UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO MATEMATICA

Statistical Method
for High Dimensional Data: Report

In Pursuit of the Perfect Valuation: Development of a Regression Model for the Market Values of European Football Players

University of Padua

Department of Mathematics

Pietro Volpato	mat. 2079419
Gianmarco Betti	mat. 2097050
Simone Bianchi	mat. 2076677

Abstract

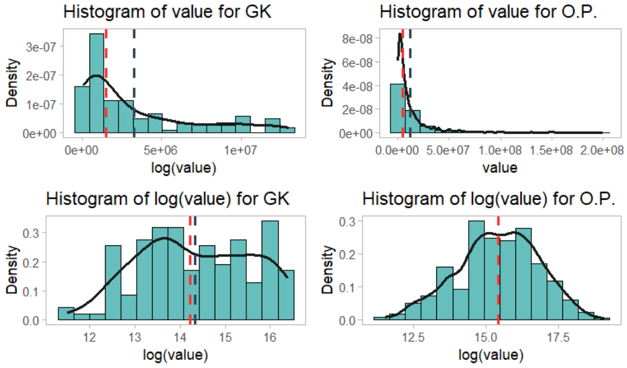
*The following project proposes a detailed analysis of a dataset obtained from **transfermarkt**, a specialized platform in football news and statistics, concerning **2232** football players from the top 5 European leagues. The dataset contains **399** features, primarily focused on on-field performances in the **2018-2019** season, supplemented by a section of **"Demographic"** data. The considered response variable is the player's market value, represented in the **"Value"** column. The main purpose of the project is to identify and analyze variables that have a **significant** impact on a football player's market value. To achieve this goal, the intention is to develop a **regression model** capable of predicting a player's market value based on the dataset's characteristics. Furthermore, this study aims to provide in-depth **insights** into the field of football player evaluations, thereby contributing to a better understanding of the determining factors in assessing market value.*

1 Data Visualization

With the aim of gaining a better understanding of the data structure, the distribution of variables and their relationships, a graphical analysis of the dataset's defining variables was conducted. The analysis took place only after a pre-processing and cleaning phase of the original dataset, where, as we will see later, it was decided to divide the initial dataset into two distinct subsets, concerning goalkeepers and outfields players. Now let's see how this graphical analysis unfolded.

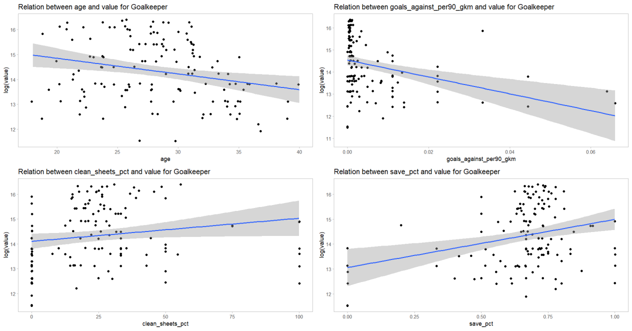
A. Response Variable Analysis

Following the creation of the two datasets related to goalkeepers and outfield players, the distribution of the variable *"value"* within these datasets was analyzed. It was observed that, in both datasets, this response variable exhibited a strongly left-skewed distribution, marked by a significant presence of outliers in the right tail. To address these issues, a logarithmic transformation was applied, and its efficacy was validated graphically.



B. Goalkeeper's Variables Analysis

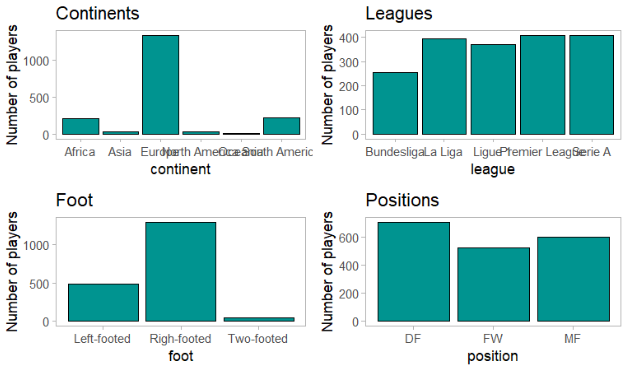
Regarding the graphical analysis of the independent variables relating to goalkeepers, we examined the distribution within the dataset of some of these variables and their relationship with the response variable, *"value"*. In fact, hypothesizing some variables that could have a more significant influence than others in predicting the market value, we attempted to visually confirm or refuse these hypotheses. From this check it was observed that variables such as *"age"*, *"goals_against_per 90_gkm"*, *"clean_sheets_pct"*, and *"save_pct"* had a significant impact on the response variable.



