BrainStation Data Science Bootcamp

Final Report

on

Capstone Project –

Predicting Football Players Market Value based on their performance

Author: Nazhad Bari

Introduction

The transfer market in Football world is considered one of the most competitive and expensive transactional activities in the world. Football club Chelsea spent around € 200 million in 2021. Top clubs spend large sums of money in the hopes of winning matches, Cups and the hearts of millions of fans. So clubs need to be very careful when they do a transaction in the transfer market. This project focuses on predicting the player market value and the factors that affect this value.

Problem Statement

This project will help answer the following two business questions:

- Using Machine Learning can we predict the Market Value of Football Players based on their on-field performance?
- What are the key predictors in determining this value?

Value Add

Clubs spend millions of dollars in buying potential players at the Transfer Market. But often after the signing, the player in question underperforms. So, the question is what truly dictates a player’s market value? Is it the on-field performance of that player, is it the player’s persona or is it their ability to generate sponsorship revenue that dictates the player’s true valuation? This model will help clubs who wants to be at the top of their leagues by efficiently spending in the transfer market by focusing on what is valuable for them – their on-field success.

Data Collection and Cleaning

Data has been collected for all players from Europe’s top 5 leagues for 10 years from 2011 to 2021. The leagues are - ***English Premier League, Italian Serie A, Spanish La Liga, German Bundesliga*** and ***French Ligue 1***. The datasets have been acquired from 2 different websites:

- Transfermarkt dataset - containing player information and their Transfer Market Value.
- fbref datasets - containing player on-field performance standard statistics data (10 datasets from the years 2011 to 2021 joined together to create one dataset).

The two datasets have been joined together to create one complete dataset that contains 3 different types of data:

- Player information like their name, age, nation, on-field position, current team, previous team, Football league, year the market value was recorded, transfer fee amount, market value, etc.
- Player performance statistics like goals scored, assists made, non-penalty goals, matches played, etc. It also contains per 90 minutes performance statistics that represent the statistics based on per game.
- Player potential performance statistics like expected goals, expected assists, etc.

The target variable of this project is the “***Market Value amount in EUR***” converted to millions. A player's Market Value is an estimate of the amount for which a team can sell the player's contract to another team.

My dataset had a lot of missing values in the columns player age, nation, transfer fee and even on my target variable, Market value. I manually inputted the missing values with the correct information from the Transfermarkt website. For the Market Value column, I manually filled hundreds of missing values from the Transfermarkt website and then for the remaining missing values I filled them with the median value of the Market Value. I removed all the duplicates from the dataset to avoid redundancy and renamed the columns for better understanding. Finally, my dataset had around **8438 rows** and **38 columns**. Among the 38 columns, there were 29 numerical variables and 9 categorical variables.

EDA and Feature Engineering

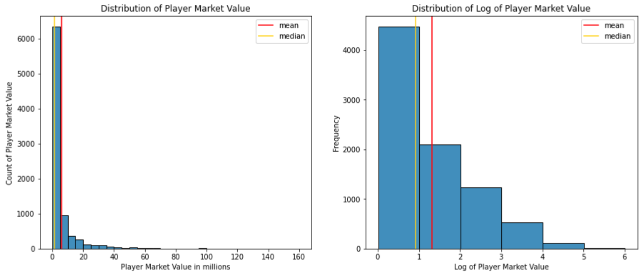
After the data cleaning I did some data exploratory analysis and feature engineered 3 new features. The first engineered feature is the "***Player position group***" where I split the player on-field positions into 4 groups - *Attackers, Defenders, Midfielders* and *Goalkeepers*. The second is the "***Player Nationality***" in which all the player nations have been grouped together by continents: *Europe, North America, South America, Africa, Australia*, and *Asia*. My last engineered feature is the "***Buying team group***" where all player current teams have been split into 4 groups with respect to their total spending in the transfer market.

Data Preprocessing and Modeling

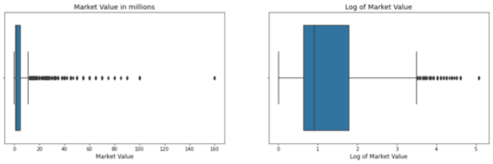
My dataset had 8438 rows and 41 columns after the feature engineering. Among the 41 columns, there were 30 numerical variables and 11 categorical variables. In order to perform modeling I would need to convert the categorical variables into numeric ones. I created dummy variables from the "**Player position group**" and converted the "***Player Nationality***" column into a binary one with '1' representing European player and '0' representing non-European players. I removed the unnecessary columns as they don't add much value. Then I instantiated my independent and dependent variables. My dependent variable is the "Market Value in millions" as previously mentioned and I had 33 features as my independent variable. I split the data into train and test sets with 67% - 33% split between them. The train set had 5653 rows and the test set had 2785 rows.

