Erasmus University Rotterdam

Erasmus School of Economics

Bachelor Thesis Econometrics and Operations Research

Differences in the top five football leagues analyzed by Shapley values

Name student: Pieter Blank

Student ID number: 484699

June 26, 2022

Supervisor: prof. dr. PHBF Franses

Second assessor: prof. dr. ir. JC van Ours

Abstract

The top five European football leagues are often thought of as having different playing styles. From the direct style of play in England to the possession-based style in Spain, each league represents a certain way of playing the game. In this paper, data from the top five European football leagues as given by Sports-reference is analyzed to analyze whether those differences are existing and if so, what the important variables are that lead to success in those individual leagues. This data is analyzed by means of standard linear regressions and Poisson regressions. Success is looked at by analyzing two dependent variables; the number of goals scored and the number of shots on target. The former is analyzed by the Poisson regression, the latter by the standard linear regression. The important variables are determined by looking at Shapley values. The main findings are that the English Premier League seems to have a more direct style of play due to the high importance of long balls. Besides that, different leagues seem to be comparable in the factors that are important for success.

Contents

1 Introduction

1.1 Research problem

In this research, I will investigate the differences in the important factors that affect the number of goals scored in football. This will be researched by means of Shapley values. The main research question is as follows:

'Are there differences in the importance of variables that lead to goals in football in the top 5 European leagues.'

Multiple sub-questions will be investigated that could help in finding an answer to the research question. The sub-questions are as follows:

1. What variables lead to goals scored and attempted shots on target in the top 5 European leagues combined?

2. What leads to scored goals and attempted shots on target in football over the different leagues individually?

3. Have the driving factors of scored goals and attempted shots on target changed over the years per league?

Data that will be analyzed is data over the seasons starting in 2017 up until the season that started in 2020 for the top 5 European competitions.

1.2 Motivation

Football, also known as soccer, is the most popular sport in the world (Giulianotti (2012)). It brings pleasure to people with all kinds of backgrounds and unites people all over the world. As with everything, differences in cultural, social and historical aspects can lead to variation in the way football is played in different countries, as stated in Yi et al. (2019). This is visible not only on the small, local level but also on the highest stage.

The traditional top 5 football leagues of Europe are those of England, Germany, Spain, Italy and France (Littlewood et al. (2011)). Different football leagues are known to have different playing styles. The English Premier League for example is known for its direct style of play leading to more importance on physical aspects whereas the Spanish La Liga is characterised by more focus on possession and thus more importance on technical abilities. The Italian Serie A is known for its focus on defending and the German Bundesliga and French Ligue 1 are combinations of those leagues(Crolley et al. (2000)).

Different styles of play lead to different qualities demanded of players. A player that is very physical and thus could do well in the English Premier League might not be successful in the Spanish La Liga due to the differences in those leagues. This research can give insight to those differences and can thus be useful for both football players and clubs. Players can use it to their advantage to see what league might fit their qualities; a technical player could prefer a certain league over another. The same applies to clubs. When clubs look for new players, they want someone that fits their needs. Knowing what is important and asked for in a certain country can thus help in finding a suitable player of which the odds are high that that player will perform well.

1.3 Literature

Data analysis is becoming more and more important in sports. This is the same for football. Most professional clubs have at least one data analyst in their team and their role becomes increasingly more important. This increase in usage of data analysis goes hand in hand with an increase in research in this field.

First of all, a lot of research is done on how leagues differ from one another. Yi et al. (2019) does so by looking at how different variables such as shots, long balls and offside differ over different leagues. It does so by looking at statistics such as the mean. Another way of looking at how leagues may differ is as done in Crolley et al. (2000). Here, the way the media portrays the leagues is looked at and analyzed to come up with different playing styles.

Not only research on the league or team level is done, but also individual qualities and characteristics are looked into. Di Salvo et al. (2013) looks at how motions of players of the two highest English leagues differ. Another research that looks at physical characteristics is as in Gardasevic & Bjelica (2020) where body composition is analyzed for different football clubs. This could say something about their playing style; teams with more physical players will most likely play more physically than teams with less physical players.

Since goals are often a good indicator of a team's performance, a lot of research is done on expected goals. Rathke (2017) is an example of how expected goals can be used in analyzing performance.

Shapley values are researched often in terms of game theory. Winter (2002) and Hart (1989) give explanation on this method applied to game theory. Research on Shapley values used to analyze the importance of various variables is scarce. Dong et al. (2020) and Yu et al. (2014) are examples of how Shapley value decomposition can be used in real-world problems.

As previously described, a lot of research has been done on football and, to a lesser extent, on

Shapley values. However, the combination of football research and Shapley values is new. This research could thus lead to renewing insights into how certain characteristics are of influence in different leagues.

1.4 Methods summary

Multiple methods are used in this research to answer the research questions. First of all, a simple linear model will be looked into and analyzed to see whether it suits the data. However, as the literature suggests, the number of goals scored in football is often analyzed by means of Poisson regression, as for example done in Karlis & Ntzoufras (2000). After performing the regressions for a variety of datasets, the main statistic that is of importance is the R^2 of those regressions. These values will namely be used in calculating Shapley values, done in STATA by use of the shapley2 package (Juarez (2012)). The Shapley values will then be analyzed to see what the driving factors in the dependent variables are and they will thus be used to answer the research questions.