C. Outfield Players' Variables Analysis

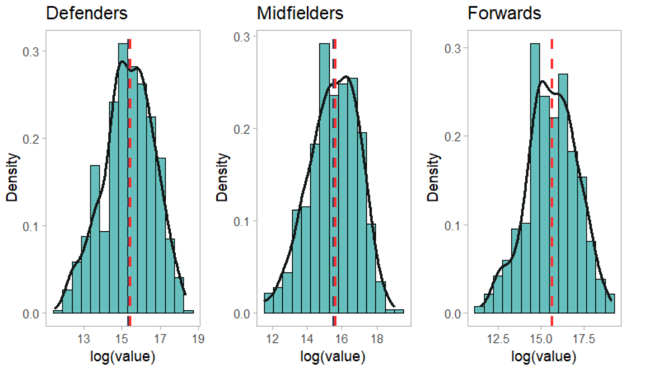
The following analysis took place in several steps:

1) **Categorical Variables:**
The distribution of the categorical variables *"leagues"*, *"continent"*, *"foot"* and *"position"* within the dataset was observed.

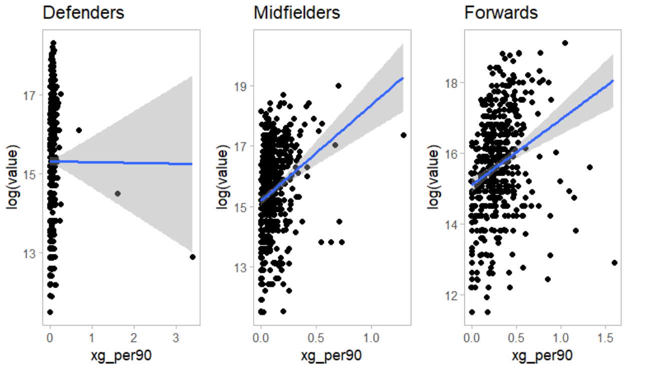


2) **log("value") Density Among the Categories:**
The distribution of the response variable among the previously mentioned variable categories was analyzed.

3) **Analysis of variables for each position:**
After further subdividing the dataset of movement players based on the roles defined by the "position" variable, we repeated the previous analyzes on the response variable and categorical variables for each of the roles.



4) **Relationships between variables and log(*"value"*) for each position:**
As previously done for goalkeepers, we analyzed the relationship between some independent variables present in the dataset and the response variable to visually understand their actual impact. This allowed us to make an initial observation of how these variables influence the response variable across the various positions into which we divided the dataset.



Subsequent statistical analyses will certify the actual influence of the variables.

2 Investigation

In order to examine the complex dynamics of soccer performance, we employ a dual approach by addressing outfield players and goalkeepers separately, recognizing the unique characteristics and demands associated with each position. This segmentation allowed us to delve the different sets of variables that influence the performance of both outfield players and goalkeepers.

2.1 Goalkeeper Analysis

A. Data Preprocessing

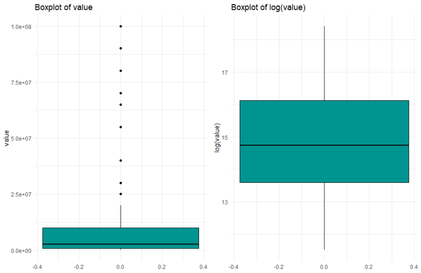
Data preprocessing plays a fundamental role in creating robust and reliable data analyses. In this chapter, we will briefly explore the various steps involved, such as handling missing data, removing outliers, and eliminating highly correlated variables.
Let's summarize the different steps in the following list:

1) **Data Cleaning and Organization Operations:**
Players with missing data in the *"value"* column are excluded. For specific players, whose missing data has been retrieved online, those values are integrated into the *"foot"* and *"height"* columns. Subsequently, the data is split into two distinct dataframes: one dedicated to goalkeepers, named *"gk_df"*, and the other focused on outfield players, called *"players_df"*. To ensure maximum clarity and relevance, some unnecessary columns are identified and removed from both dataframes (e.g., *"birth_year"*, *"nationality"*).

