Title:

Who Will be the Next Football Superstar?

---- Using FIFA game's data to predict player's potential

Name: Siyuan Peng

Date: May 21th, 2019

Abstract

The primary purpose of my project is to predict soccer player's potential. By using all numeric attributes which evaluate different aspects of a soccer player, I hope my model could predict the overall value of that player after three years.

To make it clear, the word 'potential' here means the peak of the overall value a player could ever reach in his sports career (overall value is a comprehensive index in the FIFA game to measure the general ability of a player). Figure 1 shows the variation trend of the overall value of a player in his athletics career. We could see that this player reached his peak in his 28 (which is the red dot) and we will use all attributes when he was 25 (which is the yellow dot) to explain his potential and use this trained model to find some young talents with great potential or some role players who have been underestimated by experts.

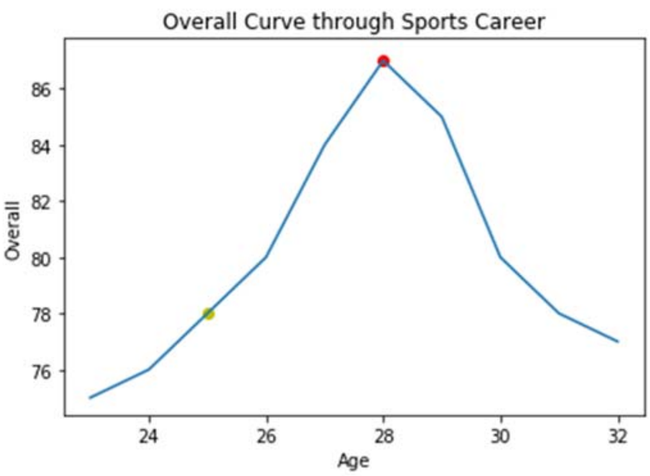


Figure 1

Data Section

Data Description and preparation

The data was scraped from the popular FIFA game website: www.sofifa.com using a python crawling script. The website contains the data of the EA Sports' game FIFA from an ancient version (FIFA 07) to its latest version (FIFA 19) and gets updated regularly with the release of new versions of the game. Through several research projects done on soccer analytics, it has been established in the field of academia that the use of data from the FIFA franchise has several merits that traditional datasets based on historical data do not offer. This data is clean but not lose credibility when compared with real-world data. Since 1995, the FIFA Soccer games provide an extensive and coherent scout of players worldwide and invite myriads of players to their lab to record their body and skill data.

Considering web crawl is hugely time consuming (it took me more than 20 hours to get my raw dataset), only the TOP 2,000 players (except for goalkeepers, considering that goalkeepers have a totally different evaluation system) in the latest version of FIFA 19 were chosen as the target players and all their 10 years data (from FIFA 19 to FIFA 10) were crawled from the website. Only 639 players show in all these ten versions and all of the numeric attributes are stored in the final candidate database. Then, the highest overall value of each particular player in these ten years was found, and this value will be our model's explained variable. This peak overall value should be viewed as the real potential of a player and the peak year of the player was recorded as well. Then, all other attributes of that player in the three years before the peak year were matched as the explaining variables. For example, if a player reached his peak at FIFA 16, his attributes in FIFA 13 will be matched. The goal of our model is to use player's attributes to predict his potential in three years later. The reason for the choice of three as the number of the gap year is that soccer player's contract is usually five years long, and the highest commercial value for that player will be in his third contract year. There are still two years before the player could become a free agent and the manager really cares about the player's performance in that year.

For each attribute in the database, we have an integer that measures how good a player is at that attribute. Attributes could be classified into seven categories: Basic information, Attacking, Skill, Movement, Power, Mentality, and Defending. All these variables build up a complete player with almost all aspects that could be quantified. A general description of all variables is shown in table 1. I convert the height from the British system to the metric system to make it easier to compare. All other variables are kept as before.

Table 1: Summary Statistics

Category	Variable	Mean	Std	Min	Max
Basic	Peak Overall	80.26	3.89	74	94
	3 years before Age	24.91	2.70	16	35
	3 years before Overall	76.10	4.86	59	94
	Height(cm.)	181.46	6.45	160.02	198.12
	Weight(lbs.)	167.76	14.87	117	209
	Expert Potential	79.09	4.85	63	94
	International Reputation	1.67	0.85	1	5
Attacking	Weak Foot	3.20	0.69	1	5
	Skill Moves	2.82	0.77	2	5
	Crossing	64.63	13.89	16	90
	Finishing	58.93	17.44	10	95
	Heading Accuracy	65.53	12.12	24	95
	Short Passing	73.21	7.73	48	92
	Volleys	58.15	16.31	13	89
Skill	Dribbling	70.46	12.45	21	96
	Curve	63.08	15.48	13	92
	FK Accuracy	58.17	16.05	10	91
	Long Passing	67.16	10.13	31	92
	Ball Control	74.54	8.20	42	96
Movement	Acceleration	73.90	10.12	34	95
	Sprint Speed	74.31	9.44	42	96
	Agility	71.86	12.06	30	95
	Reactions	74.16	6.82	54	92
	Balance	68.37	12.56	32	95
Power	Shot Power	71.52	10.52	24	94
	Jumping	70.74	10.28	33	95
	Stamina	75.96	8.20	51	94
	Strength	71.81	10.62	25	93
	Long Shots	64.02	15.08	10	93
Mentality	Aggression	69.18	12.97	23	92
	Interceptions	60.17	20.10	14	91
	Positioning	64.56	16.15	12	94
	Vision	66.95	12.37	22	93
	Penalties	60.76	13.67	13	92
Defending	Marking	54.28	22.61	10	91
	Standing Tackle	59.30	22.03	11	91
	Sliding Tackle	55.98	22.78	11	92