2 Data

The data that is used in this research is as collected by *Sports reference: Sports stats, fast, easy, and up-to-date* (n.d.). This website contains data for all kinds of sports over many years. Not only does it contain data for a lot of sports, but this data is also very extensive. The used data comes from five leagues; the English Premier League, German Bundesliga, Spanish La Liga, French Ligue 1 and the Italian Serie A. Seasons starting with the season 2017-2018 until 2020-2021 were analyzed. This thus leads to four seasons being analyzed over five different leagues. This data is used in a way that the numbers are displayed as per season data, in which a season usually is either 34 or 38 games. An exception is the year 2019-2020. Due to the outbreak of the covid-19 pandemic, not all leagues finished all their games this year meaning that a couple of games were missing. The number of teams for the Premier League, Ligue 1, La Liga and the Serie A is 20. In the German Bundesliga, there are 18 teams. These numbers multiplied by the number of seasons, which is four, gives us the number of observations that will be used.

The dataset has a large variety of variables that can be used. For football, these variables have a wide range from individual to team statistics. These variables can go into very much detail, such as the number of loose balls a team has recovered or the average length of goal kicks taken by the goalkeeper. Not all of the variables can and will be used in this research. Therefore, an outline of the used variables is given below and this selection is mainly based on pre-existing football knowledge.

Two variables are used as dependent variables in this research; the number of shots on target a team has made and the total number of goals for each team. These variables are used because they give an indication of a team's performance. A higher number of goals often leads to a higher win percentage and as the famous Johan Cruijff once said, "you have got to shoot, otherwise you can't score", meaning the shots on goal will also play a big role in the success of a team. When the total number of goals is used as the dependent variable, shots on target is used as independent variable as well. Eight other variables are used to explain the dependent variables.

Firstly, some defending stats are used, namely, the total number of tackles and the total number of times pressure is applied to the opposing team who has the ball in possession. The next variables are more offensive, namely the distance at which shots are taken and the number of attempted dribbles. Statistics that describe possession and ball movement are given by the number of touches on the ball and the number of short, medium and long passes. The distance of passes will be used since, as described in the introduction, different leagues are thought to have different playing styles. These differences can be for example a more direct style of play with more long balls as opposed to a more possession-based style of play where short passes and touches on the ball are of higher importance.

All variables and their abbreviations that will be used are as given in table 11 in the appendix. A summary of those used variables is as given in table 1. In that table, the mean is given with the standard deviation between brackets. For example, the average goals per season per team in the Bundesliga is 51 with a standard deviation of 17. It is important to note that not all these variables can be compared directly, mainly because of the fact that the Ligue 1 has fewer games per season which, if comparisons were made, would lead to problems. The table is given to provide some basic information on how the data looks.

Table 1: Mean and standard deviations of the variables

	PL	**Bundesliga**	**La Liga**	**Ligue 1**	**Serie A**
Goals	50(18)	51(17)	47(15)	46(18)	53(16)
SoT	155(40)	151(33)	146(34)	142(36)	163(39)
Dist	17(0.80)	17(0.92)	18(0.91)	18(1.02)	18(0.99)
CmpS	6089(1711)	4999(1409)	5663(1867)	5480(1509)	6076(1524)
CmpM	6281(2008)	5948(1391)	5902(1698)	5920(1353)	6598(1303)
CmpL	2245(435)	2201(356)	2209(362)	2100(352)	2279(336)
TklT	674(74)	596(50)	610(56)	640(86)	633(53)
Press	5822(545)	5451(447)	5744(576)	5380(789)	5785(488)
Touches	23470(3779)	21102(2844)	22147(3445)	21431(3418)	23233(2759)
Att	634(95)	567(110)	618(107)	631(114)	621(95)

As mentioned in the introduction, literature shows that the number of goals scored is usually modeled by means of a Poisson model. However, in this research, there have still been tests done to see whether or not a Poisson model actually suits the data. Not using a suitable model can namely lead to undesirable outcomes and results. Therefore, two tests for model goodness-of-fit have been done. The deviance goodness-of-fit and the Pearson goodness-of-fit, as both described in Pulkstenis & Robinson (2004), give similar outcomes. In table 2 the p-values of those two tests for the entire sample per league, 2017-2018 until 2020-2021, is given. The p-values for all individuals leagues and years is as given in table 12 in the appendix. These values can be interpreted in a way that if the value is equal to or higher than 0.05, the null hypothesis cannot be rejected at a 5% significance level. Since the null hypotheses, which is that of correct specification of the model, cannot be rejected in any sample, it seems that the Poisson model adequately represents the data.

Table 2: Goodness-of-fit tests Poisson model on goals p-values

	Premier League	**Bundesliga**	**La Liga**	**Ligue 1**	**Serie A**
Deviance	0.30	0.96	0.27	0.08	0.71
Pearson	0.33	0.96	0.31	0.08	0.71

Doing those same tests for the dependent variable shots on target gives significant evidence that a Poisson model does not adequately describe the data in this case. The results of the

Deviance and Pearson goodness-of-fit tests can be seen in table 3. It can be seen that in every case the null hypothesis of correct model specification can be rejected. This implies that there is significant evidence that the Poisson model does not adequately describe the data when shots on target is the dependent variable.

