Can We Use FIFA Videogame Data in Soccer Analytics?

Statistical Learning Project

Alvise Dei Rossi[1] - Lorenzo Corrado[2] - Riccardo Vinco[3]

Department of Mathematics "Tullio Levi-Civita"
MS in Data Science
University of Padua

A.Y. 2020/2021

[1]ID: 2004250
[2]ID: 2020623
[3]ID: 2005800

Overview

1. Introduction
2. Exploratory Data Analysis
3. Model Estimation
4. Applications
5. Conclusions and Future Works
6. References

Introduction

Introduction

- Soccer is the most popular sport in the world, both for the number of players and for the number of spectators. The soccer industry is worth about $471 billion in 2018 and predictions say it will be worth about $600 billion in 2025 [2];
- Despite the huge following that this sport has and the large number of spectators it is still very difficult to make predictions, both for the absence of adequate datasets and for the intrinsic difficulty in modeling events in this sport [7];
- Given the enormous success that this sport has over time, several simulation video games have been developed.

Introduction - II

- FIFA 20 is the most popular football video game, developed by EA Sports, available for the major videogame consoles. The video game is distributed all over the world and sold, in the year 2020, about 115 million copies for a billionaire turnover.

Introduction - III

- To ensure a high quality simulation, EA Sports uses a large number of scouts for the evaluation of the characteristics and attributes of players from all over the world, but this characterization of the players is a difficult job;
- The aim of this project is to establish if there is a relationship (and how strong it is) between the player's in-game statistics with the real world;
- We believe that in-game statistics can also be a very important data resource for many statistical analysis. applications in soccer.

Introduction - IV

- A professional soccer player is assigned a series of statistics, more than 20, that are supposed to be representative of his in-real characteristics and consider all the major leagues in the world;

- The quality of the estimation of these characteristics is fundamental in a soccer simulation game. Player stats are therefore expected to be consistent with the real world.

Exploratory Data Analysis

Exploratory Data Analysis

- The FIFA 21 dataset we used contains 18,944 players, on which 106 variables were measured. For example, players are represented by:

	sofifa_id	short_name	age	height_cm	weight_kg	nationality	...
1	158023	L. Messi	33	170	72	Argentina	...
...				...			...
16	202126	H. Kane	26	188	89	England	...
...				...			...
18944	257936	Song Yue	28	185	79	China PR	...

Exploratory Data Analysis - II

Quantitative variables in the dataset

Age	Height	Weight	League rank	Overall	Potential
Value	Wage	Reputation	Physic	Crossing	Finishing
Dribbling	Curve	Fk accuracy	Long passing	Ball Control	Acceleration
Sprint speed	Agility	Reactions	Balance	Shot power	Jumping
Stamina	Strength	Long shots	Aggression	Interceptions	Positioning
Vision	Penalties	Composure	Standing tackle	Sliding tackle	Gk Diving
Gk handling	Gk kicking	Gk positioning	Gk reflexes	Heading	Short passing
Volleys					

Categorical variables in the dataset

FIFA ID	Weak foot	Skill moves	Short Name
Nationality	Club name	League name	Player positions
Preferred foot	Work rate	Team position	

Exploratory Data Analysis - III

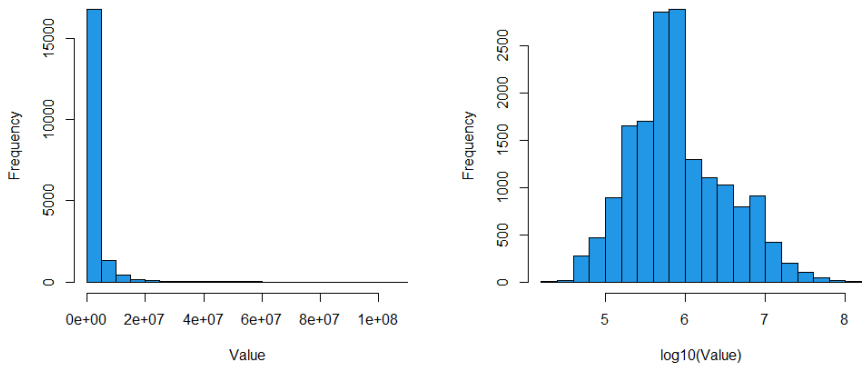


Figure: Value distribution before and after log transformation.

