



Statistical analysis of sports data

BICHET Camille
MONÉDIÈRES Emmeline

5GMM - Département Génie Mathématiques et Modélisation

Under the supervision of
Sébastien DÉJEAN
Philippe SAINT PIERRE
Javier LÓPEZ SÁNCHEZ

January 2020

Abstract

Sports analytics have been raising interest for years thanks to multiplication and automation of statistic methods. The aim of this project is to identify which actions during a football game really impact the number of points of a football team at the end of a championship. We studied data of 124 football clubs from 6 tournaments divided into 82 types of actions. We started by an exploratory analysis to visualize the distribution and interactions of features and redundancy between them. We thus removed variables related to goals, assists and key passes, directly linked to the outcome of a game. Then we used feature selection algorithms based on regression, such as LASSO, ElasticNet and BIC, to keep a reasonable number of variables and three machine learning methods, Random Forest, Support Vector Machine and eXtreme Gradient Boosting to obtain another idea of important features. By not considering features related to goals and assists we highlighted the role of shots accuracy, some shots zones, passes and especially short passes on the outcome of all tournaments. We also noticed that each tournament owns its characteristics. We could extend our study by using other statistic exploration methods, or considering more teams. We could also focus on a team during several years.

Keywords: Variable selection, Random Forest, SVM, XGBoost, PCA, clustering, ElasticNet, football

Acknowledgements

We would like to acknowledge Sébastien Déjean, Philippe Saint Pierre and Javier López Sánchez for their time and their help throughout this project.

We also acknowledge Paul Scanlan for his guidance during the writing of this paper.

Introduction

Can mathematics help your favorite football team to win? What if, by using statistics, you were able to identify performance indicators for a team? These are real questions and by using machine learning methods, it is possible to try to answer them. Major clubs already use some statistics to optimize their performances. Sports betting sites also use statistics to compute betting odds. Nevertheless very few papers are available on the subject.

When a team plays a game, it wins a number of points, according to the outcome of the meeting. The team receives three points for a win, one point for a draw, and zero for a defeat. The aim of a club is to maximize its number of points at the end of the season. The point of our project is to study what kind of actions, among **defensive, offensive and passing**, can lead to a high number of points at the end of a tournament. To do so, we created a data set based on data from the website *whoscored.com*. The variable to be explained is the number of points by the end of the tournament. It is a discrete quantitative variable.

We chose to study the number of points and not the number of goals because a high number of goals does not necessarily mean a high number of points. For example, a team can score ten goals in a game and win, and then lose the three following games, goalless. Its number of points will only be three while its number of goals will be ten. Another team, winning its four meetings, scoring only once by game, will have four goals and twelve points. The ranking at the end of a season is based on the number of points and not the number of goals, thus, the second team will have a better ranking.

We divided the study into three different parts. First we had an exploratory analysis to discover our data. For this analysis, we studied correlation between variables to start by a bi-dimensional analysis and we computed a Principal Component Analysis to project individuals and variables on planes with smaller dimensions. We also tried an unsupervised method to cluster teams and variables. Then we used variable selection methods based on regression, such as LASSO and ElasticNet, to keep a reasonable number of variables. Finally we implemented machine learning algorithms, Random Forest, Support Vector Machine and XGBoost to have another idea of the main variables, by using an importance score provided by *caret* package.

All mathematics methods presented in this report have been implemented using R.

1 Data presentation

1.1 The website: *WhoScored.com*

The website *WhoScored.com* is managed by a team of football analysts and software developers from London. It gathers football statistics, mainly from European divisions, with a lot of different variables. More than 500 tournaments, 15,000 teams and 250,000 players are represented. For some tournaments, detailed statistics are available.



When detailed statistics are available, we have around 80 variables for each tournament. These variables are the averages over a season. For example, the number of passes will not be the total number of passes during the season, but this number divided by the number of games.

1.2 The variables

Here we only study team statistics (not player or referee statistics). This data is split into three main categories: **defensive**, **passing** and **offensive**.

1.2.1 Defensive statistics

Defensive data refers to actions from the players when they are in a defensive strategy. They do not control the game anymore. They try to prevent the other team from scoring. There are 15 **defensive** data, divided into eight categories: tackles, interception, fouls, cards, off-sides, clearances, blocks and saves. They are less detailed than **offensive** or **passing** data, and thus less variables are available.

1.2.2 Passing statistics

Passing data refers to all the passing actions between players during the game. They are divided into three types of passes: passes, key-passes and assists. This data set gathers 18 actions.

Assists are passes that are immediately followed by a goal.

Key passes are important passes which lead to assists and then to a goal.

The third type of pass, just called **passes** are all the other kind of passes during the game.

1.2.3 Offensive statistics

Offensive data refers to actions from the players when they are in an offensive strategy. There are five categories of **offensive** data: shots, goals, dribbles, possession loss and aerial duels. For the aerial duels and the dribbles, we only know if the action was successful or unsuccessful. Shots and goals data are more detailed. We have statistics on the zones of the football field where shots and goals have been tempted. We can find the type of situation of these shots and goals, the accuracy of these actions and the players body parts used to shoot. There are a total of 27 offensive data.

1.3 Our data

At first, we tried to work only with the data from the French tournament, the "*Ligue 1*". We only had 20 teams to study, which meant 20 individuals only. It was not enough to study statistics on it, in front of the large number of explanatory variables. Two options were available:

- Gather data from several seasons, which did not satisfy the rules of independence necessary for the implementation of statistical methods. Most of the teams keep a similar composition and a similar playing style over the years. Using data from several successive years would have made individuals dependant.
- Use data from different tournaments.

We chose to use data from different tournaments, from six different countries: France's *Ligue 1*, Spain's *LaLiga*, the German *Bundesliga*, the Italian *Serie A*, England's *Premier League* and Argentina's *Superliga*. The data

from all six tournaments was recorded during the 2018-2019 season. The data set is composed of 124 independent teams and 86 variables, described in the previous section, Section 1.2.

To study the number of points at the end of a tournament, each team has to have played the same number of games. In the *Superliga* tournament from Argentina, each team plays 25 games, while in *Ligue 1* in France, for example, each team plays 38 games. In average, the number of points for a french team will be higher than for an Argentinean team.

Since the data set was composed of averaged values, we multiplied the values for each team by the number of games played by each team.

2 Exploratory analysis

2.1 Analysis of the correlations between variables

The aim of this section is to carry out a qualitative analysis of the data. We use the `corrplot` package ([Wei and Simko, 2017]) to display correlation matrices. Coefficients of the linear correlation matrices are computed on each data set (**defensive**, **passing** and **offensive**) and displayed in the lower triangular part of each plot for more readability.

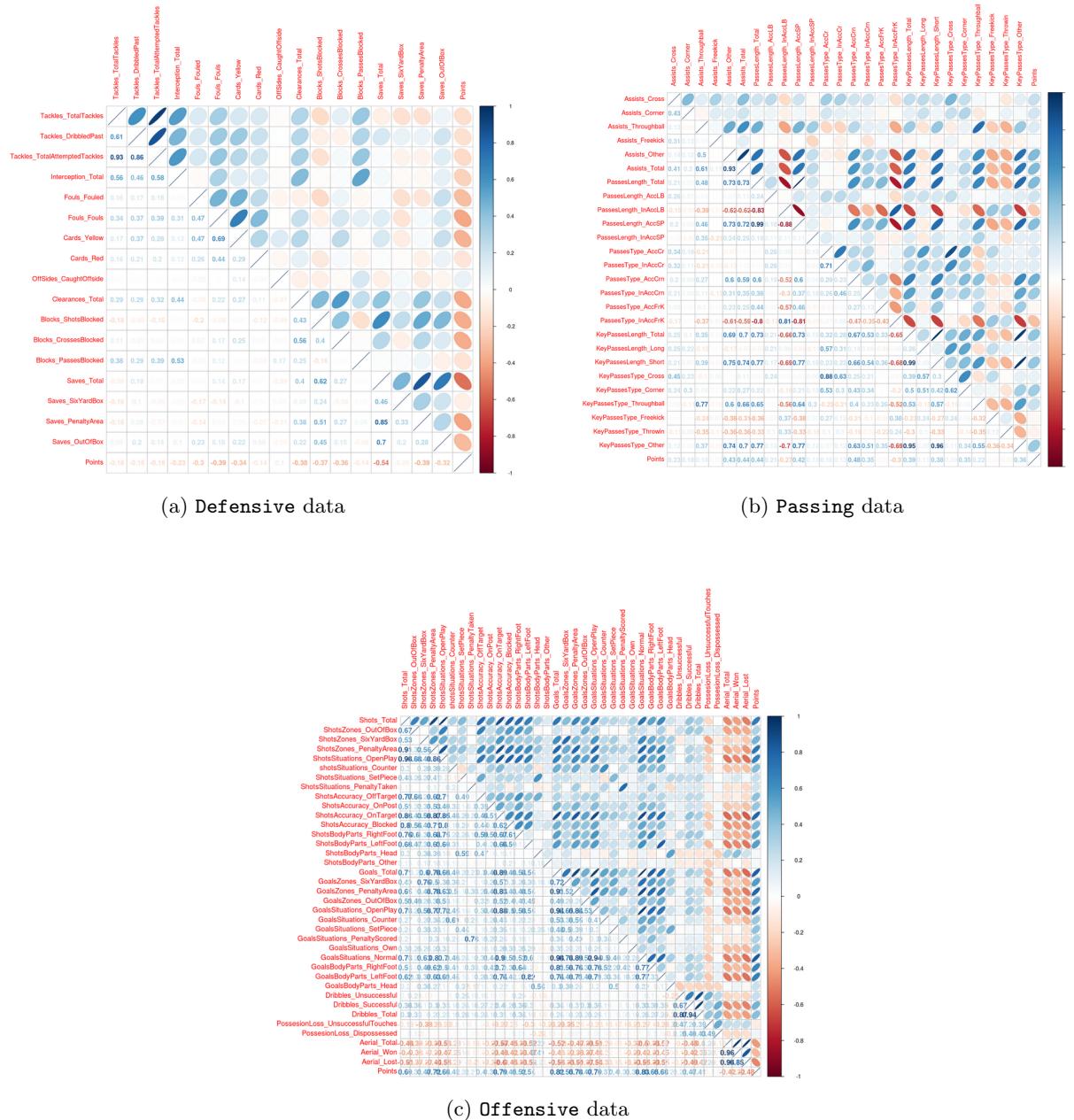


Figure 1: Correlation matrices of **defensive**, **passing** and **offensive** data

We sought to identify some correlated variables and particularly which variables are correlated with the number of points at the end of a tournament. This is the last column of each matrix represented in Figure 1. In Figure 1, we represent on the same graph the real correlation coefficient between each variable and an ellipse standing for this coefficient. A red ellipse means that two variables are negatively correlated, whereas a blue one means that they are positively correlated. The thinner the ellipses, the more correlated the variables are.

