Predictive Model of Football Players Market Value with Gradient Boosting

Daniel González Vallejo

Universitat Politècnica de Catalunya

Abstract

This work tries to obtain a model that predict football players market value. To achieve this objective, I worked with databases with some statistics and characteristics of players of some European leagues since 2018 to 2020. Then, I made an exploratory analysis of the variables against the players market value. Finally, I built a model with gradient boosting and analyzed some particular cases of players.

1. Introduction

The football industry worldwide has become more important in recent years. It is one of the sports that generates more resources and greater affinity in the general public. The large amount of data collected from the matches has generated very powerful historical databases of event and tracking data. Because of this, data science has taken part in the analysis of games. So, it made possible to use several mathematical models to improve strategies, player market decisions and team performance.

Football as any other industry needs to reduce its costs and maximize its profits. So, it's known that prices of players are given for the market according to the current players situations. Then, predict these values would be an interesting achievement for data scientists in the way of trying to have a tool to measure its potential prices through the time. This would give more certainty to clubs when they will join any player in their team.

2. State of the art Analysis

Mahadevan S. (2020) built a model to predict the players market value and used machine learning model. He uses as features some statistics, characteristics and popularity variables for each player. So, he obtained a good model with an accuracy of 85% and with a RMSE of 5.800,00. He also had some problems predicting highest values of the players and the model made better predictions with low-medium prices.

He Y. (2014) made a model with principal component analysis. To build it, he uses as output the players market value and as features some statistics values of performance. Then, he obtained that the most important variable into the explanation of the price was the value of the last year.

3. Project Definition

Aim

- Find a predictive model of the players market value and explain the behavior of its features in terms of contribution to the value.

Objectives

- Analyze the relation between features and players market value.
- Find a model to predict the players market value.
- Find some applications for the model.

4. Methods

4.1. Data Management

To build the model I used some data bases of players, values, statistics and teams. For them I made different operations to make them usable. For statistics data base I calculated de summary of some features per year. Then, for players data base I create a variable to identify the players that played for the national team with an If-else clause. I created a variable that identify the positions into just 4: goalkeeper, forwards, midfielders and defenders. For the values I extracted the year and created a sub base with averages of the values per year. The next step was to merge all the bases treated and obtained the final data base. Finally, just made a one hot encode for categorial values and treated some values like NA's, and deleted some outliers in the minutes because were negative values and some high that make no sense compared with the sample.

4.1.1. Data Source

TransferMarkt

It's a German web site that has scores, results, transfer news, league schedules, and player values in football.

4.1.2. Variables/Features Selection

In this section in the table 1 I'm going to explain every variable that I used as predictor of the players market value. They were selected with the idea to make a simple and general model with basic information of the players. Some researches before made similar models with linear regression and machine learning models as He Y. (2014) and Mahadevan S. (2020) who used some statistics and marketing values for their models.

Table 1. Variables Definition

Variable/Feature	Type	Description	Categories	Origin
Age	Continuous	Age of the player	NA	TransferMarkt
Minutes	Continuous	Cumulated number of minutes played per year	NA	TransferMarkt
Log_lvalue	Continuous	Natural logarithm of the past player market value (t-1)	NA	TransferMarkt
lmkval	Continuous	Natural logarithm of the team market value	NA	TransferMarkt
Seleccion	Categorical	If player played for national team	1: Yes 0: No	Created - TransferMarkt
Pos_Defensa	Categorical	Player plays as defender	1: Yes 0:No	TransferMarkt
Pos_MedioCentro	Categorical	Player plays as midfielder	1: Yes 0:No	TransferMarkt
Pos_Delantero	Categorical	Player plays as attacker	1: Yes 0:No	TransferMarkt
Pos_Portero	Categorical	Player plays as goalkeeper	1: Yes 0:No	TransferMarkt