Exploratory Data Analysis - IV

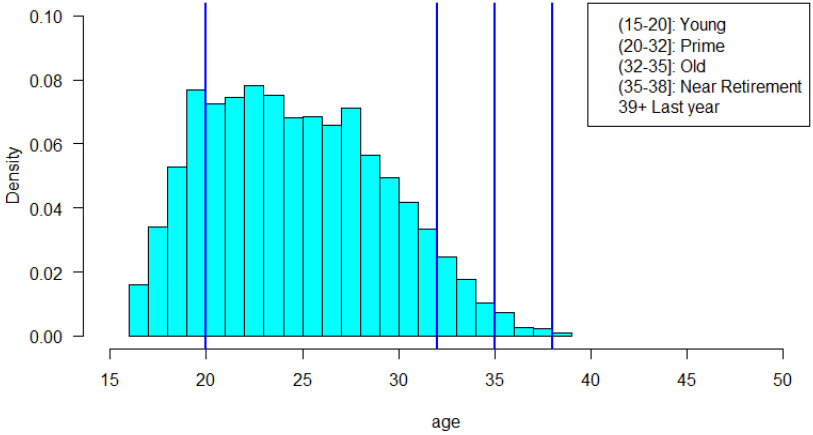


Figure: Age distribution and player career phases.

Exploratory Data Analysis - V

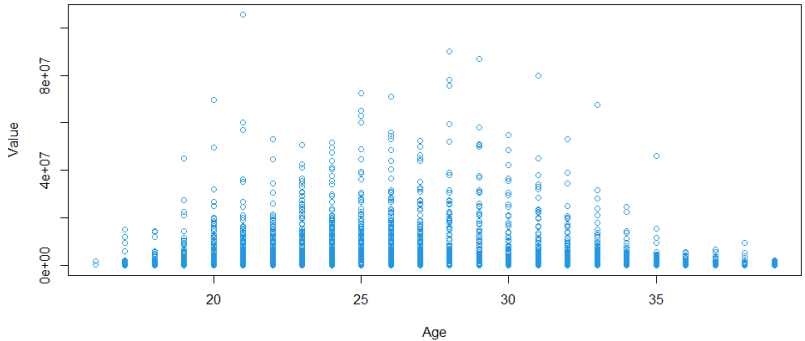


Figure: Relation between age and value of a player.

Exploratory Data Analysis - VI

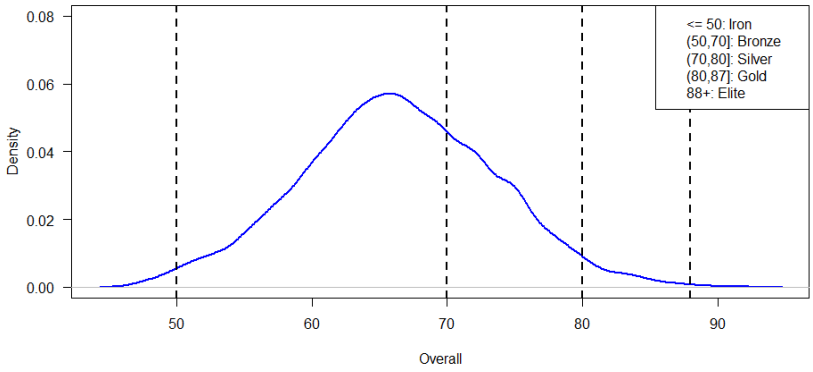


Figure: Overall distribution and categories of players.

Exploratory Data Analysis - VII

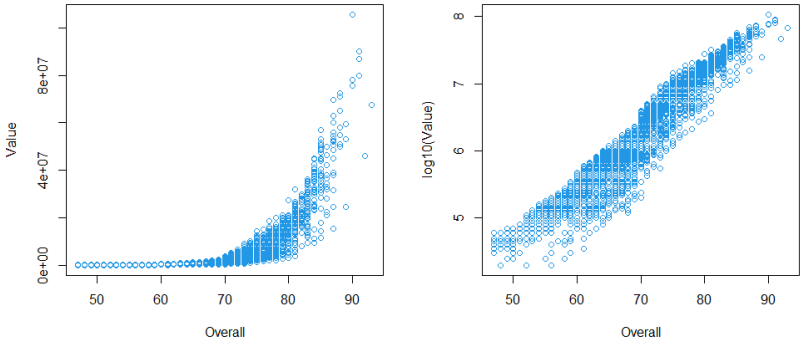


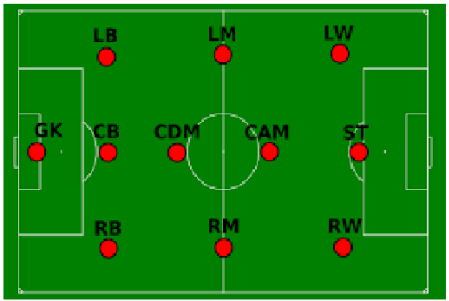
Figure: Relation between overall and value of a player.