From the description, we could see that this database is representative. The average age of the player in his peak year is 28 (24.91+3), which is a well acknowledged golden age for soccer players with mature body and sufficient experience. What’s more, the average height (181.5cm) is just the average height for FA Premier League. All these 639 players are from Europe five major league and they represent the best scorer players in the world. The distributions of height and weight are shown in Figure 2.

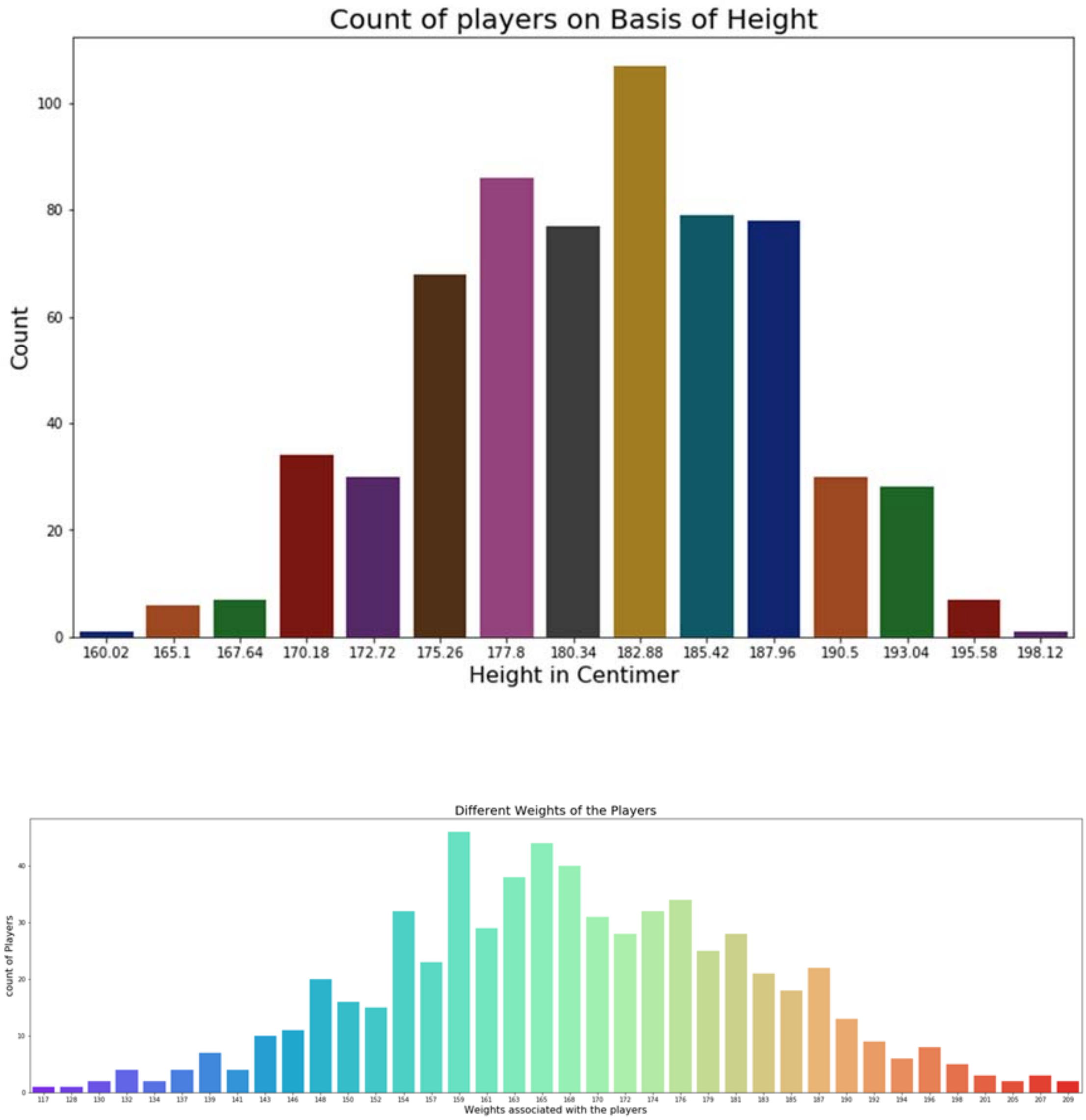
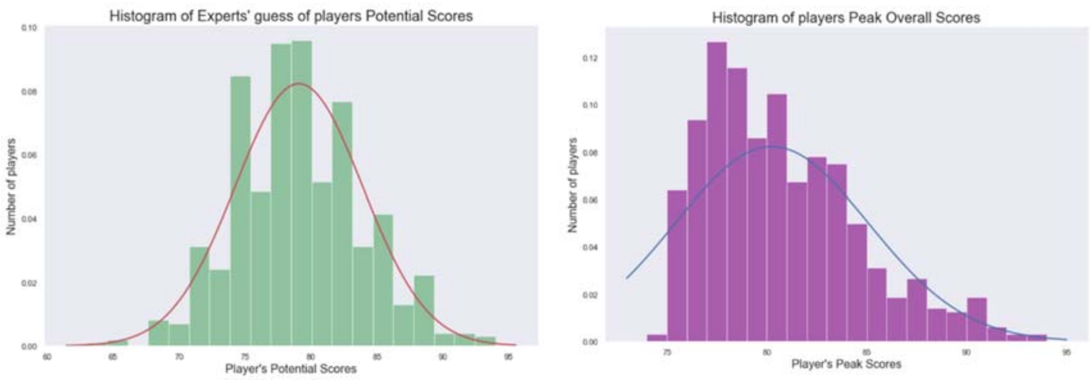