2) **Threshold-Based Discrimination:**
It is observed that the columns of the *"players_df"* dataframe where 95% of the values are zero concern the goalkeepers (e.g., *"pens_allowed*, *"pens_saved*), therefore they are transferred to *"gk_df"*. Initially, an identification attempt was made with a threshold of 90%, but this led to the inclusion of columns related to outfield players.

3) **Response Variable:**
Afterwards, we examined the response variable and we noticed that there existed a disparity between players with low values and those with high values. This creates influential points within our model that affect our predictions. To address this issue, we adjusted the response variable using the natural logarithm, a monotonic transformation that maintains the value order.

4) **Response Variable Outliers:**
A boxplot is generated to represent the distribution of goalkeepers salaries, and subsequently, outliers are removed in accordance with the Interquartile Range (**IQR**) criterion.



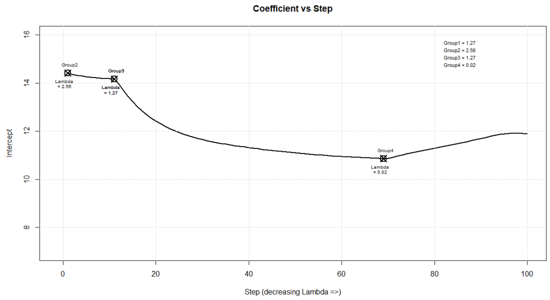
5) **Correlation Analysis:**
All variables with zero standard deviation are recognized and removed.subsequently, a correlation matrix is computed for goalkeepers-related information in which highly correlated variables are removed (with a correlation greater than 0.8).

6) **Data Finalization:**
The dataframe is further prepared for analysis, with the removal of categorical variables (Ultimately, we are left with 48 features). In addition, a vector called "group vector" is created, with the purpose of categorizing variables into specific groups in anticipation of a Grouped Lasso analysis.

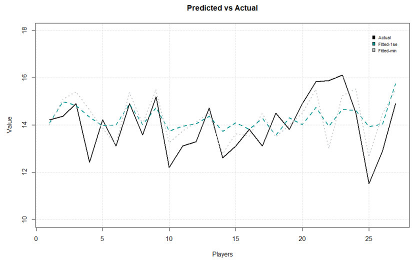
B. Implemented Models

In this section, we will provide a brief overview of the models used for conducting goalkeeper analysis. The following are the various implemented models:

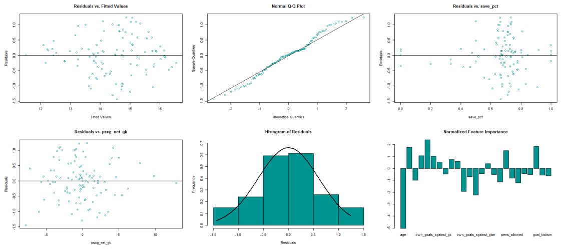
1) **Grouped Lasso:**
Implementation of the Grouped Lasso model on the data, *"gk_df"*, using the response variable *"value"* and a group structure defined by the group vector, *"group_vector"*. The next step involves searching for the indices of positions where the coefficients' rows are zero, followed by the visualization of a graph to represent the trend of the intercept in relation to those indices.



2) **Grouped Lasso Cross-Validation:**
Application of the Cross-Validation technique using the method *"cv.gglasso"* to the Grouped Lasso model with a subdivision into 10 folds. A comparison of the model results is then carried out considering the values of $\lambda.min$ and $\lambda.1se$ (one standard error).

