Predict Soccer Goalkeepers' Values Using FIFA Video Game Data

Introduction

Signing excellent soccer players is one of the most important tasks for club managers and coaches to improve the performance of their teams. Since their budgets are usually limited, they need to estimate the values of the players in order to allocate their budget wisely. This project primarily aims at building a model to predict goalkeepers' values based on their various attributes.

The data source used in the project is the soccer player information database from the latest version of FIFA video game. The dataset accurately reflects current soccer players' personal information, skills, values, etc. Therefore, it is sensible to use it to simulate the real-world situation, and to analyze the association between players' values and various characteristics of the players.

The value of goalkeepers is treated as a response, while the other 38 variables in the dataset are all potential predictors. We first combined different nations on the same continent into a new variable which has fewer categories. The resulting dataset contains one categorical variable "continent", which indicates the continent where the player is from, and the other variables in the dataset are all numerical variables. The detailed description of these variables can be found in the attached file "Variable description". Next, the dataset was randomly split into training and testing datasets (75% vs. 25%). The training dataset(Table 1) contains 1523 observations, and the test dataset contains 506 observations. In the training dataset, the distribution of the player's value is highly right-skewed. Therefore, we transformed it using quartic root(i.e., ¼-power) transformation.

Exploratory data analysis

In the exploratory data analysis for the categorical variable, we found that players from different continents tend to have different distributions in value. For numerical variables(Fig. 1 and Fig. 2), it can be seen that players' values increase with increasing diving skill, handling skill, positioning skill, and reflection speed. The relationships between players' transformed values and these attributes of players appear to be non-linear. Besides, players' transformed values also show a non-linear relationship with age. In general, players' values increase with the increase of age for players younger than 30 years old, whereas it decreases with the increase of age for players older than 30 years old. In addition, it is observable that some observations that are at the outer range of some potential predictors have unusual response values which depart the trend shown by other observations.

Models

Median was used to impute the missing values. Multiple linear regression model was first fitted to predict players' values for simplicity and easier interpretation purpose. However, given that most of the predictors are about players' abilities and skills, they might be correlated with each other. Thus, we also tried the Ridge regression, the Lasso, the Elastic net, Principal components regression and Partial least squares to deal with the potential collinearity between predictors. Multiple linear regression was fitted using all the predictors. For other linear models, the range of the tuning parameters was carefully chosen and the best tuning parameter was determined by the mean squared error of 10-fold cross-validation method.

The generalized additive model (GAM) and the multivariate adaptive regression splines model (MARS) were built to capture the non-linear relationship between players' transformed values and selected predictors.

The GAM we fitted used the smoothing spline method to build each block for every continuous predictor. The degrees of freedom of each continuous variable was obtained by the generalized cross-validation. The GAM enables nonlinearity of several variables, so it performs better than standard linear regression when there are non-linear relationships between the response and predictors.

We also tried the MARS to model non-linear relationships. There are two tuning parameters in this model building process, the degree of features (including interaction terms or not) and the number of predictors. The best tuning parameters were chosen by the mean squared error of the 10-fold cross-validation. The chosen model contains 18 terms which were made up of 12 selected predictors. The final multivariate adaptive regression splines model we chose is:

$$
\begin{aligned}
Y_{transformed} &= 0.703 + 0.041X_{Europe} - 0.024h(X_{age} - 29) - 0.057h(X_{age} - 33) + 0.128h(X_{age} - 39) \\
&-0.005h(71-X_{potential})+0.017h(X_{potential}-71)-0.002h(61-X_{agility})-0.003h(X_{balance}-47)-0.007h(68-X_{gk\_diving}) \\
&+0.016h(X_{gk\_diving}-68)-0.005h(67-X_{gk\_handling})+0.014h(X_{gk\_handling}-67)+0.009h(X_{gk\_kicking}-72) \\
&+0.008h(X_{gk\_positioning}-43)+0.015h(X_{gk\_reflexes}-71)+0.011h(X_{reactions}-63)-0.013h(41-X_{strength})
\end{aligned}
$$

$$h(x) = x_+$$

By comparing the mean squared error and R-squared of the 10-fold cross-validation, the boxplots (Fig. 3) clearly show that non-linear models perform better than linear models. Furthermore, considering the results in exploratory analyses, we know that there are non-linear relationships between the response and some variables. Thus, we chose two non-linear models (GAM and MARS) as our final models.