In Figure 1a, we note some highly correlated variables. This is the case for `TotalTackles` and `TotalAttemptedTackles`, because `TotalTackles` is included in `TotalAttemptedTackles`. The same phenomenon is observed with `SavesTotal` and `SavesPenaltyArea`, with the same explanation. About `Points`, our interest variable, represented on the right column, there is no highly correlated variable. `SavesTotal` is softly negatively correlated to the number of points.

Several passing variables are correlated negatively, as we can see in Figure 1b. Those variables are `PassesLength_AccSP` (Accurate Short Passes) and `PassesLength_InAccLB` (Inaccurate Long Balls); `PassesLength_InAccLB` (Inaccurate Long Balls) and `PassesLength_Total`; `PassesType_InAccFrK` (Inaccurate Freekicks) and `PassesLength_AccSP` (Accurate Short Passes); `PassesType_InAccFrK` (Inaccurate Freekicks) and `PassesLength_Total`. Each variable is the "opposite" of the other, which can explain why they are negatively correlated. For example, when a team spends the game doing short passes, it does not do long passes.

There are positively correlated variables as well. These include `KeyPassesType_Other` with `KeyPassesLength_Total` and `KeyPassesLength_Short`, or `Assists_Other` with `Assists_Total`.

The `Points` variable is slightly positively correlated with `PassesType_AccCrn`, `PassesLength_Total`, `Assists_Total`. It is also slightly negatively correlated to `PassesType_InAccFrK` and `PassesLength_InAccLB`. The variable `Points` is positively correlated to assits variables and lightly negatively correlated to inaccurate passes.

Finally, the last correlation graph, Figure 1c represents the correlation matrix of offensive data. As with the other matrices, some variables are positively correlated. `Aerial_Total`, `Aerial_Loss` and `Aerial_Won` are closely correlated, they must carry the same information. The same phenomenon is observed for `Dribbles_Successful`, `Dribbles_Unsuccessful` and `Dribbles_Total`. Then `Shots_Total` is positively correlated to `ShotsSituations_OpenPlay` and `ShotsZone_PenaltyArea`, which are the most common types of shots during a game. Some other variables are positively correlated, they have narrow blue ellipses. There are very few negative correlations, and they are weak.

`Points` is positively correlated with `Shots_Total`, `ShotsAccuracy_OnTarget`, `Goals_Total`, `GoalsZones_PenaltyArea`, `GoalsSituations_OpenPlay` and `GoalsSituations_Normal`. It makes sense, and it is quite trivial. Having a high number of shots or goals in a game, means that the team had the ball and had goals opportunities, which is helpful to win.

The three corrplots in Figure 1 allowed us to discover the data set and its three different parts (`defensive`, `offensive`, `passing`). Interactions with the variable to be explained, `Points`, are numerous for `passing` and `offensive` data, but are rare for `defensive` data. To win a tournament, a team must be in offensive and not defensive positions. It should be scoring goals, not trying to avoid conceding them.

2.2 Principal Component Analysis

Another way to analyse our data is to compute a Principal Component Analysis (PCA) to obtain a display of individuals and variables on the same plane. We chose this method because it provides a good representation of individuals and variables by maximizing the variance between variables ([Wikistat, 2016]). Thus it completes the previous part by moving from a bi-variate to a multivariate study.

This linear dimensionality reduction method converts correlated variables into a basis of linearly uncorrelated variables, called principal components.

In this method, variables and individuals are projected on principal components, as we can see in Figure 2. Such graphs make possible to see links between variables. Two variables which are close on a principal plane are strongly and positively correlated, in regard to the main variables carried by this plane. In contrast, two variables which are radially opposed are strongly but negatively correlated. On the contrary, two variables that form a 90° angle on a principal plane are said to be independent, in regard to the principal components of this plan.

On a factorial map, we can also see the links between individuals and variables by using a biplot display such as in Figure 2. Individuals which are close to an arrow tip have a high value of the variable associated to this arrow.

In Figure 2, we have computed the PCA on the entire data set excluding `Points`. `Points` is displayed on the factorial map but has not been used to compute the PCA, since it is a quantitative supplementary variable. That is why the arrow is a red dotted line. Even if `Points` is a supplementary variable, it is highly correlated

with the first dimension: it is almost horizontal. We can say that the first principal axis stands for the number of points and hence for the teams rank.

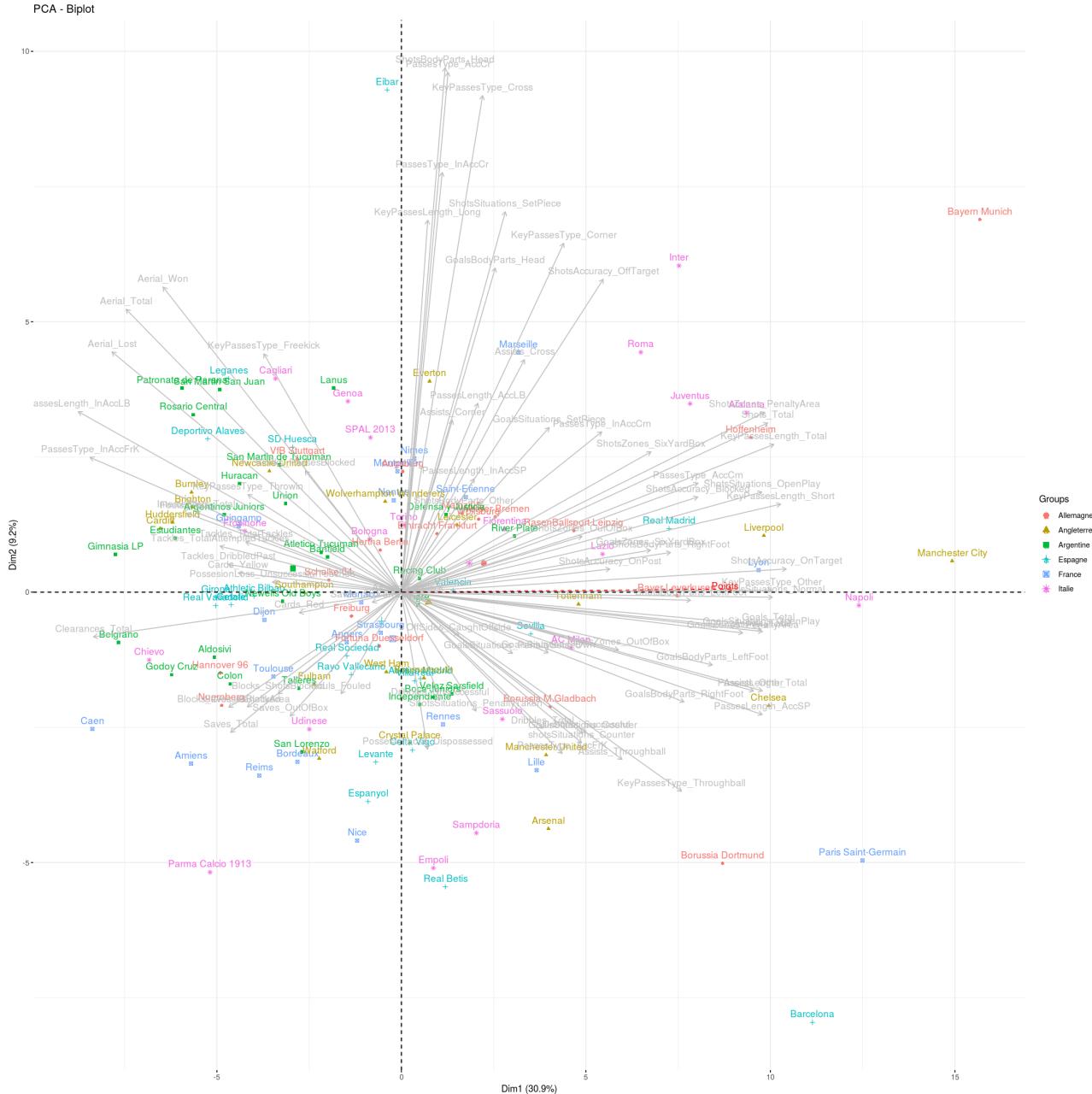


Figure 2: PCA biplot

There are two types of variables that are opposite along the first axis. Variables related to aerial duels, fouls, tackles and ball losses (among others) are displayed on the left, whereas variables related to passes length, shots and goals are displayed on the right. It should be noted that these latest variables, correlated to Points according to the PCA, are logically influential on the number of points of a team. To have a lot of points, a team needs to score goals and construct its game, especially thanks to the numerous and short passes. In contrast fouls, cards and inaccurate actions will lead to few points.