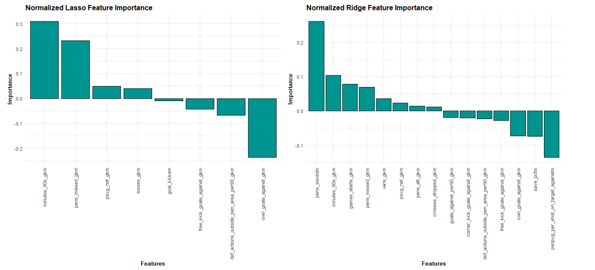


3) **Linear Regression:**
Implementation of the linear regression model on the data, *gk_df"*, using the response variable *Value"*. We proceed to generate graphs to evaluate the model and examine the residuals.



4) **Lasso Regression Cross-Validation:**
Application of the Cross-Validation technique using the *"cv.glmnet"* method to the Lasso Regression model ($\alpha = 1$), employing a suitable range for the values of λ to determine the optimal value. Subsequently, a plot is generated to evaluate the relevance of variables in the model.

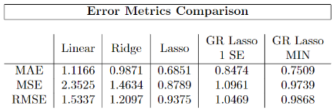
5) **Ridge Regression Cross-Validation:** Application of the Cross-Validation technique using the *"cv.glmnet"* method to the Ridge Regression model ($\alpha = 0$), employing a suitable range for the values of λ to determine the optimal value. Subsequently, a plot is generated to examine the importance of variables in the model.



For a better visualization of the barplot illustrating normalized features importance, we excluded features with a value less than 0.01 and greater than -0.01

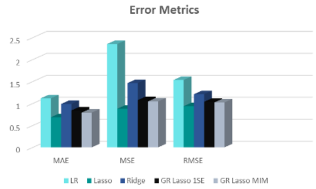
C. Results Comparison

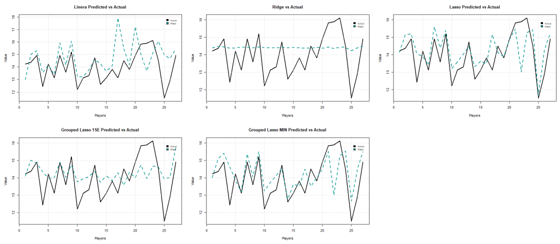
In the following section, we will examine the performance of different regression models and present the results through various error metrics, such as Mean Absolute Error (**MAE**), Mean Squared Error (**MSE**) and the Root Mean Squared Error (**RMSE**), along with scatter plots and predicted vs. actual plots.



Error Metrics Comparison

	Linear	Ridge	Lasso	GR Lasso 1 SE	GR Lasso MIN
MAE	1.1166	0.9871	0.6851	0.8474	0.7509
MSE	2.3525	1.4634	0.8789	1.0961	0.9739
RMSE	1.5337	1.2097	0.9375	1.0469	0.9868

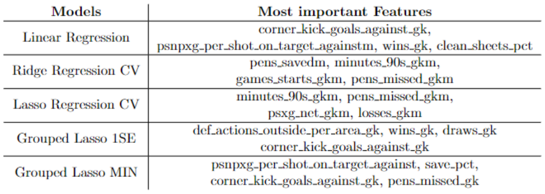




D. Features Importance

A crucial aspect of analysis projects is understanding the importance of different features in the dataset. In order to achieve this, we computed a normalized features importance, allowing us to identify which variables had the most significant impact on our models' response variable.

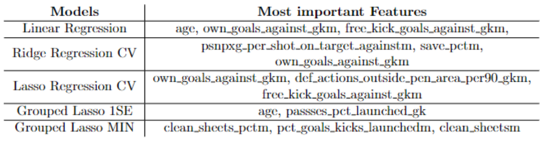
1) **Positive Impact:**



Models	Most important Features
Linear Regression	corner_kick_goals_against_gk, psnpxg_per_shot_on_target_againstm, wins_gk, clean_sheets_pct
Ridge Regression CV	pens_savedm, minutes_90s_gkm, games_starts_gkm, pens_missed_gkm
Lasso Regression CV	minutes_90s_gkm, pens_missed_gkm, psxg_net_gkm, losses_gkm
Grouped Lasso 1SE	def_actions_outside_per_area_gk, wins_gk, draws_gk corner_kick_goals_against_gk
Grouped Lasso MIN	psnpxg_per_shot_on_target_against, save_pct, corner_kick_goals_against_gk, pens_missed_gk