Comparing the non-linear models, the MARS uses fewer predictors and have a more stable prediction performance. The predictors in the MARS also make sense in determining goalkeepers' values in the real world. Using the varImp function in the caret package to calculate variable importance for the MARS, it is shown that diving skill, age, positioning skill, reflection speed, and potential are the top five most important variables among 12 selected variables. In general, partial dependence plots for each continuous feature in the MARS (Fig. 4) are consistent with the patterns in the exploratory analysis. However, regarding age, we expected the value of players older than 39 years old would decrease with increasing age, which is opposite to what was given by the model. The reason might be that there are only a few players in our training dataset whose ages are above 39 and these players have unusually high values. Hence, this model may not be generalized to players over 39 years old.

Using the varImp function, it is shown that age, vision, penalties, balance and long passing are the top five most important variables among all predictors in the GAM. The result is consistent with what we expected in general. In the real world, we know that age, vision, penalties, and long passing should be among the most important variables in determining goalkeepers' values, although penalty skill may not be really important for goalkeepers.

By calculating the test error, we further confirmed that non-linear models perform better than linear models. By comparing the training error and the test error (Table 2), as the test error is over 30% higher than the training error both in the MARS and the GAM, it is possible that our final models have a problem of over-fitting.

Conclusion

There are non-linear relationships between the response and predictors. The GAM and MARS methods are more appropriate for prediction of the players' values. The player's value tends to increase with stronger abilities and higher skills, including player's agility score, diving skill, etc.

However, the final two models we chose also have some limitations. Although the MARS captures the nonlinearity, it may not model the complete nonlinearity as it can only fit straight line within each piece. As for the GAM we built, interaction terms cannot be added in the model using the caret package. Possibly other non-linear models fit the data better and have a better performance in prediction. Thus, in the future, it is possible to try more efficient non-linear models. In addition, since there are only a few observations in the training dataset that are outliers in the value of some of the predictors like age, the model has large variance at the outer range of those predictors, and it may not do well in the prediction for observations that have outlying values in those predictors.

Table 1. Descriptive statistics of continuous variables in the training dataset

variable	Min	1st Q	Mean	Median	3rd Q	Max	Std Dev	missing
acceleration	11	31	38.91	40	46	65	10.87	0
age	16	22	26.11	26	30	47	5.4	0
aggression	11	21	26.62	25	32.5	46	7.9	0
agility	14	32	40.6	38	48	74	11.53	0
balance	11	35	43.22	43	51	70	10.86	2
ball_control	8	16	20.07	20	23	54	5.71	3
composure	11	27	36.45	33	45	71	12.71	1
crossing	5	12	14.42	14	17	41	3.9	0
curve	6	12	14.84	14	17	65	4.57	0
dribbling	2	11	14.14	14	17	33	4.27	1
finishing	2	10	12.49	12	15	34	4.02	1
free_kick_accuracy	4	12	14.59	14	16	66	4.41	0
gk_diving	39	60	65.28	65	70	91	7.86	18
gk_handling	43	58	62.94	63	68	91	7.84	18
gk_kicking	39	56	61.56	61	66	95	7.91	13
gk_positioning	38	57	63.09	63	69	91	8.79	23
gk_reflexes	37	61	66.18	66	72	90	8.33	16
heading_accuracy	4	12	14.68	14	17	45	4.16	0
interceptions	4	13	17.6	18	22	55	5.8	2
jumping	13	52	58.05	59	66	85	11.43	4
long_passing	9	20	25.54	24	30	73	7.95	1
long_shots	3	10	13.21	13	16	40	4.39	1
marking	4	10	12.88	13	16	35	4.27	1
penalties	5	16	20.47	20	24	73	6.83	0