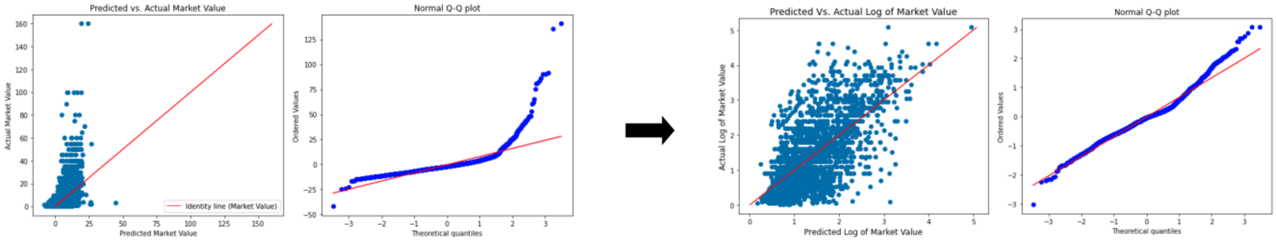
I created a baseline model for my analysis by running a Linear Regression model on the data. This would act as a measuring matrix and would help me in comparing and finding the best model. For model evaluation, ***R-squared*** metrics has been used as a relative measure of goodness of fit to measure the model accuracy and the ***Mean Absolute Errors*** is used to calculate the average absolute value of errors between the true Market Value and the predicted Market Value to measure the model performance.

My target variable is very imbalanced and skewed. I had a lot of outliers in the data which was affecting my model performance. So I transformed my target variable by taking the logarithm of the "Market Value in millions" in an attempt to linearize the data. This would help me eliminating any negative value predictions as well as help address the issue of overfitting and skewness.



After the transformation the data was less skewed than before and the trend also looked more consistent. My model prediction looked much better than before and even the Q-Q plot was linear.

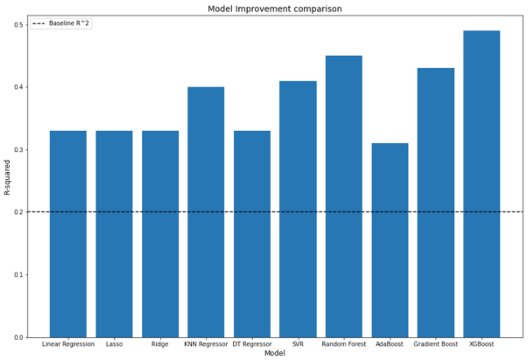


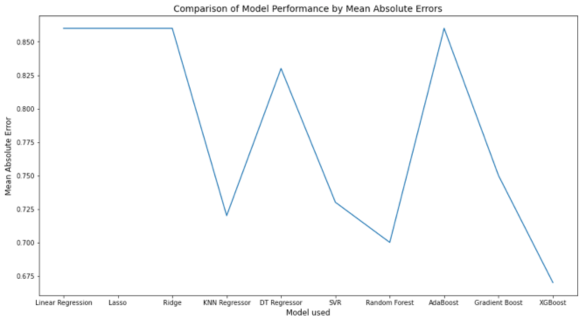


10 different models have been used to help predict the player Market Value. The results for each model have been compared to identify which model worked best to predict player Market Values with the highest R^2 and the lowest Mean Absolute Errors. Gridsearch was used and pipelines were run for hyperparameter tuning.

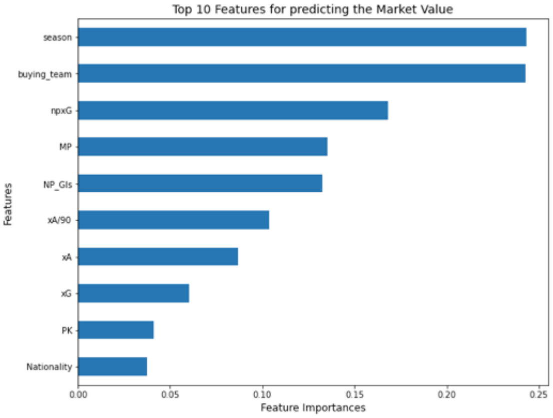
Findings

Among all the models, the XGBoost Regressor had the best accuracy score and performance with the highest R^2 and lowest MAE. The model had an error of around € 0.67 million with the model able to predict the Market Value best. Therefore, this model is good for predicting the Market Value of elite players who tend to have a higher Market value.





I also found out the key predictors that determines the Player Market Value. The features that had the most positive effect are season/year the Market Value was recorded, current team and Nationality the player belongs to. This is because the teams add a certain degree of brand value to the players and Europeans tend to have a higher market value than other nationalities. Player on-field performance like the number of matches played, non-penalty goals, penalty kicks made and player potential performance statistics like non-penalty expected goals, expected assists and expected goals had a strong positive effect on the market value. Features that had the most negative effect include player age, number of yellow cards and red cards received. As the player gets older, their stamina decreases, their performance deteriorates and since they can't perform well the demand to have them play in a club decreases along with their Market Value. A higher number of red and yellow cards could imply a disciplinary issue of character.



Conclusion and Next Steps

There were some limitations to my project. First, I had data for only a limited number of players as I only used the players from 5 leagues. Secondly, I only had a limited number of features as my independent variables. Also my model is good for predicting the Market value of elite players who have a higher Market value. Moreover, based on my findings, I believe that to accurately predict a player market value, only their on-field performance data may not be enough. According to the Transfermarkt website, they use information on the player’s skill, potential, recent performance, injury record, contract duration, position and others to estimate an accurate transfer fee. Factors like media image, social media coverage and what is called the “superstar status” or popularity of the player play a crucial role in determining the Market value. Players like Neymar Jr. and Cristiano Ronaldo have a big media presence and can attract more sponsorship deals than others. Therefore, a player’s image outside of the Football world and their crowd-pulling power have a huge impact on the market value.

For my next steps, I would like to go deeper and include more features like contract length of the player, media coefficient and increase the number of players by taking data for all leagues.