Football Market Predictor

Ege Demir
ITU - AI and Data Engineering
150200319
demireg20@itu.edu.tr

***Abstract*—The aim of the project is to predict the market values of football players by using a ridge regression model. A tabular dataset [1], which is taken from the game FIFA 23, that consists of general information about players (age, team, position etc.) and, various scores that gives insights about strong and weak sides about their styles.**

***Index Terms*—ridge regression, market, football, score**

I. Introduction

Background: Because of the complication of determining a football player's market value, teams can waste their capital on overpriced players (Ousmane Dembele to FC Barcelona, and Harry Maguire to Manchester United are some famous examples.), or they can make underpriced sellings and lose their most valuable resources without knowing.

So, every team has to accurately predict the values of their players and transfer targets.

Overall Purpose: Predicting the market values of football players is the main purpose of this project. After data cleaning and feature extraction processes, ridge regression is used to train the model.

The proposed solution, ridge regression model makes sense with this dataset, because the correlation is high between the features.

II. Related Work

In this section, I will mention 3 papers:

Firstly, Müller et. al. (2017) [2], try to predict football players' market value, same as my project; but they only use real data, rather than artificial scores of players from a game, unlike myself. They combine 3 clusters of data: player characteristics (age and height), player performance (goals , yellow cards per season; and passes, interceptions, dribbles per game etc.), and player popularity (Wikipedia page views, Google trends search index, Reddit posts, and YouTube videos). And they use market values from Transfermarkt (https://www.transfermarkt.com/) as labels. In the modelling process, multilevel regression analysis, which fits their hierarchically structured data.

On top of the data they have, they add another data cluster as random effects, which includes information about the player's league, position, team, nationality etc. and I think this cluster is quite influential. The main problem of their data, and their project in general, is the limitedness of the player performance data. I think this problem is inevitable when working with real football statistics, because they don't provide much information, and that is the main reason I chose scores from a game which probably will acquire better results. Nevertheless, this project is significant in my opinion, because it brings league and position information's strong effect to light; also it shows the complexity of the problem.

Secondly, Stanojevic and Gyarmati (2016) [3], aprroach the same problem with a similar dataset as the previous project. Player information and performance features are almost the same, but they don't include the popularity attribute. Another difference is the handling of the team information. They add average score of the team players, and average score of the opponent's players and use them as coefficients of the performance indicators for each game; rather than simply using the team name.

For the model, after experimenting with various supervised learning methods, they decide on linear regression, and I think that is a mistake. As Müller et. al. (2017) [2] stated, the features of the dataset are unlikely independent, and linear regression, which assumes independent features, can't be used. For example, assists per minute is most likely affected by key passes per minute. In general, this paper is not significant as the paper of the previous one, in my opinion.

Last but not least, Behravan and Razavi (2020) [4], attracted me the most, because their dataset is nearly same as mine. They work on the data from FIFA 20, predecessor of FIFA 23, which is the origin of my data. They divided their data to 4 clusters: goalkeepers, defenders, midfielders, and forwards; and I don't think that is an efficient way. For example central defensive midfielders are more alike to central backs (which classified as defenders), rather than central attacking midfielders. 4, is not enough, or too much clusters. I think there should be 8 (gk, rb/lb/rwb/lwb, cb, cdm, cm, cam, rm/lm/rw/lw/rf/lf, st/cf), or 2 (goalkeeper or not). Goalkeepers must be handled as a separate case, because their performance scores is nothing like a player from another positions.

At the modelling part, they use particle swarm optimization, where they start with random parameters, and update them at the end of every iteration, and repeating this process until stopping criteria is met. I think that's an interesting idea, but a simple ridge regression model would give better results, but I still think this is a crucial paper for one reason: it proves that artificial scores from FIFA, and any game in general, can be accurate enough to work on.

III. The Dataset

- The dataset, is a matrix, and it's from a .csv file. It has 18540 rows and 90 columns.
- Each row represents a football player, and each column carries a piece of information about that player. While some columns are about general information about players (age, team, position etc.), others are various scores (agility, strength, positioning etc.) from FIFA 23, one of the most popular football games of all time.
- The market value column which is called "ValueEUR" will be the label, and the other columns will be features.
- Source: https://www.kaggle.com/datasets/cashncarry/fifa-23-complete-player-dataset (players_fifa23.csv)

Head of the dataframe can be seen below:

	ID	Name	FullName	Age	Height	Weight	PhotoUrl	Nationality	Overall	Potential	...	LMRating
0	158023	L. Messi	Lionel Messi	35	169	67	https://cdn.sofifa.net/players/158/023/23_60.png	Argentina	91	91	...	91
1	165153	K. Benzema	Karim Benzema	34	185	81	https://cdn.sofifa.net/players/165/153/23_60.png	France	91	91	...	89
2	188545	R. Lewandowski	Robert Lewandowski	33	185	81	https://cdn.sofifa.net/players/188/545/23_60.png	Poland	91	91	...	86
3	192985	K. De Bruyne	Kevin De Bruyne	31	181	70	https://cdn.sofifa.net/players/192/985/23_60.png	Belgium	91	91	...	91
4	231747	K. Mbappé	Kylian Mbappé	23	182	73	https://cdn.sofifa.net/players/231/747/23_60.png	France	91	95	...	92
5	209331	M. Salah	Mohamed Salah	30	175	71	https://cdn.sofifa.net/players/209/331/23_60.png	Egypt	90	90	...	90
6	192119	T. Courtois	Thibaut Courtois	30	199	96	https://cdn.sofifa.net/players/192/119/23_60.png	Belgium	90	91	...	34
7	167495	M. Neuer	Manuel Neuer	36	193	93	https://cdn.sofifa.net/players/167/495/23_60.png	Germany	90	90	...	47
8	20801	Cristiano Ronaldo	C. Ronaldo dos Santos Aveiro	37	187	83	https://cdn.sofifa.net/players/020/801/23_60.png	Portugal	90	90	...	87

IV. Proposed Work

For the problem of predicting market value of players, I am proposing a ridge regression method which I will explain in detail at the modelling section but firstly, I have to talk about data preprocessing steps.

A. Data Preparation

First of all, I separated the goalkeepers and the players to 2 different datasets, since they can be thought as a factor that increases noise to each other's model, for the reasons I explained above. (The model for the goalkeepers was not present when I wrote the intermediate report, that's the main improvement.) Then I removed the unnecessary columns, such as: name, growth ([potential - overall], remove cause is redundancy), kit number, goalkeeping skills.

Secondly, I converted categorical columns, to numeric ones. I divided 'nationality' column to 4: top 5 nationalities, nationalities from 5 to 10, nationalities from 10 to 20, and others(0,1,2,3). Then I did exactly the same thing for 'national team' column (converted 'not in a national team' to -1). After that, I created 17 new columns (1 for each position), to numerize the features: positions, best position, club position, nationality position. I increased every position column by 1 for each position mentioned at the original columns (they were 0 at first). Then I numerize the binary columns: 'Preferred Foot', 'On Loan' (made them 1 and -1); and 3 class features 'Attacking Work Rate', 'Defensive Work Rate' (made them 0,1,2). For the last numerization, I handled the 'Club' feature, by dividing the teams to 4 categories (elite teams, sub-elite teams, average teams, others) according to their UEFA rankings (I took initiative for some clubs I thought as under-ranked or over-ranked). Then I simply convert them to 0,1,2,3 like other cases.

Lastly, I smoothed out 'Contract Until' and 'Club Joined' columns by subtracting the min year from them (min is 2022 for the 1st and 2002 for the 2nd one). And after the normalization by mean and standard deviation, the data is ready for the model.

B. Modelling

As I mentioned at the project proposal, my first choice was polynomial regression; but after examining the data thoroughly, I changed my mind. I noticed that some features highly depend on other ones. For example, features like 'Shooting Total', 'Defending Total', and 'Passing Total'; are just the average of some other features. And 'Overall' feature is average of all of them, multiplied by some coefficient according to player's position. Not to mention the correlation between 'nationality' and 'national team'. For that reason, I abandoned polynomial regression, which assumes independent features.

As R. Hoerl (2020) [5] mentions, expansion of regression coefficients' variances, is the main issue in collinear regression problems when usage of least squares estimation is on display. A. Hoerl, Kannard and Baldwin (1975) [6], proposes to solve this problem with a biasing parameter 'k'.
In R. Hoerl's words [5],"Ridge Regression shrinks the estimated coefficient vector towards the origin along a path (...) The user then selects the appropriate value of k." Ridge Regression does all that with the formula:

$$\widehat{\boldsymbol{\beta}_R} = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\boldsymbol{y},$$

The formula is taken from the paper of R. Hoerl (2020) [5].

I should also declare that I used cross validation to maximize the efficiency of my limited dataset.

I used RidgeCV function of scikit-learn library for acquiring the R^2 score, with this parameters: alphas=[0.001, 0.01, 0.1, 1, 10].

And I used RepeatedKFold function of scikit-learn library for acquiring the MSE and MAE scores, with this parameters: n-splits = 10, n-repeats=5.

