Statistical Learning Project: How can football clubs cope with overpaid players?

Loris Bartesaghi, 967378

July 6, 2021

Abstract

The starting point of the analysis is the observation of the increasing salaries of football players in the last years. The main goal of this project is to provide a statistical analysis on football players' salaries and performance, an attempt to evaluate the ability of teams' management to correctly evaluate the performance of players and the correct level of the relative salaries and an approach to decrease the number of overpaid players, hiring athletes with the same characteristics but with low salary. Random forest and forward stepwise regression are used to select most relevant variables. Linear regression is applied on the selected variables to understand the impact of variables on gross salary. Residuals are considered as a measure to comprehend if a player is underpaid or overpaid, and a logistic regression is applied to compute the probability of being overpaid if an athletes had signed to a particular team. Clustering is applied to group players with the same characteristics and to provide meaningful substitute to overpaid players.

Keywords— Random Forest, Forward Stepwise Selection, Linear Regression, Logistic Regression, Hierarchical Clustering, Football players, Salary

1 Introduction

Professional athletes' fields performance and salaries is a topic that has attracted the interest of researchers: the general question of interest is whether players deserve their salaries based on their performance statistics. The aim of this project is to study this topic relying on data of "Serie A" players during the season 2020/2021. Machine learning algorithms will be used to understand the relationship between salaries and players' performance during the season and to extract meaningful information to set up an evaluation method for the teams' managers administration.
Hierarchical clustering will be used to group football players, based on the statistics relative to the performance on the field.
Due to large amount of information available on Internet, Web Data Extraction is becoming a fundamental technique to create suitable datasets for data analysis. The first step of the project is to create a dataset by extracting information from different websites related to football players' performance and salaries.
A preliminary analysis is fundamental to comprehend variables and observations, to avoid incompleteness or anomalies of data. High correlation is the main problem to deal with. Highly correlated variables, which express similar aspects of an athlete's performance, are removed from the starting dataset. This is done to increase the reliability of results provided by the algorithms implemented for features selection. Moreover, some variables express redundant information, others are very low correlated with the variable of interest and for these reasons are removed from the analysis.
Two different algorithms are used for features selection: forward stepwise selection and random forest. Linear regression is applied on selected variables to interpret the influence that these features has on football players' gross salary. Correlated variables and outliers are removed to have more reliable results from the regression output. A comparison between the accuracy of the two linear models is done to choose the algorithm that fits better data.
Residuals computed implementing the model selected are considered as the measure of overrecompense/underpayment for each player. The summation of the residual (both positive and negative) grouped by teams are computed. These values and logistic regression are useful to understand the goodness of decision of teams' managers.
Hierarchical clustering is used to group players based on theirs statistics on the pitch. Clustering could be a solution for overpaid players: in fact, they could be substituted by athletes in the same cluster with lower salaries. This strategy could lead to decrease the costs of football societies for football players, preserving (or even increasing) the quality of the team.

2 Description of data

The starting point of the entire process is to compile all the required information about the players' performance on the field (which are provided by the web site "https://fbref.com/en/"), and on their salaries (data are available on the the website "https://salarysport.com/football/"). The data acquired were narrowed down the Seria A season 2020/2021 for 525 players.
Throughout the player statistics data accumulated, a multitude of 123 features provided a plurality of information about each players' performance over the season and per game, as the position and players' team.
Next on data collection process, "salarysport.com" is consulted to attain the fundamental information about the players' gross income expressed in pound for the 2020/21 season.
The volume of data accumulated needed to be merged into a unified dataset that would serve the purpose of the analysis. The objective of this process was to associate each player's wage with their field's statistics.
As said in the introduction: high correlated regressors, low correlated regressors with the response variable and not relevant variables are excluded from the analysis.
Goalkeeper are excluded from the analysis, the available statistics are not suitable for the evaluation of goalkeeper performance. Players' misleading statistics are substituted with the mean of the anomalous feature. This is done because during the season some athletes played only few minutes, but producing a key action for the match, resulting in unrealistic information (e.g. "Federico Santander" create "9.73" score action per match, which is obviously a misleading information).
Finally, the ultimate data set is composed by 477 observation and 49 features. Looking at the distri-

bution of the athletes' salary is clear that the observations are not normally distributed.

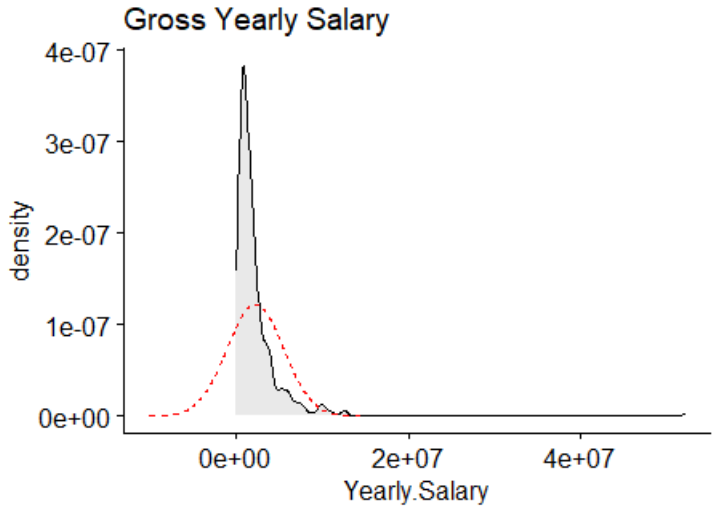

Figure 1: Gross Yearly Salary

The hypothesis is confirmed by the result of the Shapiro-Wilk normality test (p-value = 2.2e-16), which lead to a very strong rejection of normality of the response variable.

