Statistical Learning Project Report

Moneyball Serie A

William Biondi

Abstract

Association football is a popular sport, but it is also a big business. From a managerial perspective, the most important decisions that team managers make is to decide which players to buy to improve the team while staying in a given budget. Market values can be understood as estimates of transfer fees—that is, prices that could be paid for a player on the football market—so they play an important role in transfer negotiations. The project starts with the dataset of players from last Serie A season and their evaluation (source:Transfermarkt.com) playing statistics are combined from another source (fbref.com)

In this project I am interested to discover:

- Which football statistics affects more the evaluation of a player
- How can players be grouped without considering roles, squad or evaluations

This will be done with decision trees and clustering methods combined with dimensionality reduction.

Data Gathering

Starting from Transfermarkt platform player evaluations are obtained, also I joined detailed football metrics from fbref.com. At the end the dataset is composed by 513 players with these characteristics:

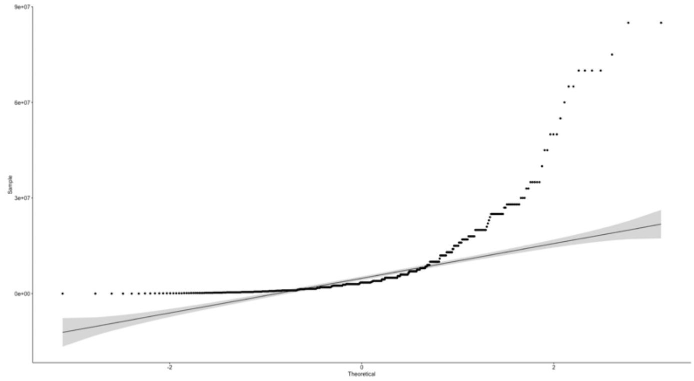
- Team
- Player Name
- Age
- Height
- Playing time (6 indicators)
- Attacking skills (6 indicators)
- Passing skills (7 indicators)
- Defending skills (8 indicators)

- Goalkeeping skills (6 indicators)
- Market evaluation in EUR

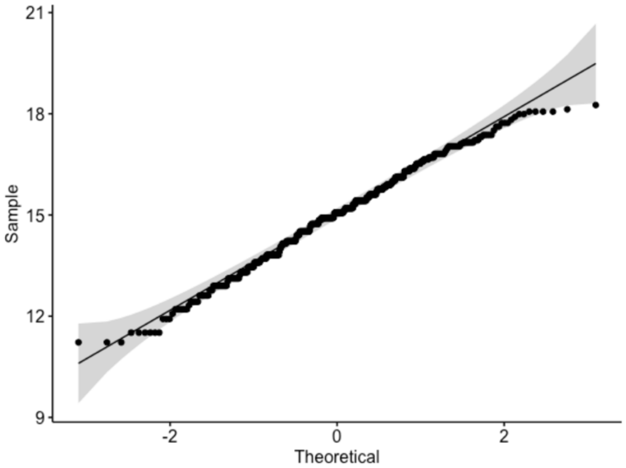
Market value analysis

Response variable

To start with we do a normality test to the response variable with the Shapiro-Wilk test since regression trees need the normality assumption, since the test gives us a p-value < 0.05 we reject the null hypothesis