We can notice that the most frequently occurring features are: *"corner_kick_against _gk"*, *"wins_gk"*, *"minutes_90s_gkm"* and *"pens_missed _gkm"*

2) **Negative Impact:**



Models	Most important Features
Linear Regression	age, own_goals_against_gkm, free_kick_goals_against_gkm,
Ridge Regression CV	psnpxg_per_shot_on_target_againstm, save_pctm, own_goals_against_gkm
Lasso Regression CV	own_goals_against_gkm, def_actions_outside_pen_area_per90_gkm, free_kick_goals_against_gkm
Grouped Lasso 1SE	age, passses_pct_launched_gk
Grouped Lasso MIN	clean_sheets_pctm, pct_goals_kicks_launchedm, clean_sheetsm

We can notice that the most frequently occurring features are: *"own_goals_against_gkm"*, *"age"* and *"free_kick_goals_against_gkm"*

2.2 Outfield Players Analysis

A. Data Preprocessing

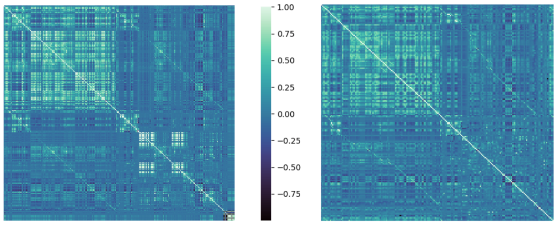
In order to analyze outfield players we created 3 different datasets, each containing only players belonging to a specific role **[Df, Mf, Fw]**. This was done after plotting the response variable and observing that each role had a different distribution of the value. This makes sense also from a "domain expertise" point of view, since what influences the value of a stricker is different from what influences the value of a defender.

1) **Data Cleaning:**
The first step on the *"players_df"* was to analyze the correlation between variables, their *"per_minute"* and *"_per90"* version. We kept both variables only if the correlation between the original and the redundant variable was lower than 0,7, otherwise we remove the latter.

2) **Correlation Issue:**
The subsequent significant problem in the *"players_df"* was related to the high correlation among variables inside the dataset. This correlation is natural, since we expect variables belonging to the same domain but carrying slightly different information (*"passes"* and *"passes_completed"*) to be highly correlated. The issue was solved mainly in three ways:

   - **Delete**: If variables presented redundant information already present in other features, we removed them.
   - **Merge:** Sometimes we created ratios between highly correlated variables to remove their dependency (e.g., *"passes_completed"*/ *"passes"*= *"passes_pct"*).
   - **Average:** If there were more than two correlated variables within the same domain, we calculated a weighted average to generate a score for that variable.

The correlation map before (left) and after (right) the data cleaning process is plotted below.



4) **Response Variable:**
Later we studied the response variable and observed that for each role there was a discrepancy between the number of players with a low value and players with a high one. This generates leverage points in our model and influences our predictions. To remedy this problem we transformed the response variable using the natual logarithm, a monotonic transformation that preserves the order of values.

5) **Categorical Variables:**
The dataset presented five categorical variables: *"nationality"*, *"foot"*, *"league'"*, *"team"* and *"position"*. The latter was used to create the three datasets, *"team"* was removed because it contained too many levels, while the other three were kept and transformed. From *"foot"* we create a binary 0-1 variable, for the five leagues present in *"league"* we created four dummies, while the 72-levels *"nationality"* variable was converted into the five levels *"continent"* variable (they converted into a dummy).

B. Implemented Models

The models implemented in the second part of the analysis were consistent with those used for goalkeepers, with three additional implementations:

1) **PCA:**
Given the higher dimensionality of this problem (more than 190 features) it made sense for us to use the principal component analysis to reduce the computational complexity of the problem. The threshold for determining the number of principal components was set at the cumulative explained variance of 90% and 95%, resulting in approximately 58 and 82 components, respectively. These numbers exhibited slight variations based on each role.