3 Methods and Empirical Results

The Box-Cox transformation is applied to the response variable, so that it could closely resemble a normal distribution. For regression, that is applied in the next steps, normally distributed errors are an assumption that allow to construct confidence intervals and conduct hypothesis tests. By transforming target variable, we can (hopefully) normalize errors.
The transformation depend on the lambda value in the following formula, which in our case is 0.25:

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

Figure 2 shows the distribution of the transformed 'Gross Yearly Salary'. The p-value of Shapiro-Wilk test is 5.197e-07, thus another rejection of normality of the variable. For the analysis, the Box-Cox transformation of the Gross Salary will be considered the dependent variable.

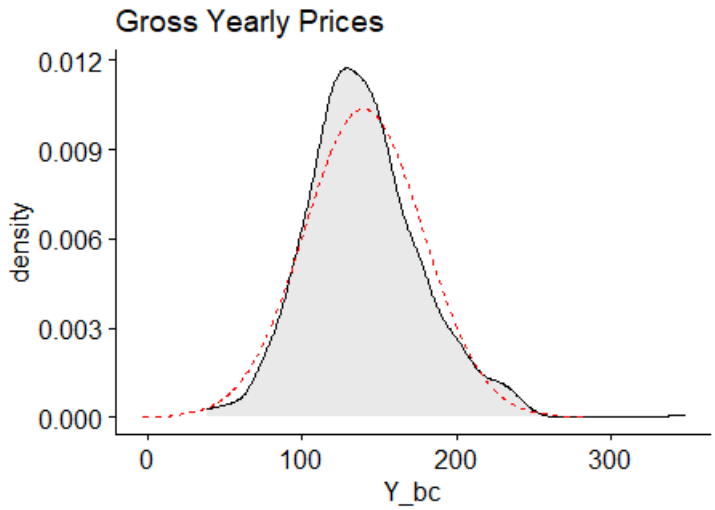

Figure 2: Box-cox transformed Gross Yearly Salary

3.1 Forward Stepwise Selection

Best subset selection is not suitable for features reduction due to the number of regressors in the analysis. Forward stepwise regression, which is more appropriate for the purpose, is a step-by-step iterative construction of a regression model that involves the selection of independent variables to be used in a final model. It involves adding potential explanatory variables in succession and testing for statistical significance after each iteration.
The purpose of using features selection is to reduce the number of features that the computer must process. Furthermore, features reduction removes multicollinearity resulting in improvement of the machine learning model in use. Another benefit of feature reduction is that it makes data more interpretable.
For the purpose of selecting the best number of features, a 10-fold cross-validation is used to estimate the average prediction error (RMSE) of each models, containing from 1 to 25 variables. The models with lowest root-mean-square-error is selected.

The variable selected by the algorithm are: Age, team, shot per match, percentage of successful pass, number of completed pass and number of goal creating actions per season.
A regression model is computed on the selected features and removing 20 players, that are considered outliers. The coefficients are shown in the following table.

Regression on features selected by Forward Stepwise			
Variable	Coefficient	Standard Error	p-value
(Intercept)	33.444995	13.067252	0.01082 *
Team			
A.C.F. Fiorentina	-15.018937	5.174288	0.00389 **
A.S. Roma	2.807761	5.347597	0.59982
Atalanta Calcio	-36.643170	5.431967	4.90e-11 ***
Benevento Calcio	-41.086034	5.372152	1.34e-13 ***
Bologna F.C. 1909	-34.682381	5.097418	3.41e-11 ***
Cagliari Calcio	-25.620541	5.217964	1.29e-06 ***
F.C. Crotone	-45.220858	5.290971	2e-16 ***
F.C. Internazionale	15.329211	5.390772	0.00467 **
Genoa C.F.C.	-25.687350	5.295031	1.72e-06 ***
Hellas Verona F.C.	-37.506626	5.248566	3.83e-12 ***
Juventus F.C.	29.270896	5.808857	6.89e-07 ***
S.S. Lazio	-14.607946	5.601488	0.00943 **
S.S.C. Napoli	6.530737	5.367986	0.22442
Parma Calcio 1913	-31.620479	5.037614	8.41e-10 ***
Spezia Calcio	-50.276006	5.089421	2e-16 ***
Torino F.C.	-21.306035	5.317597	7.25e-05 ***
U.C. Sampdoria	-34.326042	5.307406	2.70e-10 ***
U.S. Sassuolo	-43.024729	5.186129	1.38e-15 ***
Udinese Calcio	-42.560949	5.247071	5.23e-15 ***
Age	2.920820	0.195163	2e-16 ***
Sh.90	10.460377	1.147045	2e-16 ***
X.total..completed.	0.436960	0.146282	0.00298 **
total.completed	0.007368	0.002286	0.00136 **
GCA	0.381198	0.265371	0.15159

From the table it is possible to understand that the team for which the players have signed is statistically significant to determine the level of the Gross Salary. Age is statistically significant and positive, so more experience lead to an increase in income (as any other job). Percentage of successful passes, number of completed passes and and number of shots per match are positive and significant. Goal creating actions per season has positive coefficient, but p-value shows that the estimate is not significant. The next table shows the accuracy on training (75 % of the entire data set) and test set (25 % of the entire dataset).