Exploratory Data Analysis - VIII

Name	Value	Age	Overall
K. Mbappé	105500000	21	90
Neymar Jr	90000000	28	91
K. De Bruyne	87000000	29	91
R. Lewandowski	80000000	31	91
S. Mané	78000000	28	90
M. Salah	78000000	28	90
V. van Dijk	75500000	28	90
R. Sterling	72500000	25	88
H. Kane	71000000	26	88
P. Dybala	71000000	26	88

Table: Value and basic stats of the Top10 valued players.

Exploratory Data Analysis - IX



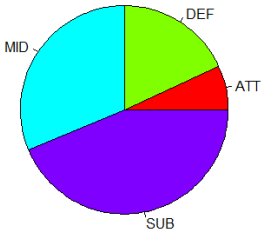


Figure: Positions of players and summary of the in-game positions.

Exploratory Data Analysis - X

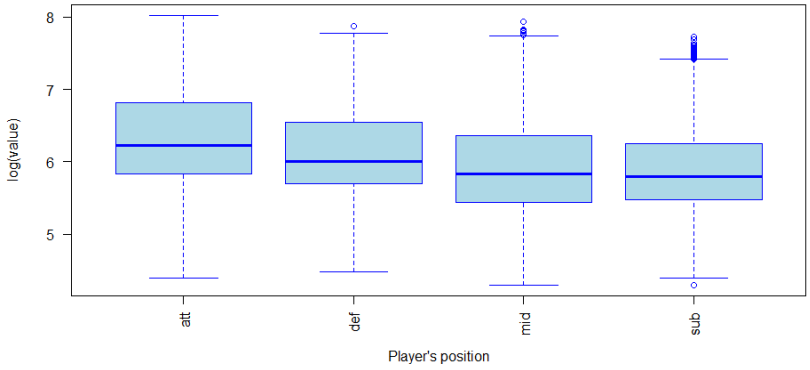


Figure: Players value by main position.

Exploratory Data Analysis - XI

- Is the league in which an athlete plays important? Usually the best players compete in Europe:

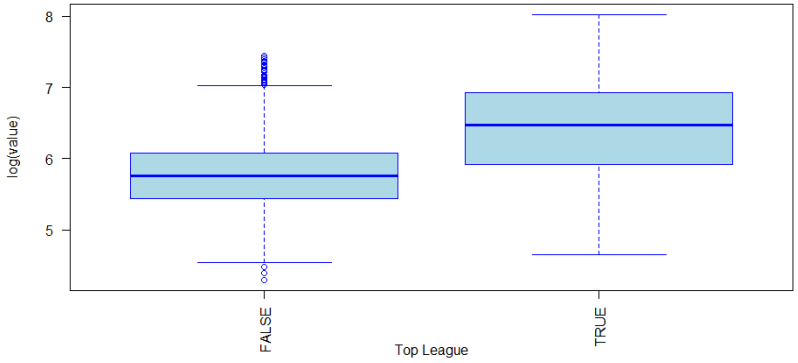


Figure: Influence of the league played on the value of a player.

Model Estimation

Models Estimation

- Can we model a player's value based on his FIFA statistics? If so, what's the effect of the different stats?
- We'll consider simple linear models:

$$Y = \alpha_0 + \alpha_1 X_1 + ... + \alpha_n X_n$$

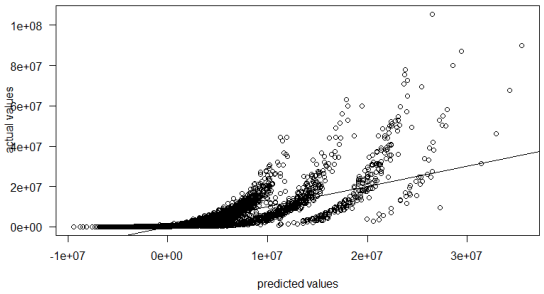
- At first only intuitive variables are considered for estimation:

Quantitative variables
Overall
Potential
Age
International Reputation

Table: Predictors in the first model.

