What factors influence the overall rating of soccer players in the FIFA video game*

Yizhuo Liu, Leo Cheng, Haobo Ren

October-02-2024

Contributions

Yizhuo Liu: Data cleaning, introduction
Leo Cheng: Preliminary result
Haobo Ren: Data description

Introduction

In the long history of soccer, there have been many ways of evaluating the performance of soccer players. Whether by the number of goals they score or the number of trophies they earn, but these are very unilateral methods that only favor a certain kind of player, not everyone, especially when it comes to unknown players. Certainly, the video game franchise FIFA comes to mind. Every year FIFA evaluates professional soccer player's previous year's performance and gives them a rating from 0 to 100. From 1993 to now, there has been a new FIFA game every single year, yet no one knows what factors influence the rating from FIFA that the majority of the people agree on. Therefore, the main objective of this report and the research question is going to be "What factors influence the overall rating of soccer players in the FIFA video game". If we can accomplish this objective, then we can use it to quantify players' performance and predict further performance and other stuff related to performance. With our research question in mind, we proposed a hypothesis: factors wage, international reputation, age, whether substitution, club, value in EUR, height, weight and release clause in EUR will affect the response variable - FIFA overall rating. These variables are chosen to be our predictors in the preliminary model. To back up our hypothesis, the article "Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques" (Al-Asadi and Tasdemır, 2022), which is on using FIFA to get a data-driven approach to player valuation, also uses very similar predictors that we have chosen, and the base model they end up using is also linear regression. In addition, in "PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach" (Pappalardo et al., 2019) they use real-world data to quantify players' performance with a reasonable success. Last but not least, the article "Predicting the Future Performance of Soccer Players" (Arndt and Brefeld, 2016), uses a combination of linear regression and multitask regression, to evaluate current player performance and predict the outcome of a soccer game. Two out of the three articles have chosen linear regression, therefore it shows that linear regression is commonly used in this area. Also, by looking at the scatter plot for our response variables and predictions, there is a linear relationship. By using linear regression method we are looking to get an accurate prediction of FIFA rating.

*Code and data are available at: https://github.com/HaoboRrrr/FIFA20_Player_Potential_Rating_Analysis/tree/main

Methods

We outline our methods as follows: First, we check the linear assumption for each of our predictors. Second, we fit our preliminary model and plot the corresponding diagrams to address the potential problem of our preliminary model. Third, we use the diagrams plotted to decide whether to perform transformations or add interaction terms to the preliminary model. Fourth, on the newly developed model we will perform a hypothesis test to test the significance of the coefficients as well as AIC backward selection. Finally, we will validate our model using VIF multicollinearity and Cook's distance and address our research question. The following sections give a detailed explanation.

Linear Assumption Check

We use residuals against fitted values scatter plot to check linearity, independence of error, and constant variance assumptions. - Curves or trends suggest non-linearity; Box-Cox transformation is applied. - Residual clustering indicates dependence issues; we add interaction terms. - Funnel-shaped residuals suggest heteroscedasticity; we use Box-Cox transformation.

After the check, the residuals will randomly scatter around 0 and show no pattern. Lastly, on the Q-Q plot, residual points deviating from the diagonal line suggest violation of the normal error assumption. If this is the case, we apply power transformation on the response variable.

Hypothesis Test & Variable Selection

t-test and F-test can be used to further check and improve our model. First of all, T-test is used to determine whether individual predictors in the model are statistically significant in explaining dependent variables. To begin with, we choose our null and alternative hypothesis $H_0 : \beta = 0$, $H_1 : \beta \neq 0$. Then, we conduct a t-test on each predictor, we get p-values for each predictor and if the p-value is smaller than 0.05, our chosen significant level, reject H_0. Otherwise, drop the variable along with all interaction terms containing it. For categorical predictors, we will drop all terms. After the t-test, the F test evaluates whether the independent variables, as a group, explain a significant portion of the variation in the dependent variable. If the F-test fails, we need to reconsider our preliminary model and choose a new one.

Validations and Interpretations

After we get the final model, we will validate our model. The first step is to check whether there exists multicollinearity between predictors by using variance inflation factor (VIF). If the VIF for some predictor is too big, showing that this variable has a high collinear relationship with others, we will drop the predictor and get a new model. Next, we will check which data points are influential using the Cook's distance. After choosing a threshold, we will drop all extreme data points and fit a more robust model. The final step is to perform backward selection along with AIC (Akaike criterion) to choose our final model with the most effective predictors. The final model can answer our research question by telling us which predictors have the greatest effect on the overall rating by comparing the relative magnitude of the coefficient, and what is the linear relationship after performing transformation on the dataset. We will calculate the confidence interval and interpret the most impactful predictor.

