Who Will be the Next Soccer Superstar?

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

Siyuan Peng[1]

June 2019

Abstract

The potential of a player is a rather less popular topic in the sports science domain. Even though scouts have their judgment about players' potential, their prediction is somewhat inaccurate. This paper aims to build a machine learning model to predict a young soccer player's potential using FIFA game data. Thirty-five attributes which descript the physical, skill and psychological level of the player have been implemented in the model to illustrate the peak of a soccer player. All ML models' performances are much better than experts' and the Backpropagation Multilayer Perceptron (BP MLP) model stands out. The model's prediction is much more accurate than experts' prediction, especially on role-players. Therefore, the model could give much help in the judgment of the role-player transfer decisions.

Keywords: Potential, BP MLP, Role-player

[1] University of Chicago, Computational Social Science, siyuanpeng@uchicago.edu

1 Introduction

The evaluation of match result and athlete performance is a widely discussed topic in sport science. Due to the tremendous amount of money on sports lottery, lots of attention and research resource have been put into the prediction of the competition result in prior years. However, with the emphasis of the quantitative research in the work of sport managers and the sharp rise of players' market value, the number of studies about the player's performance has also increased.

In my research, I hope to build a machine learning model to predict the potential of soccer players using game data. To make it clear, the word 'potential' in my paper 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). My project will be the first one to discuss the potential of a player. Different from real-world data, game data has more attributes which I could select from. The model in previously sports analysis paper[2] usually has 5 to 8 parameters while I have 35 characteristics which descript the physical, skill and psychological level of the player. Besides, I admit that the prediction of the player potential might not give you an instant profit just like the result of the match or the underestimated market value of the player. However, it could help the manager of the team to make the correct judgment which would benefit the future. A considerable gap between the potential and current performance could also be viewed as an underlying investment, and the rate of return could be huge as well.

In this paper, I will propose a different approach about player scouting, one that purely relies on data. 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. Figure 1 shows the variation trend of the overall value of a player in his athletics

[2] A brief literature review of sport analysis papers could be found at the Appendix.

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. With the application of this trained model, I hope to find some young talents with great potential or some role players who have been underestimated by experts.

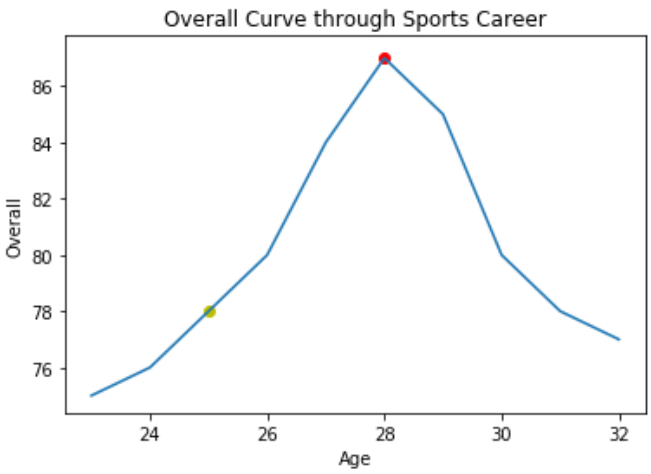


Figure 1

All selected models are machine learning based and all their predictions have higher accuracy compared with experts. To be specific, the Back Propagation Multilayer Perceptron model has the lowest MSE and its prediction is much more accurate than experts' prediction for role players. That would be the most apparent practical significance of my model. With limited budget and time, expert's prediction for those noteless role-players is rather inaccurate and subjective. In that case, given some crucial numeric attributes about those players body index and skill level, my model could show a somewhat reliable prediction for that player's potential. This result would be a valuable supplement material for the player transfer decision made by the manager.

2 Data Section

2.1 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 test and 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 10 to FIFA 19) 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 overall peak 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 soccer players in the world. The distributions of height and weight are shown in Figure 2 and Figure 3.

