Predicting the Market Value of FIFA Players

By Luiz Augusto Carneiro

Sydney - May 2020

Photo by Leonard von Bibra on Unsplash

Objective

Using the soFIFA.com dataset, predict a **player's market value** based on his physical and technical characteristics.

Neymar da Silva Santos Junior (ID: 190871)   FIFA 20 APR 29, 2020
LW CAM 27y.o. (Feb 5, 1992) 5'9" 150lbs

92	92	€105.5M	€290K
Overall Rating	Potential	Value	Wage

ATTACKING
- 87 Crossing
- 87 Finishing
- 62 Heading Accuracy
- 87 Short Passing
- 87 Volleys

SKILL
- 96 Dribbling
- 88 Curve
- 89 FK Accuracy
- 81 Long Passing
- 95 Ball Control

MOVEMENT
- 94 Acceleration
- 89 Sprint Speed
- 96 Agility
- 92 Reactions
- 84 Balance

POWER
- 80 Shot Power
- 61 Jumping
- 81 Stamina
- 49 Strength
- 85 Long Shots

MENTALITY
- 51 Aggression
- 36 Interceptions
- 87 Positioning
- 90 Vision
- 92 Penalties
- 94 Composure

DEFENDING
- 35 Defensive Awareness
- 30 Standing Tackle
- 29 Sliding Tackle

GOALKEEPING
- 9 GK Diving
- 9 GK Handling
- 15 GK Kicking
- 15 GK Positioning
- 11 GK Reflexes

TRAITS
- Injury Prone
- Flair
- Speed Dribbler (AI)
- Outside Foot Shot
- Technical Dribbler (AI)

Summary

ETL / EDA

- 2019 sofifa.com
- The target variable is the **market value**
- The data set has 18,147 observations, with 39 numerical features (normalised) and 3 categorical features (193 dummies generated): nationality, preferred foot and position.
- There are no missing values

Modelling

Models=RF and GBM

Top 1 characteristc = REACTIONS

Most important features are selected by using *random forest* and *gbm* models:

reactions, ballcontrol, standingtackle, dribbling, shortpassing, shotpower, finishing

Insights

- reactions is by far the most important feature
- The linear models (Linear regression and GLM) do not perform well.
- polynomial model has the best performance:
  - R^2 (train) = 86.5%
  - R^2 (test) = 85.8%

The Most Important Features

Random Forest	GBM

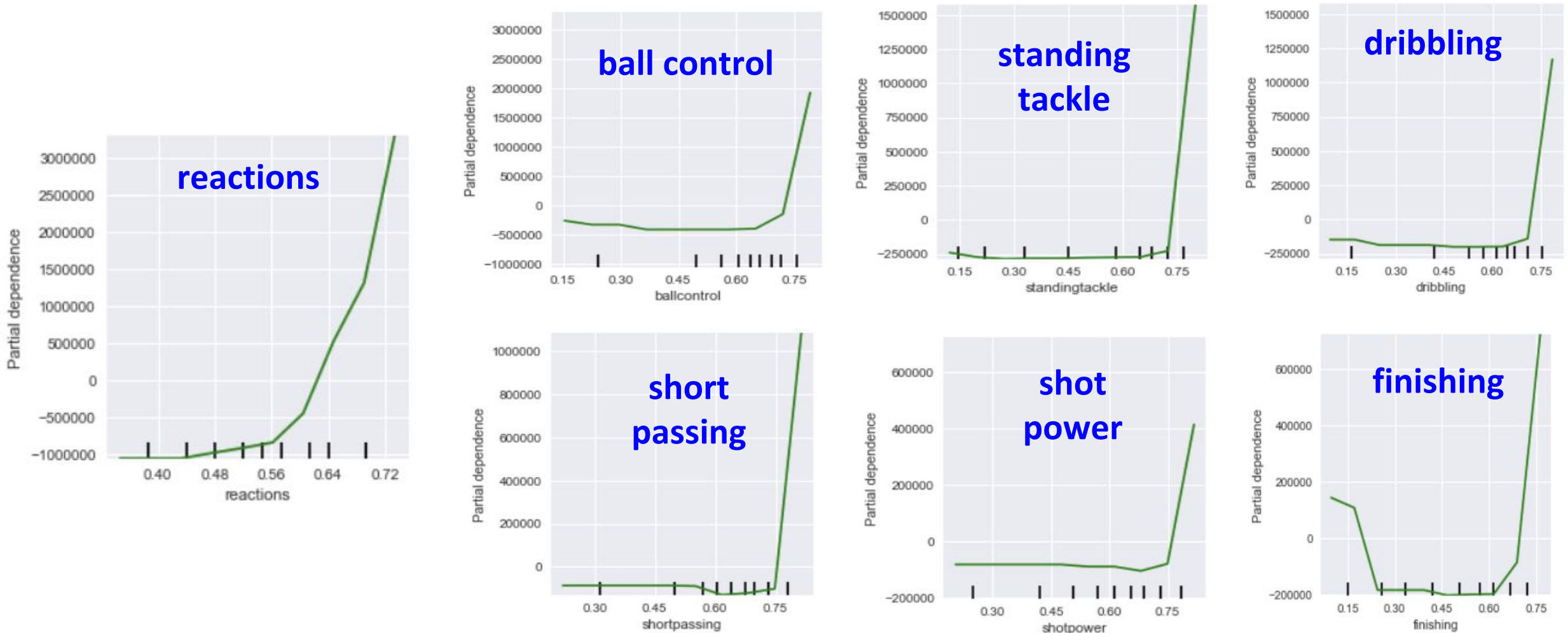The seven most important features are the same for random forest and GBM.