Then, teams have been plotted by color according to their countries. We can see that teams of each country, excepting Argentina, are spread all along the horizontal axis. On the right of the map, we find the best team (or at least the second best team) of each tournament (Bayern Munich, Manchester City, Napoli, Paris Saint-Germain and Barcelona) excepting the Argentina. On the left of the graph, low-performing teams, excepting for the Argentinian, are plotted. The center of the biplot gathers teams in the middle of the ranking, but they are rearranged compared to the real ranking.

We notice that all the Argentinian teams are displayed on the left of this graph, maybe because they do not play the same football as in Europe. It can be explained by the fact that it is the only non European team in our data set.

By seeing this graph, the best teams of the tournaments can be considered as outliers. Eibar and Parma Calcio 1913 are also outliers from the "middle" teams.

2.3 Clustering

In this part we try to gather teams or variables in homogeneous clusters. Two elements of the same cluster have more similarities than two elements of two different clusters. Similarity criterion can be computed using distances, dissimilarities or homogeneity criterion.

An easy way to interpret method is Hierarchical Cluster Analysis (HCA). This is an iterative method based on a distance measure between an individual and a cluster, as explained in [Bridges Jr, 1966]. Starting with one cluster per individual, at each step of the algorithm the two closest clusters are grouped together. At the end we obtain only one cluster containing all the individuals and we can visualize all the "sub-clusters" on a dendrogram, as in Figure 4.

2.3.1 Teams clustering

To create individual clusters, we started from the results of the PCA. Thanks to the PCA, the dimensions of individuals coordinates are already reduced, and then, it is easier to cluster the football clubs. To create clusters, we need to know how many clusters we want to keep. To do so, we use Figure 3 which represents the mean decrease of distance between two clusters for each new grouping. An optimal number of clusters maximizes the jump between two consecutive bars on graph 3. Here two or three clusters are optimal but clustering football teams into only three clusters will not be precise enough. We notice that 5 clusters also provide a large jump. Hence we choose 5 as the optimal number of clusters.

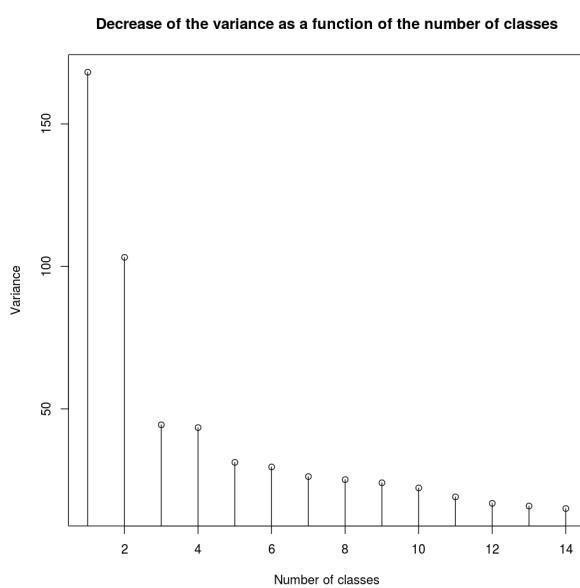


Figure 3: Decrease of the variance

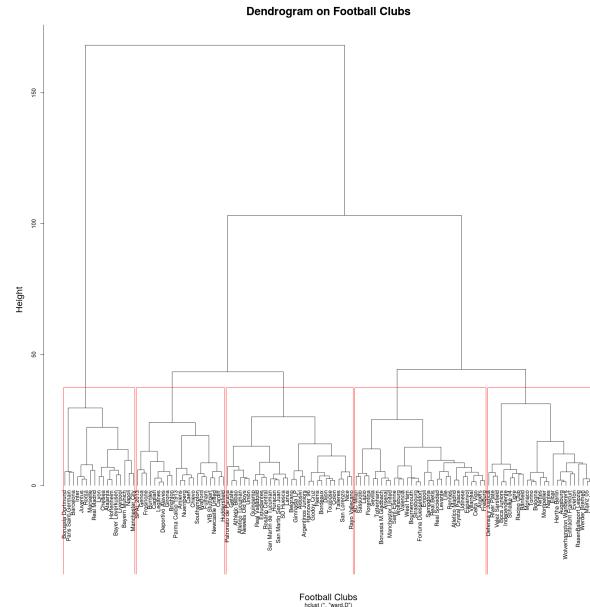


Figure 4: Dendrogram on individuals

Knowing the optimal number of clusters, we are able to divide individuals in 5 clusters, as it is shown in Figure 4. Football clubs seem to be gathered according to their performances. On the left of the dendrogram are the best clubs and on the right the weaker. To have a more precise idea on the characteristics of each cluster, we plot the PCA biplot where we add the clusters. It is Figure 5.

Individual classes on PCA

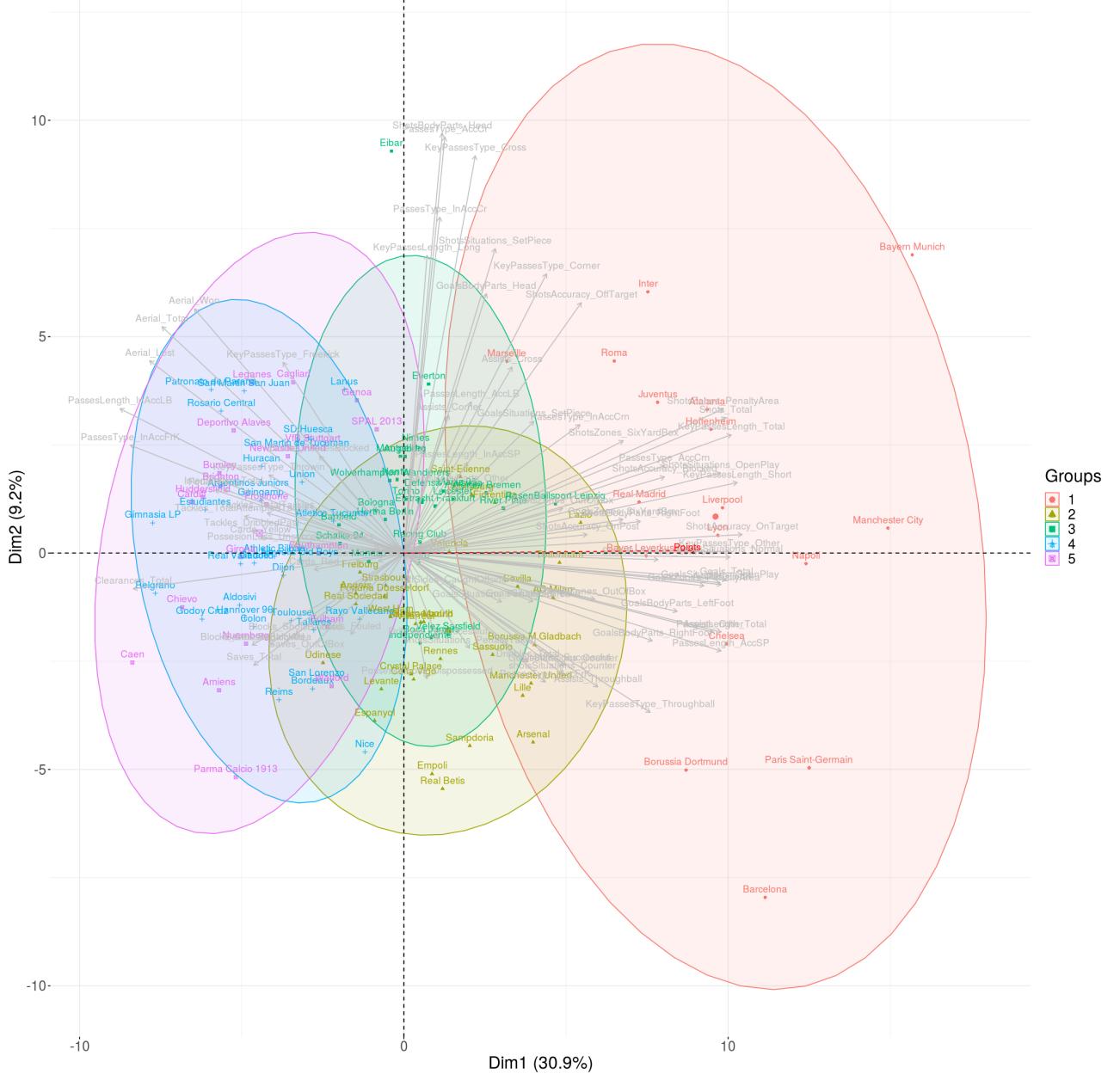


Figure 5: Individuals clusters represented on the biplot

Since the first dimension is highly linked to the number of points of a team, as mentioned in Part 2.2, we can clearly visualize the first cluster, in red, composed of the best teams. This cluster is the most spread out because each team seems to have its own characteristics. Good teams seem to often be good, but in different ways. The separation is obvious for the top of the ranking, as evidenced by the PCA biplot and the dendrogram (Figure 4) because this cluster is linked to the others only at the end. The other clusters are more or less mixed up. The divisions are fuzzy.

Other principal components are not relevant because they explain only a very small part of the variance. This is why we do not show projections of individuals and variables on these components.