Models Estimation - II

- A simple linear model based on these stats isn't enough:



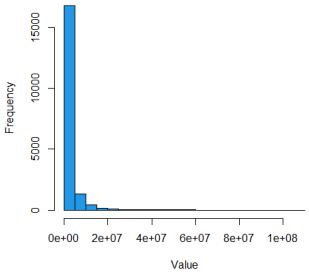
- The errors of the model show clearly that it's not able to predict well the value of the players and that there are other factors to consider to evaluate them. Some players are even evaluated with a negative value which is absurd;
- So how can we change the model?

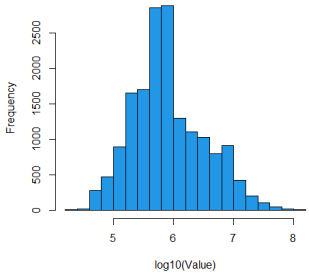
Models Estimation - III

- The huge difference in scale for the value of the players considered (starting from 20k€ to over 100M€!) made it really hard for the previous model to be accurate;
- It's much better to consider a logarithmic scale for the problem!

$$log_{10}(Y) = \alpha_0 + \alpha_1 X_1 + ... + \alpha_n X_n$$

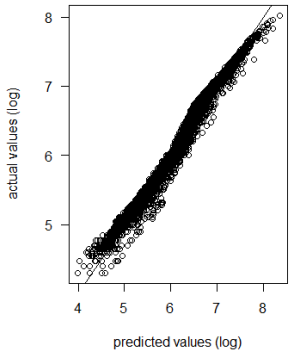
So the value distribution will change like:

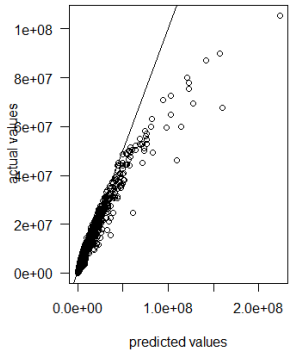




Models Estimation - IV

- Just by applying the logarithmic transformation, using the same intuitive variables as before, the model improves substantially:





Models Estimation - V

- It's clear from the residuals that the errors are not distributed uniformly. There are clearly other patters in the data we were not able to capture adequately;

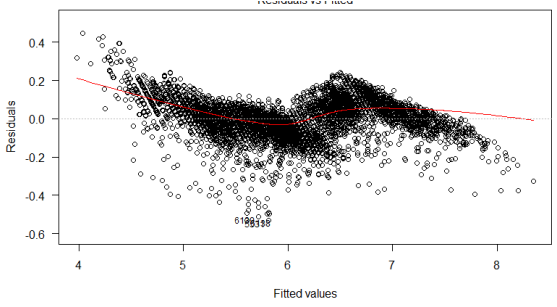
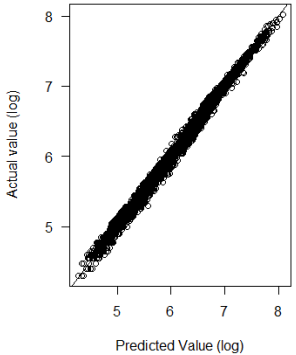


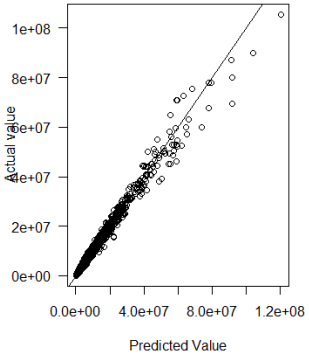
Figure: Residuals of the model.

- Players that have a low evaluation are usually underestimated while important players are consistently overestimated from this model.