Table 3: Goodness-of-fit tests Poisson model on shots on target p-values

	Premier League	**Bundesliga**	**La Liga**	**Ligue 1**	**Serie A**
Pearson gof	0.00	0.00	0.00	0.00	0.00
Deviance gof	0.00	0.00	0.00	0.00	0.00

Therefore, a linear regression model is tried and tested for. In this case, tests for normality and tests for heteroskedasticity are performed. Normality is tested for by a skewness-kurtosis test as in Bai & Ng (2005). Heteroskedasticity is tested for by means of the Breusch-Pagan test as in Breusch & Pagan (1979). The results are as given in table 4. Here, a value equal to or higher than 0.05 can be interpreted in a way that the null hypothesis, that of normality of the error terms in case of the skewness-kurtosis test and homoskedasticity in the case of the Breusch-Pagan test, cannot be rejected at a 5% significance level. For the data to be adequately represented by a linear model, both null-hypothesis should not be rejected. In table 4 it is seen that this is not always the case. However, it still seems to generally hold. When looking at the individual leagues per season, none of the null-hypothesis can be rejected, as shown in the appendix table 13. Therefore a linear model is used to describe the dependent variable shots on target.

Table 4: P-values for tests for linearity

	Premier League	**Bundesliga**	**La Liga**	**Ligue 1**	**Serie A**
Normality test	0.51	0.68	0.00	0.51	0.18
Heteroskedasticity test	0.45	0.94	0.00	0.00	0.93

3 Methodology

In this research, the following research question is investigated:

'Are there differences in the importance of variables that lead to goals in football in the top 5 European leagues.'

Multiple sub-questions will be investigated that could help in finding an answer to the research question. These sub-questions are as previously given in the introduction.

The first sub-question will help in getting a general idea of what variables have an important influence on the dependent variables by looking at the full dataset consisting of four seasons and five different leagues. The second sub-question will then look at the leagues individually and tell us more about potential differences. The last sub-question looks at each competition individually to see whether there have been any changes within those countries. All this combined will lead to an answer for the central research question as stated above.

To answer those research questions, multiple methods will be used. Both a linear regression model and a Poisson regression model will be used to analyze the data. The linear model will be used for the analysis of the dependent variable shots on target. Poisson regression is used for the dependent variable goals scored. Since standard t-values based on coefficients do not give any insight into how important certain variables are, Shapley values will then be used to give an insight into the relative importance of individual variables in the explanation of the variance. An explanation of the differences in t-values that show significance of the coefficient values and Shapley values will also be given.

3.1 Linear model

Due to the nature of the shots on target data, linear regression as described in Weisberg (2005) will be used for this variable. Reasons for this are as given in the data section.

A simple multiple regression model consists of a dependent variable that is explained by multiple independent or explanatory variables. The regression looks as follows, with y_i as the dependent variable and x_i as the vector of independent variables and an intercept α:

$$y_i = \alpha + x_i'\beta + \varepsilon_i$$

In the previous equation, the error terms ε_i are assumed to be normally distributed. The coefficients β will be estimated by means of ordinary least squares as also described in Weisberg (2005).

In this research, this linear model will look as follows.

$$SoT = \alpha + \beta_1 Dist + \beta_2 CmpS + \beta_3 CmpM + \beta_4 CmpL + \beta_5 TklT + \beta_6 Press + \beta_7 TouchesT + \beta_8 Att + \epsilon \tag{1}$$

3.1.1 Validations linear model

There are two assumptions that have been tested with the use of linear regression. The first assumption is that of the normality of the error terms. To test whether or not this assumption

holds, a skewness-kurtosis test as described in Bai & Ng (2005). This test has the null hypothesis of normality of the error terms and thus, if this null hypothesis cannot be rejected, the residuals show normal distribution.

The second test is a test for heteroskedasticity. This has been tested using the Breusch-Pagan test as described in Breusch & Pagan (1979). This test has a null hypothesis of homoskedasticity and thus if this cannot be rejected, there is no significant sign of heteroskedasticity.

3.2 Poisson model

The Poisson distribution is a non-negative discrete probability function and, as mentioned in the introduction, is often used for the number of goals scored in football. As supported by tests, mentioned in the data section, it adequately describes the data used in this research. A Poisson regression model can be described by the following two equations. Here, y_i is the dependent variable and takes values 0,1,2,... and x_i is the set of explanatory variables. A more extensive explanation can be found in Coxe et al. (2009).

$$f(y_i|x_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!},\ y_i = 0,1,2,... \tag{2}$$

$$\lambda_i = e^{x_i^T \beta} \tag{3}$$

The coefficients of the variables are then estimated by means of maximum likelihood as given in Gourieroux et al. (1984). This leads to values of coefficients that can be interpreted in a slightly different way than how they are interpreted in linear models. A one-unit change in the explanatory variable leads to a change of the value in the coefficient in the log of the expected dependent variable, keeping everything else constant.

The estimated Poisson model in this research looks as follows.

$$log(Gls) = \alpha + \beta_1 SoT + \beta_2 Dist + \beta_3 CmpS + \beta_4 CmpM + \beta_5 CmpL + \beta_6 TklT + \beta_7 Press + \beta_8 TouchesT + \beta_9 Att + \epsilon \tag{4}$$

3.2.1 Validations Poisson model

As previously mentioned, Poisson models are often a good fit when count data with non-negative values are to be described. This is the case with the dependent variable goals scored. To see whether the Poisson model adequately describes the data, two tests as described in the data section were performed. The tests are goodness-of-fit tests; the deviance and the Pearson

goodness-of-fit test.

3.3 Shapley values

Shapley value decomposition is a relatively new field of research. The values indicate how much a certain independent variable explains the variation in the dependent variable. This can be seen as a measure of importance that can vary over different explanatory variables. A high Shapley value leads to much of the variation being explained by a certain variable. This means that this certain variable is of high importance in influencing the dependent variable. Shapley value decomposition is performed after regressing and it makes use of R^2 values. There are multiple steps in calculating the Shapley values.