2.3.2 Variable clustering

We can use the same method as previously on variables: by clustering them, we can obtain another point of view about correlations between features. To create a dendrogram for variables, the same kind of calculation is made, but before, we compute the Spearman correlation coefficients between each variables. Then the dendrogram is built using the Ward distance method applied to the correlation coefficients. Note that a high correlation coefficient means that two variables are really close, so this is an adapted version of Ward distance.

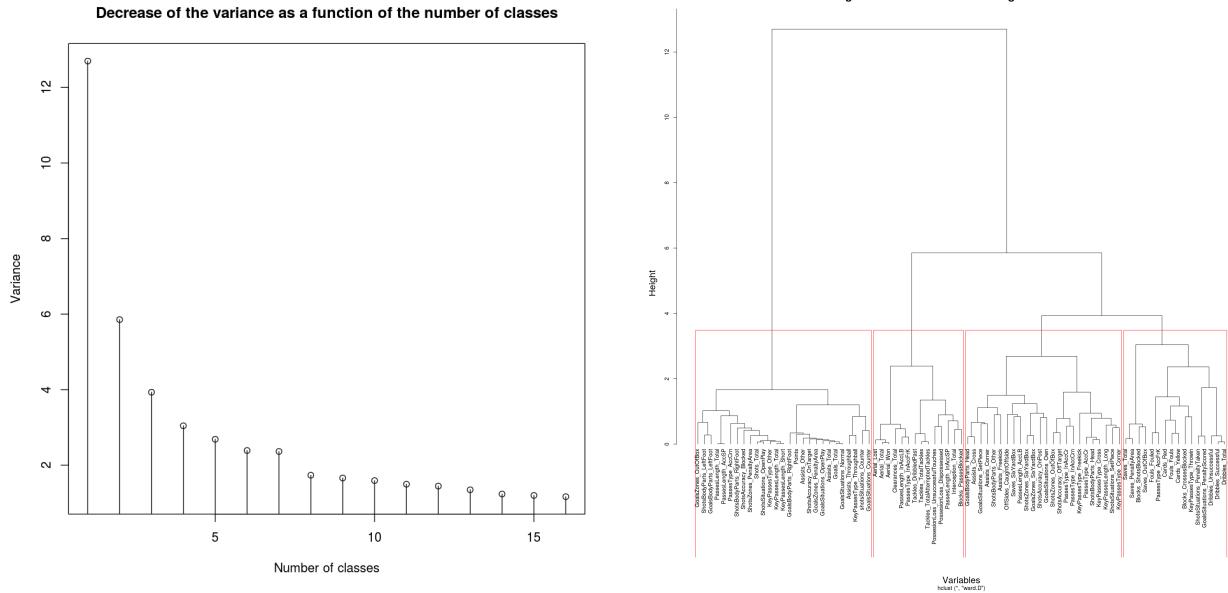


Figure 6: Decrease of the variance

Figure 7: Dendrogram on variables

Analysing Figure 6, we can choose to split the variables into 3, 4 or 8 clusters, according to the same criterion as for the teams. Since we want to obtain a precise gathering of variables, we choose to keep 4 clusters, as shown in Figure 7.

We can see a cluster gathering aerial duels and long passes (clearances for example). Another cluster gathers actions happening after fouls like corners and free kicks. These clusters contain variables negatively correlated with the number of points.

Another cluster, the first rectangle on the left, is relatively different from the others because it is separated very early on the dendrogram. It is related to actions that are important to score goals (key passes, assists...) and hence to goals themselves.

The other clusters do not allow us to draw any useful conclusions.

In Figure 8 we see the 4 clusters of variables perfectly separated. Each variable is colored according to the cluster it belongs. This is therefore the main interest of the PCA, namely to separate the variables as much as possible in order to have the greatest possible variance, while reducing the dimension of individuals.

Variable classes on PCA

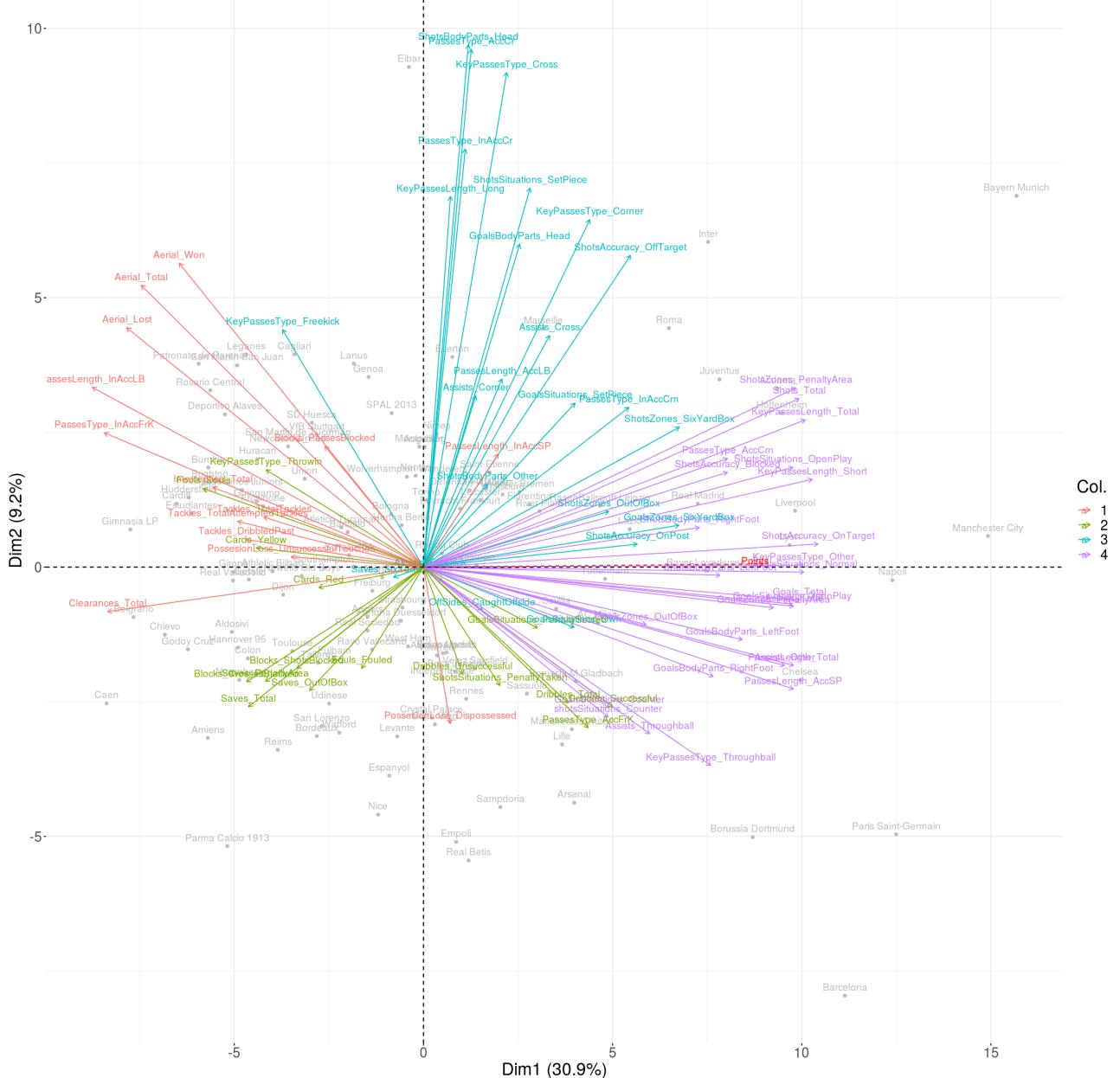


Figure 8: Variables clusters PCA biplot

2.3.3 Total classification

It is possible to visualize clusters on variables and individuals. To do so, we can use a **heatmap**. In Figure 9, each cell stands for the contribution of an individual to a variable. The greener a cell, the more positively influenced by this variable the individual is, whereas a dark pink cell represent an individual highly negatively influenced by a variable. For instance, we can see a dark green cell between **ShotsBodyParts_LeftFoot** and **Barcelona**. It makes sense because Messi plays for Barcelona, he is left-footed, and he scores a lot.

On the bottom left corner are the greener cells. It corresponds to good ranked clubs, hence, variables associated with these cells lead to a good ranking. Next to those variables, there are deep pink cells. The corresponding variables are negatively influential on the ranking. There are aerial duels and inaccurate actions, for example.