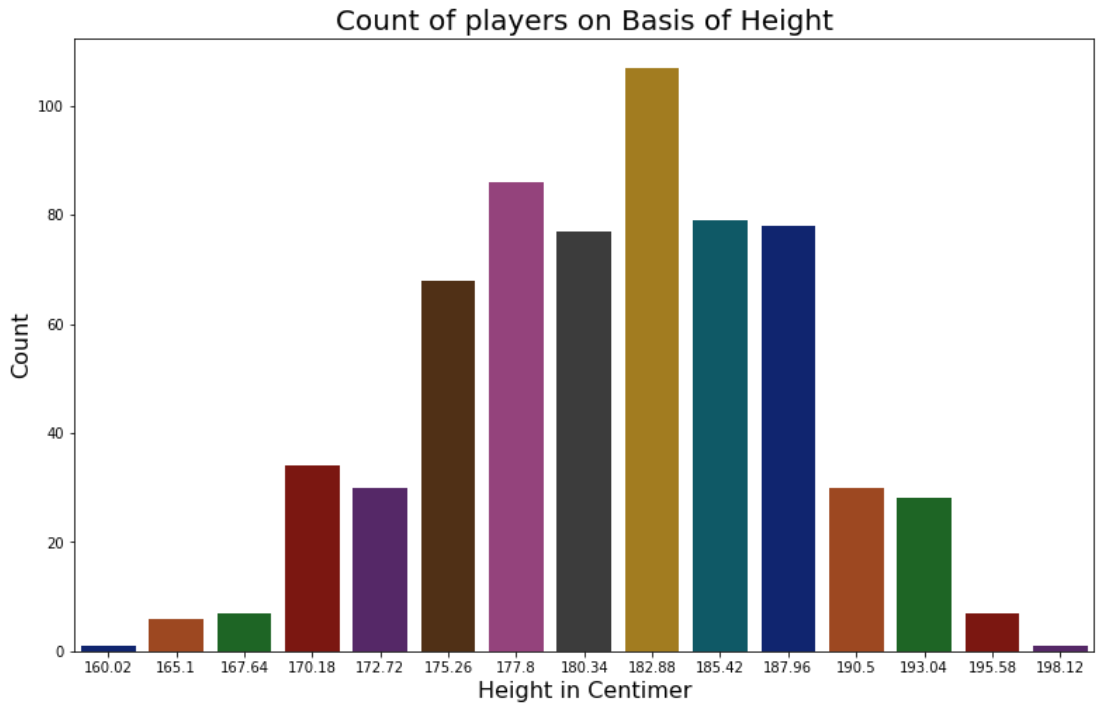


Figure 2

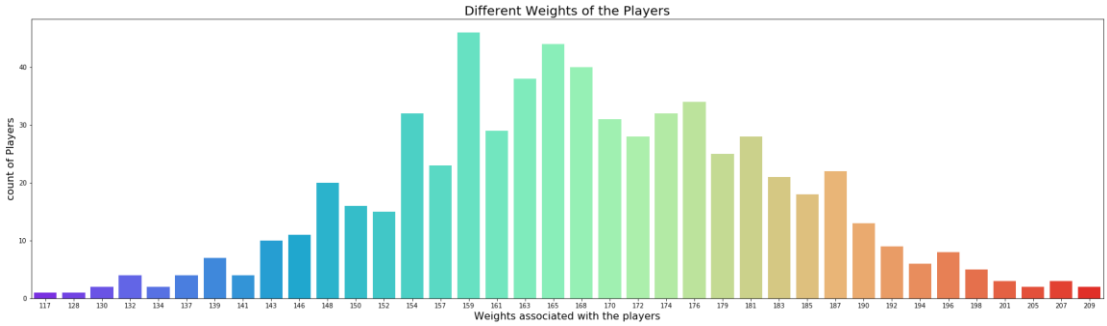
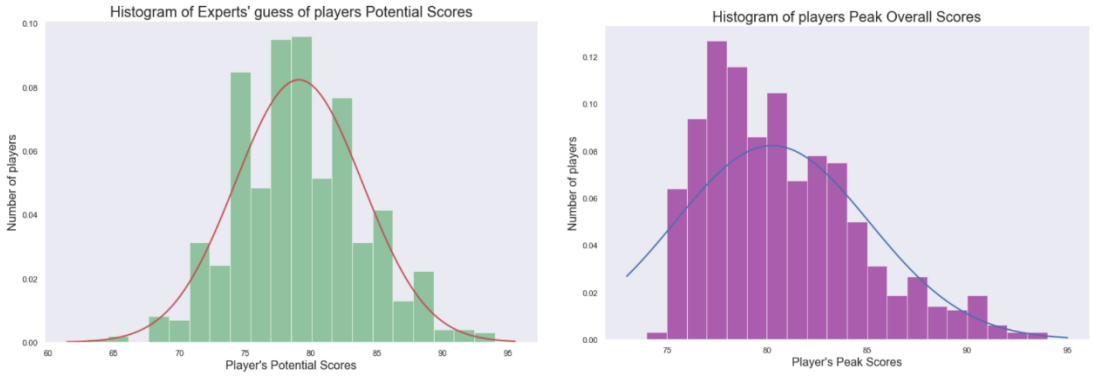