Models Estimation - VI

- If instead we use all variables available in FIFA plus the factors added by us, we get a much better response from the model:

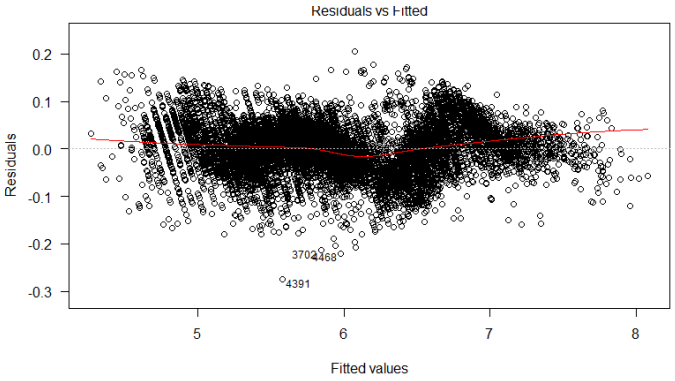




- Compared to before, important players aren't consistently overvalued anymore and predictions are much closer to what they should be in general.

Models Estimation - VII

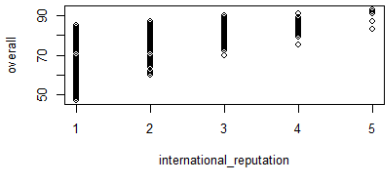
- Residuals still have some patterns in the distribution but less than before:

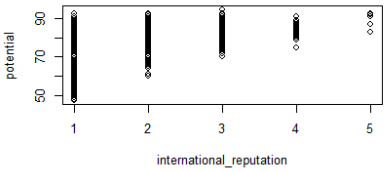


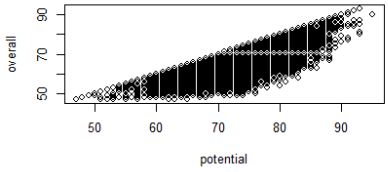
- What else can we do to try to improve the model?

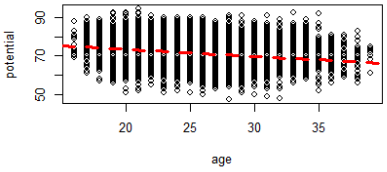
Models Estimation - VIII

- We can expect some of the variables to influence one another. Interactions effects of the 4 basic intuitive variables are added.









Models Estimation - IX

- Not all variables used are useful to predict the value of the players. Through techniques of model selection, the final model to predict the value of the players use the following variables:

Categorical variables	Quantitative variables
Career Phase	Overall
Position	Potential
League Importance	Age
Category	International reputation
	Weight
	Pace
	Passing
	Defending (general, tackle)
	Attacking (crossing, finishing)

Table: Predictors of final model

Final Model interpretation

- How do the variables influence the model in determining the value of a player? The coefficients of the model can help us explain how to give an interpretation to the results;

Categorical Variables

- *Career Phase*: The model is calculated using as default the level Young, with respect to which every other level has a negative coefficient. In particular it holds true that:

$$0 > \alpha_{prime} \gg \alpha_{old} > \alpha_{near\_retirement} > \alpha_{last\_year}$$

Final Model interpretation - II

- *Position*: The model consider as standard the level ATK for position;

$$0 \approx alpha_{MID} \gg \alpha_{DEF} \approx \alpha_{SUB}$$

- *League Importance*: As expected it's relevant for the value of a player to be playing in an important European league. The coefficient related to this factor is positive and highly significant;
- *Category*; Interpretation of this factor isn't as trivial as the other factors. Acts as a balancing variable for *Overall*. A model with only *Category* as predictor does indeed predict that the higher the category, the higher the value.

Final Model interpretation - III

Quantitative variables

- *Overall*: Most important predictor, recap of all variables; in all models $\alpha_{\text{overall}} > 0$
- *Age*: As expected significant and $\alpha_{\text{age}} < 0$
- *International Reputation*: The more a player is known, the more is valued, $\alpha_{\text{Interational\_reputation}} > 0$
- *Potential*: Odd behavior, positive coefficient if interaction with overall isn't included, negative coefficient if included. $\alpha_{\text{overall*potential}} > 0$

Final Model interpretation - IV

Other significant variables selected

- *Pace*: Speed and acceleration of a player has a positive effect on his value: $\alpha_{\text{pace}} > 0$
- *Passing*: Accuracy in the passing of a player has clearly also a positive effect on his value: $\alpha_{\text{passing}} > 0$
- *Defending variables*: Defending variables (defending/defending tackle..) tend to have negative coefficient, most likely reflecting that defensive players are generally valued less.
- *Attacking variables*: For the same reason the selected attacking variables (attacking_finishing, attacking_short_passing), when significant, tend to assume positive coefficient, especially when it's a measure related to scoring goals.