Results

Our model is fitted following the methods described above, and we will discuss our results in each of the subsections. Throughout, we will denote our model at different stages with M_i.

Linear Assumptions

The preliminary model we fit on the initial predictors introduced in the introduction is denoted as M_0 and is shown in Table 2. The following chart shows the comparison of the original series of graphs of assumption checks and transformed graphs. First of all, we drop weight and height since those two variables does not exhibit any linearity with other variables (c.f. appendix, Figure.).

The results of the assumption checks before and after applying the Box-Cox transformations demonstrate significant improvements in meeting the linear regression assumptions. Prior to transformation, the residuals versus fitted values plot showed clear patterns, indicating non-linearity and heteroscedasticity. Besides, the Q-Q plot revealed deviations from the diagonal line, particularly at the tails, suggesting non-normal residuals. Using the exact Box-Cox lambda values for the response variable ($\lambda = 2.4$) and predictors (e.g., $\lambda = 0.4646$ for age, $\lambda = 2$ for value and wage), $overall$, age, $wage$, $value$ and $release$ are transformed to $overall^{2.4}$, $(age^{0}.465 - 1)/0.465$, $wage^{3.7}$, $value^{5.1}$ and $release^{8}$ respectively. After the transformations, the residuals versus fitted values plot almost exhibits a random scatter around the horizontal axis, resolving the earlier patterns and indicating that both linearity and homoscedasticity have been addressed. Similarly, the Q-Q plot shows the residuals closely following the diagonal line, reflecting a significant improvement in the normality of residuals. These results confirm that the exact power transformations have successfully aligned the data with the assumptions of linear regression, enhancing the model's reliability. The transformed model M_1 is shown in the table.

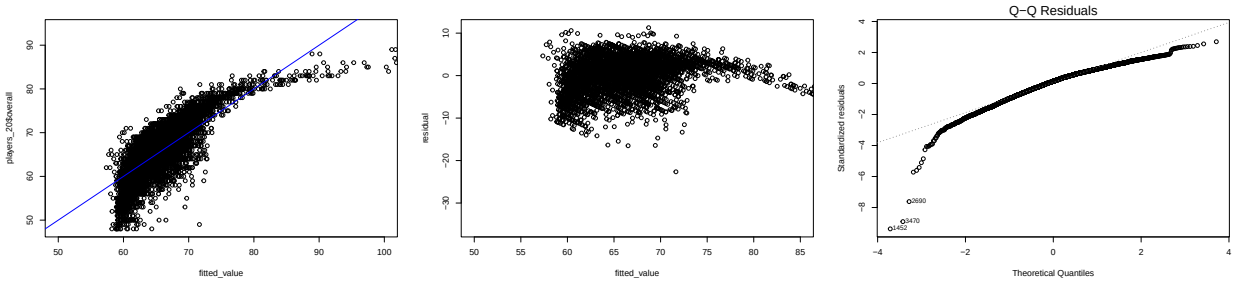

Figure 1: The reponse v.s. fitted (left), residual v.s. fitted (middle) and Q-Q plot (right) of our preliminary model M_0

Variable Selections

t-Test and F-test

We performed t-Test and F-test on our predictors. According to Table 1, all of the predictors have a very small p-value except for predictor value_eur, therefore rejecting the null hypothesis for all predictors but value_eur. So we are keeping all the existing predictors minus value_eur and its interaction term. This means the majority of our predictors are statistically significant in explaining the overall rating. This gives us a new model M_2 After fixing our predictors, we have done a F-test to determine the overall significance of the regression model. All of our predictors left have a high F value and as a group, which shows that our model successfully rejects the null hypothesis. This confirms that the model as a whole explains a significant portion of the variability in the data. So all of our predictors except value_eur are factors that may potentially influence overall rating in FIFA2020 according to t-test and F-test.