The relative importance of the remaining features is very low.

	importance
reactions	0.614950
ballcontrol	0.141138
standingtackle	0.030285
dribbling	0.022470
shortpassing	0.011981
shotpower	0.011932
finishing	0.011815
marking	0.009748
gkreflexes	0.009563
weight_kg	0.009108
sprintspeed	0.008370
slidingtackle	0.007923
stamina	0.007714
headingaccuracy	0.007538
gkdiving	0.007282

	importance
reactions	0.598664
ballcontrol	0.131013
standingtackle	0.047197
dribbling	0.030048
shortpassing	0.023687
shotpower	0.018867
finishing	0.015246
gkdiving	0.013471
headingaccuracy	0.011920
sprintspeed	0.010603
gkreflexes	0.009966
gkhandling	0.009236
marking	0.008714
slidingtackle	0.008636
stamina	0.007831

Partial Dependence Plots

The plots below show the isolated impact of each feature on the target

Predicting the Market Value

The seven most important features are used to predict the market value

The **polynomial model (degree 4)** has the best performance:

- R^2 (train) = 86.5%
- R^2 (test) = 85.8%

The **linear regression model** has a much worse performance:

- R^2 (train) = 30.6%
- R^2 (test) = 31.0%

The **generalised linear model (GLM)** does not perform well, based on its MSE.

Polynomial Regression
Predicted Values vs. Actual Values of y
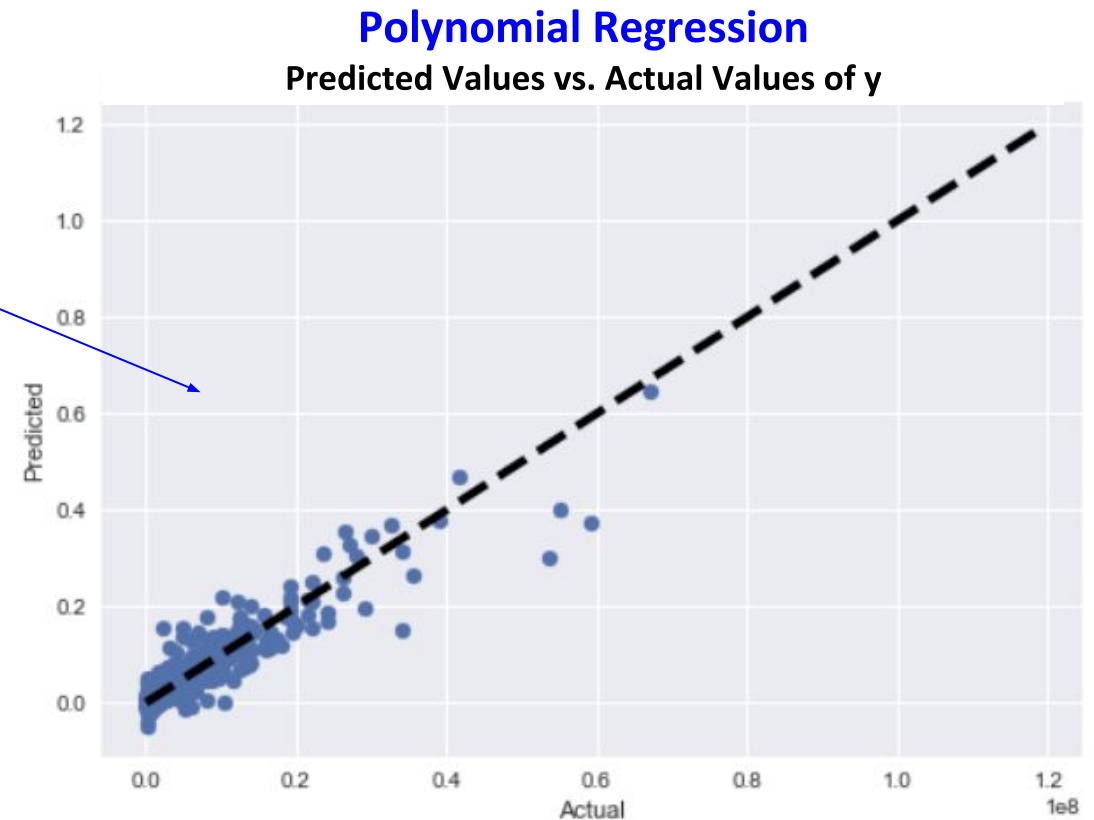
The End

Any questions?!!

Thank you!

Luiz Augusto Carneiro
carneiro_aus@yahoo.com.au