Figure 2

Justification of the variables selection

From the above data, we could find that EA sport's experts have their prediction of the player's potential. The average value of their speculation (79.1) is quite close to the real number (80.3). However, from the below graph, it is clear that the distributions of these two data are unlike each other, which again emphasize the importance of our research.



(a) Experts' prediction of player potential (b) Players' Real potential

Figure 3

The FIFA database provides a wide range of attributes that could select from; however, only the above attributes have been chosen. Below are some reasons why I abandon the rest of them.

To begin with, there has been a long-time debate about the influence of preferred foot toward the performance of scorer players. McMorris [1] has done detailed research about the different strategies' goalkeepers will use when facing right- or left- foot penalty kickers. In my final dataset, the majority are right-foot. However, from my perspective, the reason for that is mainly because most population is a right-hander. From figure 4 (lmpot of the preferred foot with ball control and dribbling), we could not reach to a conclusion that the difference in the preferred foot has an influence of the skill for a player.

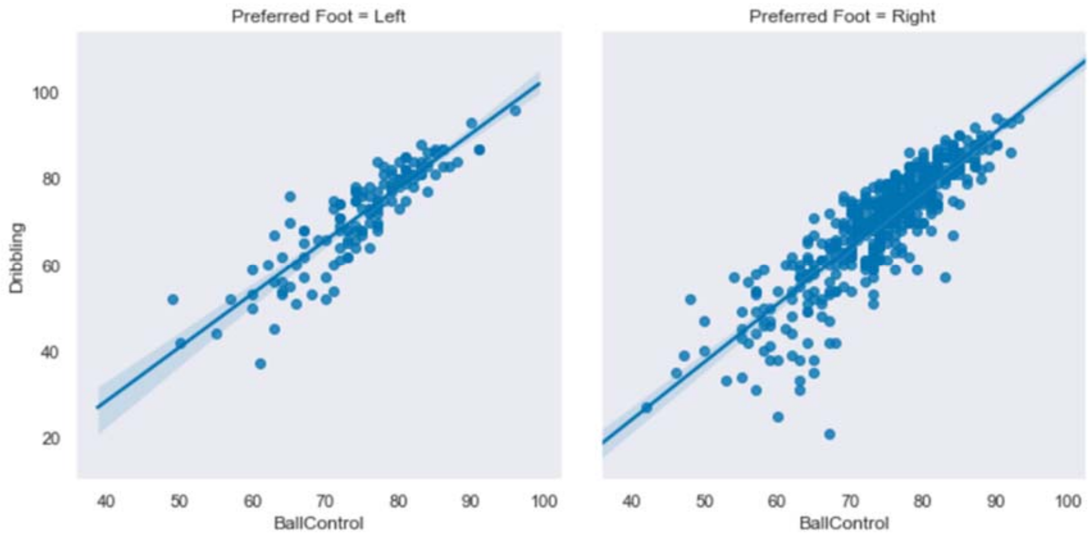


Figure 4

As for the rest unchosen variables, I didn't find a reasonable explanation of why EA employ them in the FIFA game. Considering that they might lack practical significance, I abandoned all these variables.

The chosen variables heatmap is shown in Figure 5. From the heatmap, it is clear that defending variables have a high correlation between each other. However, considering that these three variables evaluate different aspects for a defender, the model keeps them all to make a comprehensive judgment to that player.