Regression on features selected by Forward Stepwise			
	RMSE	R^2	MAE
Training Set	22.83277	0.6061068	17.27525
Test Set	23.4397249	0.6947975	17.6120034

3.2 Random Forest

To assess the goodness of the variables selected by forward stepwise algorithm, the same approach is followed using Random Forest.
To select the best number of nodes and the depth of the trees in Random Forest Algorithm, root-mean-squared-error will be computed for different number of trees (100, 250, 500, 1000) and for different depth (2, 8, 16, 32). The following figure allow to understand the best combination to choose.

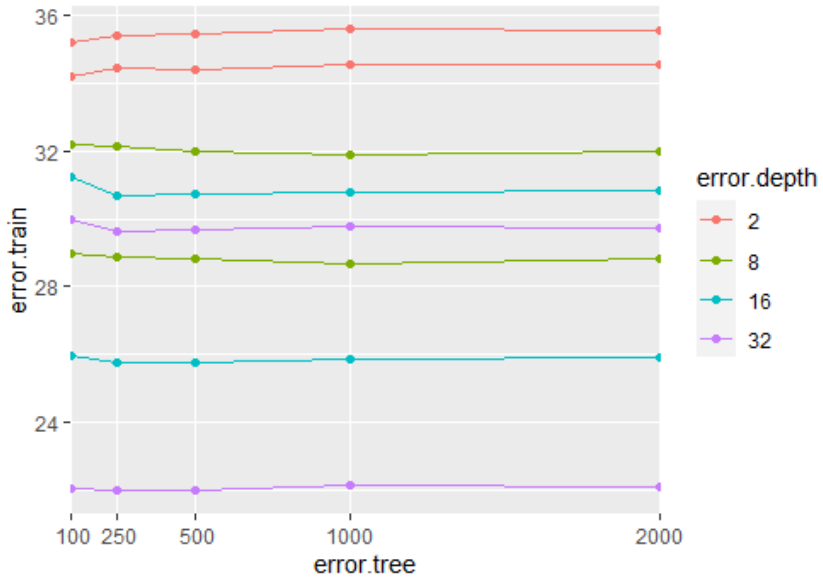

Figure 3: Training and test error random forest

Random Forest run on 500 trees and with depth of 8 is used to extract the importance of the variables. The percentage increase in MSE is the measure that allow to select the most important variables.

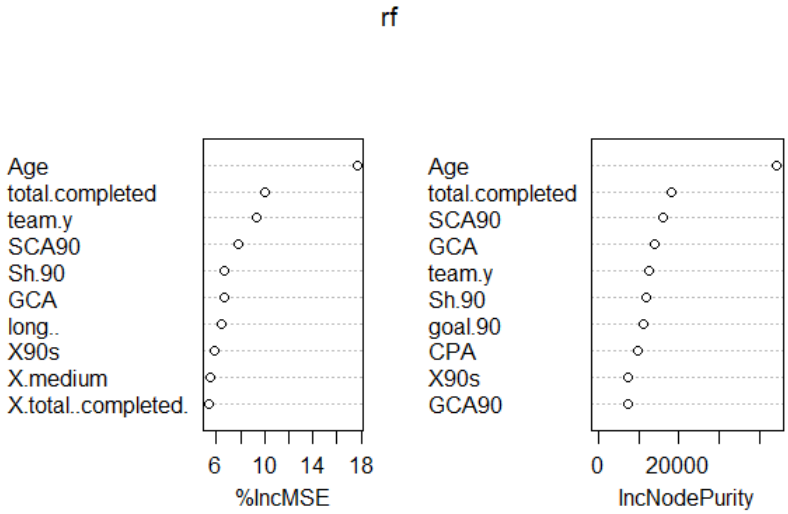

Figure 4: Features Importance

The features selected are Age, number of total completed passes, team, shot creating action per match, shot per match and goal creating action per season. Linear regression is applied to data,

removing from the analysis 23 outliers.