Table 1: The ANOVA table of performing t-test and F-test.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age_tran	1	3.604509e+10	3.604509e+10	2037.8491514	0.0000000
wage_tran	1	1.994841e+09	1.994841e+09	112.7805664	0.0000000
international_reputation	1	2.737795e+10	2.737795e+10	1547.8430428	0.0000000
is_sub	1	8.551313e+09	8.551313e+09	483.4580408	0.0000000
Top_Club	3	9.040277e+09	3.013426e+09	170.3673750	0.0000000
value_tran	1	2.878984e+07	2.878984e+07	1.6276657	0.2020872
release_tran	1	3.868536e+08	3.868536e+08	21.8712020	0.0000030
value_tran:release_tran	1	5.169354e+05	5.169354e+05	0.0292255	0.8642665
age_tran:value_tran	1	1.062276e+07	1.062276e+07	0.6005694	0.4383986
wage_tran:value_tran	1	4.203376e+07	4.203376e+07	2.3764256	0.1232442
Residuals	4859	8.594506e+10	1.768781e+07	NA	NA

Backward selection - AIC

The backward selection using AIC criteria gives an unchanged model. The model starts with an AIC of 15460. Wage is dropped using backward selection since dropping them will result in a lower AIC. This gives us the model M_3.

Validations

We checked if there is any relationship between our predictors, as well as what datapoint may influence our model prediction a lot to validate the effect of our linear model.

Multicollinearity

We calculate the VIF for each predictor. We regard predictors with VIF greater than 5 as the multicollinear variable showing an explosion of variance, except for the inherently multicollinear interaction term. The result is that we have dropped the variable release since it has a VIF of 157, and give us the new model M_4. Other variables except for the interaction term and variable involved in the interaction term have a VIF below threshold.

Influential points

The Cook's distance is shown in the diagram below. There are several points that are very outstanding on the diagram. We use the cutoff rule of $\frac{4}{n}$, where n is the number of observations we have. In this case, the cutoff is 8.33×10^{-4}. These outliers are the values that significantly affect our inference on the coefficients since some of them are not representative enough for other players. For example, the player with the highest Cook's distance is L. Messi, who deviates significantly from other players. Therefore, we remove them and refit the model to make the model more robust. The refitted model is M_5.

Model Scoring and Assumption Check

The model has residual standard error=624.6, $R^2 = 0.472$ value and adjusted $R^2 = 0.471$. These are our criterions for assessing the model. We also plot the final response v.s. fitted value plot and other residual plots in Figure. 3. The final plot basically meets all our expectations on linear assumption as stated above, with a minor violation for the residual v.s. fitted plot, which will be discussed in the Conclusions section. After this, we get our final model M_5.

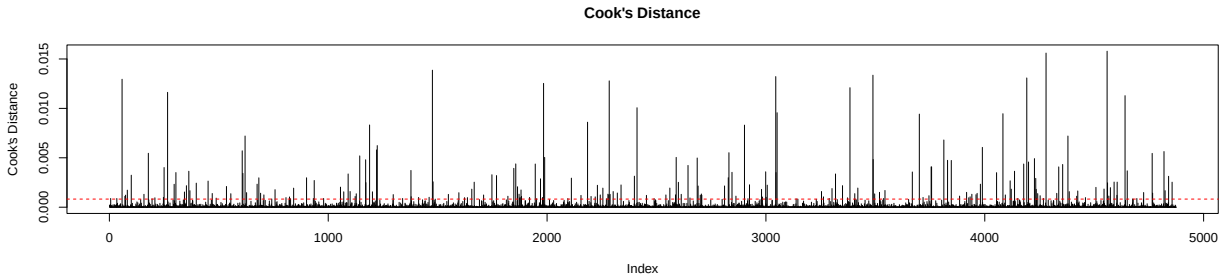

Figure 2: The Cook's distance of model M_4. Note that there are many extreme observations in our 4800 observations. The red horizontal line shows the cutoff we use. We will drop all data points above the cutoff line and fit the new model agian.

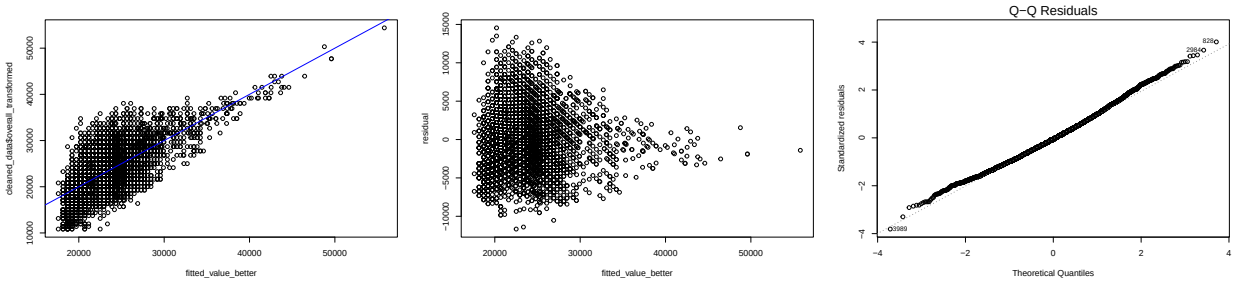

Figure 3: The reponse v.s. fitted (left), residual v.s. fitted (middle) and Q-Q plot (right) of our final model M_1. We can see that compare to Fig. 1, all plots are improved in the sense of assumption checking discussed in the paragraph.