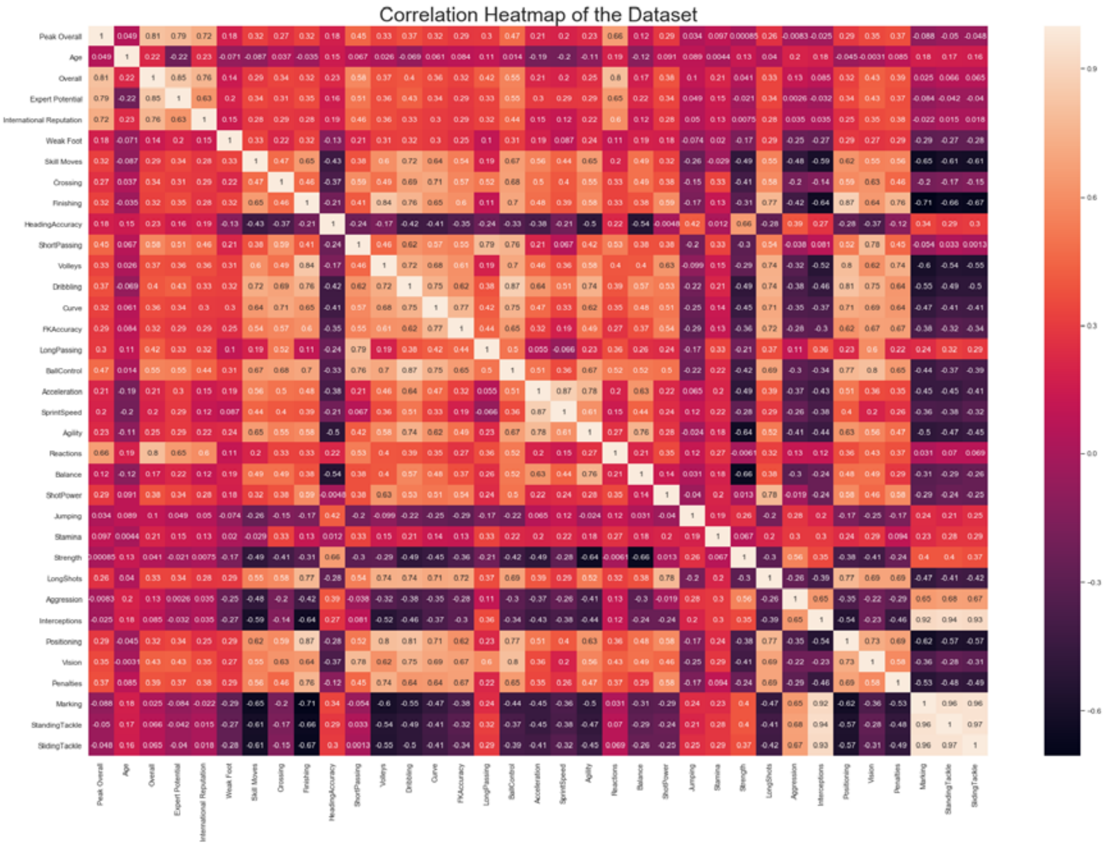


Figure 4

Method Section

Model Selection

The Problem I hope to address in my project is to lift the veil of:

$$\text{Potential} = \Phi(X_i)$$

in which X_i stands for all variables in table 1 except for the Expert Potential. Followed by the general instruction I have made in the Literature review, below are the models I would like to use to solve this problem and do a horse racing on them.

Ordinary Least Squares

The first and most intuitive approach is always the basic linear regression. The above problem could be further written as:

$$\text{Potential} = \beta_0 + \beta_1 * Overall + \beta_2 * Age + \cdots$$

The potential of the player is expected to be a linear combination of all the 35 attributes.

This linear regression aims to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. From the heatmap, we could see that, besides the defending variables, the multicollinearity between each variable is rather low. Therefore, the coefficient estimates for OLS is reliable.

Stochastic Gradient Descent

Stochastic gradient descent is a simple yet very efficient approach to fit linear models. In my case, it is particularly useful considering that the number of explaining variable is somewhat large (35 explaining variables). SGD regression implements a plain stochastic gradient descent learning routine which supports different loss functions and penalties to fit linear regression model. I hope this model could outperform the basic OLS.

Random Forest regressor

Random Forest is a representative of ensemble methods, and such method aims to combine the predictions of several base estimators built with a given learning algorithm to improve generalizability or robustness over a single estimator.

The goal of my Random Forest regression is to create a model that predicts the value of potential by learning simple decision rules inferred from the data features. By using all 35 provided attributes, the model should find a route which leads to an accurate value of the potential.

Support Vector Regression

Support Vector Machine is an ML approach which is effective in high dimensional spaces. It uses a subset of training points in the decision function (called support vectors) because the cost function for building the model ignores any training data close to the model prediction, so it is also memory efficient. Considering that the relationship between the potential of a player and his attributes is likely to be highly non-linear, SVR might provide a reliable prediction.

Back-Propagation Multi-Layer Perceptron

BP MLP network is a supervised Artificial Neural Network. This model is the most popular and successful model in the sports analysis domain. Myriads of papers have used this model to predict a wide range of sports events, such as the result of a match or the constitution of a team. The success of this model makes sense that too many modes influence the result of a sports event and exposure to a large number of possibilities under supervised learning conditions would yield a network that had the better predictive capability.

This model learns a function $f(\blacksquare): R^{35} \rightarrow R^{1}$ by training on a dataset, where 35 is the number of dimensions for my model's input and 1 is the number of dimensions for the output – player's potential. In this model, there can be one or more non-linear layers, called hidden layers. I hope I could find the best structure for this model.