Regression on features selected by Random Forest			
Variable	Coefficient	Standard Error	p-value
(Intercept)	68.695507	6.089755	2e-16 ***
Team			
A.C.F. Fiorentina	-15.803935	5.125197	0.00218 **
A.S. Roma	5.666923	5.378193	0.29262
Atalanta Calcio	-36.404935	5.390916	4.71e-11 ***
Benevento Calcio	-44.146018	5.224798	4.57e-16 ***
Bologna F.C. 1909	-35.973984	5.037440	3.97e-12 ***
Cagliari Calcio	-27.134578	5.145830	2.13e-07 ***
F.C. Crotone	-46.767117	5.218849	2e-16 ***
F.C. Internazionale	16.421324	5.339503	0.00224 **
Genoa C.F.C.	-28.177380	5.176894	8.82e-08 ***
Hellas Verona F.C.	-39.778398	5.142742	7.43e-14 ***
Juventus F.C.	30.054590	5.767566	2.92e-07 ***
S.S. Lazio	-14.866586	5.554868	0.00773 **
S.S.C. Napoli	7.201700	5.320960	0.17662
Parma Calcio 1913	-34.624688	5.019592	1.90e-11 ***
Spezia Calcio	-50.859271	5.049992	2e-16 ***
Torino F.C.	-22.767655	5.246459	1.78e-05 ***
U.C. Sampdoria	-37.100060	5.163941	3.00e-12 ***
U.S. Sassuolo	-42.237238	5.139101	2.45e-15 ***
Udinese Calcio	-44.166958	5.169559	2.28e-16 ***
Age	2.939657	0.192971	2e-16 ***
SCA90	0.984075	1.048711	0.34858
total.completed	0.009806	0.002086	3.49e-06 ***
GCA	0.092684	0.278536	0.73948
Sh.90	8.673342	1.298205	7.37e-11 ***

As in the previous regression, teams are very significant to predict the gross income of a football players. Shot per game, age and total completed passes are again positive and significant.

Regression on features selected by Forward Stepwise			
	RMSE	R^2	MAE
Training Set	25.03849	0.5596128	19.06796
Test Set	20.0003333	0.7054512	15.4869715

From the two regressions output, it is clear that the predominant factor to determine gross salary for football players is the team for which they had signed, followed by experience (measured with the variable Age) and the attitude to shot on target per match. The last characteristics are the number of passes and the relative precision.
For the features mentioned, the estimated coefficients are very similar in both regression output. So, the results are consistent, due to the fact that two different model give very similar results.

3.3 Teams's Managers Evaluation

Nowadays, football clubs' cost are increasing due to the growth of football players salaries and athletes' prices on the market. For this reason, a set of rules must be followed by the club to increase their financial sustainability, and an evaluation on the clubs' managers performance (based not only on the ability to win trophies, but also on their ability to cut cost and still be competitive on the field) has to be introduced. An attempt to find a way to evaluate mangers is using the previous models. The previous machine learning algorithm are compared, and linear regression with variable selected by Random Forest has lower error and an higher R^2. Using this model, an attempt to understand if a player is overpaid or not is done. The following plot shows the sum of the residual for all the players grouped by teams.

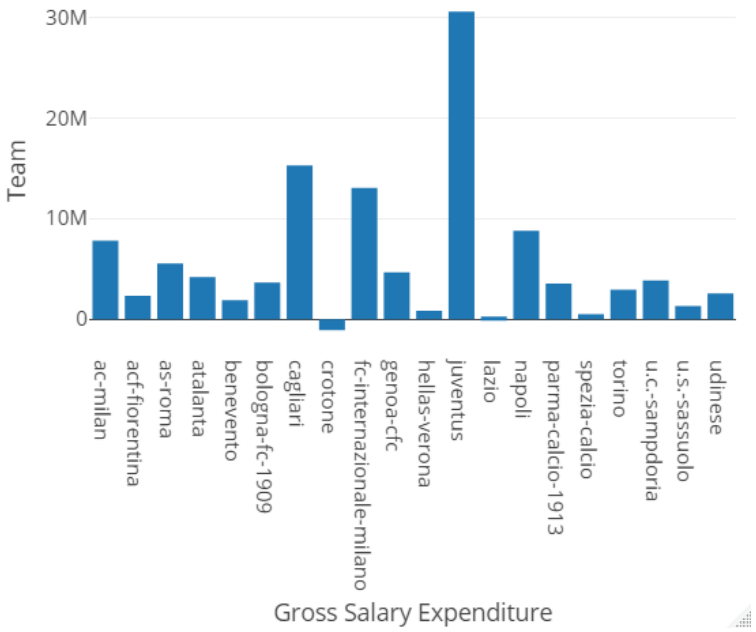

Figure 5: Difference between expected and real gross salary

From the previous plot it is possible to understand that Juventus F.C. and F.C. Internazionale-Milano management tend to overpay athletes. This problem arise because clubs which are competing to win trophies need to sign the best players available and, as a result, an increase of the purchase prices and the salaries for the best players occur.
Society as S.S. Lazio, F.C. Hellas Verona and U.S. Sassuolo management are able to pay players based on theirs contribution on the pitch, despite the good final positions in the championship.
Instead, for teams that compete to avoid relegation, as Spezia and Crotone, players' salary are based on theirs expected performances.
Cagliari Calcio, despite the final position in the 'Serie A' championship (16°), is ranked as the third team in the leader-board of 'big-spender' football clubs. This fact could be a proof of bad management for a football society in that season.

In the following table the numerical values of the previous plot are shown. Furthermore, with a logistic regression is possible to define the probability to be in the 20% most overpaid players in the season 2020/21.