Figure 3

2.2 Justification of the variable's 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 4

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 [12] 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 5 (lmplot 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.

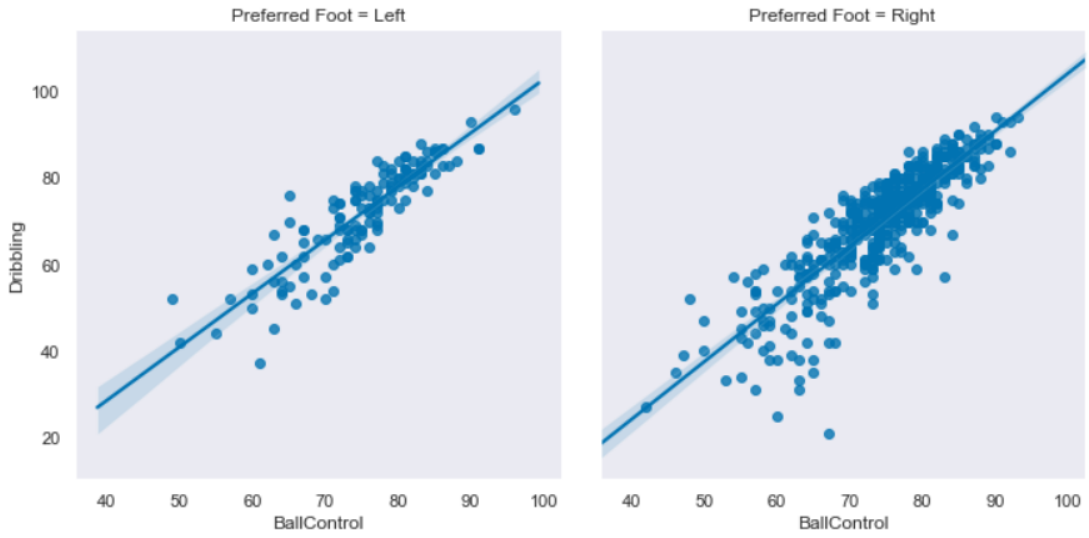


Figure 5

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 6. 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 all of them to make a comprehensive judgment to that player.