Final Model interpretation - V

Interactions

- The model was given freedom to choose through model selection also interactions between *Age*, *Overall*, *International Reputation*, *Potential*;
- Interpretability isn't as trivial as simple variables in this case but the techniques of model selection have indeed shown that in some cases considering interactions between the stats is beneficial;
- The final model in fact uses as predictors also:

Age*Overall	Age*International Reputation
Age*Potential	Overall*Potential
Potential*International Reputation	

Table: Highly significant interactions included in the final model.

Application 1

Undervalued and Overvalued Players

Undervalued and Overvalued Players

- The aim of this application is to use the our model to highlight the undervalued and overvalued players in the market;
- For a club it is of fundamental importance to be able to find the players correctly;
- Several works have been done in the literature that take into account the influence of social media on the player's price, for example, and very few on the actual qualities of the player.

Undervalued and Overvalued Players - II

- As mentioned in [6] we know that a player's value is influenced not only by his in-game skills but also by popularity with respect to the general public;
- This is especially true for players who are in the Top10% of the price;
- Fortunately, the dataset also has an explanatory variable that takes into account the popularity of the player, that is *International reputation*;
- In fact, this explanatory variable is highly significant in all the estimated models and its coefficient assumes a strictly positive value, this indicates that the price is positively influenced by increasing the value of this variable.

Undervalued and Overvalued Players - III

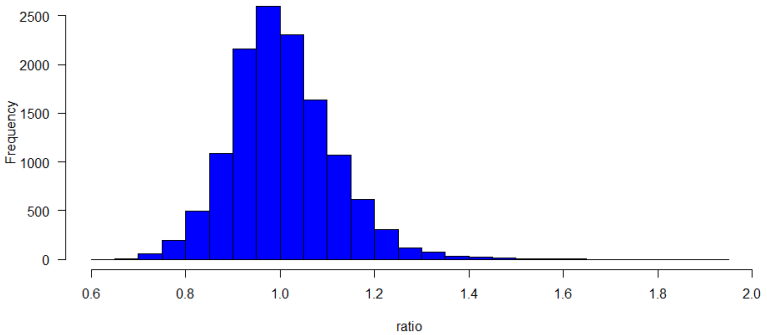


Figure: Distribution of the price ratio.

Undervalued and Overvalued Players - IV

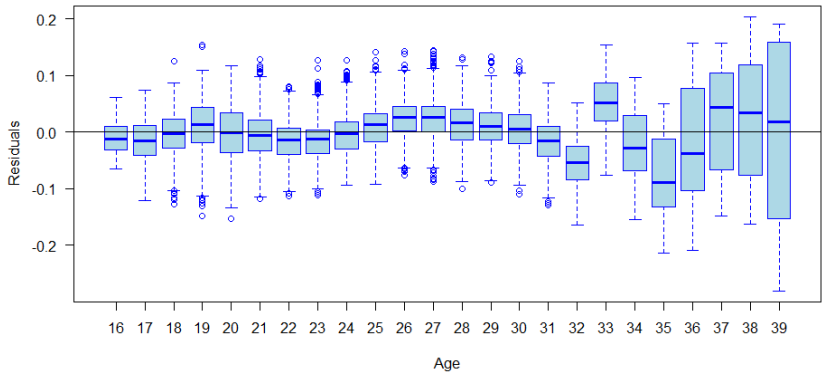


Figure: Prediction errors by age.

Undervalued and Overvalued Players - V

Names	Age	Overall	Actual Value	Predicted Value
J. Sancho	20	87	69500000	93648749.71
K. Mbappé	21	90	105500000	123278258.65
T. Alexander-Arnold	21	87	60000000	77240304.31
Bruno Fernandes	25	87	63000000	74992910.26
P. Aubameyang	31	87	45000000	56480129.33
R. Lewandowski	31	91	80000000	91186122.02
H. Son	27	87	52500000	63520590.07
E. Haaland	19	84	45000000	55795301.99
Bernardo Silva	25	87	60000000	70662598.88
C. Immobile	30	87	48500000	58994440.32

Table: Top undervalued players.

Undervalued and Overvalued Players - VI

Name	Age	Overall	Actual Value	Predicted Value
T. Kroos	30	88	55000000	45808130.55
H. Kane	26	88	71000000	62344168.17
P. Dybala	26	88	71000000	62504189.10
J. Kimmich	25	88	65000000	58697221.85
R. Sterling	25	88	72500000	66354323.75
K. Koulibaly	29	88	50000000	43984297.90
L. Suárez	33	87	31500000	26038853.47
E. Hazard	29	88	58000000	52788084.58
E. Can	26	82	26500000	22601423.02
S. de Vrij	28	84	30500000	26817201.63

Table: Top overvalued players.