Evaluation table for teams		
Team	Probability overpaid player	Difference of real and predicted value by team
A.C. Milan	0.17	7791480
A.C.F. Fiorentina	0.04	2313849
A.S. Roma	0.17	5543261
Atalanta Calcio	0.14	4174914
Benevento Calcio	0.12	1903203
Bologna F.C. 1909	0.19	3656670
Cagliari Calcio	0.31	15255899
Crotone Calcio	0.04	-1083400
F.C. Internazionale	0.23	13029043
C.F.C. Genoa	0.18	4646387
Hellas Verona F.C.	0.12	838462
Juventus F.C.	0.23	30540504
S.S. Lazio	0.00	257156
S.S.C. Napoli	0.23	8784446
Parma Calcio 1913	0.21	3550938
Spezia Calcio	0.12	515166
Torino F.C.	0.09	2924234
U.C. Sampdoria	0.08	3851268
U.S. Sassuolo	0.04	1331635
Udinese Calcio	0.12	2569521

Looking at the probabilities, U.S. Sassuolo and S.S. Lazio are the best-run football clubs due to the ability to correct estimate the players' salary. Both clubs spent in wages more than the expected, but with values lower than 1 million, which could be see as a good result. Furthermore, the probabilities to be in the top 20% of overpaid players are below 5%.
As before, Cagliari Calcio could be seen as the worst managed football club of the last season. The probability to be in the top 20% of overpaid players is over the 30%, a very bad result considered tha final position in the championship (16°)

3.4 Clustering Algorithm applied to Serie A football players

With clustering algorithms it's possible to group players into the same subgroups so that observations within a subgroup are quite similar to each other and observations in different subgroups are quite different from each other. The purpose of this part, is to create group of players which are very similar for statistics on the field. Based on the following classification of positions on the pitch of football players:

- Full back
- Centre back
- Central midfielder
- Wide midfielder
- Attacking midfielder
- Wing
- Centre forward
- Striker

hierarchical clustering algorithm is run on 8 cluster The following plot allows to understand how the algorithm divided the entire set of players.

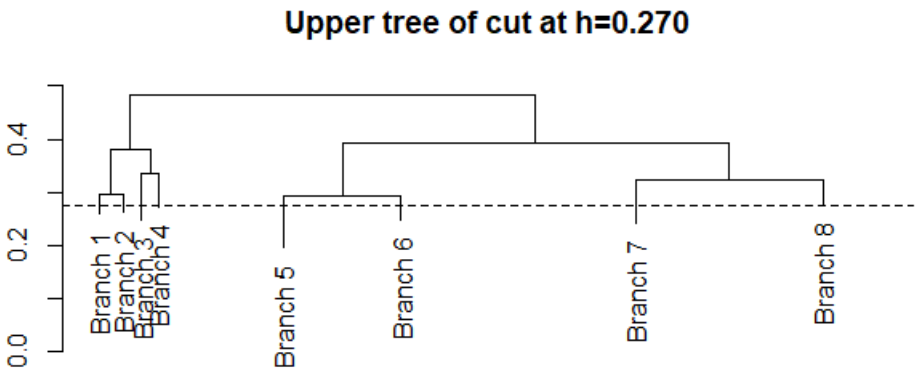

Figure 6: Dendrogram

A simple analysis is done looking at the clusters' mean with respect to the remaining observations' mean of the following variables: age, goals per season, salary, matches played, number of dribbled players per season, number of crosses made during the season, the percentage of tackles won during the season, the goal creating actions for the season, completed passes during the season and finally the proportion of defenders, midfielders and strikers on the total players on the cluster and on the entire dataset. Furthermore, in some cases plots of distribution of some crucial statistics are needed to have a better comprehension of the cluster.

The group formed by the algorithm are the following:

- **Centre-back and Full-back**: 152 players belong to this cluster.