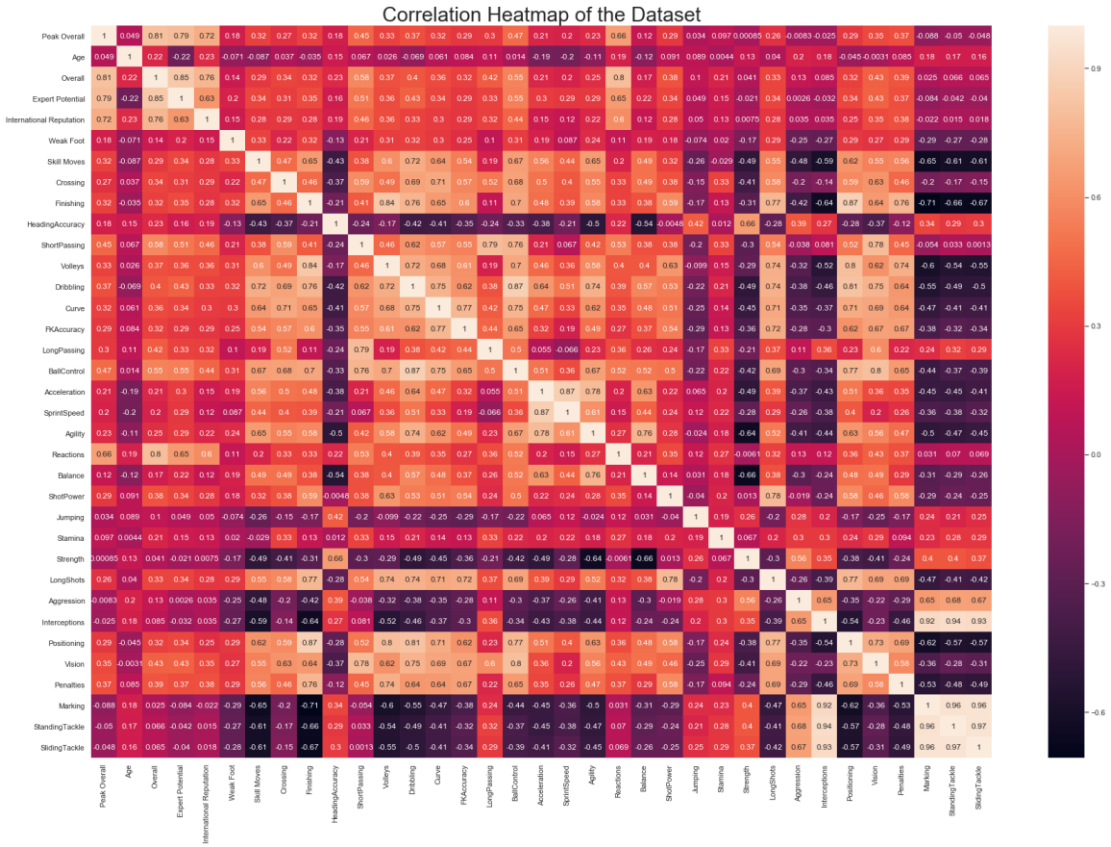


Figure 4

3 Method Section

3.1 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.

3.1.1 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.

3.1.2 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.

3.1.3 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.

3.1.4 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.

3.1.5 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.

3.2 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.

4 Result Section

4.1 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 6, 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.