The first step is calculating all individual R^2 values for individual regressions. In the case of three explanatory variables, seven regressions will be performed. Three with all explanatory variables individually, three with all possible combinations of two explanatory variables and one with all variables.

The next step is to calculate the contributions of individual explanatory variables. This is done by looking at differences in R^2 values with and without a certain variable and weighing those differences. The weights are dependent on the number of variables and are given by a so-called Pascal triangle.

The individual parts of the Shapley value are calculated by subtracting two R^2 values. The first value is that of the regression with a certain number of regressors included, including the one in which contribution is calculated. The second value is that of the regressors excluding the one whose contribution is calculated of. The intuition behind this is that in this way, the extra explanation of the variance that is due to a certain regressor is calculated. Calculating the differences and weighing them before adding them all together is what gives the Shapley value.

Before a general formula for the Shapley values is given to support the previous, more intuitive explanation, some definitions have to be introduced. T is the set of all individual regressors consisting of x_1 to x_k. Q_i is the set of all combinations of the regressors excluding x_i. S_i is the set of all possible combinations of regressors, including the regressor x_i. Using those sets of combinations of regressors, the following notation for the calculation of the Shapley values can be introduced. SV_i is the Shapley value for regressor x_i. The value in the denominator is the weighing factor.

$$SV_i = \sum_{j \in N_k} \frac{R^2(S_j) - R^2(Q_j)}{\binom{k_j - 1}{s_j - 1} k} \tag{5}$$

An example of such a calculation is as follows. Here, X_1, X_2 and X_3 are the independent variables. R_i^2 is the R^2 value for a regression with regressors i. For example, R_1^2 is the value for

the regression with only X_1, R^2_{12} has both X_1 and X_2 included and R^2_{123} has all three regressors in the regression. The contribution of X_1 is then calculated in the following way.

$$Contribution_1 = \frac{1}{3}R^2_1 + \frac{1}{6}(R^2_{12} - R^2_2) + \frac{1}{6}(R^2_{13} - R^2_3) + \frac{1}{3}(R^2_{123} - R^2_{23}) \tag{6}$$

After all individual contributions are calculated, Shapley values can be found. This is simply the individual contribution divided by the total contribution of all variables added. For a regression with n variables, this looks as follows.

$$SV_i = \frac{Contribution_i}{\sum_{i=1}^{n} Contribution_i} \tag{7}$$

3.4 Poisson R^2

Standard R^2 values as calculated in Miles (2005) are what is calculated in standard linear regressions. A limitation to Poisson regression is that standard R^2 values are not correct. This is due to the fact that a Poisson model is non-linear, and thus it does not meet the requirements that are desired for R^2 values. Therefore, the same research suggests a number of pseudo R^2. The R^2 as used by the statistical package STATA and thus this research is one that meets most criteria. It is calculated by means of the likelihood ratio index. This is calculated as follows and it compares the log-likelihood of the fitted and intercept-only models.

$$R^2_{LRI} = 1 - \frac{l(\hat{u})}{l(\bar{y})} \tag{8}$$

The only limitation is that it has an upper limit of less than unity (Cameron & Windmeijer (1996)). In comparing different R^2 and calculating the Shapley values, this will not be of much influence.

3.5 Shapley vs t-values

In determining what factors are important in influencing certain dependent variables, t-values are often used. They show whether or not an explanatory variable is significantly different from zero. This is clearly different from Shapley values because Shapley values do not worry about significance but simply tell something about the percentage of variance that is explained. A shortcoming of t-values is that due to high collinearity, certain variables can be interpreted as insignificant. Shapley values do not have this problem.

4 Results

4.1 Full Dataset

At first, the full dataset including the five leagues over the full period from 2017 until 2021 is analyzed. These results will give an indication of the overall importance of variables within football. Both dependent variables, the number of goals and the shots on target, are analyzed. The former, the number of goals, is analyzed by means of a Poisson regression. For the latter, a linear regression is looked into. For both models, Shapley values are calculated after running the regressions. In table 5 the individual contribution of the variables calculated by means of Shapley values is given for this full dataset. The percentages show how much of the variance is explained by each independent variable. For example, the variable TouchesT, the total number of touches, explains 26.18 percent of the variance in the dependent variable shots on target. It is clear that for the dependent variable goals, the variable shots on target explains a big part of the variance. This is in agreement with intuition and with, as mentioned in the introduction, what Johan Cruijff said. Other variables that seem to be of high importance are the total number of touches, the passes of middle length and the passes of short length. For the linear regression on the dependent variable shots on target, the total number of touches seems to explain a big part of the variance, together with the short and medium passes.

Table 5: Percentage of variance explained by variables

	Poisson regression (Goals)	**Linear regression(Shots on target)**
SoT	36.71%	-
Dist	7.48%	4.31%
CmpS	12.54%	22.69%
CmpM	14.72%	23.90%
CmpL	8.35%	13.44%
TklT	0.35%	0.59%
Press	1.01%	0.69%
TouchesT	14.17%	26.18%
Att	4.68%	8.21%

4.2 Full period individual leagues, shots on target

Next, the leagues are analyzed individually over the full period from 2017 until 2021. In table 6 the results for the Shapley value calculations are given for the independent variable shots on

target. This table is based on a linear regression. From this table, there seem to be a lot of similarities between the different leagues. First of all, the number of touches explains the most variance compared to the other variables in all leagues. Therefore, it seems that this variable is an important factor in the number of shots a team makes. At the same time, the number of short and medium passes are in all leagues the second and third most important variables in explaining the variance. Both the number of touches and those variables are variables related to possession. It can thus be concluded that possession is an important factor in the number of shots on target a team makes. Another result that can be seen is how the influence of the number of completed long balls, CmpL is highest in the Premier League. This can be supported by the idea of the direct style of play in England, as mentioned in the introduction. What is also something that appears is that in the Bundesliga and the Ligue 1, the distance at which shots are taken is of higher importance in explaining the variance than in the other leagues.