4.2. Data Analysis

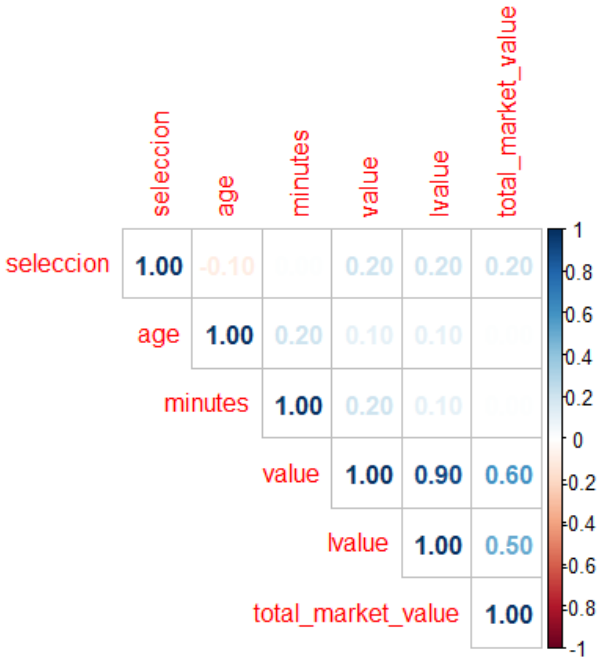
This part of the work I will show some statistics of the continuous variables to understand them little bit more. And, I will show some characteristics of the variables that make me thing that they are useful as predictors of the players value.

Table 2. Statistics of Variables

Variable	Mean	Median	Standar Deviation
Age	25	25	4.45
Minutes	1854	1684	1509
Players value	7.46M	2.25M	14.05M
Team market value	250M	163M	263M
lvalue	14.40	14.44	1.73
lmkval	18.74	18.91	1.21

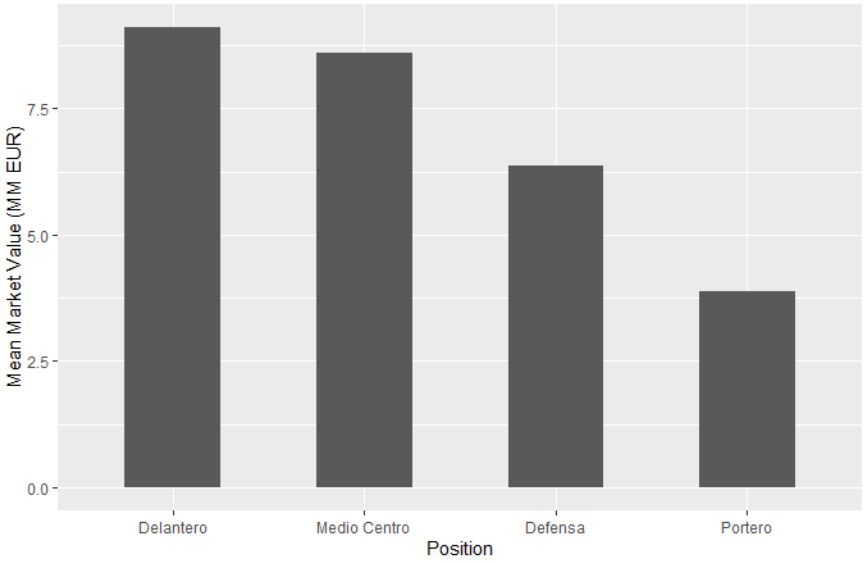
For age we can see that the mean value and median value are the same 25 and it has a standar deviation of 4.45. The mean of the minutes (1854) is higher than the median (1684) so it means that we have a bias to the right tail. For the players and teams values I converted them into logarithms because they had huge quantities. Also, values has the same kind of fistribution concentrated in lower prices as the table 2 shows with the values of the mean and median.