positioning	2	9	11.82	12	15	27	4.13	1
potential	46	65	69.7	69	74	94	6.41	0
reactions	28	53	59.31	60	67	86	10.21	6
short_passing	11	22.25	26.97	26	31	59	7.28	1
shot_power	3	19	22.53	22	24	70	7.1	1
sliding_tackle	4	12	14.22	13	16	35	3.47	0
special	728	974.5	1051.11	1063	1137	1493	125.85	0
sprint_speed	11	32	39.38	41	47	70	10.84	0
stamina	12	25	30.81	31	36	48	7.54	2
standing_tackle	6	12	14.3	14	16	33	3.39	0
strength	12	54	61.2	63	69	85	11.27	2
vision	10	27	36.06	35	45	71	12.73	1
volleys	4	10	13.07	13	16	40	4.57	1
transformed_value	0	0.61	0.86	0.78	0.99	2.79	0.38	0

Table 2. train errors and test errors of different models

Model	**Train error**	**Test error**
Multiple linear regression	0.0306	0.0399
Ridge regression	0.0317	0.0393
The Lasso	0.0315	0.0392
The Elastic net	0.0314	0.0392
PCR	0.0332	0.0401
PLS	0.0306	0.0399
MARS	0.0195	0.0260
GAM	0.0168	0.0272

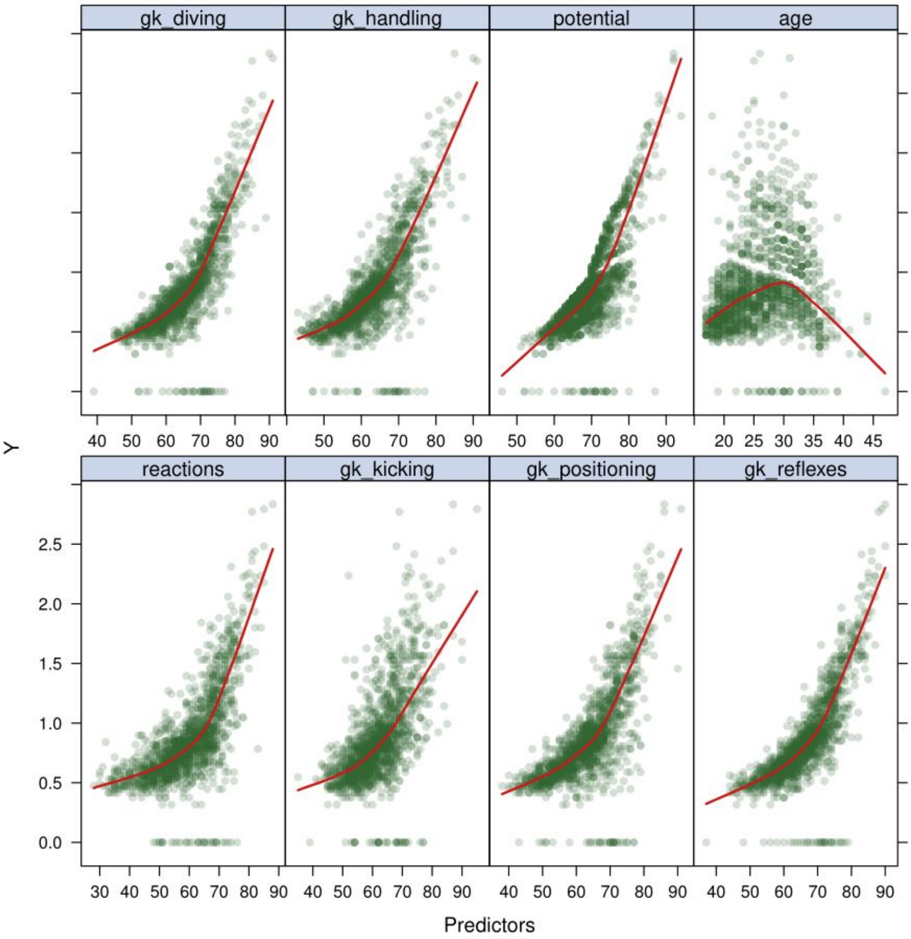

Figure 1. Plots of the response (transformed player's value) vs. some of the continuous variables