Table 6: Percentage of variance explained by variables for leagues individually

Full period(shots on target)	**Premier League**	**Bundesliga**	**La Liga**	**Ligue 1**	**Serie A**
Dist	0.21%	13.04%	2.65%	12.56%	1.82%
CmpS	21.90%	20.03%	24.34%	19.45%	21.37%
CmpM	23.43%	19.46%	18.48%	18.43%	27.15%
CmpL	15.80%	11.43%	9.30%	11.53%	10.78%
TklT	1.10%	1.40%	0.70%	3.53%	0.26%
Press	5.07%	4.24%	5.25%	0.79%	1.61%
TouchesT	25.78%	23.20%	24.51%	26.96%	27.77%
Att	6.71%	7.20%	14.77%	6.76%	9.25%

4.3 Full period individual leagues, goals

In table 7 the percentages of the variance of the dependent variable goals explained by the individual variables is displayed. A clear result is that in all leagues, shots on target explains a big part of the variance in the number of goals scored. The number of short and medium passes and the number of touches on the ball are other variables that explain a large amount of the variance. This is similar for all leagues and shows, just as in the previous part, that the different leagues might be quite comparable. Again, the part of the variance explained by the independent variable long balls completed is highest for the Premier League. This supports the idea of a more direct style of play once more. Here, as in the previous part, the distance at which shots are taken seems to be of higher importance in the Bundesliga and the Ligue 1.

Table 7: Percentage of variance explained by variables for leagues individually

Full period(Goals)	**Premier League**	**Bundesliga**	**La Liga**	**Ligue 1**	**Serie A**
SoT	33.27%	32.86%	32.83%	33.71%	34.27%
Dist	2.24%	12.69%	5.78%	12.12%	4.72%
CmpS	13.50%	13.67%	14.97%	13.26%	10.85%
CmpM	14.48%	11.97%	13.60%	13.88%	17.75%
CmpL	10.51%	6.54%	6.89%	6.34%	7.91%
TklT	1.82%	0.99%	1.32%	1.28%	0.62%
Press	4.08%	2.92%	1.35%	0.31%	4.21%
TouchesT	15.34%	13.80%	14.93%	14.38%	14.49%
Att	4.77%	4.56%	8.34%	4.71%	5.16%

To illustrate that there are differences between Shapley values and the outcomes of normal regressions, the regression results for the individual leagues over all seasons have been estimated. In this case, the dependent variable the number of goals scored is analyzed by means of a Poisson regression. The values in table 8 are the coefficient values for the variables over the different leagues for all seasons combined. It can be seen that in these regressions, the possession statistics (touches, short, medium & long passes) often are not significantly different from zero. This is an interesting result because when the Shapley values for the same dataset were analyzed, the number of touches and the short& medium passes often explained an above-average part of the variance. This shows that Shapley values and regression outcomes can lead to different outcomes.

Table 8: Regression results (coefficient values) Poisson model for Goals. * means significance at 5%.

	PL	**Bundes**	**La Liga**	**Ligue 1**	**Serie A**
SoT	0.00724*	0.00670*	0.00607*	0.00649*	0.00567*
Dist	-0.06668*	-0.06567*	-0.04775*	-0.05221*	-0.04755*
CmpS	0.00002	0.00015*	0.00005	0.00005	0.00012*
CmpM	0.00005	0.00004	0.00008	0.00010*	0.00014*
CmpL	0.00007	0.00010	0.00004	-0.00001	0.00014
TklT	0.00028	0.00054	-0.00016	0.00072*	0.00091*
Press	0.00002	0.00006	-0.00006	0.00001	-0.00011*
TouchesT	0.00003	0.00007	-0.00006	-0.00003	-0.00013*
Att	0.00030	0.00021	0.00000	-0.00056*	0.00027

4.4 Differences over the years

When comparing the different Shapley values for subsequent years for the individual leagues with the goals as the dependent variable, it appears that only minor changes take place. The four highest values are as given in table 14 in the Appendix. In the Bundesliga, after the first season, the distance at which shots were taken becomes more important. In the Ligue 1 in the season 2017-2018 the distance at which shots were taken seemed to be of more importance than in the subsequent seasons. Another interesting result is how the variable attempted dribbles in the Serie A explains a large part of the variance in the season 2018-2019. This variable is only of importance in that specific season.

Except for other minor variations in the Shapley values, it seems that the important factors for the number of goals scored stays relatively consistent over time when looking at the different leagues.

In the case when the dependent variable is shots on target as given in 15, the differences are only minor too. An example is that in the Bundesliga in the season 2018-2019 the variable that indicates the distance at which shots were taken seems to be of relatively much influence. Another result is that in the season 2018-2019 in the Serie A, the variable attempted dribbles is the one that explains most of the variance whereas, in the other years, its influence is a lot less. Besides these minor differences, it is clear that in most years and leagues the most important factors are once again those related to possession; the number of touches and the number of short and medium passes.