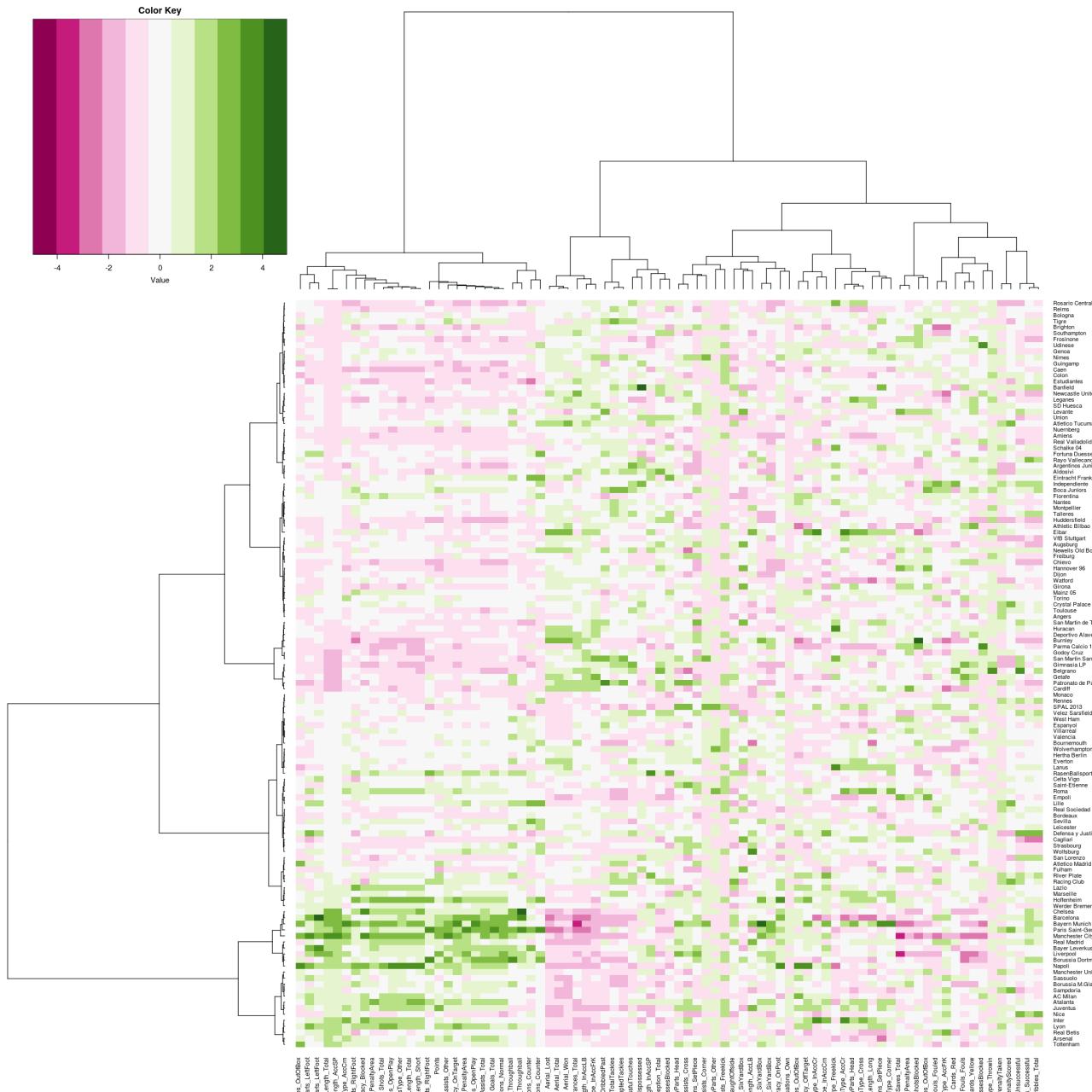


Figure 9: Heatmap on the two dendograms

For the two next sections, we have decided to remove variables related to goals, assists and key passes. They are directly linked to the number of points and thus to the ranking. We want to identify which actions during the game are the most relevant to the outcome of a championship.

3 Variable selection in regression

3.1 Bayesian Information Criterion

The Bayesian Information Criterion, also named *BIC*, is a criterion used to select a model among a finite set of models. The finite set of models is composed of the complete model and every possible combination with different numbers of variables ([Zhang, 2016]). A good model minimizes the BIC, defined as

$$BIC = \ln(x)k - 2\ln(\hat{L})$$

where

- \hat{L} is the maximized value of the likelihood function;
- x is the observed data;
- n is the number of individuals;
- k is the number of parameters estimated by the model.

To select variables, we worked stepwise. At each iteration, the models tries to add a variable to the model and checks if one variable can be removed.

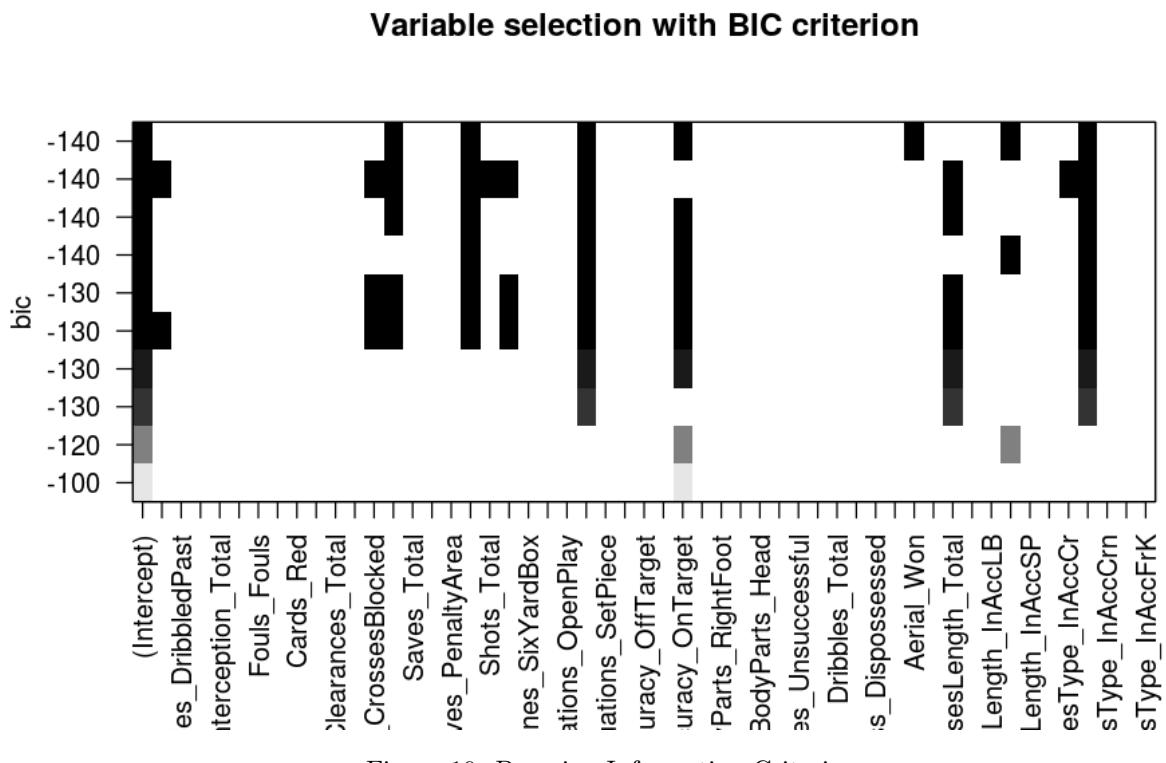


Figure 10: Bayesian Information Criterion

The variables selected are those allowing the lower BIC. The best model according to this criterion includes the black variables at the top. There are 7 variables left. These variable are detailed at the end of this part, and compared to the ones retained by other selection methods.

3.2 Lasso and ElasticNet

Our data set is composed of a number of explanatory variables (actions) really close to the number of individuals (football teams). A classical regression will lead to estimators with high variance.

A solution can be to use regularization methods, by introducing penalization criterion in the optimisation problem. Let us consider a linear model with n observations:

$$Y_i = \theta_0 + \theta_1 X_i^1 + \theta_2 X_i^2 + \dots + \theta_p X_i^p + \epsilon_i$$

where $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^p$. We assume that the variables ϵ_i are independent and identically distributed, centered and with a variance σ^2 .

We can rewrite this equation using matrix \mathbf{X} with general term X_i^j (we set $X_i^0 = 1 \forall i$) and the vectors θ and \mathbf{Y} . We want to minimize $\|\mathbf{Y} - \mathbf{X}\theta\|^2$. The classical least squares estimator, in the case of the matrix \mathbf{X} is of full rank, is:

$$\hat{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Since we noticed redundancy between some features, it can be possible that the matrix \mathbf{X} is not of full rank. That is why we introduce the LASSO regression, where the optimisation problem can be written as:

$$\min_{\theta} L(\theta) = \|\mathbf{Y} - \mathbf{X}\theta\|^2 + \lambda \sum_{i=1}^p |\theta_i| \quad (1)$$

With this method we obtain sparse solutions, which means that several coefficients θ_i are set to 0. In other words, some variables are not used to explain the target variable Y . That is why we can say that LASSO regression is a variable selection method.

The parameter λ in equation 1 has to be chosen by us. To help us find the better value of λ , we used a cross-validation method.

For example, in Figure 11, some coefficients are put to 0, and when $\lambda = 0.018$, we have less than 10 coefficients left.