Parameter Preprocessing

Considering that some of the Machine Learning algorithms are sensitive to feature scaling, the data needs to be scaled to achieve high accuracy. The majority of the parameters are centesimal and some of them are five-point scale or even without a clear scale standard (like age, weight, and height). To make sure all these models work properly, I standardized all of them to have mean 0 and variance 1.

The dataset has been split, in which 75% of it is the training set, and 25% of it is the test set.

Result Section

Horse Racing Result

I used Mean Squared Error as the leading judgment between the different result of all these models. The predicted potential calculated by each model was compared with the real potential of that player in the test set and the MSE between these two values were recorded. In addition, considering that experts have their prediction of the potential of the player, the MSE between their speculation and the truth was calculated as well.

All machine learning models (except for OLS) have been optimized by cross-validated search over parameter settings. For SGD Regression, the optimized penalty term is L2 and the penalty parameter is 0.0659. For Random Forest Regressor, the maximum depth is 3, the maximum features is 3, the minimum samples leaf is 3, the minimum samples split is 11 and the number of estimators is 96. For Support Vector Regression, the penalty parameter is 1.3697, Kernel type is 'rbf' and its coefficient is set to auto, and the model doesn't use shrinking heuristic. As for the BP MLP model, the hidden layer size is 52, the activation method is the rectified linear unit function (returns f(x) = max(0, x)) and the L2 penalty parameter is 7.9094.

In Figure 5, it is clear that the BP MLP model with a structure of 35-52-1 (35 input neurons, one hidden layer with 52 neurons and one outcome) has the lowest MSE. SGD Regression follows. It is somewhat surprising to see that the random forest regressor and support vector regression's performance is even worse than the basic OLS. However, we could not conclude that the ML model's performance doesn't outperform the traditional linear regression model's. Even though the difference of MSE between the best two of them is only 0.03 (0.2959 – 0.2653), as for percentage, it is a 10.14% difference. If I could use that advantage in investment, it is a considerable gap to make a super profit.

Even though the error rate of my best model is still as high as 26.53%, I believe that rate could be lowered when I implement more training data. If I have more time, I could download the whole dataset from FIFA 07 to FIFA 19 and find players who appear in these datasets at least ten times. Then, followed by the aforementioned data processing steps, I could use this enriched dataset to retrain my model and the MSE value could be lowered.

As for the expert prediction, considering that experts may have personal preference or emotion to some of the players, their judgments are rather subjective and may have personal bias. Thus, their predictions' MSE is quite high compared with my models.

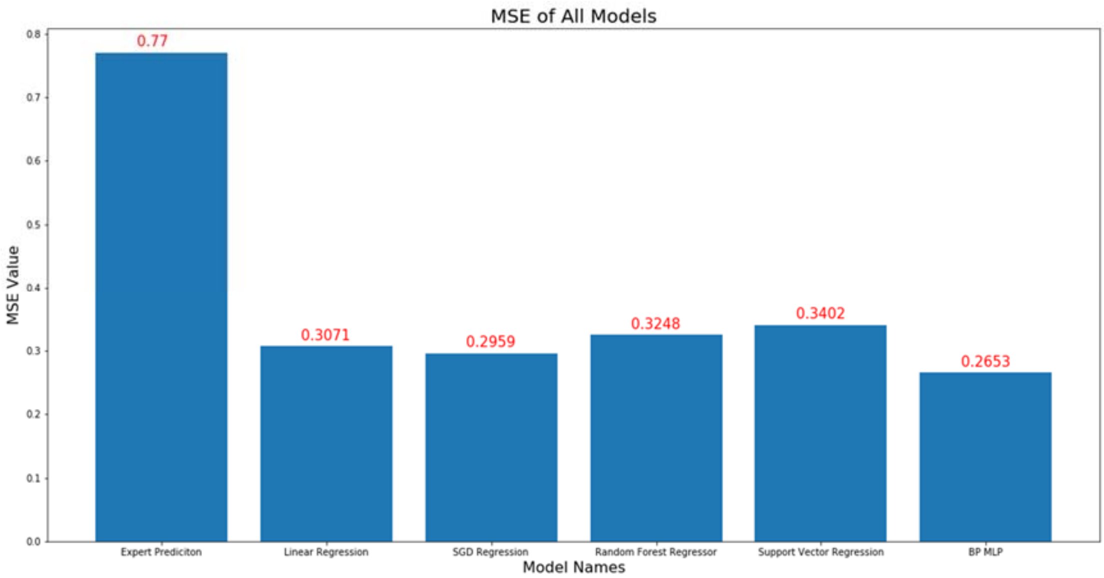


Figure 5

Weights of All Variables

In this part, I want to show and compare the weights of all variables calculated by my five models. To make it comparable to each other, I would show the weights of **permutation importance** of all these five models.

The idea of permutation importance is quite straightforward: feature importance can be measured by looking at how much the score (accuracy, R^2) decreases when a feature is not available. The detailed instruction of permutation importance could be found at its library instruction [2].