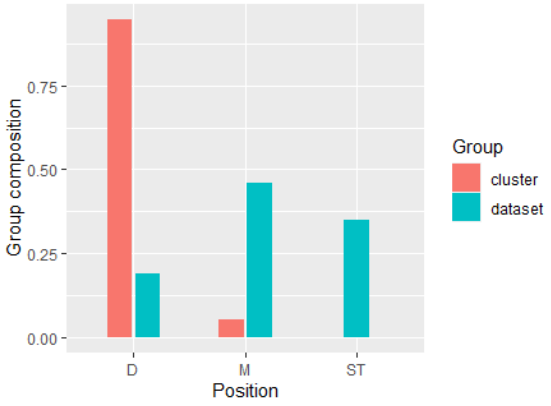

Figure 7: Position proportion

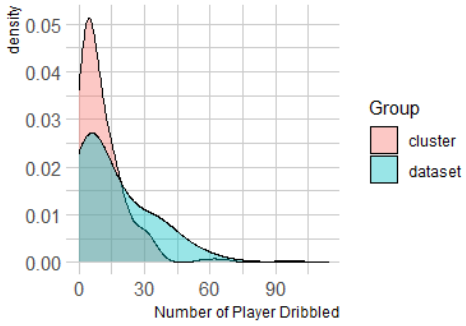

Figure 8: Number of player dribbled

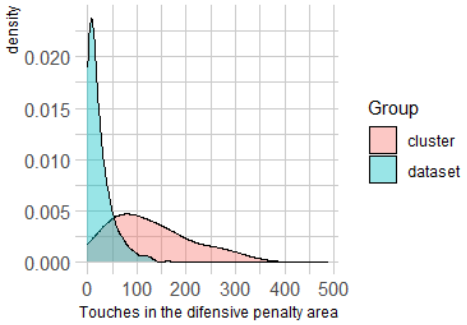

Figure 9: Touches in defensive penalty area

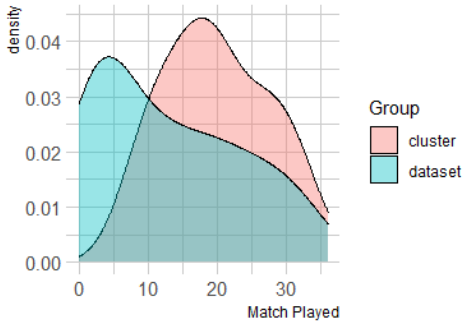

Figure 10: Matches played

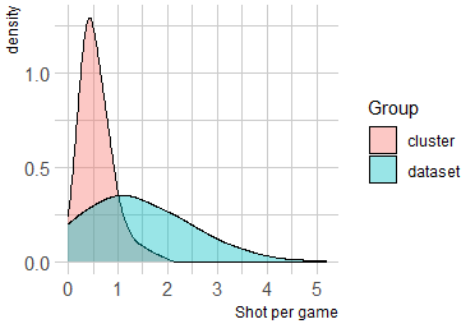

Figure 11: Shot per game

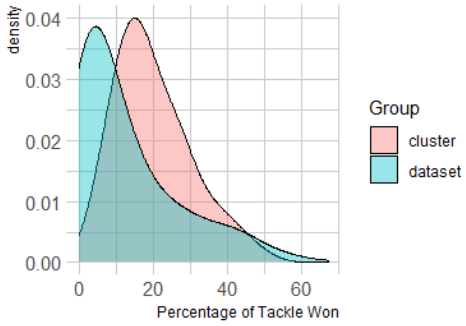

Figure 12: Percentage of tackle won

From 'Figure 7' is possible to understand that more than 75% of players in first cluster are defender, less than 25% are midfielder and no strikers are presents. From 'Figure 9' it's possible to notice that the cluster distribution have an higher mean of touches in defensive penalty area,

with a lot of players that played the ball in that part of the field more than 100 times. 'Figure 10' is plotted to understand how many games the athletes in this cluster are on the field, and an higher mean of matches played are capable from this plot. 'Figure 11' and 'Figure 12' allow to understand the ability to win tackles during the season and the ability to shot on target during the match. Clearly, this cluster contain defensive players, which are capable to win tackles during the match, they are able to set up actions from the defensive part of the field and provide good reliability over the season (high number of presences).

- **Strikers(forwards)**: 5 players in the cluster. Strikers are able to shoot confidently with either foot, possess great power and accuracy, and have the ability to link-up with teammates and pass the ball under pressure in breakaway situations and have good dribbling skills.

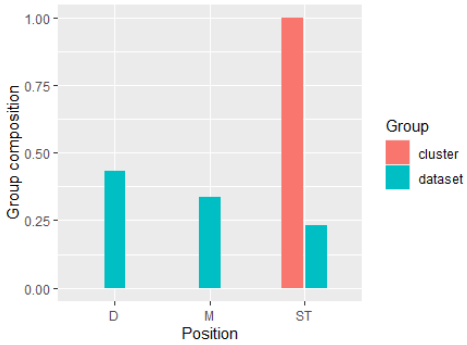

Figure 13: Position proportion in cluster

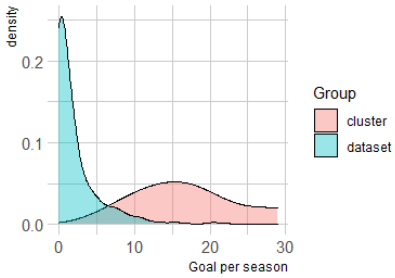

Figure 14: Goals per season

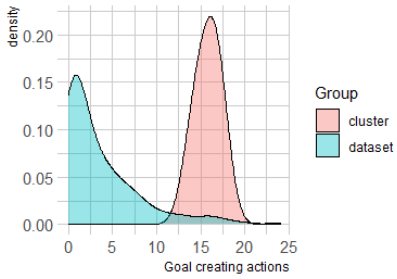

Figure 15: Goal creating actions

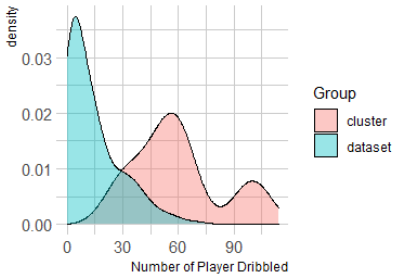

Figure 16: Number of players dribbled

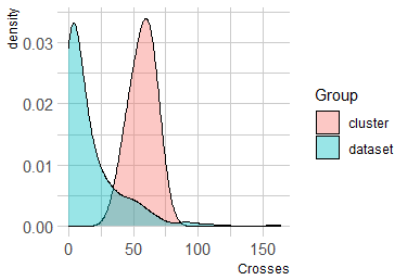

Figure 17: Number of crosses

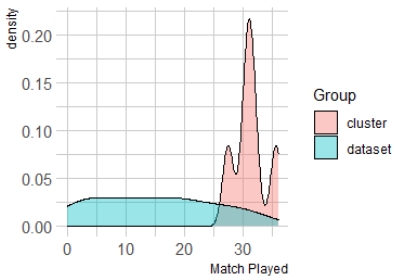

Figure 18: Matches played

From plots it's possible to understand that all players in the cluster are strikers. They played a huge number of games, they score more than the mean of the others clusters' players, they

dribbled on average more than 60 players during the season, and all of them played more than 27 games (about 75% of the matches of the season). They can be classified as complete forwards, which are able to score but also to assist, dribble opponents and create dangerous situations.