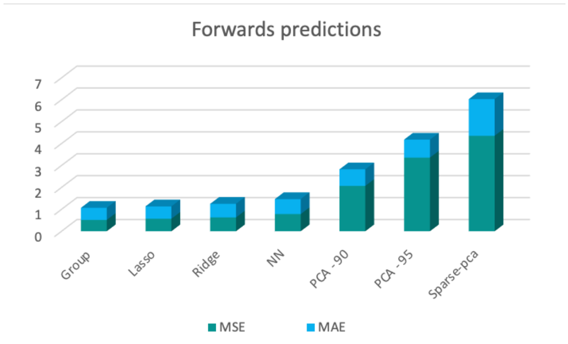
2) **Sparse-PCA:**
After testing traditional PCA, we introduced Sparse-PCA as a preprocessing technique to build our predictive models. We followed an explained variance Cross-Validation process to choose the most suitable α for our dataset. The range of tried values was between 0 and 0.15, with a value of 0.023 yelding the best results.

3) **Neural Network:**
We conducted experiments with a basic neural network architecture, featuring 192 neurons as input, 64 in the first hidden layer, 32 in the second, and 1 in the output layer. Our strategy involved incorporating both L1 and L2 regularization into the layers, along with a 10% dropout rate, and employing a ReLU activation function for all neurons.

C. Results Comparison

The criteria applied to goalkeepers were also maintained for outfield players, using the same set of four metrics: Mean Absolute Error (**MAE**), Mean Squared Error (**MSE**) and the Root Mean Squared Error (**RMSE**). However, in this case, we had three different datasets for training our models and three corresponding test sets for evaluation. It's important to note that the metrics are not comparable among different datasets.

1) **Forwards:**



The Graph plotted above represents the Mean squared error and the Mean absolute error of all the seven models we trained and tested on the forwards dataset. The process has been repeated also for the other two dataset but we will avoid plotting them here for redundancy purposes.

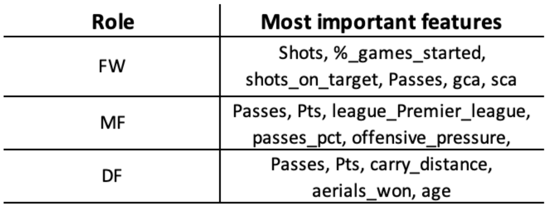
We can observe how the Grouped lasso and lasso regression outperform the others in the two metrics of interest.

2) **Midfielders:**
For the midifielders the situation is a bit different: the grouped lasso still outperforms the other algorithms but in this case the other five algorithms (excluded the sparse-PCA) do not show significant differences in performance. It is worth pointing out that for midfielders the PCA with 90 less components outperforms the one with more componenets.

3) **Defenders:**
When predicting defenders the group lasso looses its crown as the best predictive algorithm in favour of the PCA with more components. The ridge and the sparse-PCA are the algorithm performoing worsly while among the other four there is no significant difference.

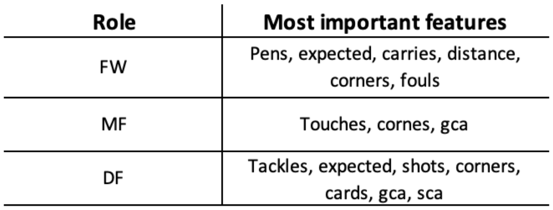
D. Features Importance

After training our model we can finally grasp which are the most important aspects building up the value of a player. As expected, for each role there are different coefficients contributing with different magintudes to the response variable. In the table below we can see the most important coefficients in the Lasso-regularized regression, divided per role.



Role	Most important features
FW	Shots, %_games_started, shots_on_target, Passes, gca, sca
MF	Passes, Pts, league_Premier_league, passes_pct, offensive_pressure,
DF	Passes, Pts, carry_distance, aerials_won, age

For Fw and Mf the results are coherent with the expectations, while for defenders the most important varaibles are not intuitive.
Now we can take a look at the groups that the grouped lasso selected as non-relevant in the regression per role.



Role	Most important features
FW	Pens, expected, carries, distance, corners, fouls
MF	Touches, cornes, gca
DF	Tackles, expected, shots, corners, cards, gca, sca

Here the different positions on the field show different groups with coefficient eqaul to 0, which are not directly connected with the role of interest.