Table 2 shows the weights of permutation importance of all variables calculated by these five models. The deeper the green, the more influential of that variable. Transparent color or even light red means that variable has little or even negative effect on the player's potential. Figure 6 shows all these weights in one plot to make it easier to compare. From the plot, it is evident that the potential of a player after three years is still mainly determined by that player's current overall value. Besides, international reputation also plays an important role, which somewhat shows a positive attitude towards the work which has been done by scouts(experts). What's more, Reaction, Age, Volley Ability, Ball Control, and Long Shot Ability also have a significant impact on the potential value. Therefore, all these attributes deserve to be paid more attention to the evaluation system of a player.

Some of the variables, such as the ability of weak foot and height have little or even negative influence of the player's potential. It is reasonable for the case of soccer. The upper bound of a player is determined by how good he is on using his preferred foot. The influence of the weak foot is negligible, considering that soccer players could choose not to use it when they make a pass or shot. As for the height, different from basketball or football which hugely emphasize the ability of physical confrontation and the control of high balls, soccer pays more attention to speed and agility. If a soccer player is too high, he might lose these merits.

As for the rest attributes, they did not quite show off in my models (weight is less than 1%). However, this finding doesn't mean that those attributes are less important than others in the evaluation system of a soccer player. Considering that I mixed all soccer players from various positions, those attributes could still play an important role when I further divide my dataset into different parts (forward, mid, guard) and build separate models on them to improve accuracy. I could not further divide my current dataset, otherwise the sample size for each sub-dataset would be too small. However, if I could obtain a much larger dataset, the model specifically designed for different positions would have a more accurate prediction and different attributes would stand out in different positions (like attacking attributes in the forward dataset and defending attributes in the guard dataset). As for now, the attributes have a large weight in my model might indicate that these variables are essential for all soccer players regardless of their position.

Table 2

Weight	Feature
0.6735 ± 0.1327	Overall
0.1283 ± 0.0601	International Reputation
0.1150 ± 0.0410	Marking
0.0367 ± 0.0283	LongShots
0.0366 ± 0.0095	Volleys
0.0341 ± 0.0193	Reactions
0.0337 ± 0.0229	Age
0.0220 ± 0.0215	ShortPassing
0.0148 ± 0.0080	Weight
0.0103 ± 0.0110	Vision
0.0100 ± 0.0238	FKAccuracy
0.0074 ± 0.0084	SlidingTackle
0.0059 ± 0.0091	LongPassing
0.0058 ± 0.0087	Interceptions
0.0057 ± 0.0103	Aggression
0.0036 ± 0.0073	Agility
0.0028 ± 0.0164	Positioning
0.0024 ± 0.0011	Stamina
0.0019 ± 0.0016	Dribbling
0.0019 ± 0.0030	Balance
0.0018 ± 0.0028	Strength
0.0018 ± 0.0047	Finishing
0.0018 ± 0.0098	Acceleration
0.0016 ± 0.0054	HeadingAccuracy
0.0015 ± 0.0011	SprintSpeed
0.0013 ± 0.0048	StandingTackle
0.0007 ± 0.0025	Skill Moves
0.0002 ± 0.0104	BallControl
0.0001 ± 0.0004	Curve
-0.0003 ± 0.0011	Jumping
-0.0003 ± 0.0047	Weak Foot
-0.0005 ± 0.0021	Penalties
-0.0027 ± 0.0042	ShotPower
-0.0034 ± 0.0045	centimeter height
-0.0037 ± 0.0059	Crossing

Linear Regression

Weight	Feature
0.3389 ± 0.0819	Overall
0.1411 ± 0.0440	International Reputation
0.0774 ± 0.0178	Reactions
0.0227 ± 0.0198	Age
0.0116 ± 0.0021	Volleys
0.0106 ± 0.0161	LongShots
0.0075 ± 0.0097	HeadingAccuracy
0.0063 ± 0.0026	Stamina
0.0061 ± 0.0094	Skill Moves
0.0052 ± 0.0075	Marking
0.0037 ± 0.0026	Weight
0.0036 ± 0.0061	Agility
0.0029 ± 0.0019	SprintSpeed
0.0022 ± 0.0050	Vision
0.0018 ± 0.0139	FKAccuracy
0.0016 ± 0.0076	BallControl
0.0015 ± 0.0018	Curve
0.0013 ± 0.0019	Strength
0.0012 ± 0.0025	LongPassing
0.0007 ± 0.0101	Aggression
0.0004 ± 0.0010	Finishing
0.0003 ± 0.0006	Positioning
0.0002 ± 0.0005	Balance
0.0001 ± 0.0005	StandingTackle
0.0001 ± 0.0007	Interceptions
0.0000 ± 0.0008	Acceleration
0.0000 ± 0.0001	Jumping
-0.0000 ± 0.0029	Weak Foot
-0.0000 ± 0.0011	Penalties
-0.0003 ± 0.0003	Dribbling
-0.0005 ± 0.0008	SlidingTackle
-0.0005 ± 0.0006	Crossing
-0.0010 ± 0.0043	centimeter height
-0.0010 ± 0.0016	ShotPower
-0.0030 ± 0.0024	ShortPassing