Figure 1. Correlation Plot



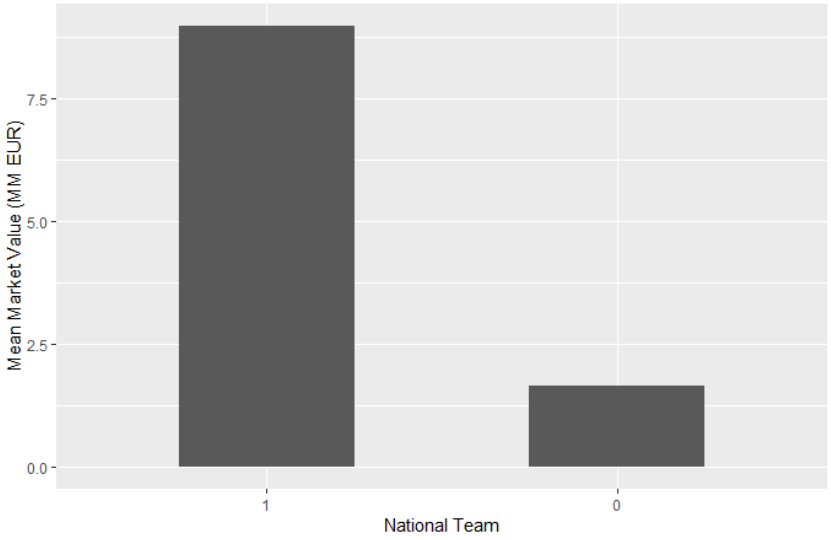
The correlation plot shows how variables are related with the pearson way. So, we can see that the most corelated variable with player value is total market value with a value of 0.6. Then, the other variables are corelated with the players prices with values of 0.2 and 0.1. It is a firs aproach to the relation between variables.

Figure 2.Players Market Value vs Positions



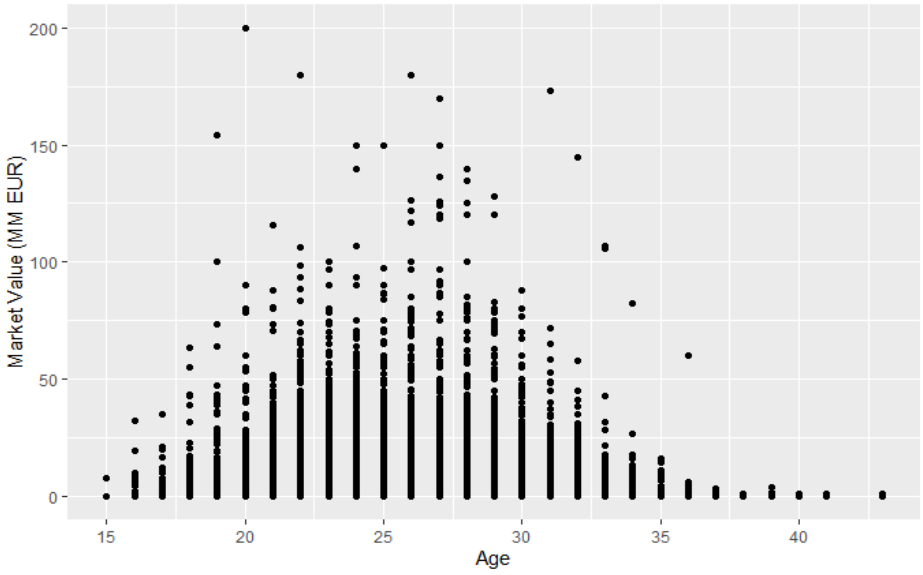
In the figure 2 we can see that forwards have the highest price over all the other positions with a mean price of 9.11M EUR, followed for midfielders with 8.6M EUR. These two positions have more mean value than the general mean value of the players (7.4M EUR). Finally, the lowest mean values are with defenders and goalkeepers with 6.3M EUR and 3.8M EUR respectively.

Figure 3. Players Market Value vs National Team



In the figure 3 we can see the difference between the mean prices of players that play for their national teams and who don't do it. Players that participate in international teams earn in mean 8.6M EUR and in the other side player who don't do it earn in mean 1.6M EUR.

Figure 4. Players Market Value vs Age



In the figure 4 we can appreciate that most of younger and older players tend to earn less than the player that are in a range of 20 to 30 years old.

In fact, with these figures we could understand that the variables selected could contribute for the explanation and prediction of the players values. We found that forwards are the players that earn more in average, players who play for their national team earn more against who don't and younger and older players tends to has lower prices.