Application 2

Match Outcome Prediction

Match Outcome Prediction

- The aim of this application is to see if we can use the data provided by EA Sports to estimate a model and predict the outcome;
- There are many models in the literature that attempt to predict the outcome of matches using in-real match statistics (i.e. number of goals, number of assists, etc.) but very few works that attempt to do so from player attributes;
- We will build a simple model that attempts to predict the match outcome using the statistics provided by EA Sports and see if it will match the real data.

Match Outcome Prediction - II

- To do so we use we use data from the Italian Serie A in the 2020/21 season. This dataset contains all the 380 matches played during the season, on which 105 variables have been measured;
- For example, the matches are represented by:

Div	Date	Time	HomeTeam	AwayTeam	FTHG	FTAG	
I1	19/09/2020	17:00	Fiorentina	Torino	1	0	...
I1	19/09/2020	19:45	Verona	Roma	0	0	...
...	...	...	...	...	...	...	...
I1	23/05/2021	19:45	Torino	Benevento	1	1	...

- For the estimation of this model we will only use the columns related to the number of goals of the home team and the away team, to obtain the result of the match.

Match Outcome Prediction - III

- These datasets are very informative and provide the basis for estimating different types of soccer prediction models;
- Among other things, these datasets also contains the odds estimated by the major bookmakers for each possible result;
- We will try to use the odds provided by complex bookmakers' models to compare our simple model to theirs.

Match Outcome Prediction - IV

- To establish the strength of a team, it is necessary to obtain a summary statistic;
- In the models present in the literature [8] generally the strength of a team is represented by assigning it an attack parameter (i.e. number of goals scored) and a defense parameter (i.e. number of goals conceded);
- In our model instead we will use a single parameter to represent the strength of the team that is the average of the *Overall* calculated among all the players of that team.

Match Outcome Prediction - V

- Using only one statistic for a team's strength for the entire season is a bit restrictive as we know a team's performance varies throughout all the season;
- In order to have a more realistic parameter, we decided to make it dynamic throughout the time;
- The idea that we have considered is to vary the parameter of the team's strength in a similar way to what happens with its Elo score, a parameter that is assigned to each professional team.

Match Outcome Prediction - VI

- The Elo is a score that is assigned to each team which measures its strength, based on the results of previous matches;
- This rating system comes from the world of Chess;

- The basic idea is to provide a summary measure of each team's strength, updating its score based on the result of each match;
- The variation depends also on the strength of the team faced, if the victory takes place at home and on the difference in goals scored.

Match Outcome Prediction - VII

- Considering I_0^H, I_0^A respectively the pre-match scores of the home team and the away team. On average, for the match in question it is assumed that the home and away teams mark:

$$\gamma^H = \frac{1}{1 + 10^{\frac{I_0^A - I_0^H}{400}}} \qquad \gamma^A = 1 - \gamma^H$$

- Considering the possible results, with reference to the home team:

$$\alpha^H = \begin{cases} 1, \text{ if the home team won} \\ 0.5, \text{ if the match war drawn} \\ 0, \text{ otherwise} \end{cases} \qquad \alpha^A = 1 - \alpha^H$$

- Team Elo ratings are updated after the end of each match via:

$$I_1^H = I_0^H + 20(\alpha^H - \gamma^H) \qquad I_1^A = I_0^A + 20(\alpha^A - \gamma^A)$$

Match Outcome Prediction - VIII

- This rating system has been extensively studied and used to predict the outcome of many matches [9];
- We therefore decided to derive the state of form of the team starting from the Elo score through:

$$r_t^H = \frac{l_t^H}{l_1^H} \qquad\qquad r_t^A = \frac{l_t^A}{l_1^A}$$

So we used these coefficients to rescale the *Overall* of each team and take into account the state of form during the season.