SGD Regression

Weight	Feature
0.1395 ± 0.0381	Overall
0.0459 ± 0.0175	International Reputation
0.0446 ± 0.0077	Reactions
0.0260 ± 0.0083	BallControl
0.0236 ± 0.0132	Dribbling
0.0154 ± 0.0087	Vision
0.0115 ± 0.0018	StandingTackle
0.0113 ± 0.0016	SlidingTackle
0.0108 ± 0.0014	Curve
0.0098 ± 0.0022	ShortPassing
0.0096 ± 0.0072	Positioning
0.0083 ± 0.0048	Marking
0.0080 ± 0.0030	Volleys
0.0079 ± 0.0029	Interceptions
0.0078 ± 0.0016	Finishing
0.0077 ± 0.0056	LongShots
0.0075 ± 0.0023	HeadingAccuracy
0.0063 ± 0.0027	Penalties
0.0035 ± 0.0045	LongPassing
0.0031 ± 0.0013	Skill Moves
0.0026 ± 0.0019	ShotPower
0.0025 ± 0.0018	Acceleration
0.0020 ± 0.0006	Crossing
0.0015 ± 0.0008	FKAccuracy
0.0009 ± 0.0008	Strength
0.0009 ± 0.0011	SprintSpeed
0.0006 ± 0.0008	Agility
0.0004 ± 0.0008	Aggression
0.0004 ± 0.0006	Stamina
0.0003 ± 0.0008	centimeter height
0.0001 ± 0.0000	Weak Foot
-0.0000 ± 0.0002	Age
-0.0000 ± 0.0005	Weight
-0.0002 ± 0.0005	Balance
-0.0003 ± 0.0004	Jumping

Random Forest Regression

Weight	Feature
0.1529 ± 0.0695	Overall
0.0494 ± 0.0165	Reactions
0.0301 ± 0.0307	BallControl
0.0289 ± 0.0255	International Reputation
0.0236 ± 0.0090	LongShots
0.0196 ± 0.0318	Age
0.0172 ± 0.0077	Volleys
0.0171 ± 0.0102	HeadingAccuracy
0.0140 ± 0.0148	Interceptions
0.0139 ± 0.0167	Positioning
0.0135 ± 0.0066	Skill Moves
0.0130 ± 0.0083	Marking
0.0128 ± 0.0165	LongPassing
0.0121 ± 0.0188	Dribbling
0.0104 ± 0.0111	Agility
0.0096 ± 0.0119	Strength
0.0090 ± 0.0106	SprintSpeed
0.0072 ± 0.0056	Weak Foot
0.0072 ± 0.0074	Weight
0.0064 ± 0.0102	Finishing
0.0055 ± 0.0035	Curve
0.0054 ± 0.0082	centimeter height
0.0043 ± 0.0129	ShortPassing
0.0043 ± 0.0133	ShotPower
0.0042 ± 0.0112	Penalties
0.0039 ± 0.0163	FKAccuracy
0.0038 ± 0.0076	Aggression
0.0036 ± 0.0135	StandingTackle
0.0034 ± 0.0092	Acceleration
0.0034 ± 0.0106	Crossing
0.0033 ± 0.0091	Vision
0.0031 ± 0.0076	Jumping
0.0027 ± 0.0066	SlidingTackle
0.0009 ± 0.0078	Balance
-0.0007 ± 0.0126	Stamina

Support Vector Regression

Weight	Feature
0.3211 ± 0.1007	Overall
0.0753 ± 0.0121	Reactions
0.0303 ± 0.0171	International Reputation
0.0184 ± 0.0169	Age
0.0151 ± 0.0031	Volleys
0.0139 ± 0.0109	LongShots
0.0091 ± 0.0114	HeadingAccuracy
0.0074 ± 0.0063	Marking
0.0063 ± 0.0201	BallControl
0.0061 ± 0.0012	Stamina
0.0060 ± 0.0057	Interceptions
0.0059 ± 0.0064	SprintSpeed
0.0035 ± 0.0024	Strength
0.0031 ± 0.0039	Vision
0.0029 ± 0.0029	Jumping
0.0021 ± 0.0030	Agility
0.0018 ± 0.0016	Acceleration
0.0018 ± 0.0058	LongPassing
0.0011 ± 0.0051	Skill Moves
0.0009 ± 0.0030	Weight
0.0008 ± 0.0100	FKAccuracy
0.0007 ± 0.0031	Aggression
0.0004 ± 0.0019	Balance
0.0002 ± 0.0030	Penalties
0.0001 ± 0.0019	Dribbling
0.0000 ± 0.0036	Finishing
-0.0002 ± 0.0044	Weak Foot
-0.0002 ± 0.0029	SlidingTackle
-0.0003 ± 0.0048	StandingTackle
-0.0007 ± 0.0049	centimeter height
-0.0007 ± 0.0030	Curve
-0.0010 ± 0.0071	Positioning
-0.0016 ± 0.0031	ShotPower
-0.0027 ± 0.0028	Crossing
-0.0058 ± 0.0035	ShortPassing