5. Modeling

CatBoost is an algorithm that seeks the prediction of a variable through simple models in the first instance. Also, to build a more robust model it uses optimization, in this case the gradient descent model. From each model made, the algorithm takes into account the error of each model and minimizes it in its next prediction. This model works with non-linear models in this case the "symmetric decision trees", the final result is the aggregate of all the predictions. This algorithm was performed for the use of categorical values (Prokhorenkova, et al., 2018).

Objective Function:

$$\mathrm{L}(\phi) = \sum \zeta(\hat{y}_i - y_i) + \sum \Omega(f_k)$$

Where $\hat{y}$ is the predicted value and "y" is the real value, and

$$\Omega(\mathrm{f}) = \gamma \mathrm{T} + (\frac{1}{2})\, \lambda \|\mathrm{w}\| 2$$

The term ζ is the prediction error, Ω punish the complexity of the model to avoid overfitting.

The data was divided into 2 subsets, 80% for training and 20% for test. Also, I made a grid search for hyperparameters and obtained the best model with the following configuration in the table 3.

Table 3. Hyperparameters and configuration used in the model

Iterations	20000
Learning_rate	0.01
Depth	8
Loss_function	RMSE
Eval_metric	RMSE
Random_seed	0
Metric_period	200
Od_type	ITER
Od_wait	20
Verbose	TRUE
Use_best_model	TRUE

5.1. Results

To understand the performance of the model we are going to use the root mean squared error (RMSE) which was defined by Hyndman (2006) by the following equation:

$$RMSE = \sqrt{\frac{\sum(\hat{y} - y)^2}{T}}$$

Where $\hat{y}$ is the predicted value and "y" is the real value.

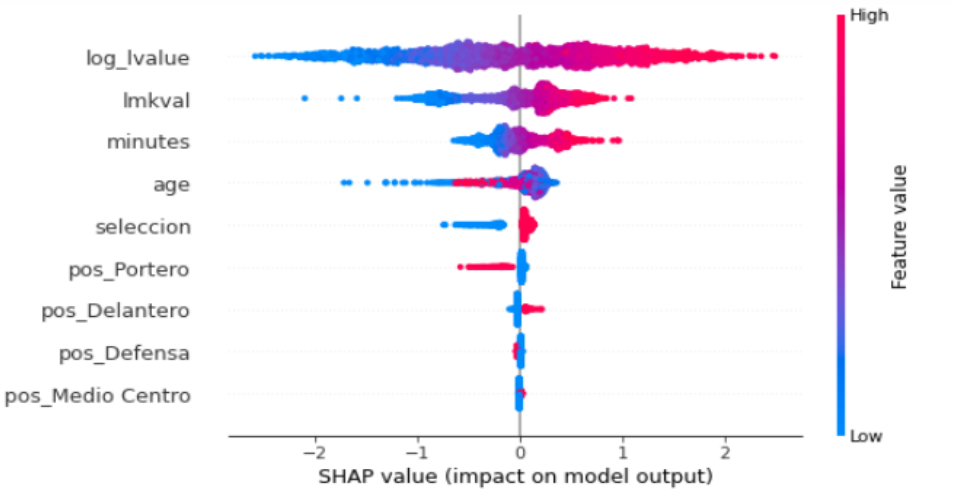
RMSE	Value LN	Value EUR
Test	0.65	5.6M
Train	0.59	5.1M

The RMSE obtained was 0.65 for the values in logarithms. To understand the amount, I transformed it into the original scale in euros and calculated again. The RMSE was about 5.6 million for test set. It could be taken as a measure of dispersion and accuracy of the data. Therefore, this error could be big for players with low values but not for players with high values. The model could be improvable for sure. But it's a good first approximation.

5.1.1. Shap Values

Shap value is a tool that help to quantify the contribution of every feature in the model. They were taken from the game theory from the cooperative games. The idea is that there are some players (features) that participate in a game (prediction). So, each feature that joins in the coalitions generate a change in the prediction. Then, the mean of this changes is considered as the shaply value (Monlar, 2022).