4.5 Extra part of variance explained

In this section, there are combinations of different leagues analyzed. A dataset of a combination of two leagues is created with all standard variables. In addition to this, new columns of variables are added which have the value of zero for all observations for a specific league. The values for the other league are how they were before. This indicates the use of a dummy variable and is used to explain the extra variance explained by an added dummy variable. Regressions were performed for each dummy variable individually. This analysis is performed for the dependent variable goals by means of Poisson regression. The regressions look as follows. Here the dots represent the other variables as in 4 and SoTD is the dummy variable for shots on target that is filled with all zeros for the second league.

$$Goals = \beta_1 SoT + ... + \beta_9 Att + +\beta_{10} SoTD + \epsilon \tag{9}$$

In the name PLxBundes this would be the Bundesliga. The Premier League values are in

this case how they originally were. In table 9 and table 10 all the Shapley values are shown for each of those dummies. For example, in the combination of the Premier League (PL) and La Liga, the Shapley value of the dummy variable for the total touches(TouchesT) is 2.17%. It can be seen that none of these values exceed 5%. In fact, the highest value is 4.30%. This indicates again that the leagues barely differ from this point of view.

Table 9: Explanation variance by dummy variable

	PLxBundes	**PLxLaLiga**	**PLxLigue1**	**PLxSerieA**	**BundesxLaLiga**
SoT	2.04%	2.29%	2.67%	0.67%	4.30%
Dist	1.87%	0.30%	1.01%	0.74%	1.24%
CmpS	2.71%	1.95%	2.74%	0.92%	3.67%
CmpM	2.84%	2.30%	3.10%	1.14%	3.11%
CmpL	2.13%	1.15%	2.02%	0.69%	2.29%
TklT	1.90%	0.27%	1.00%	0.82%	1.43%
Press	1.88%	0.23%	0.94%	0.84%	1.38%
TouchesT	2.17%	1.14%	1.97%	0.64%	2.46%
Att	1.74%	0.61%	1.28%	0.57%	2.11%

Table 10: Explanation variance by dummy variable

	BundxLigue1	**BundxSerieA**	**LaLigaxLigue1**	**LaLigaxSerieA**	**Ligue1xSerieA**
SoT	3.32%	1.44%	1.17%	0.13%	0.24%
Dist	1.19%	0.82%	0.20%	1.09%	1.31%
CmpS	2.94%	1.09%	1.56%	0.42%	0.20%
CmpM	2.55%	0.82%	1.24%	0.43%	0.21%
CmpL	1.97%	0.67%	0.58%	0.59%	0.41%
TklT	1.37%	0.85%	0.23%	1.04%	0.83%
Press	1.35%	0.92%	0.21%	0.95%	0.92%
TouchesT	2.10%	0.73%	0.70%	0.47%	0.31%
Att	1.87%	0.63%	0.52%	0.48%	0.52%

5 Conclusion

In this research, I tried to answer the following research question:

'Are there differences in driving factors that lead to goals in football in the top 5 European

leagues.'

To answer this question, multiple analyses were performed, leading to many results.

First of all, it became clear that there were not many differences between the leagues individually for the different time periods. This means that it seems that the leagues do not change a lot in the important factors for success over the different time periods. This became clear because, besides minor exceptions, the variables explaining most of the variance were very similar over different years. One other result seemed to be of major interest. after analyzing different dependent variables using different models, it became clear that the variable that indicates the number of long balls was of more importance in the English Premier League when compared to the other countries. This shows that in England, a more direct style of play could lead to more success. Interesting is to note that when standard t-values that indicate significance show different results. Variables that seem to be of high importance when analyzing Shapley values can be found to not be significantly different from zero. This shows that Shapley values can give a new insight in what variables are of importance in the analysis of dependent variables.

In a linear model describing the dependent variable shots on target it became clear that possession statistics such as the number of touches and short passes were the variables that explained most of the variance and thus were of most importance. This was the same for all leagues with little variation in the percentages. One result that is found is as supported by the literature; the more direct style of play in the English Premier League. This is due to the Shapley values for the number of completed long balls. Besides that, only a handful of minor variations could be detected but the general trend seems to be that the different football leagues do not differ a lot in this aspect.

When looking at a Poisson regression modelling the number of goals, the results were similar. The number of shots on target explained most of the variance in all leagues. This variable was then in most cases followed mainly by possession statistics as previously described. Here again, the result of the more direct style of play in the English Premier League based on the high Shapley values for long balls is found. This seems to be a particularly interesting result because it is as described in previously done research.

To answer the research question, two main findings are of particular interest. The first one is that the direct style of play in the Premier League as expected from the literature has been supported by this research. The second finding is that it seems that the factors that lead to goals in the top five European football leagues seem to be comparable when analyzed by Shapley values.

Since previously done research and popular opinion in football suggests differences in various

leagues, it can be interesting to do more research on what exactly it is that makes those differences. Future research can thus for example look at physical aspects of the game of football; are there differences in the physicality of players, is there more running involved in a certain league, etc. It can also be interesting to see how and why significance of variables and Shapley values differ widely in showing what variables are of importance in explaining the dependent variable.

References

Bai, J., & Ng, S. (2005). Tests for skewness, kurtosis, and normality for time series data. *Journal of Business & Economic Statistics*, *23*(1), 49–60.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the econometric society*, 1287–1294.

Cameron, A. C., & Windmeijer, F. A. (1996). R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics*, *14*(2), 209–220.

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to poisson regression and its alternatives. *Journal of personality assessment*, *91*(2), 121–136.