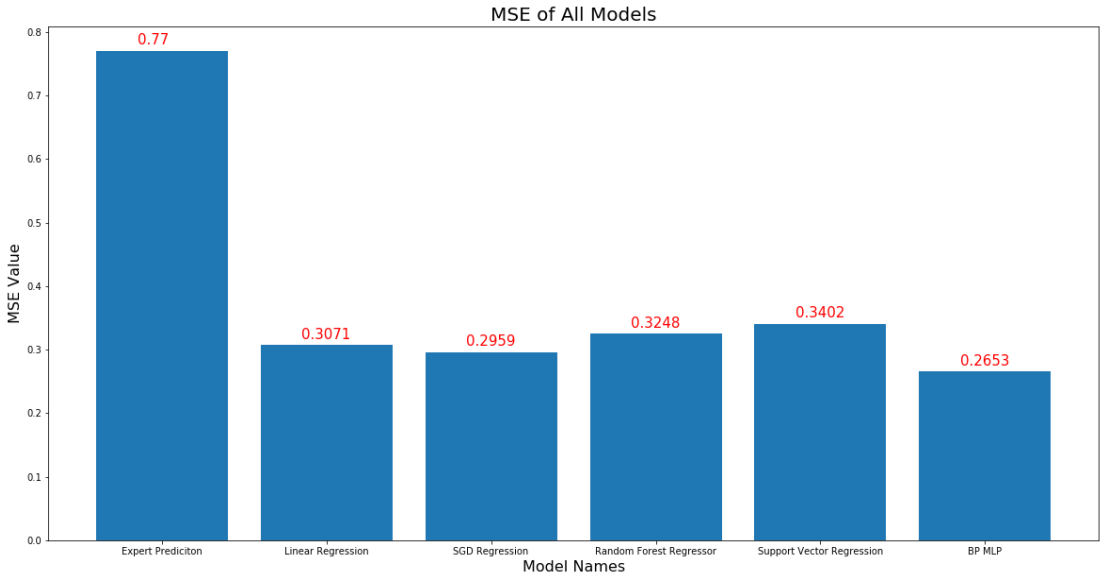


Figure 6

4.2 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 [13].

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.

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

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.

Weights of Variables

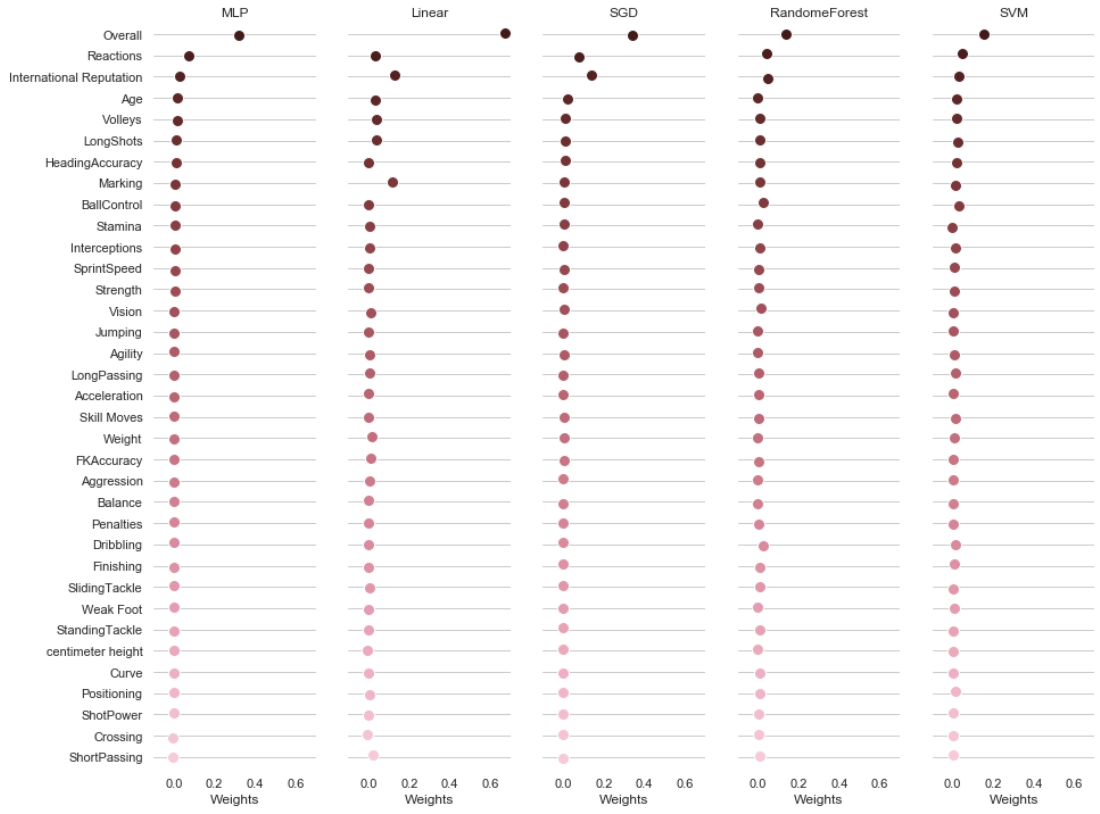


Figure 6

5 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 (whose peak overall value is between 75 to 80).

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. The main reason for the underestimation of these role players might due to their low international reputation.