- **Low used offensive players**: the players in the cluster are 109. 2% of them are defenders, 39% are midfielders and 60% are strickers. They played 10 games per season on average, and no particular features describe this cluster.
- **Wing Back**: the role is a variation of the full-back, with an heavier emphasis on attack. Players belonging to this cluster are 16, and the defenders are 50% of the cluster, midfielders are 38% and forwards are the 12% of the dataset.

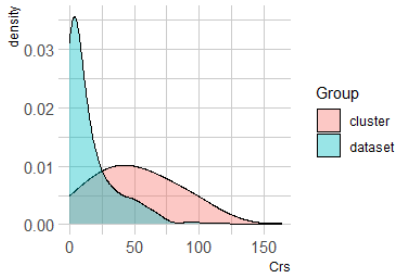

Figure 19: Number of crosses

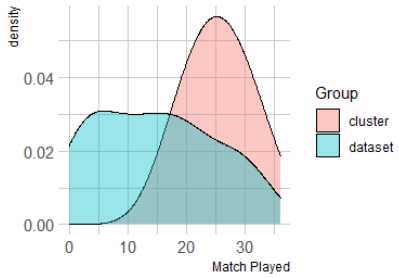

Figure 20: Matches played

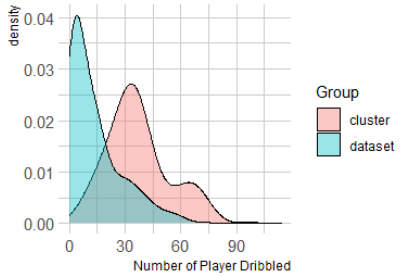

Figure 21: Players dribbled

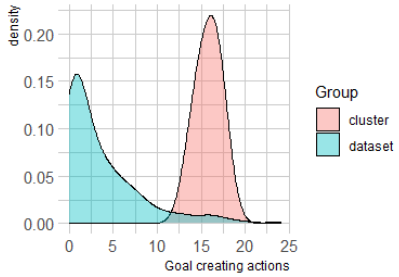

Figure 22: Goal creating actions

From the previous plots it's possible to understand that players in this clusters are highly used, they shown ability in dribbling and crossing to assist others players, despite theirs position on the field this players are able to create dangerous action during the season.

- **Very low used players**: The mean of games played by players of this cluster is 2.5, with respect to 18.77 games of players that belong to others clusters. 104 players belong to this cluster and the composition is: 50% are defenders, 34 % are midfielder and 16 % are forwards. No significant information could be extracted, given the low number of matches played.
- **Attacking midfielder**: the attacking midfielder is an important position that requires the player to possess superior technical abilities in terms of passing and dribbling, as well as, perhaps more importantly, the ability to read the opposing defence in order to deliver defence-splitting passes to the striker.
  the cluster is composed by 10 players, 80% of them are midfielders and the remaining are strikers. On average they played more than 26 matches in the last season, they have over the mean statistics for goal creating actions, number of players dribbled during the season, number of passes during the season, number of crosses during the season, but they have lower mean for

goals scored with respect the clusters "Centre Forward" and "Strikers". So, this kind of players are able to set up the offensive maneuver and to create dangerous actions in the offensive half of the pitch, but leaving the duty of scoring to more offensive players.

- **Defensive/ Box-to-Box midfielder**: defensive midfielders are midfield players who focus on protecting their team's goal. These players may defend a zone in front of their team's defence, or man mark specific opposition attackers. While the term box-to-box midfielder refers to central midfielders who are hard-working and who have good all-round abilities, which makes them skilled at both defending and attacking. These players can therefore track back to their own box to make tackles and block shots and also carry the ball forward or run to the opponents' box to try to score.
  From the first figure, it's possible to notice that near the 100 % of the 61 players in this cluster are midfielders. They are able to win tackles and even to create dangerous action when they control the ball and the majority of them played over than 20/25 games in the last season.