Crolley, L., Hand, D., & Jeutter, R. (2000). Playing the identity card: Stereotypes in european football. *Soccer & Society*, *1*(2), 107–128.

Di Salvo, V., Pigozzi, F., Gonzalez-Haro, C., Laughlin, M., & De Witt, J. (2013). Match performance comparison in top english soccer leagues. *International journal of sports medicine*, *34*(06), 526–532.

Dong, F., Yu, B., Pan, Y., & Hua, Y. (2020). What contributes to the regional inequality of haze pollution in china? evidence from quantile regression and shapley value decomposition. *Environmental Science and Pollution Research*, *27*(14), 17093–17108.

Gardasevic, J., & Bjelica, D. (2020). Body composition differences between football players of the three top football clubs. *International Journal of Morphology*, *38*(1).

Giulianotti, R. (2012). Football. *The Wiley-Blackwell encyclopedia of globalization*.

Gourieroux, C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods: Applications to poisson models. *Econometrica: Journal of the Econometric Society*, 701–720.

Hart, S. (1989). Shapley value. In *Game theory* (pp. 210–216). Springer.

Juarez, F. C. (2012, November). *SHAPLEY2: Stata module to compute additive decomposition of estimation statistics by regressors or groups of regressors.* Statistical Software Components, Boston College Department of Economics. Retrieved from `https://ideas.repec.org/c/boc/bocode/s457543.html`

Karlis, D., & Ntzoufras, I. (2000). On modelling soccer data. *Student*, *3*(4), 229–244.

Littlewood, M., Mullen, C., & Richardson, D. (2011). Football labour migration: an examination of the player recruitment strategies of the 'big five'european football leagues 2004–5 to 2008–9. *Soccer & Society*, *12*(6), 788–805.

Miles, J. (2005). R-squared, adjusted r-squared. *Encyclopedia of statistics in behavioral science*.

Pulkstenis, E., & Robinson, T. J. (2004). Goodness-of-fit tests for ordinal response regression models. *Statistics in medicine*, *23*(6), 999–1014.

Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise*, *12*(2), 514–529.

Sports reference: Sports stats, fast, easy, and up-to-date. (n.d.). Retrieved from `https://www.sports-reference.com/`

Weisberg, S. (2005). *Applied linear regression* (Vol. 528). John Wiley & Sons.

Winter, E. (2002). The shapley value. *Handbook of game theory with economic applications*, *3*, 2025–2054.

Yi, Q., Groom, R., Dai, C., Liu, H., & Gómez Ruano, M. Á. (2019). Differences in technical performance of players from 'the big five'european football leagues in the uefa champions league. *Frontiers in psychology*, 2738.

Yu, S., Wei, Y.-M., & Wang, K. (2014). Provincial allocation of carbon emission reduction targets in china: An approach based on improved fuzzy cluster and shapley value decomposition. *Energy policy*, *66*, 630–644.

6 Appendix

Table 11: Variable abbreviations

Abbreviation	**Variable**
Gls	The number of goals scored
SoT	The number of shots on target in yards
Dist	The average distance of which a shot is taken in yards
CmpS	The number of completed short (5-15 yards) passes
CmpM	The number of completed medium (15-30 yards) passes
CmpL	The number of completed long (30+ yards) passes
TklT	The number of tackles made
Press	The number of times pressure was applied to the opposing team
Touches	The number of touches on the ball
Att	The number of times a dribble was attempted

Table 12: Deviance and Pearson test p-values

League	**Test**	**2017-2018**	**2018-2019**	**2019-2020**	**2020-2021**
PL	Deviance	0.78	0.22	0.36	0.10
	Pearson	0.77	0.24	0.37	0.10
Bundes	Deviance	0.81	0.53	0.97	0.77
	Pearson	0.82	0.52	0.97	0.77
La Liga	Deviance	0.05	0.93	0.16	0.68
	Pearson	0.04	0.93	0.17	0.68
Ligue 1	Deviance	0.22	0.11	0.79	0.82
	Pearson	0.23	0.12	0.79	0.82
Serie A	Deviance	0.06	0.82	0.74	0.89
	Pearson	0.06	0.83	0.73	0.89

Table 13: P-values tests for normality

League	**Test**	**2017-2018**	**2018-2019**	**2019-2020**	**2020-2021**
PL	Skewness	0.92	0.94	0.85	0.53
	Hetereogeneity	0.33	0.43	0.53	0.27
Bundes	Skewness	0.19	0.76	0.63	0.95
	Hetereogeneity	0.87	0.63	0.05	0.65
La Liga	Skewness	0.23	0.77	0.48	0.24
	Hetereogeneity	0.90	0.24	0.62	0.53
Ligue 1	Skewness	0.83	0.27	0.99	0.34
	Hetereogeneity	0.75	0.69	0.13	0.33
Serie A	Skewness	0.03	0.31	0.85	0.95
	Hetereogeneity	0.56	0.46	0.10	0.68

Table 14: Four highest Shapley values for all leagues and years for a Poisson model on goals