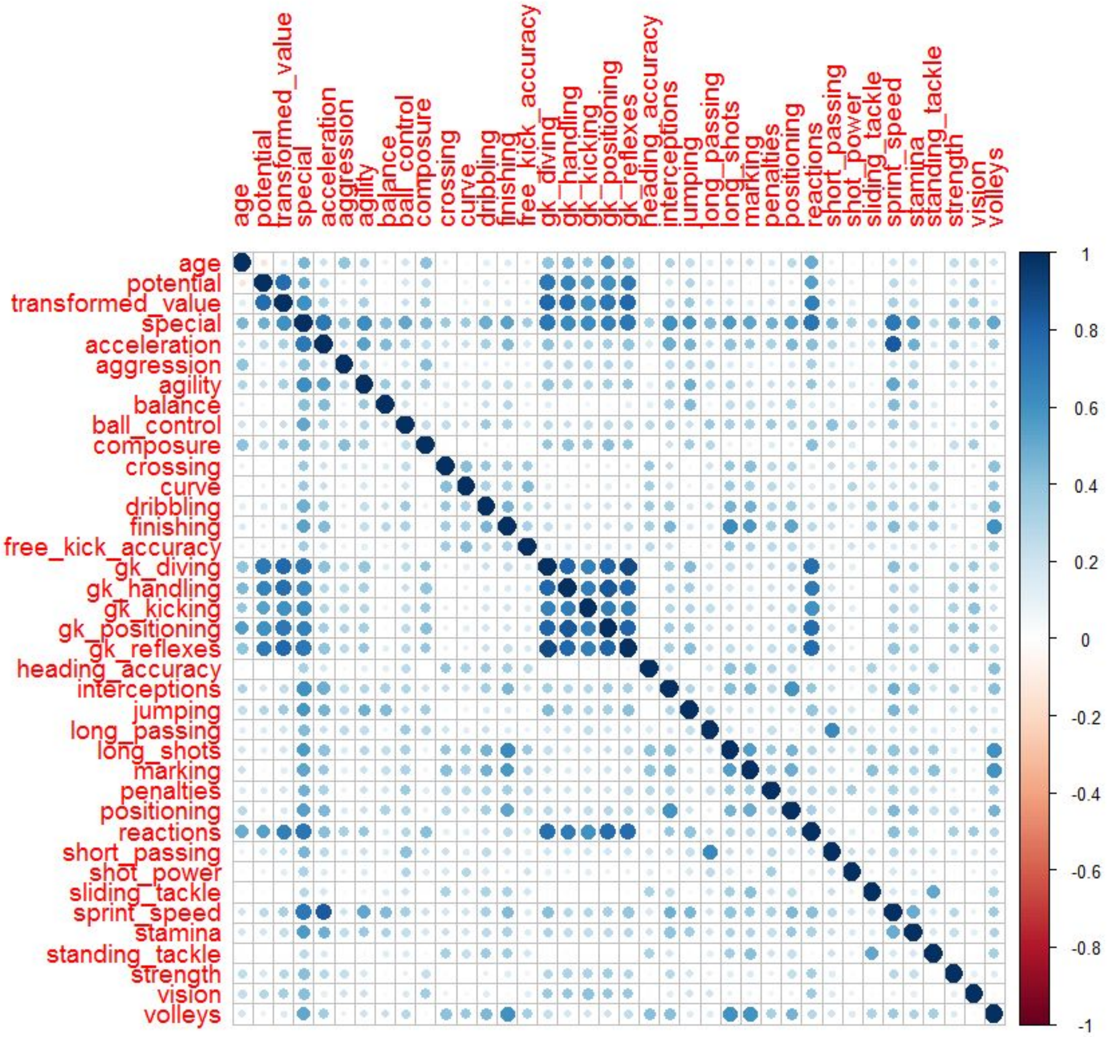

Figure 2. Correlation plot of the response and all the continuous predictors based on the training dataset