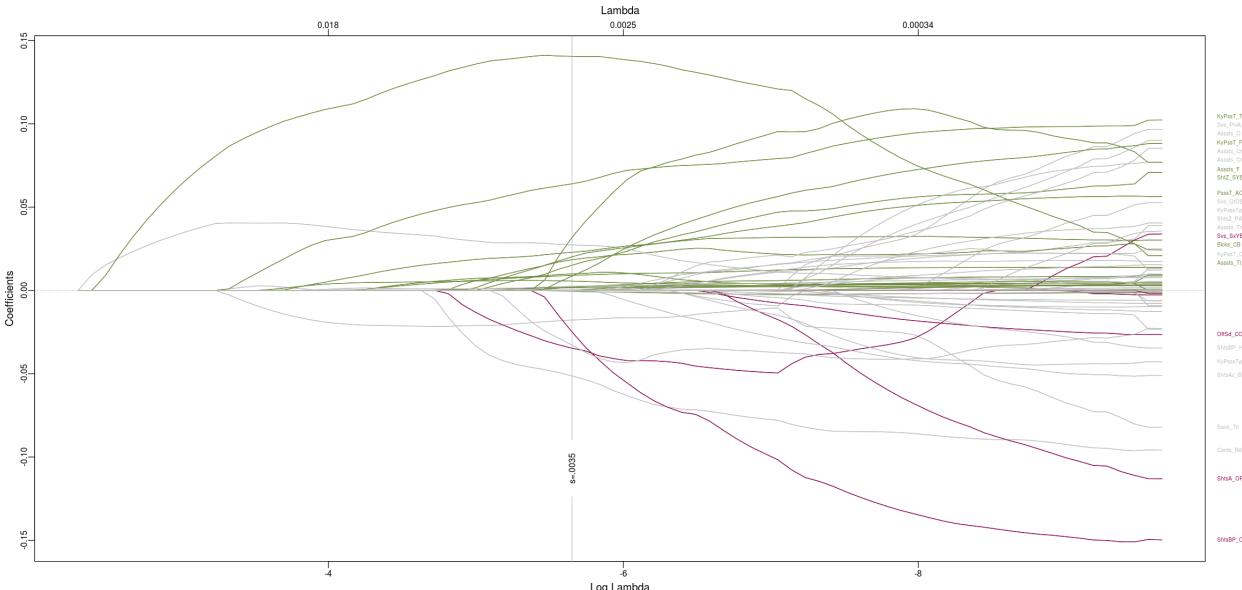


Figure 11: Lasso regression on the data set

Another regularization method in regression, called Ridge regression consists in minimizing:

$$\min_{\theta} L(\theta) = \|\mathbf{Y} - \mathbf{X}\theta\|^2 + \lambda \sum_{i=1}^p \theta_i^2 \quad (2)$$

This method do not lead to sparse solutions but it prevents parameters θ_i from taking high values. This is not a variable selection method.

LASSO and Ridge regressions combined form the ElasticNet method, which can be written, as explained in [Zou and Hastie, 2005]:

$$\min_{\theta} L(\theta) = \|\mathbf{Y} - \mathbf{X}\theta\|^2 + \lambda \left(\alpha \sum_{i=1}^p |\theta_i| + (1 - \alpha) \sum_{i=1}^p \theta_i^2 \right) \quad (3)$$

ElasticNet method is a good compromise between sparse solutions and solutions with enough non-zero parameters. There is not as much null coefficients as in the LASSO regression, but we keep a reasonable number of explanatory variables.

3.3 Results comparison

3.3.1 Methods comparison

Using different methods, we obtained different results. The best method between ElasticNet and LASSO seems to be ElasticNet because it is smoother by mixing LASSO and Ridge. It keeps around 20 variables, which is a reasonable number. If we want to reduce the dimension, we can remove the variables with a coefficient close to 0, but keeping all of them allows a larger study.

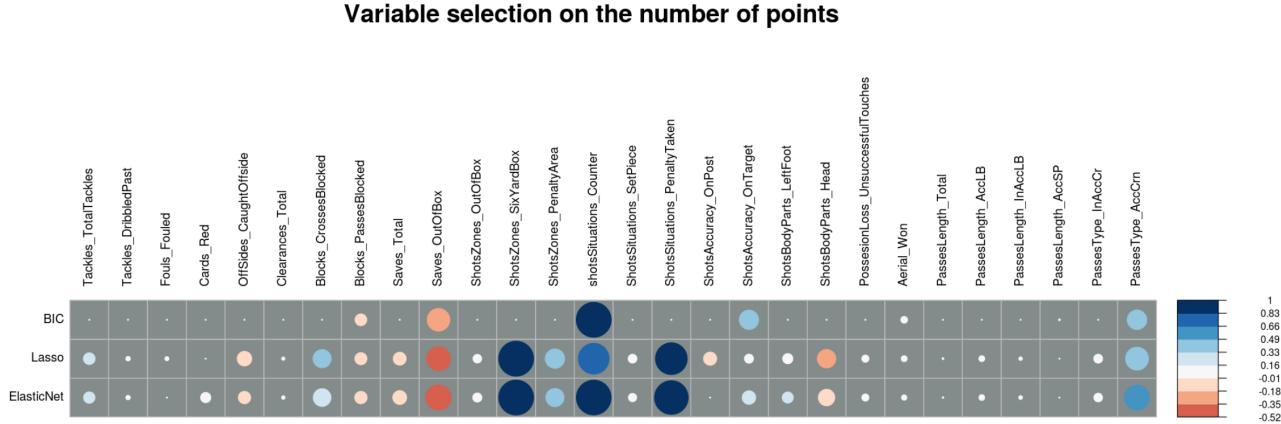


Figure 12: Variables kept according to the method

In Figure 12, a large circle stands for a variable highly correlated to the number of points. If it is blue, it is positively correlated, and if it is red, it is negatively correlated. We normalized coefficients to compare these three methods on the same basis.

Based on ElasticNet, shots (`ShotsZones_SixYardBox`, `ShotsSituations_Counter` and `ShotsSituations_PenaltyTaken`) are highly positively influential on the number of points at the end of the tournament. `PassesType_AccCrn` has also a strong positive link with the target variable. `Saves_OutOfBox` and `ShotsBodyParts_Head` are negatively influential on the number of points. We had already seen the importance of shots in part 2.1, and it is confirmed. LASSO highlights the same variables.

Based on the BIC results, we keep 5 variables: `ShotsSituations_Counter`, `ShotsAccuracy_OnTarget` and `PassesType_AccCrn` are positively influential on Points; `Saves_OutOfBox` and `Blocks_PassesBlocked` are negatively linked to the number of points. These variables have already been chosen with ElasticNet and LASSO. Their number is smaller. It is interesting to observe that `ShotsZones_SixYardBox` and `ShotsSituations_PenaltyTaken` are really important in LASSO and ElasticNet regression, but BIC does not choose them.

3.3.2 Differences between tournaments

The data set is composed of data from six different countries. We have already seen in Part 2.2 that Argentina's teams play differently from European teams. We applied the ElasticNet variable selection to data from each country to observe possible differences. This methods, with LASSO was the more adapted to the problem because we have 20 individuals each time, with more than 80 variables. Since LASSO and ElasticNet gave the same results earlier, we have decided to focus on ElasticNet.

Variable selection on the number of points for different countries

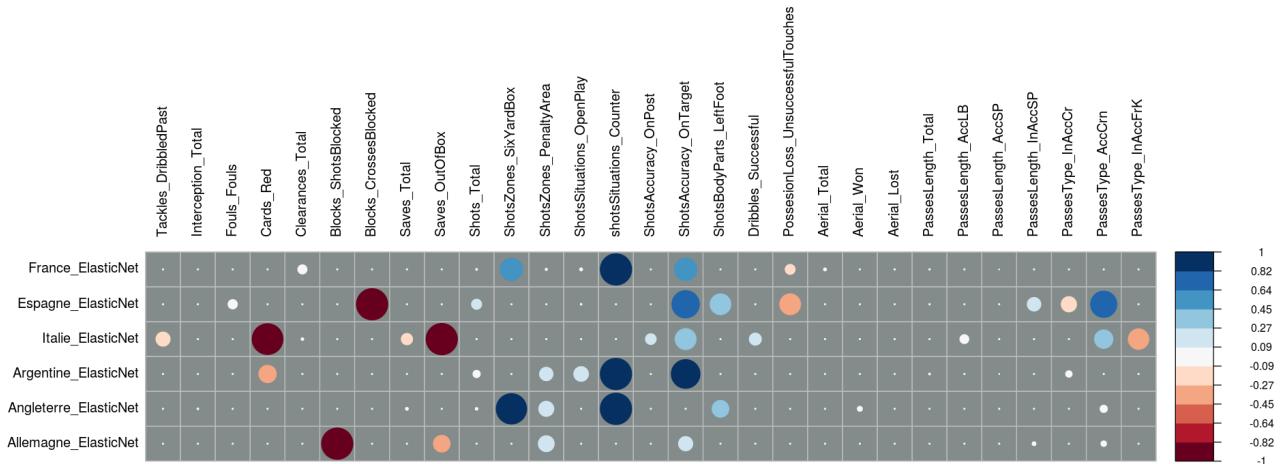


Figure 13: Variables kept according to the country

In Figure 13, we can note some differences. No variable emerges in the 6 countries. `ShotsAccuracy_OnTarget` is the only one present in 5 over 6 countries (every country except England). Variables selected for Argentina are not really different from the ones for the other countries. Since those variables are the main ones, some other "small" variables may be responsible of the specificity of Argentina seen in Figure 2.

4 Machine learning methods

In the next subsections, we will compare three machine learning algorithms (Random Forest, Support Vector Machine and eXtreme Gradient Boosting) in variable selection. To exploit results these methods we will use the importance of features, computed by the `caret` package ([Kuhn, 2008]). Classical SVM and XGBoost methods do not provide variable importance but `caret` package computes it, based on the area under the ROC or the value of the t -statistic of each variable, depending on the form of the regression method.

4.1 Random Forest

4.1.1 Classification and Regression Tree

A Classification and Regression Tree (CART) is a non parametric algorithm which aims in constructing a binary decision tree, easy to interpret.

Figure 14 is an example of a tree computed with the number of points as target variable.

Its construction is detailed in [Loh, 2011]. For each node, the algorithm chooses the explanatory variable which minimizes the sum of heterogeneities of the two son nodes, and the threshold which best divides these two son nodes. The heterogeneity of the node κ is defined by

$$D_\kappa = \sum_{i \in \kappa} (y_i - \bar{y}_\kappa)^2$$

On the tree represented in Figure 14, we can see several nodes. Each node represents a variable which splits individuals according to a threshold. For example, the root (first node on top of the tree) in Figure 14 stands for `ShotsAccuracy_OnTarget` with its threshold 165.