Figure 5. Shap values and Features



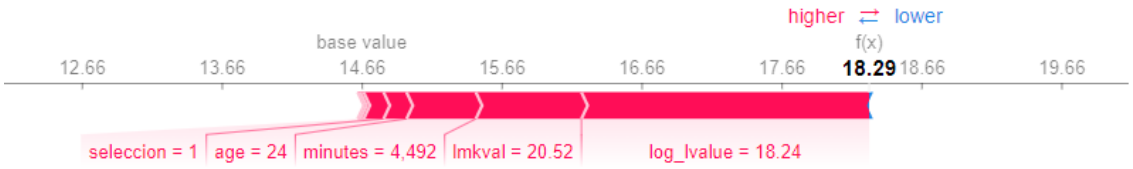
The figure 5 shows the most important features in the SHAP values criteria. First, we have the logarithm of the previous player's price. It has a positive relationship, when the values in this feature are high, they have a positive impact over the model output. And when the previous value was low it has a negative impact over the output. The same relationship with the team's market values and minutes. Then, the age has a different behavior because it's not too obvious to generalize the impact. But the impact could be negative with high values of age, but more negative impact with youngest players. For dummy variables "seleccion" and "pos_Delantero" this analysis shows that high values (1: Yes) have a positive impact in the output. Then, Goalkeepers and defenders have a negative impact. Finally, for midfielders the impact is not clear, but it tends to have a positive relation.

5.1.2. Single Applications

In this section I will show some examples of predictions two specific players.

Frenkie de Jong:

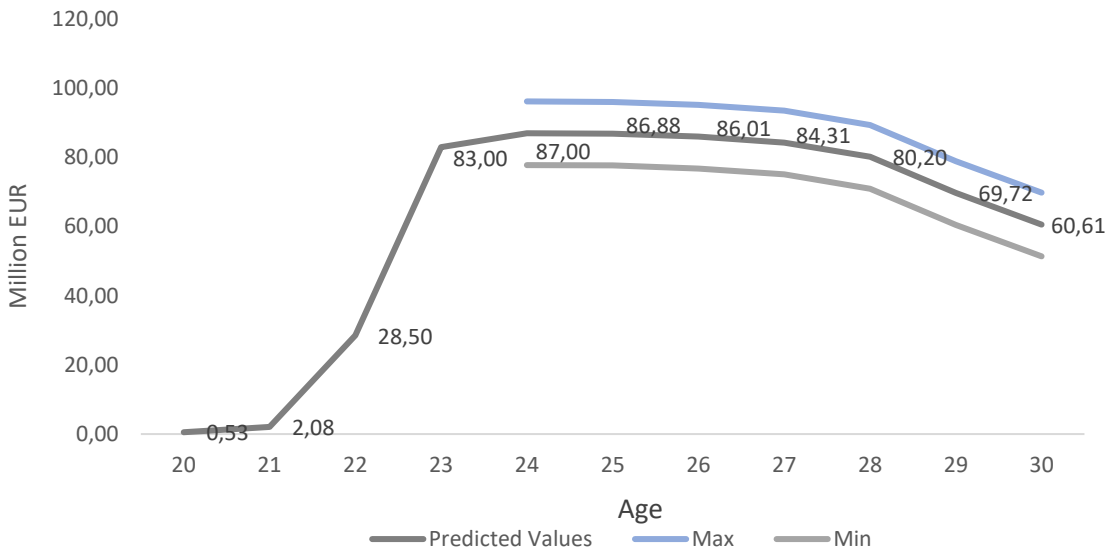
Figure 6. Shap Value Explanation FDJ



Variable	Mean
National Team	Yes
Age	24
Minutes	4492
Team Value	20.52 (8.1 Bill.)
Last Value	18.24 (83 Mill.)
Value	18.29 (87 Mill.)

The figure 6 describe the contribution of each feature to the player market value. So, the features that contribute more in this case are the previous market value and the team value. At the end he obtained a prediction of 87 Mill. EUR.