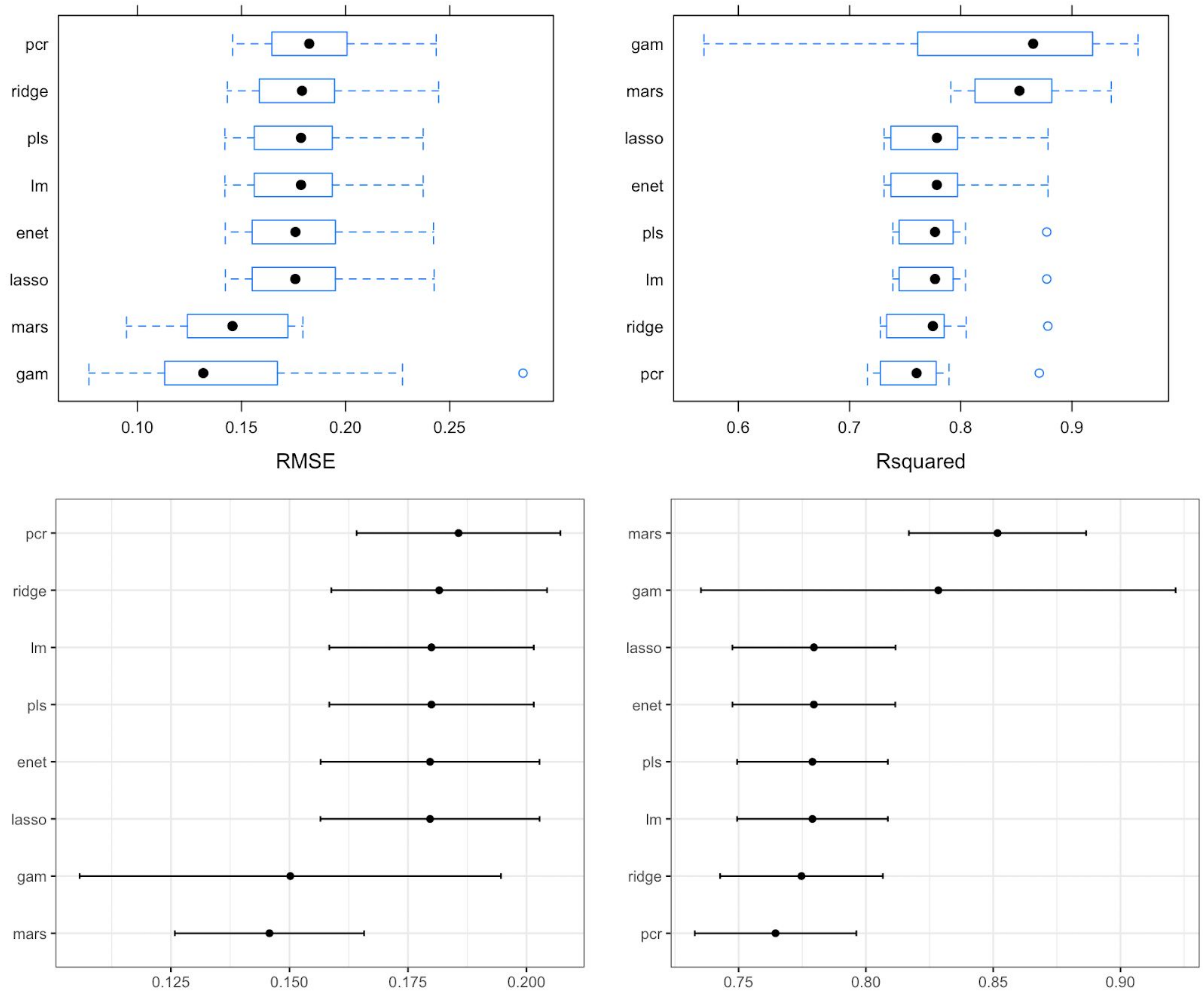

Figure 3. Boxplots of the RMSE and R-squared values for different models by 10-fold cross-validation (above). The average value (with two-sided confidence limits) of RMSE and R-squared in cross-validation for each model (below).