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

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 4 shows some valuable findings based on the latest data. All players showing in

this table are under 25 years old and they represent the future of soccer. While the first column is the current top 10 players, the second and third column are experts' predicted future rank and the MLP model's predicted future rank, respectively. The fourth column is players who have the biggest divergence between the MLP model's predicted potential and the experts' predicted potential. The common point for these players is that they all have a very low international reputation (1 out of 5).

Table 4

Predictions Comparision (2019)				
Rank	Current Overall	Expert Potential	MLP Potential	Most Undervalued
1	H. Kane	K. Mbappé	H. Kane	J. Damm
2	P. Dybala	P. Dybala	P. Pogba	F. Boulaya
3	K. Mbappé	M. de Ligt	K. Mbappé	L. Krejčí
4	P. Pogba	M. Škriniar	P. Dybala	S. Skrzybski
5	S. Umtiti	H. Kane	H. Son	T. Goiginger
6	R. Sterling	S. Umtiti	J. Kimmich	Hwang Ui Jo
7	M. Icardi	O. Dembélé	R. Lukaku	O. Hernández
8	H. Son	L. Sané	D. Alli	P. Ntep
9	A. Laporte	Vinícius Júnior	R. Sterling	S. Kittel
10	M. Verratti	P. Pogba	M. Icardi	Carles Gil

6 Conclusion

6.1 Summary

In this paper, I build an ML model to predict the potential of soccer players. From the above model section, we could see that the BP MLP model stands out from the horse racing. Considering that the relationship between the player's potential and all other numeric attributes are highly non-linear, this is not a surprising result. Compared with the second-best method, the MSE of the BP MLP model is still 10% lower, which again emphasize the dominant position of this model in the current sports analysis domain. What's more, if I could obtain more training data, the MSE could be further decreased.

According to the permutation importance, Current overall value, international

reputation, Reaction, Age, Volley Ability, Ball Control, and Long Shot Ability have a significant impact on the potential value. Considering that I mixed all players from different positions, this might indicate that these variables are essential for all soccer players regardless of their position. Still, if I could obtain a bigger dataset, I would further divide it according to the position of players and build separate models. Different attributes would stand out in different positions and the accuracy for these models would be improved as well.

Finally, as for the practical significance of my model, the MLP model's prediction is much more accurate than experts' prediction for role players. I won't say that my finding shows that machine learning outperforms human in the prediction of soccer players. However, it might indicate that the ML method for analyzing data would be a great auxiliary tool in the sports management domain. With limited budget and time, experts or scouts could not spend the same time on those noteless role players as for well-known future stars. In that case, given some crucial numeric attributes about those players body index and skill level, my model could show a somewhat reliable prediction. The manager could refer to these results to make their final decisions.

6.2 Future Work

In the future work of my research, besides the enrich of the dataset and subdivide of positions, I hope the model could show some future trend of soccer teams. While the transfer of players is frequent in soccer clubs, the composition of national teams is somewhat stable. Therefore, based on my model, I hope I could predict which team would be the winner of the 2022 Qatar FIFA World Cup. Since the sum of the potential of all players in a team is too arbitrary, I might build a comprehensive index to illustrate the different influence of different positions. Then, based on this evaluation of a whole team and the potential of players, I might show some interesting findings.

7 Acknowledgment

Some of the web crawl codes were modified from the original version of amanthedorkknight's project: https://github.com/amanthedorkknight/fifa18-all-player-statistics

Some of the model codes were modified from our perspective class, credited to Dr. Evans: https://github.com/UC-MACSS

Last, but not least, thanks for all the help and guidance from Dr. Evans in my project. Especially for the concept of the potential for a player and the career overall value curve, these concepts gave me great inspiration to build my models.