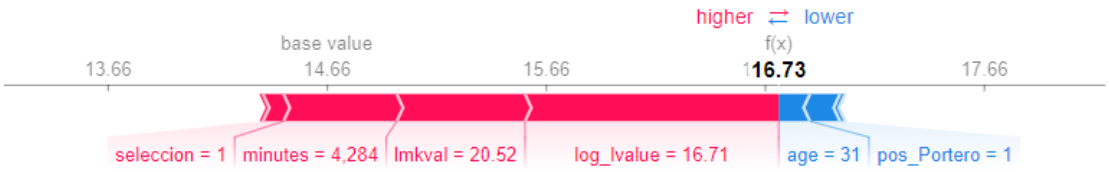
Figure 7. Market Value Projection FDJ



In the figure 7 we can see the evolution in the time of the predicted market value. The highest value he will achieve will be at age of 24 (87 M) and then it will decrease until 60 million of euros at age of 30.

Neto:

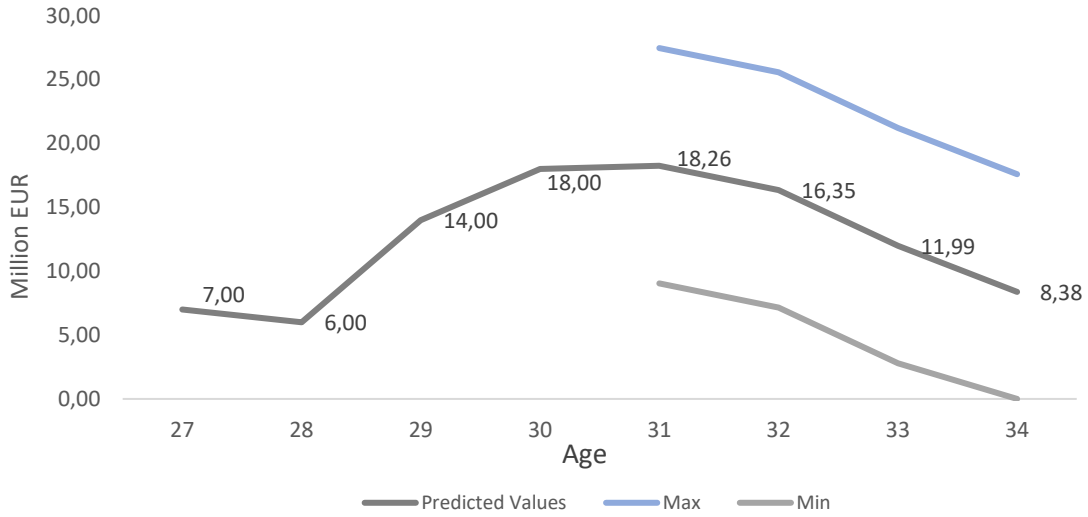
Figure 8. Shap Value Explanation Neto



Variable	Mean
National Team	Yes
Goalkeeper	Yes
Age	31
Minutes	4284
Team Value	20.52 (8.1 Bill.)
Last Value	16.71 (18 Mill.)
Predicted Value	16.73 (18.4 Mill.)

The figure 8 shows that previous value and team value are the most important features to predict the price. Besides, the age and the position (goalkeeper) reduce his predicted value.

Figure 9. Market Value Projection Neto



In the figure 9 we can see the evolution of the predicted prices of Neto. For the model, he will achieve the highest value at the age of 31 (18.26 M.) and it will decrease until 8.4 million of euros at the age of 34.

6. Limitations

The model has problems to predict highest values, and with younger player it tends to assign low values. So, it is what the model learned of the database but it could be fixed adding other features. The value of the RMSE was high and could be a problem for player with low values. Some features like the player popularity, interactions and other statistics that might help to improve the model. Finally, could be interesting using other methodologies like neural networks of deep learning.

7. Reference

Hyndman, Rob J.; Koehler, Anne B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting 22 (4): 679-688.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, Andrey Gulin. NeurIPS, (2018). CatBoost: unbiased boosting with categorical features. 1Yandex, Moscow, Russia. Moscow Institute of Physics and Technology.

Mahadevan S. (2020). Predicting Market Value of Football Players using Machine Learning Algorithms. Bournemouth, United Kingdom. Bournemouth University.

Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/

Hen Y. (2014). Predicting Market Value of Soccer Players Using Linear Modeling Techniques.