With a logarithmic transformation, response variable is closer to normal

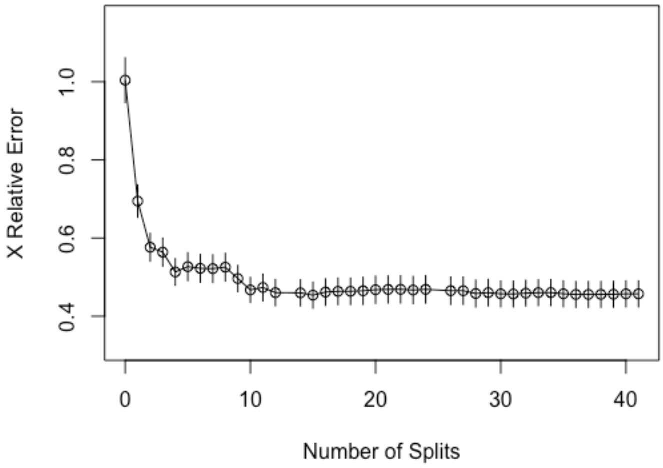


Decision tree

The algorithm of decision tree models works by repeatedly partitioning the data into multiple sub-spaces, so that the outcomes in each final sub-space is as homogeneous as possible. This approach is technically called *recursive partitioning*. I selected this method because the goal of this study is oriented on the interpretation rather than accurate prediction. The resulting tree is composed of *decision nodes*, *branches* and *leaf nodes*. The tree is placed from upside to down, so the *root* is at the top and leaves indicating the outcome is put at the bottom. Each decision node corresponds to a single input predictor variable and a split cutoff on that variable. The leaf nodes of the tree are the outcome variable which is used to make predictions.

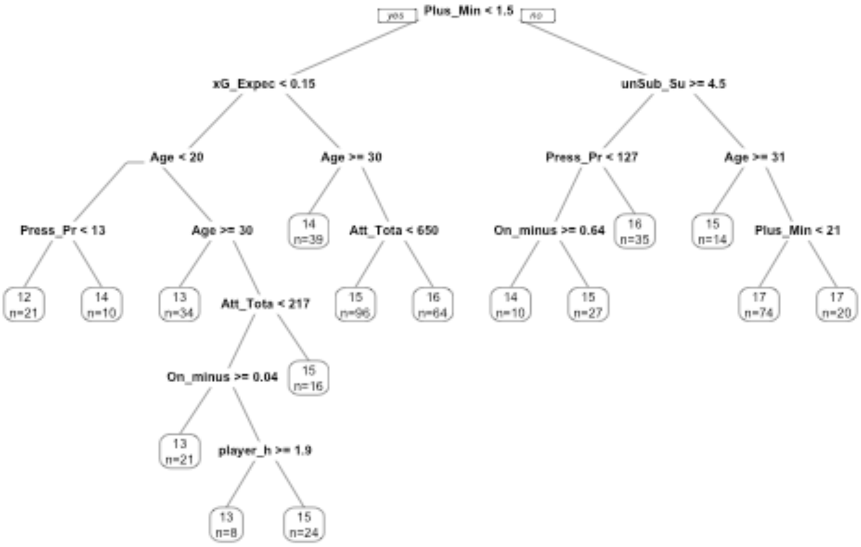
The tree grows from the top, at each node the algorithm decides the best split that results to the greatest purity in each partition.

So I started with a fully grown tree which uses 32 variables, then prune this tree to overcome overfitting with cross validation error, so I extract the best complexity parameter

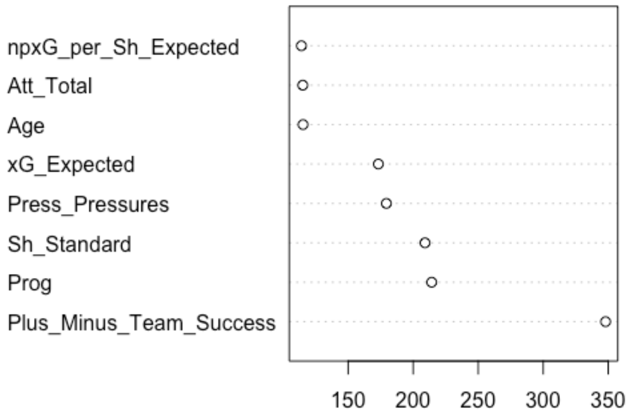


Results

The final tree uses 8 variables

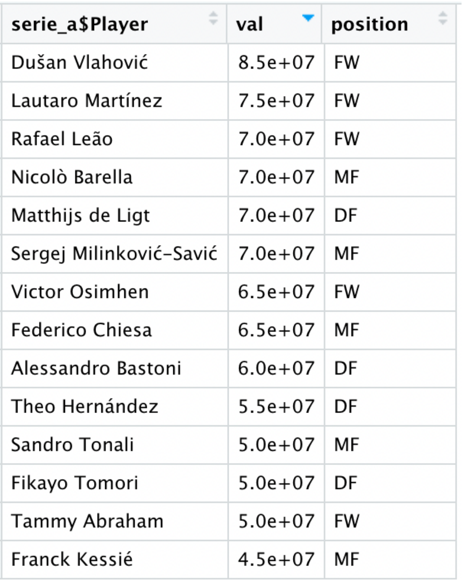


In this chart is reported the importance of each



So according to this model the impact of the player on team success is the most important, then progressive passes (passes towards attacking side) and shots. Then pressures (important to gain ball possession) and expected goals which is the probability of a shot given the position to be a goal (more on: https://towardsdatascience.com/modeling-expected-goals-a756baa2e1db). The final splitting are made by the Age, attempted passes and the last one is non-penalty expected goals per shot.