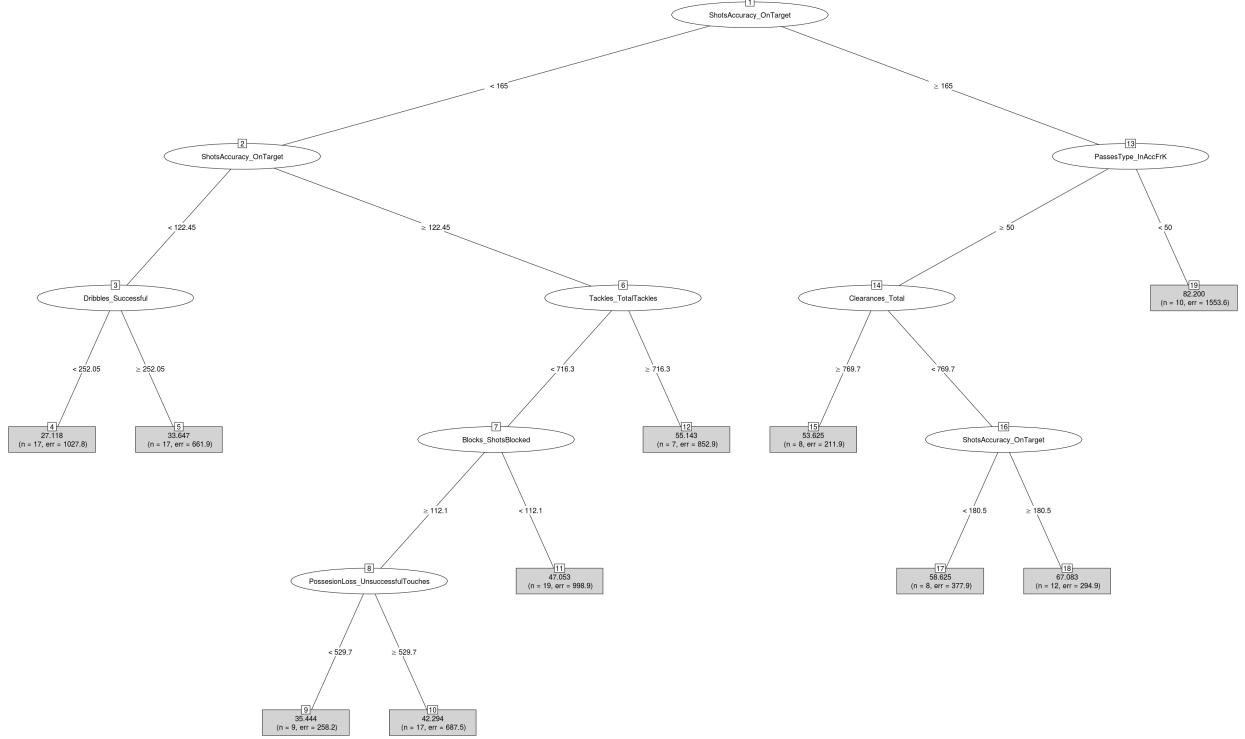


Figure 14: Example of a tree used in Random Forest algorithm

A variable can be selected to split several nodes. In Figure 14, the variable `ShotsAccuracy_OnTarget` is used three times to split three nodes. This variable is also the variable used to split the first node. Thus, this variable has an important effect on the number of points.

4.1.2 Feature selection using Random Forest

Random Forest is an aggregation method based on *bagging* applied on CARTs ([Breiman, 2001]). To construct each CART, the algorithm starts by sampling individuals by a *bootstrapping* method, i.e randomly with replacement. The CART model is then adjusted to this bootstrap sample. In order to reduce variance of the model, before splitting each node according to heterogeneity criterion, algorithm selects randomly a set of explanatory variables among all of them. This procedure reduces correlation between trees.

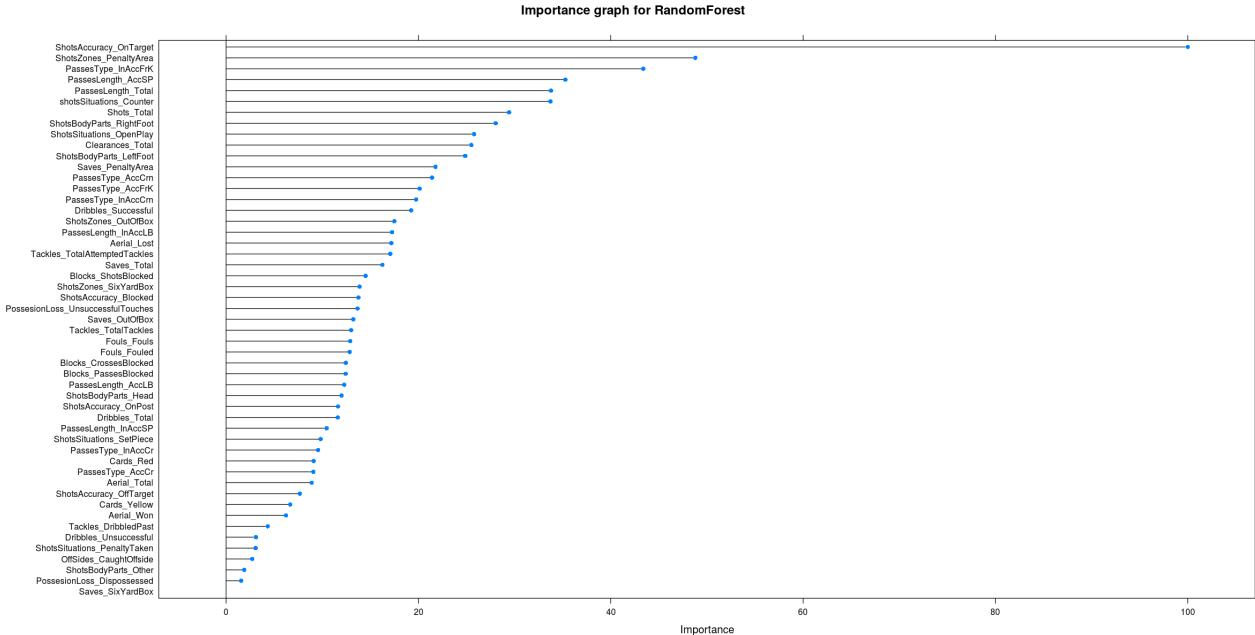


Figure 15: Variable importance based on Random Forest

Figure 15 shows the decrease of importance of the 30 most important features of our data set, for the prediction of the number of points.

We chose to keep 6 variables that seems to stand out from the others: `ShotsAccuracy_OnTarget`, `ShotsZones_PenaltyArea`, `PassesType_InAccFrk`, `PassesLength_AccSP`, `PassesLength_Total`, `shotsSituations_Counter`.

The root mean squared error of this model is 9.95.

4.2 Support Vector Machine

A support Vector Machine (SVM) constructs one or a set of hyperplanes in a high dimensional space. Hyperplanes are chosen to maximize the margin. The margin is the separation between the hyperplane and its closest value. It was first defined for classification, but can be adapted easily to regression problems. `caret` provides an importance measure calculated as explained earlier.

In Figure 16, we can pick the 7 most important variables. It is arbitrary, we could have chosen 8 as well. They are `ShotsAccuracy_OnTarget`, `ShotsZones_PenaltyArea`, `PassesLength_Total`, `PassesLength_AccSP`, `Shots_Total`, `shotsSituations_OpenPlay`, `PassesType_AccCrn`.

The root mean squared error of this model equals to 12.68.

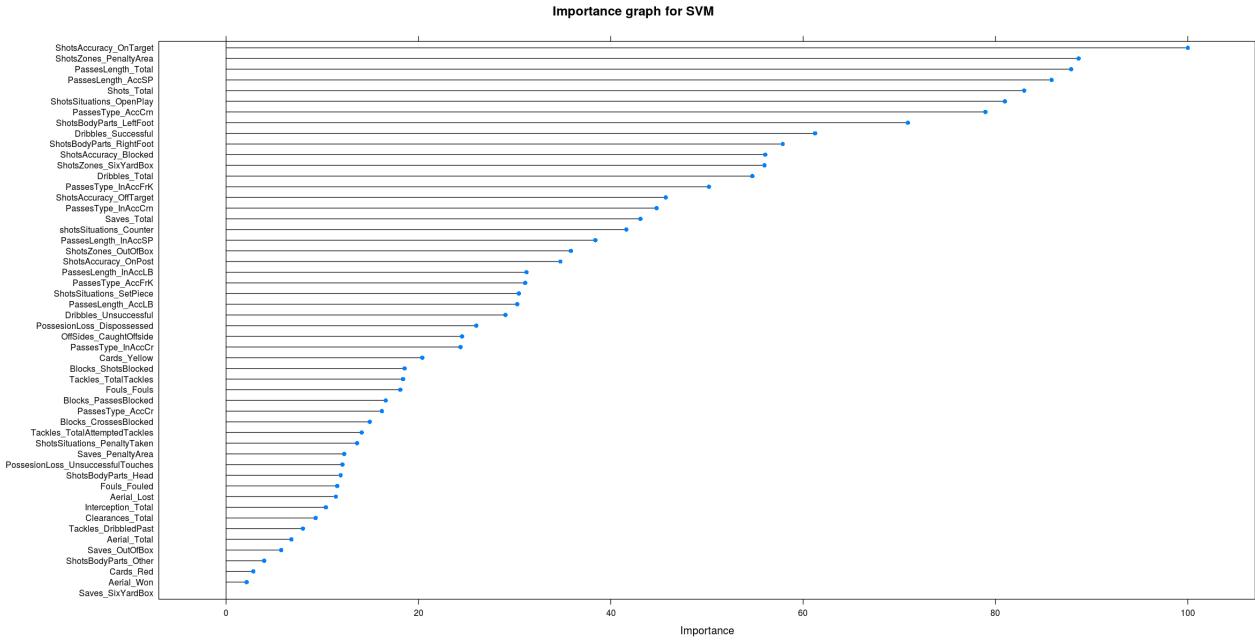


Figure 16: Variable importance based on SVM

4.3 XGBoost

eXtreme Gradient Boosting is an optimized version of **Gradient Boosting** algorithm. **Boosting** is a supervised learning method in which several methods from a model set are sequentially aggregated.