		2017-2018	**2018-2019**	**2019-2020**	**2020-2021**
Premier League	**1**	SoT(20.07%)	SoT(24.92%)	SoT(29.65%)	SoT(49.65%)
	2	TouchesT(16.03%)	CmpM(12.33%)	TouchesT(15.29%)	CmpS(13.46%)
	3	CmpM(15.12%)	TouchesT(11.78%)	CmpM(15.27%)	TouchesT(11.86%)
	4	CmpS(14.33)	CmpS(9.89%)	CmpS(13.05%)	CmpM(10.73%)
Bundesliga	**1**	SoT(29.34%)	SoT(27.16%)	SoT(26.96%)	SoT(32.34%)
	2	CmpS(15.59%)	Dist(15.82%)	Dist(14.63%)	Dist(13.85%)
	3	TouchesT(15.29%)	TouchesT(13.71%)	CmpS(14.22%)	TouchesT(13.76%)
	4	CmpM(14.63%)	CmpS(12.21%)	TouchesT(13.10%)	CmpM(13.25%)
La Liga	**1**	SoT(33.34%)	SoT(32.18%)	SoT(15.13%)	SoT(22.05%)
	2	CmpM(13.92%)	CmpS(15.74%)	TouchesT(14.35%)	CmpM(19.55%)
	3	TouchesT(13.37%)	TouchesT(13.19%)	CmpS(13.38%)	CmpS(17.87%)
	4	CmpS(13.20%)	Att(11.74%)	CmpM(11.74%)	TouchesT(17.68)
Ligue 1	**1**	SoT(22.93%)	SoT(30.08%)	SoT(21.33%)	SoT(40.23%)
	2	Dist(20.14%)	CmpS(18.74%)	TouchesT(18.42%)	TouchesT(11.50%)
	3	CmpM(13.53%)	TouchesT(14.64%)	CmpM(16.99%)	CmpS(10.01%)
	4	TouchesT(12.95%)	Dist(13.05%)	CmpS(15.71%)	CmpM(9.47%)
Serie A	**1**	SoT(24.09%)	Att(22.19%)	SoT(34.09%)	SoT(36.28%)
	2	CmpM(19.84%)	SoT(21.46%)	CmpM(14.99%)	TouchesT(17.16%)
	3	TouchesT(15.69%)	TouchesT(14.09%)	TouchesT(14.49%)	CmpM(16.94%)
	4	CmpL(12.08%)	CmpS(12.73%)	CmpS(11.36%)	CmpS(15.49%)

Table 15: Four highest Shapley values for all leagues and years for a linear model on Shots on target

Linear model		**2017-2018**	**2018-2019**	**2019-2020**	**2020-2021**
Premier League	**1**	TouchesT(22.28%)	TouchesT(20.56%)	TouchesT(24.12%)	TouchesT(24.72%)
	2	CmpM(22.26%)	CmpM(19.86%)	CmpM(21.28%)	CmpS(21.87%)
	3	CmpS(18.48%)	CmpS(18.66%)	CmpS(19.71%)	CmpM(21.33%)
	4	CmpL(16.05%)	CmpL(12.14%)	CmpL(16.19%)	CmpL(14.37%)
Bundesliga	**1**	TouchesT(22.53%)	TouchesT(17.97%)	Dist(20.39%)	TouchesT(24.96%)
	2	CmpS(21.77%)	Dist(16.79%)	TouchesT(19.55%)	CmpM(23.13%)
	3	CmpM(17.13%)	CmpS(16.30%)	CmpM(17.06%)	CmpS(20.31%)
	4	CmpL(9.24%)	CmpL(15.15%)	CmpS(16.84%)	CmpL(11.55%)
La Liga	**1**	CmpS(24.57%)	CmpS(27.09%)	TouchesT(22.09%)	CmpS(26.96%)
	2	TouchesT(20.19%)	TouchesT(23.33%)	CmpS(19.52%)	TouchesT(26.42%)
	3	CmpM(16.65%)	CmpM(16.88%)	CmpL(16.30%)	CmpM(20.45%)
	4	Att(13.92%)	Att(14.31%)	CmpM(15.89%)	Att(9.39%)
Ligue 1	**1**	TouchesT(19.58%)	Dist(22.03%)	CmpS(23.44%)	TouchesT(23.20%)
	2	CmpS(18.13%)	TouchesT(21.53%)	TouchesT(20.34%)	CmpS(19.78%)
	3	Dist(17.68%)	CmpM(17.55%)	Dist(19.42%)	CmpM(17.42%)
	4	CmpM(17.15%)	CmpS(17.54%)	CmpM(16.82%)	CmpL(12.71%)
Serie A	**1**	CmpM(28.47%)	Att(24.60%)	CmpM(26.37%)	TouchesT(27.20%)
	2	TouchesT(27.59%)	TouchesT(19.05%)	TouchesT(26.14%)	CmpS(25.74%)
	3	CmpS(19.25%)	CmpM(16.83%)	CmpS(20.31%)	CmpM(25.68%)
	4	CmpL(16.09%)	CmpS(14.22%)	CmpL(10.17%)	Att(9.08%)

7 Code explanation

Four STATA do-files are used in this research. A brief explanation is given below. Important is to note that before this code is run, the shapley2(Juarez (2012)) package should be installed.

7.1 SoT_Poisson

This file performs a Poisson regression on the dependent variable shots on target. After this regression is performed, the goodness-of-fit tests are performed followed by the calculations of the Shapley values.

7.2 Poi_gls

This file performs a Poisson regression on the dependent variable goals. After the regression, the goodness-of-fit tests are performed and the Shapley values are calculated.

7.3 lin_shap

In this file, a linear regression on the dependent variable shots on target is estimated. After the regression, the tests for linearity are performed. Lastly, Shapley values are calculated.

7.4 Poi_Dummy

In this file, multiple Poisson regressions are performed with dummy variables. After each regression, Shapley values are calculated and goodness-of-fit tests are performed.