Since the most highest value players in the dataset are in the majority forwards this gives us a validation of the tendency to value more attacking skills (3 indicators in this model) rather than defending skills (pressures) and goalkeeping skills (no indicators)



serie_a$Player	val	position
Dušan Vlahović	8.5e+07	FW
Lautaro Martínez	7.5e+07	FW
Rafael Leão	7.0e+07	FW
Nicolò Barella	7.0e+07	MF
Matthijs de Ligt	7.0e+07	DF
Sergej Milinković-Savić	7.0e+07	MF
Victor Osimhen	6.5e+07	FW
Federico Chiesa	6.5e+07	MF
Alessandro Bastoni	6.0e+07	DF
Theo Hernández	5.5e+07	DF
Sandro Tonali	5.0e+07	MF
Fikayo Tomori	5.0e+07	DF
Tammy Abraham	5.0e+07	FW
Franck Kessié	4.5e+07	MF

Unsupervised section

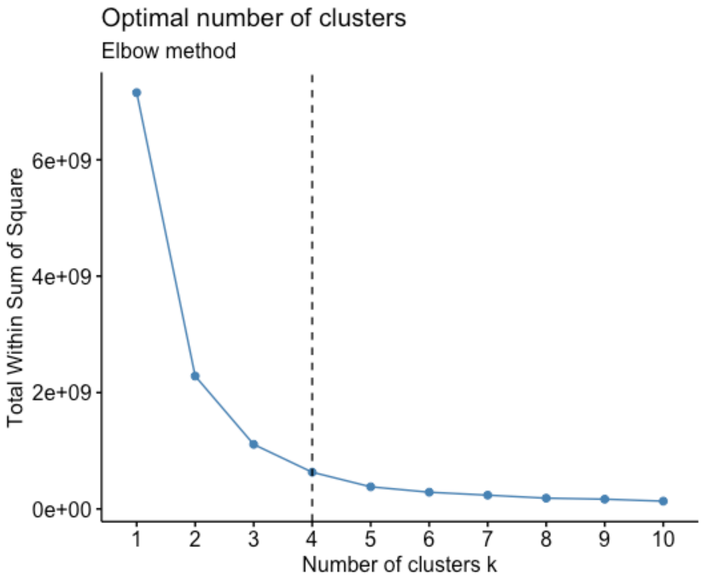
To start we delete target variables and other explicit clustering variables such as position and team. Then we apply Principal Component Analysis to project data into two dimensions.

The resulting components are strongly correlated with the following features

- PC1
    - % Playing Time (+)
    - Progressive Pass (+)
    - Pressures (+)
    - Blocks (+)
- PC2
    - Shot On Target Against (+)
    - Pass Distance (+)
    - % Saves (+)
    - Substitutions (-)

Clustering

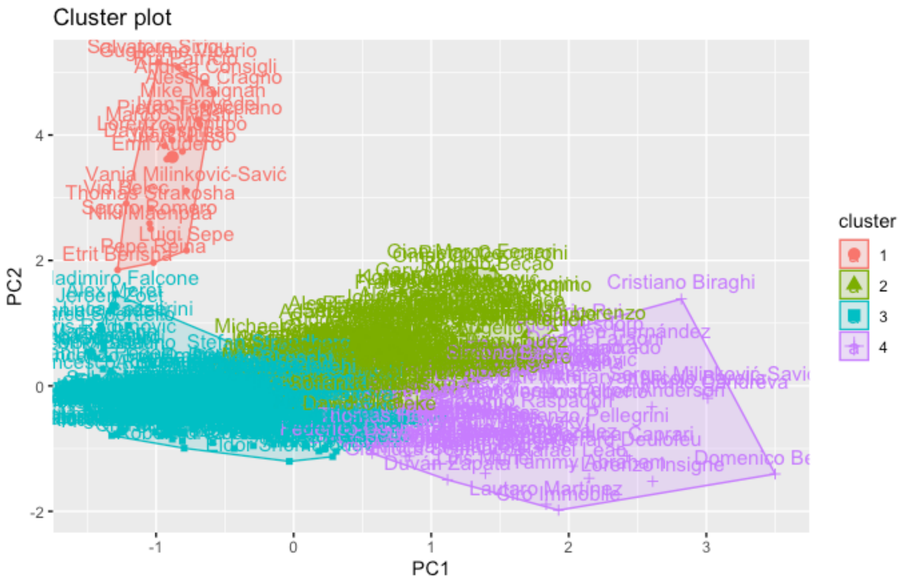
I used both k-means and hierarchical clustering methods to make comparison on how they split data of the players. Before applying both algorithms I compute the within sum of square error as a function of the number of possible cluster