V. Experimental Results

The hypothesis we are testing is this: scores from FIFA 23 can be used to evaluate football players' market value, and a ridge regression model is an appropriate way to do so. In order to test this hypothesis, we have to compare the predictions we made using our model, with the ground truth. For this, I evaluated my model with 3 metrics: mean squared error (MSE), mean absolute error (MAE), coefficient of determination (R^2 method)

A. Score Tables

Let's see the results for the players dataset:

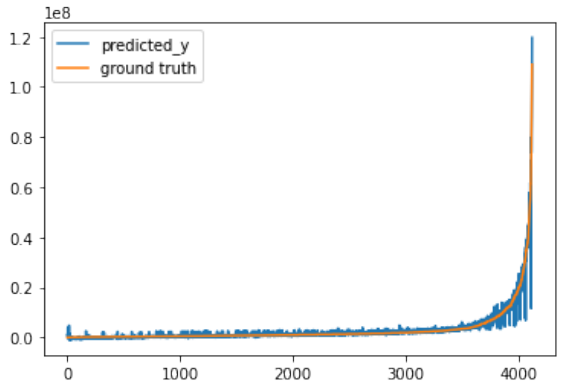
Method	Score
MSE	0.98
MAE	0.42
R^2	0.2

Let's see the results for the goalkeepers dataset:

Method	Score
MSE	0.98
MAE	0.35
R^2	0.28

B. Graph

Let's try to visualize the results:



This graph includes the predicted value, and the actual value of every player at the test set. Blue plot is predicted ones and orange plot is actual ones. We can say that, more the 2 plots overlap more accurate results we got.

C. Prediction Table

Here is a table, where the predictions and the ground truth values of the top 10 players of the game (Keep in mind that footballers are sorted on overall score):

Name	Predicted Value (Million Euros)	Ground Truth (Million Euros)
Lionel Messi	54	56,8
Karim Benzema	64	77,1
Robert Lewandowski	84	93,7
Kevin De Bruyne	107,5	105,9
Kylian Mbappé	190,5	178,5
Mohamed Salah	115,5	107,8
Cristiano Ronaldo	41	46,3
Virgil van Dijk	98	93,3
Harry Kane	105,5	103
Neymar Jr	99,5	92,8

For comparison, here is a table of predictions and ground truth values of top 10 players at the paper of Stanojevic and Gyarmati [3], like the one I did to evaluate my model:

Player name	TMVE (M £)	PDMVE (M £)
neymar	60.0	69.55
eden hazard	52.5	57.72
cesc fabregas	37.5	49.54
sergio aguero	45	48.81
lionel messi	90	48.53
nolito	7.5	46.14
luis suarez	60.0	40.65
thomas muller	41.25	38.08
marco verratti	30.0	36.07
diego costa	37.5	35.79

Table III
Top 10 players according to PDMVE. TMVE and PDMVE are in Millions of GBP (£).

As we can see clearly, results of our model is slightly better.

VI. Conclusion and Future Work

The experimental results indicate that our model works fine, and the fact that our predictions are better than the previous project, supports that idea.

I think it's safe to say that, a ridge regression model with cross validation gives satisfactory results for mean squared error, mean absolute error, and R^2 evaluation metrics for indicating football players market value problem.

However, there still exist a lot of work to do in the future for this problem. Other regression methods and classification models can and should be used to make sure which method is the most suitable one.

REFERENCES

[1] https://www.kaggle.com/datasets/cashncarry/fifa-23-complete-player-dataset

[2] O. Müller, A. Simons, ve M. Weinmann, “Beyond crowd judgments: Data-driven estimation of market value in association football”, European Journal of Operational Research, vol. 263, no 2, pp. 611-624, 2017, doi: 10.1016/j.ejor.2017.05.005.

[3] R. Stanojevic and L. Gyarmati, ”Towards Data-Driven Football Player Assessment,” 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016, pp. 167-172, doi: 10.1109/ICDMW.2016.0031.

[4] Behravan, I., and Razavi, S. M. (2020). A novel machine learning method for estimating football players’ value in the transfer market. Soft Computing. doi:10.1007/s00500-020-05319-3

[5] R. W. Hoerl, “Ridge Regression: A Historical Context”, Technometrics, vol. 62, no 4, pp. 420-425, 2020, doi: 10.1080/00401706.2020.1742207.

[6] Arthur E. Hoerl , Robert W. Kannard Kent F. Baldwin (1975) Ridge regression:some simulations, Communications in Statistics - Theory and Methods, 4:2, 105-123, DOI: 10.1080/03610927508827232