All project related codes are in: https://github.com/SiyuanPengMike/persp-research-econ_Spr19/tree/master/Paper's%20Code

8 Appendix

8.1 Literature Review

With the evolution of technologies and statistical tools, sports scientists are intensively using data mining and machine learning skills. At first, the technique is rather simple and straight-forward. Purucker(1996) [1]'s paper is one of the earliest research which used ML to predict sports result. He tried to predict the score of the National Football League (NFL) using an Artificial Neural Network (ANN) model. Instead of using the supervised neural network which is the mainstream method of now, he used unsupervised neural networks based on clustering to distinguish between good and weak teams. He used the best four teams and the worst four teams as the training set and conducted score prediction based on this ANN. His method gives me some hint that I might make an arbitrary prediction on my player data to judge whether a player would be a good or bad player in the future. However, this classification is slightly rough, and the evaluation of the player is not that straight forward as the team. Thus, this method could only be used as a pre-test to the data.

Seven years later, Kahn (2003) [2] extended Purucker's work with the same data source. He switched the model to back-propagation multi-layer perceptron (BP MLP) network, which is a supervised ANN. This switch makes sense that too many modes influence the result of a football game and exposure to a large number of possibilities under supervised learning conditions would yield a network that had better predictive capability. After carefully selection, Kahn used Total yardage differential, rushing yardage differential, Time of possession differential, Turnover differential and Home or away as model parameters and found the best model structure with a learning coefficient (α) of 0.01, momentum (μ) of 0, and a network structure of 10-3-2. As for the prediction accuracy, while Purucker's model is around 60%, Kahn's is about 75% which is even higher than the ESPN experts' prediction accuracy (70%). Kahn's

model's success further proved the advantages of supervised ANN in sport science domain to dealing with the highly non-linear relationship between game result and sports statistical data.

McCabe (2008) [3] further improved the model, and he attempted to predict results in four different sports: NFL (Rugby League), AFL (Australian Rules football), Super Rugby (Rugby Union), and English Premier League Football (EPL). Besides back-propagation, he also used the conjugate gradient as the learning algorithm. The inputs of his model are shared with all these sports, such as points-for, points-against, home performance, away performance and so on. With his three-layer MLP with nineteen input units, ten hidden units and a single output unit, the average performance of his model prediction was around 65%. A highlight in his model is one parameter called Player Availability. That parameter shows the availability of the critical player in the team, and it just gives me some hint that there might be some indispensable parameter that the full status (like a whole team's performance) could be influenced hugely with or without it. In the case of the player potential, that parameter might be the height of the goalkeeper or the speed of the forward.

As for prediction in single player's performance, in Iyer's (2009) [4] paper, he used three different ANN to suggest the coach on whether a player should be invited to the national team or not. The model classified players as "Performer," "Moderate," and "Failure" and the standard for this classification is different within different positions. He used ANN to evaluate the performance of each player in each year, and those who are above the threshold (a specific number of "performer" or "moderate") are qualified for the national team. As for data processing, the author used two different approaches. He cleaned the data and kept all the years in the first database. As for the second one, he segregated the data into different years and further modified the threshold to make sure it is fair for those new face (the data ends at 06/07 season, and some young players join the league at 05/06 or even 06/07). Thus, his model could also recommend

promising rookies. Finally, he used MLP, linear neural network and Radial basis function (RBF) neural networks to predict the performance of the player and his model's average accuracy is about 80% (in which MLP performs the best). The most inspiring part of his paper is not the method like RBF he introduced, but different ways of data processing and segregation. In my model, I will also try to segregate my data in different ways (like years, position, age group and so on) and find the best way with the highest accuracy.

Maszczyk's (2014) [5] further compared the neural networks and non-linear regression to predict the distance of Javelin throws. The data he used involved a group of 116 javelin throwers, aged 18±0.5 years. Using correlation matrix and regression analysis, they found four significant predictors of javelin throw and used them to build the non-linear regression and MLP neural network. The results of the investigation into the group of 18-year-olds javelin throwers show that the created neural models offer a much higher quality of prediction than the nonlinear regression model (absolute network error 16.77m versus absolute regression error 29.45 m), which again shows the superiority of ANN towards other methods on sport analysis.