From this chart I set K=4 which is the corresponding elbow point.

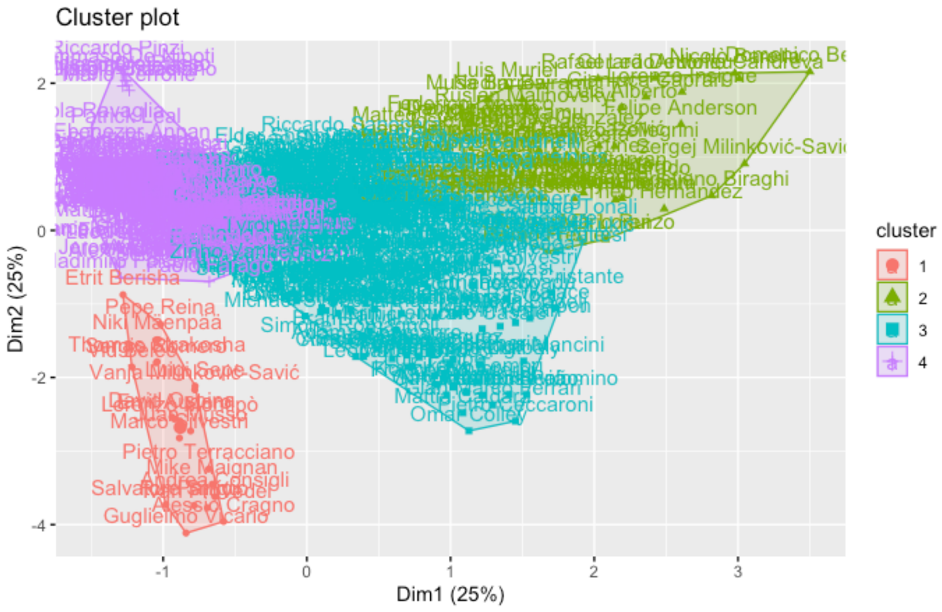
K-Means clustering

This clustering method results in an efficient distinction between goalkeepers and non-goalkeepers, this may be due to the fact that principal component explains 40% of variance. For the other three clusters, cluster 2 is made of defenders and midfielders with defensive characteristics, cluster 3 with low appearances players (14% of minutes played on average), cluster 4 with the majority of forwards the rest with midfielders and defenders with attacking skills.



Hierarchical clustering

This clustering method results similar to the k-means clustering with a good distinction between goalkeepers and non-goalkeepers, other clusters are mixed with majority of forwards in cluster 2, majority of defenders in cluster 3 and low appearances players in cluster 4



Conclusions

This study on football market confirms that player evaluations are driven by the impact that a player has on the pitch with his team and also with his attacking skills which could provide goals to win games rather than defensive skills which prevent the opponent team to score. However some players are evaluated by a contract clause which is manipulated by the club to insert a price barrier and tends to be higher than previous evaluations, on the other hand contract situations can manipulate prices down due to the possibility of the player to refuse a renewal and let the contract expire to be free-agent. Futhermore this model is not handling physical information in detail: some features like injuries, weight, speed, resistance, etc. may have a relevant impact on the evaluation of a player.

In the second part we can confirm that the position is not absolute, the only position that has clear differences with the others is the goalkeeper; football teams are not considering positions anymore they rather consider if the characteristics of a player are matching with the position they want to improve (ex. a midfielder with good defensive numbers can be used as a defender)