly this.

FIFA Score Dependent Variable (figure 2b): Overwhelmingly, recoveries are the most important factor in this model. The second positive factor is duels won, which also makes sense. Additionally, we see club teams such as Liverpool and Manchester City listed as positive factors, while Norwich and QPR are negative factors. What this means is that as a team, Liverpool and Manchester City have stronger defenses than average. This is a result of player development, recruitment and other factors at these clubs. On the flip side, Norwich and QPR have weaker defenses. These are correlation factors.

Goals Conceded Dependent Variable (figure 2c): These factors are similar to the player salary model factors. Interestingly, passes backward are a large negative factor. The takeaway here is that this is a factor that any viewer can spot while watching soccer on television - a good defender will rarely pass backwards unless it is absolutely safe to! This model shared many similar factors as the other models, so we can conclude that these shared factors, especially recoveries, final third passes and touches, unsuccessful passes, and clearances are strong indicators of a defenders level of quality.

4. Conclusion

While there were differences between the three models based on different dependent variables, there were enough similarities in the most important features to suggest that all three dependent variables are highly correlated. Additionally, our algorithm has shown that it is possible to accurately predict the rankings of defenders. This has application not only to fans who are struggling over which defender they need for their fantasy team, but also to managers who are looking to increase the value of their team by snatching up defenders with talent who are underappreciated. In a game where strikers with their flashy moves and keepers with their heroic saves steal the spotlight from the defenders, it is time to bring the defenders the attention they deserve.

5. References

1) ‘Premier League Player Salaries (Club by Club).’ TSM PLUG RSS. N.p., n.d. Web. 04 Dec. 2014.

2) Opta. ‘UK Premier League Match-by-Match 2011/2012 - XLS by Opta - the Datahub.’ UK Premier League Match-by-Match 2011/2012 - XLS by Opta - the Datahub. N.p., n.d. Web. 04 Dec. 2014.