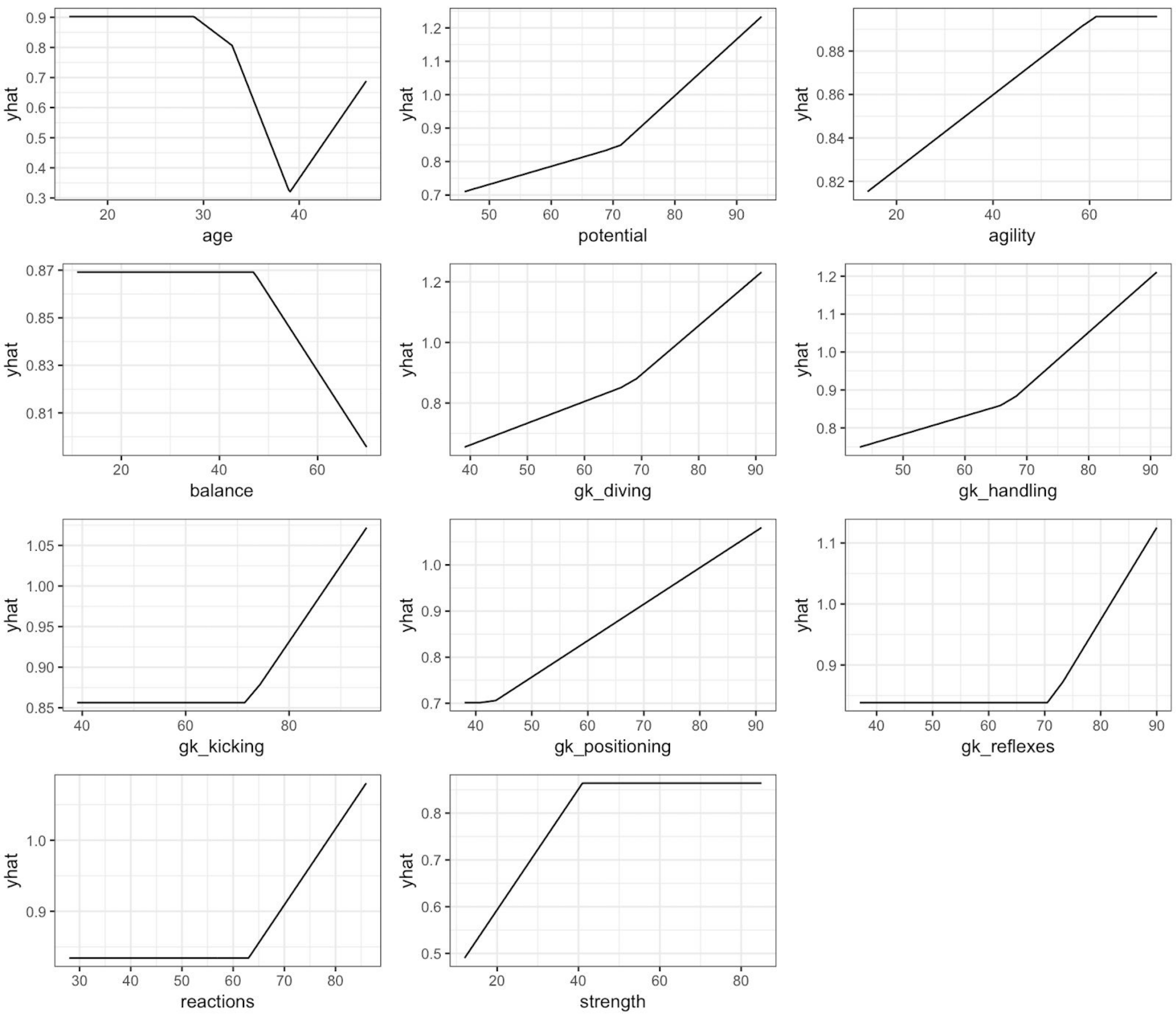

Figure 4. Partial dependence plots for each continuous feature in the multivariate adaptive regression splines model, reflecting the marginal effects of predictors.