Match Outcome Prediction - IX

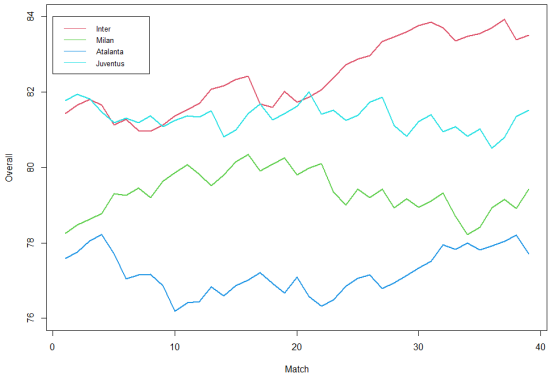


Figure: Overall trend for the Top4 ranking teams.

Match Outcome Prediction - X

- Finally, to estimate a probability on the match result we use logistic regression in which we use only one covariate variable:

$$x = Overall^H - Overall^A$$

Then, the model is:

$$log(\frac{p}{1-p}) = \alpha + \beta x + \epsilon$$

- To have a comparison term, use the probability estimates provided during the championship by the bookmakers, in this case Betfair365 (B365).

Match Outcome Prediction - XI

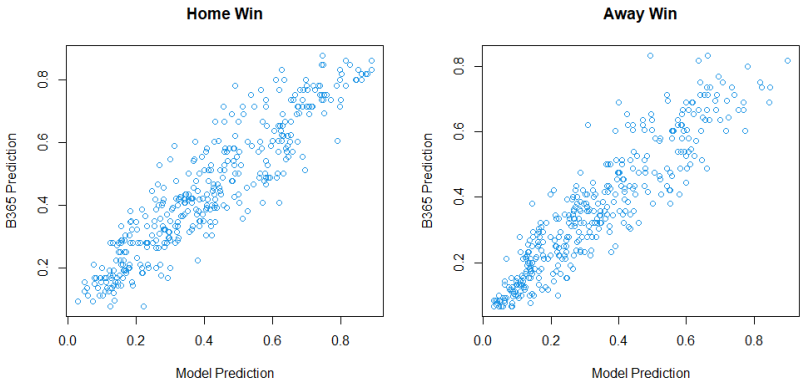


Figure: Predictions of our model and predictions provided by B365.

Match Outcome Prediction - XII

- We only used the data related to the 2020/2021 season, we could obtain significant improvements by including data related to more seasons;
- We only used data from a single league, but by including data from multiple leagues the model estimates could improve considerably;
- We only used the *Overall* as a parameter to estimate the team strength, we could use a different criterion for assigning the team's strength or a different criterion for the update;
- Use different data sources, also including statistics from real matches of the championship.

Conclusions and Future Works

Conclusions and Future Works

- In this project we wanted to establish if there was a relationship between the football data provided by FIFA 21 with the real world, to do so we tried to predict the market value of active professional players;
- We have also provided two useful applications that are dealt very frequently in statistical analysis in the soccer field;
- There are numerous works in this area because soccer is a difficult sport to model because it contains a lot of randomness.

Conclusions and Future Works - II

- In order to improve the forecasting capabilities of the models, the real in-game statistics could be considered as well;
- Furthermore, one could try to use different modeling and feature extraction techniques, apply dimensionality reduction or include also personal opinions;
- Until now the problem with this type of literature was the scarcity of data but, as we have seen, given the quantity and quality of available datasets we have many possibilities for interesting applications.

References

References

[1] https://github.com/SunPy-FIS/Statistical-learning-2021. *Repository Github*

[2] https://www.statista.com/statistics/1087391/global-sports-market-size/

[3] http://clubelo.com/. *Club Elo Data*

[4] https://www.kaggle.com/stefanoleone992/fifa-21-complete-player-dataset. *Kaggle Dataset*

[5] https://www.football-data.co.uk/italym.php *Serie A 2020/21 Dataset*

[6] Behravan, Iman and Razavi, Seyed Mohammad (2021). *A novel machine learning method for estimating football players' value in the transfer market*. Soft Computing. 25(3), 2499–2511.

[7] Cotta, Leonardo and de Melo, POV and Benevenuto, Fabrício and Loureiro (2010). *Using fifa soccer video game data for soccer analytics*. International Journal of forecasting. 26(3), 460–470.

[8] Dixon, Mark J and Coles, Stuart G (1997). *Modelling association football scores and inefficiencies in the football betting market*. Journal of the Royal Statistical Society: Series C (Applied Statistics). 46(2), 265–280.

[9] Hvattum, Lars Magnus and Arntzen, Halvard (2016). *Using ELO ratings for match result prediction in association football*. Workshop on large scale sports analytics.

The End