Appendix

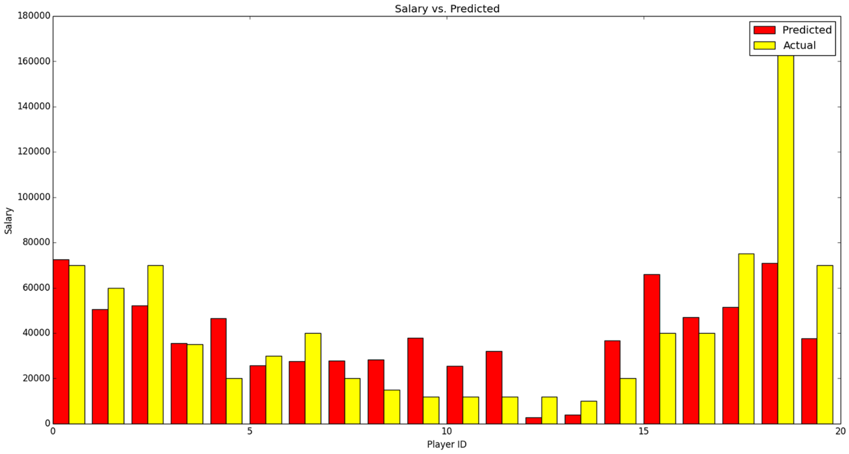

Figure 1a

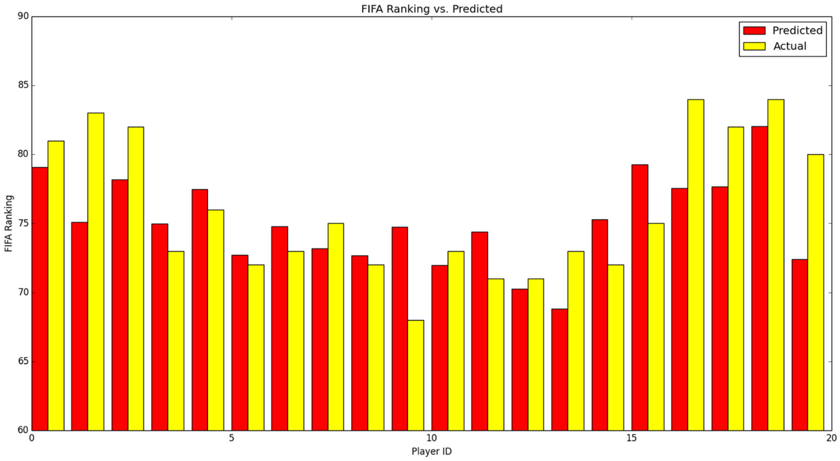

Figure 1b

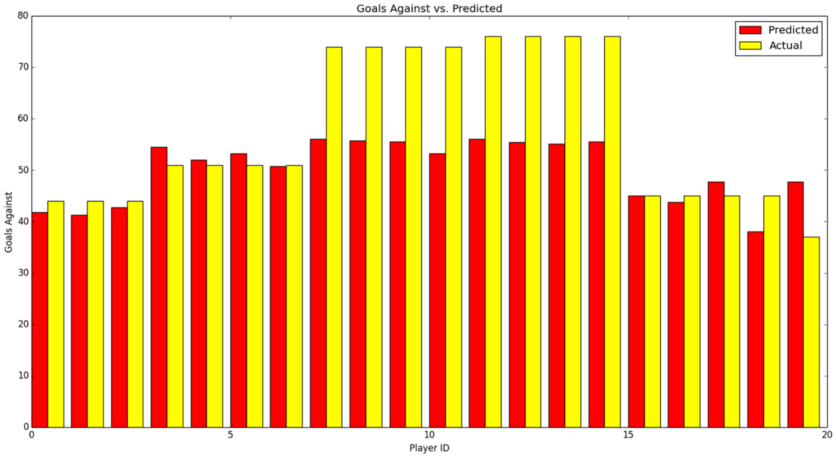

Figure 1c

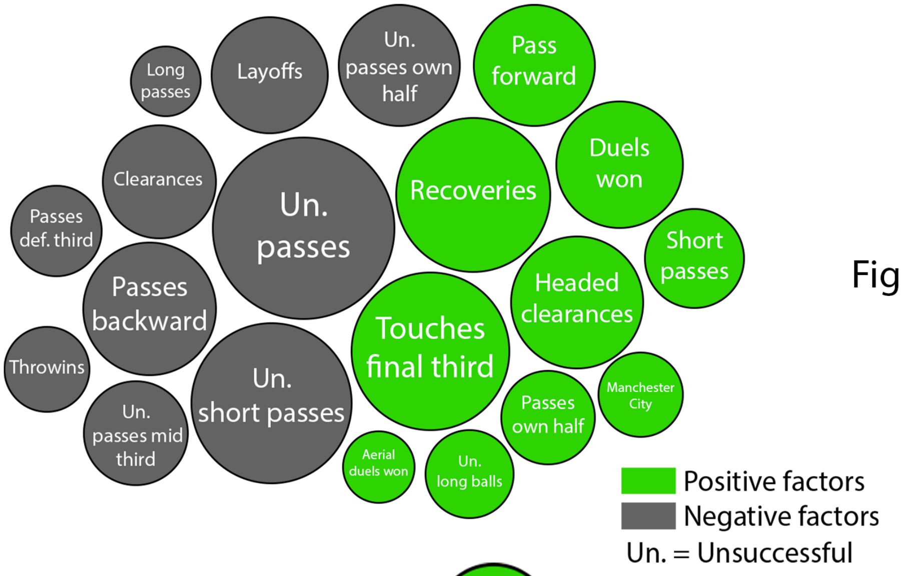

Figure 2a

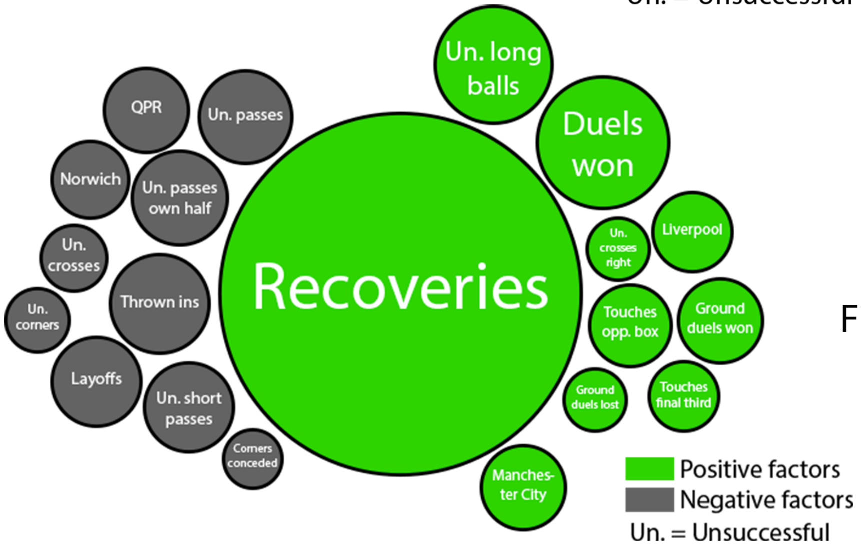

Figure 2b

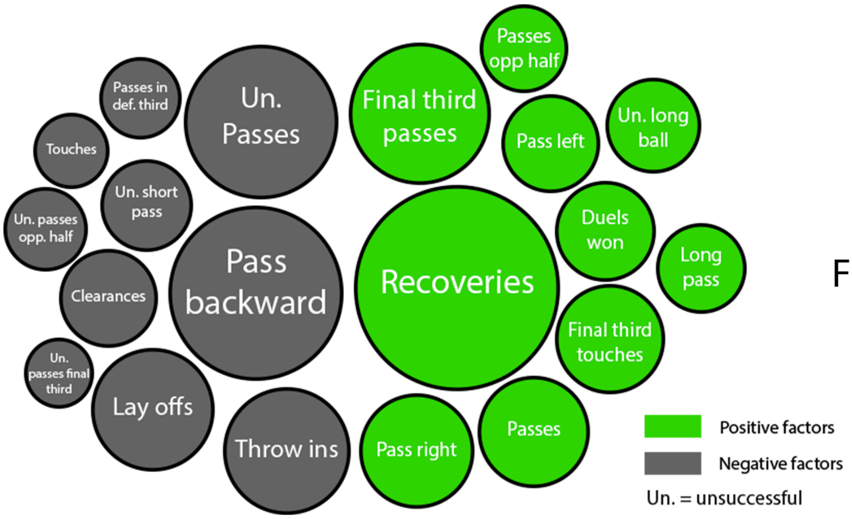

Figure 2c