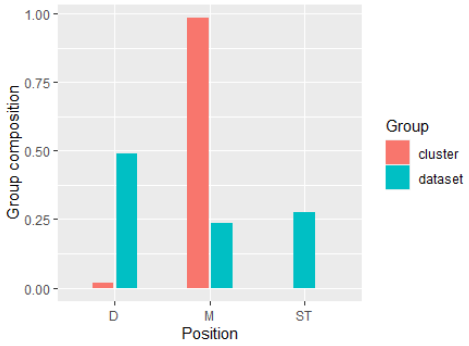

Figure 23: Position proportion in cluster

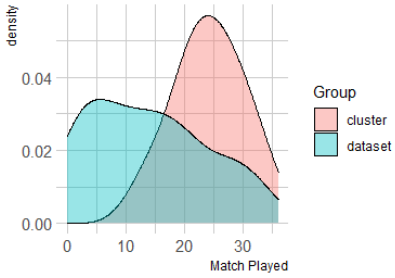

Figure 24: Matches played

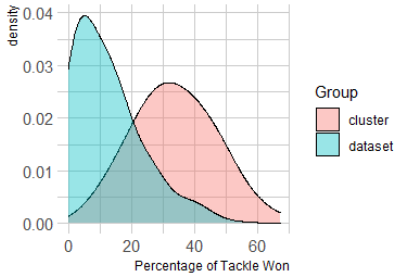

Figure 25: Tackle won

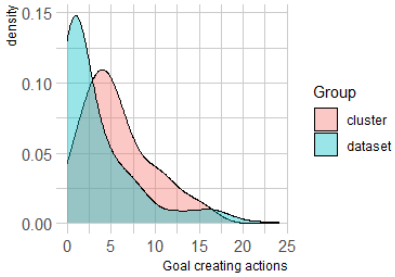

Figure 26: Goal creating actions

- **Centre-Forward**: The traditional role of a centre-forward is to score the majority of goals on behalf of the team. All 23 players in the cluster are forwards, the majority of athletes played more than 20 games in the last season. They score goals, but without helping the game-play of the team (looking at the number of crosses, it's possible to understand that they stay in the offensive penalty area, without moving on the board of the pitch to create dangerous situations or to assist teammates.

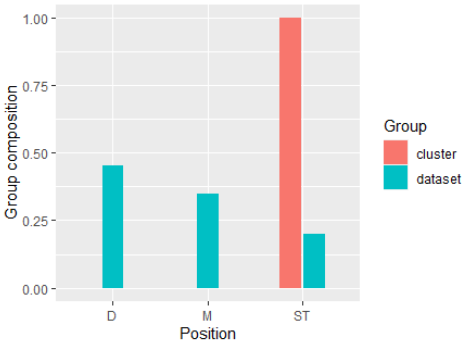

Figure 27: Position proportion in cluster

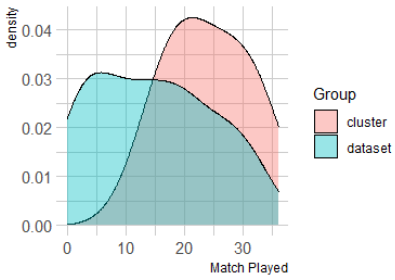

Figure 28: Matches played

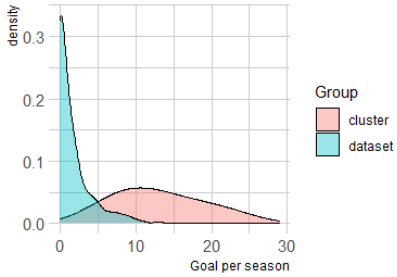

Figure 29: Goals per season

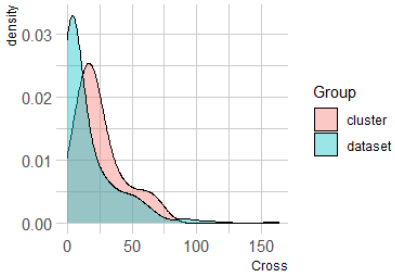

Figure 30: Number of crosses

In conclusion, grouping players based on 8 different cluster, allow to find very defined clusters. As expected, the classification created by the algorithm is similar to the one introduced before.
In fact, it is possible to observe the cluster "Full backs and Centre backs" which is a unification of the expected clusters "full backs" and "centre-backs".
This cluster is differentiated from "wings backs", in which is present a predominance of defensive players, but with inclination to create dangerous situation.
Two clusters for midfielders are presents: defensive/box-to-box midfielder and offensive midfielder. For the attacking positions, as expected, centre-forward and striker are the two created groups.
New groups are created with respect to the starting classification: low used offensive players and very low used players. In these cases, the algorithm splits players based on matches played, instead of the position on the pitch or others statistics.

4 Conclusions

The project results allow to conclude that the most relevant variables that determine the salary of a football player are the team for which he had signed, age, the ability to shoot on target, the number of completed passes during the season and the relative precision.
Grouping the residuals (difference between real and predicted salaries with regression) of players for each team, it is possible to understand if the teams' managers successfully managed the clubs and correctly evaluating the players' wages based on their expected performance. Moreover, a logistic regression is applied to data to compute the probability of finding a player in the 20 % of overpaid athletes for the season 2020/21, given the team for which that player had signed. U.S. Sassuolo and S.S. Lazio were the best managed clubs of the previous seasuon, while Cagliari Calcio was the worst one.
Hierarchical clustering is applied to create groups of similar players. This algorithm could provide a solution for the overpaid players and for poorly managed clubs. In fact, given the information extracted by clustering, a football club can easily find players that are very similar to overpaid players of the team, but with lower salaries.
The 8 clusters found are: centre backs/full backs, strikers, wing backs, low used offensive players, very low used players, attacking midfielder, defensive/box-to-box midfielder and centre-forward.
The accuracy of regression could be improved by considering also the number of individual trophies and team trophies won by a player. A winning/experienced player is, obviously, more paid. Furthermore, football players have high influence on the audience, and for this reason variables such as followers on social media must be introduced in the analysis.
To discover some needed players, clustering algorithms could be used, as in the previous analysis. The analysis done in this paper could be improved considering data of all European players and increasing the number of clusters, improving the quality of the research of players.