At each step, the algorithm adds to the combination a new model adapted from the previous one. In the new model, more importance is given to observations that were maladjusted on the previous one.

There are several *boosting* methods but the most famous is **Gradient Boosting** [Friedman, 2002]. At each step the new model is constructed in the direction of the gradient of the loss function. This gradient is approximated by a Regression Tree.

XGBoost [Chen and Guestrin, 2016] is an improved version of Gradient Boosting in particular because it requires the adjustment of more parameters. The major changes are an additional penalization in the loss function and an approximation of the gradient by a Taylor development.

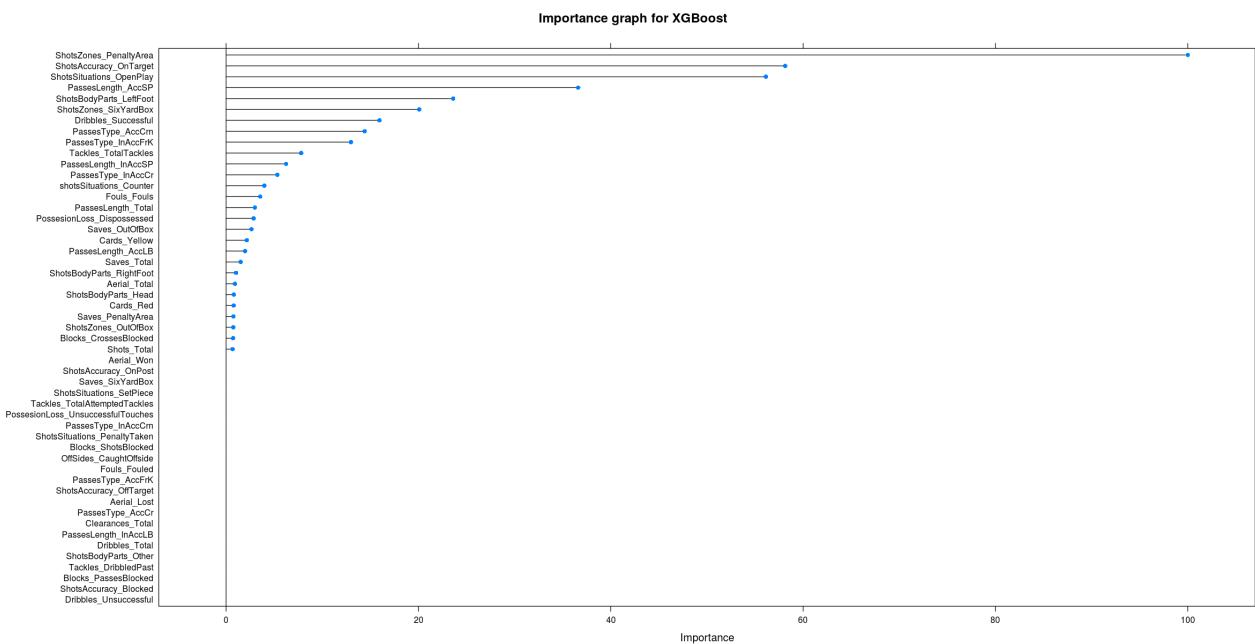


Figure 17: Variable importance based on XGboost

We computed importance of features for XGBoost, as we can see in Figure 17. We notice 6 variables really important: `ShotsZones_PenaltyArea`, `ShotsAccuracy_OnTarget`, `ShotsSituations_OpenPlay`, `PassesLength_AccSP`, `ShotsBodyParts_LeftFoot` and `ShotsZones_SixYardBox`.

The model fitted has a root mean squared error of 10.37.

4.4 Comparison of the machine learning results

The three machine learning methods highlighted different main variables, but three of them are common.

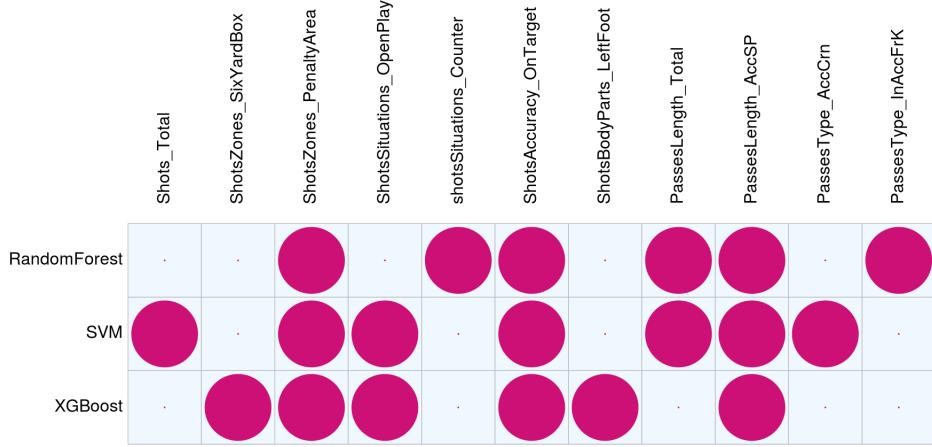


Figure 18: Comparison of the main variables

In Figure 18 we can identify `ShotsAccuracy_OnTarget`, `ShotsZones_PenaltyArea` and `PassesLength_AccSP` as the three main variables, identified by all three methods. They represent the number of shots on target, the number of shots from the penalty area and the number of accurate short passes. The total pass length and the number of shots from the open play also seem important.

We recognise some variables selected with methods based on regression (Figure 12). `ShotsZones_SixYardBox`, `ShotsSituations_Counter`, `ShotsZone_PenaltyArea` were selected by ElasticNet regression or BIC, and by some of the machine learning methods.

Regarding root mean squared error as a performance indicator of these three methods, we can say that Random Forest is better than XGBoost, which is better than SVM. Moreover, Random Forest is the fastest algorithm.

Conclusion

The aim of the study of the data set from *whoscored.com* was to highlight the main variables leading a team to the head of a tournament. It was also to find out the characteristics of the best teams and see if they are specific to the country.

The first step of this study, exploratory analysis, was necessary to discover our data set and to have a first sight of interactions and redundancy between variables. We noticed that the number of points of a club was highly correlated to **offensive** and **passing** statistics.

Then we computed a PCA to obtain another representation of our variables and individuals, by maximizing variance. According to the PCA biplot, we identified the best teams of each championship as outliers. First axis of this PCA was highly linked to the number of points. Teams spread along this axis : good teams on the one hand and bad teams on the other hand. Teams of middle ranking are rearranged along this axis.

For the second and third parts of this project, we decided to remove variables directly linked to goals, assists and key passes. Variable selection methods applied on the entire data set have highlighted these variables, but it is obvious that a team needs to score goals to win matches. By not considering these features, we emphasized the number of accurate short passes, of shots attempted from the penalty area, and of shots on target. Short passes testify a built-up game and interaction between players. When a team is able to construct its game, it is leading the game. The importance of the number of shots from the penalty area reflects the fact that a player has more chances to score goal when he shots from the penalty area, because it is the closest playing zone to the goal. Obviously, to score goals, it is necessary to shoot on target.

LASSO and ElasticNet selected main variables from **defensive** data, negatively influential, and variables related to shots, impacting positively the number of points. BIC selected a small amount of data, from every kind.

Machine learning methods highlighted features related to **offensive** and **passing** data, mostly short passes, shots accuracy and data from shots zones. The variables selected by machine learning methods seem more interesting to keep. They could be useful to implement prediction algorithms, because it has been shown that it is more important to attack than to defend.

The study showed that every tournament has its characteristics. It depends on the playing style and strategies of each country. The main difference is observed for Argentina. We were able to observe it on the PCA graph, when we clustered the teams and finally with the variable selection according to countries.

To continue this study, it could be interesting to implement these methods on players data, available on *whoscored.com*. Concerning the data set, another approach could be to focus on a team over several years, paying attention to dependency. Finally, the data set could be more homogeneous and regular by adding other tournaments. Argentina could be less of an outlier with other Latin American countries because, for this project, we considered only European tournaments, except for Argentina.

References

- [Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- [Bridges Jr, 1966] Bridges Jr, C. C. (1966). Hierarchical cluster analysis. *Psychological reports*, 18(3):851–854.
- [Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794. ACM.
- [Friedman, 2002] Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378.
- [Kuhn, 2008] Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, Articles*, 28(5):1–26.
- [Loh, 2011] Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23.
- [Wei and Simko, 2017] Wei, T. and Simko, V. (2017). *R package "corrplot": Visualization of a Correlation Matrix*. (Version 0.84).
- [Wikistat, 2016] Wikistat (2016). Analyse en Composantes Principales — Wikistat. [En ligne; Page disponible le 23 septembre 2019].
- [Zhang, 2016] Zhang, Z. (2016). Variable selection with stepwise and best subset approaches. *Annals of translational medicine*, 4(7).
- [Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.