BP MLP

Weights of Variables

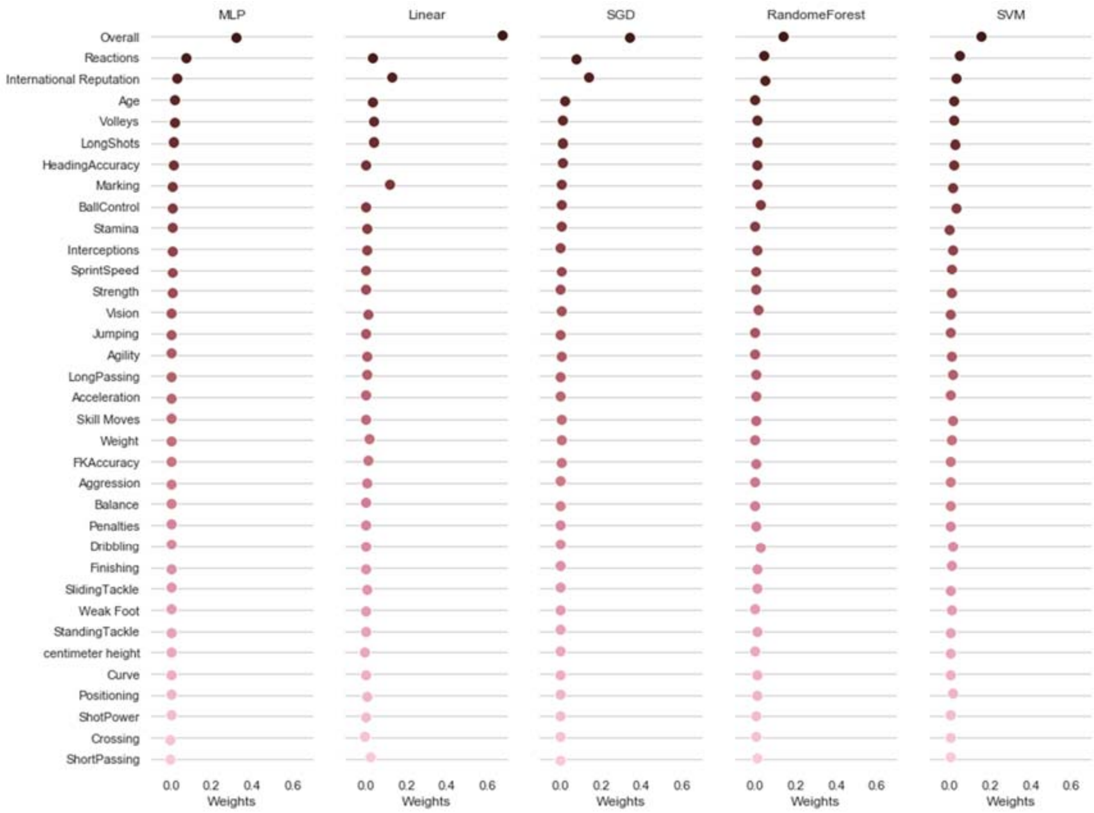


Figure 6

Prediction Section

In this part, I will use my model on the latest data of FIFA 19. I will rank the players by their predicted potential calculated by my model and hope to find the future superstar and candidates for the Golden Ball Award in the next three years.

Besides, a valuable finding would be some future soccer star who doesn't have a high expectation from those experts. I have made a precheck on my testing dataset. A quite interesting result is that, the top 10 players with the highest prediction gap between my model and experts' expectation are all role players.

In table 3, while gap 1 means the difference between my model's prediction and the real potential of that player, gap 2 means the difference between experts' prediction and the real potential and gap 3 means the difference between my prediction and experts' prediction. We could see that the highest peak value among them is just 79 while the average is 76.5. While my model made a somewhat accurate prediction (gap 1 is small), the experts tend to make an excessive low evaluation. A main reason for the underestimation of these role players might due to their low international reputation.

This finding might indicate that my model is surprisingly useful when searching for noteless role players. With a limit budget, a manager could not fill a team full of Messi (whose overall is higher than 90). Therefore, the selection of role players is crucial for the success of that team and my model might give some valuable support to their work.

Table 3

Underestimated Players							
Peak Overall	Age	Current Overall	Expert Potential	International Reputation	gap_1	gap_2	gap_3
76	26	62	63	1	0.10	-3.31	3.41
76	27	69	69	1	0.33	-1.78	2.12
76	24	65	69	1	0.21	-1.78	2.00
75	29	69	69	1	0.23	-1.53	1.76
79	25	68	70	1	-0.57	-2.29	1.73
75	25	69	71	1	0.67	-1.02	1.69
77	26	70	71	1	0.01	-1.53	1.54
77	25	69	71	1	-0.02	-1.53	1.51
76	27	71	71	1	0.20	-1.27	1.48
78	26	74	74	2	0.45	-1.02	1.47

Additional Reference (Not included in Literature Review) and bibliography

[1] McMorris, T., & Colenso, S. (1996). Anticipation of professional soccer goalkeepers when facing right-and left-footed penalty kicks. Perceptual and motor skills, 82(3), 931-934.

[2] https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html