In Rory's (2019) [5] paper, authors give a machine learning framework for sports result prediction. This paper intensively focused on the classification of the result of the sports match (win, loss or tie) using Artificial Neural Network (ANN). Author decomposes the framework into six steps: Domain Understanding, Data Understanding, Data Preparation & Feature Extraction, Modelling, Model Evaluation and Deploy Model. I will follow this 'SRP-CRISP-DM' framework in my project.

As for other method used in the sports analysis, Miljković's (2010) [7] paper used the Naïve Bayes model to predict the results of NBA matches. The classifiers are host win or visiting win. Besides, they used multivariate linear regression to examine the relationship between the point difference and all other attributes on the spreadsheet.

Finally, using K fold cross-validation as the evaluation system, their model predicts the result with a 67% accuracy, which is not quite ideal. Agarwal's (2017) [8] paper shows the necessity of using Map-Reduce Jobs when facing big data. They used Hive Framework over Hadoop to deal with a large cricket sports dataset. Zdravevski's (2009) [9] and Trawinski's (2010) [10] paper used WEKA (Waikato Environment for Knowledge Analysis, a non-commercial and open-source data mining system) to predict the match result. With Fuzzy system, decision trees, random forest, and other methods, their prediction accuracy is around 70%.

Stanojevic's (2016) [8] paper is the one that most similar to my project. In his paper, he experimented with several ensembles supervised learning methods including random forests, gradient boosting trees regression (GBT) as well as generalized linear models for assessing the player's market value using players' performance data. He then compared the transfermarkt.com (most widely used market value estimates website) market value estimate (TMVE) and his performance-driven market value estimate (PDMVE). The median difference between the TMVE and PDMVE is around 34%, while the mean difference is about 60%. He then used these two market value indicators as the performance predictor and tried to find which one of the two metrics is a better predictor of the game outcome. His result showed that PDMVE (the market value calculate by his model) has higher Pearson correlation, lower median error, and lower RMSE. The whole structure of his paper is inspiring.

Reference

[1] M.C. Purucker, Neural network quarterbacking, IEEE Potentials 15 (1996) 9–15.

[2] Kahn, J. (2003). Neural network prediction of NFL football games. World Wide Web electronic publication, 9-15.

[3] McCabe, A., & Trevathan, J. (2008, April). Artificial intelligence in sports prediction. In Fifth International Conference on Information Technology: New Generations (itng 2008) (pp. 1194-1197). IEEE.

[4] Iyer, S. R., & Sharda, R. (2009). Prediction of athletes performance using neural networks: An application in cricket team selection. Expert Systems with Applications, 36(3), 5510-5522.

[5] Maszczyk, A., Gołaś, A., Pietraszewski, P., Roczniok, R., Zając, A., & Stanula, A. (2014). Application of neural and regression models in sports results prediction. Procedia-Social and Behavioral Sciences, 117, 482-487.

[6] Iyer, S. R., & Sharda, R. (2009). Prediction of athletes performance using neural networks: An application in cricket team selection. Expert Systems with Applications, 36(3), 5510-5522.

[7] Miljković, D., Gajić, L., Kovačević, A., & Konjović, Z. (2010, September). The use of data mining for basketball matches outcomes prediction. In IEEE 8th International Symposium on Intelligent Systems and Informatics (pp. 309-312). IEEE.

[8] Agarwal, S., Yadav, L., & Mehta, S. (2017). Cricket Team Prediction with Hadoop: Statistical Modeling Approach. Procedia Computer Science, 122, 525-532.

[9] Zdravevski, E., & Kulakov, A. (2009, September). System for Prediction of the Winner in a Sports Game. In International Conference on ICT Innovations (pp. 55-63). Springer, Berlin, Heidelberg.

[10] Trawinski, K. (2010, July). A fuzzy classification system for prediction of the results of the basketball games. In International conference on fuzzy systems (pp. 1-7). IEEE.

[11] Stanojevic, R., & Gyarmati, L. (2016, December). Towards data-driven football player assessment. In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW) (pp. 167-172). IEEE.

[12] 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.
[13] https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html