Table 2: A summary of all regression models we developed from M_0 to M_4. Except for M_0, all *overall* refers to overall$^{2.4}$. age_trans=age^2, value_tran=value_eur^8, release_tran=release_clause_eur$^{5.1}$, wage_tran=wage_eur$^{3.7}$.

Model	Formula
M_0	overall = 56.442 + 0.549 * age + 0 * wage_eur + -0.32 * international_reputation + -2.08 * is_sub + -2.126 * Top_ClubTop_10 + 3.341 * Top_ClubTop_20 + 3.264 * Top_ClubTop_30 + 0 * value_eur + -0.053 * height_cm + 0.074 * weight_kg + 0 * release_clause_eur
M_1	overall = 1129.482 + 368.482 * age_tran + 0 * wage_tran + 724.868 * international_reputation + -451.703 * is_sub + 1091.352 * Top_ClubTop_10 + 943.987 * Top_ClubTop_20 + 940.517 * Top_ClubTop_30 + 0 * value_tran + 0 * release_tran + 0 * value_tran:release_tran + 0 * age_tran:value_tran + 0 * wage_tran:value_tran
M_2	overall = 1129.482 + 368.482 * age_tran + 0 * wage_tran + 724.868 * international_reputation + -451.703 * is_sub + 1091.352 * Top_ClubTop_10 + 943.987 * Top_ClubTop_20 + 940.517 * Top_ClubTop_30 + 0 * release_tran
M_3	overall = 1139.68 + 367.313 * age_tran + 725.527 * international_reputation + -455.116 * is_sub + 1131.296 * Top_ClubTop_10 + 958.418 * Top_ClubTop_20 + 950.124 * Top_ClubTop_30 + 0 * release_tran
M_4	overall = 1139.68 + 367.313 * age_tran + 725.527 * international_reputation + -455.116 * is_sub + 1131.296 * Top_ClubTop_10 + 958.418 * Top_ClubTop_20 + 950.124 * Top_ClubTop_30
M_5 (final)	overall = 831.556 + 396.197 * age_tran + 796.439 * international_reputation + -424.303 * is_sub + 971.125 * Top_ClubTop_10 + 998.555 * Top_ClubTop_20 + 972.84 * Top_ClubTop_30

Conclusions

After all the process of assumption checks, variable selections and validations, we arrive at our final model. The term with the highest coefficient is the categorical variable Top Club with level top 20. Among the others, the most important variables are age and international reputation. This gives us a direct answer to the research question: the factors that influence the overall rating the most are the clubs the players are in, the age of the player and the international reputation.

Take the most impactful variable Top Club as an example, we can interpret it as: if a player has all other predictors fixed and its club changes from the quantile above 30 to the top 20 club, then its average overall rating^2.4 will increase by 998 (i.e. the average overall rating increases by 17.77. Also, we have a 95% confidence to conclude that the true coefficient is between 816 and 1181 since if we were to repeatedly randomly sample from the population and compute a 95% confidence interval, then 95% of the intervals would include the true coefficient. This shows that the club a player is in is very important in predicting the overall rating, namely a higher ranking club will result in a much higher rating than the others.

Another surprising observation is that age is a very important factor in predicting a player's overall rating in our model compared to literature, (Al-Asadi and Tasdemır, 2022).

The overall model meets our expectation on accuracy (reaching a R^2 of 0.) and on linear assumptions. However, due to the complexity of the dataset, there are still limitations for this model. First of all, there are still assumption violations. The residual v.s. fitted model shows a slightly converging trend, which may imply a violation of homoscedasticity. In addition, extreme observations are common due to the nature of soccer players, shown in Figure 2. The model is still sensitive to extreme values. The third problem is on the dataset itself. Our ultimate goal is to quantify and predict a player's performance. Since the dataset is collected from a video game, there may be bias and may not reflect the real-world scenario. This may be fixed by combining some real-world data with the synthetic one.

References

Appendix

We will include several plots in the appendix.

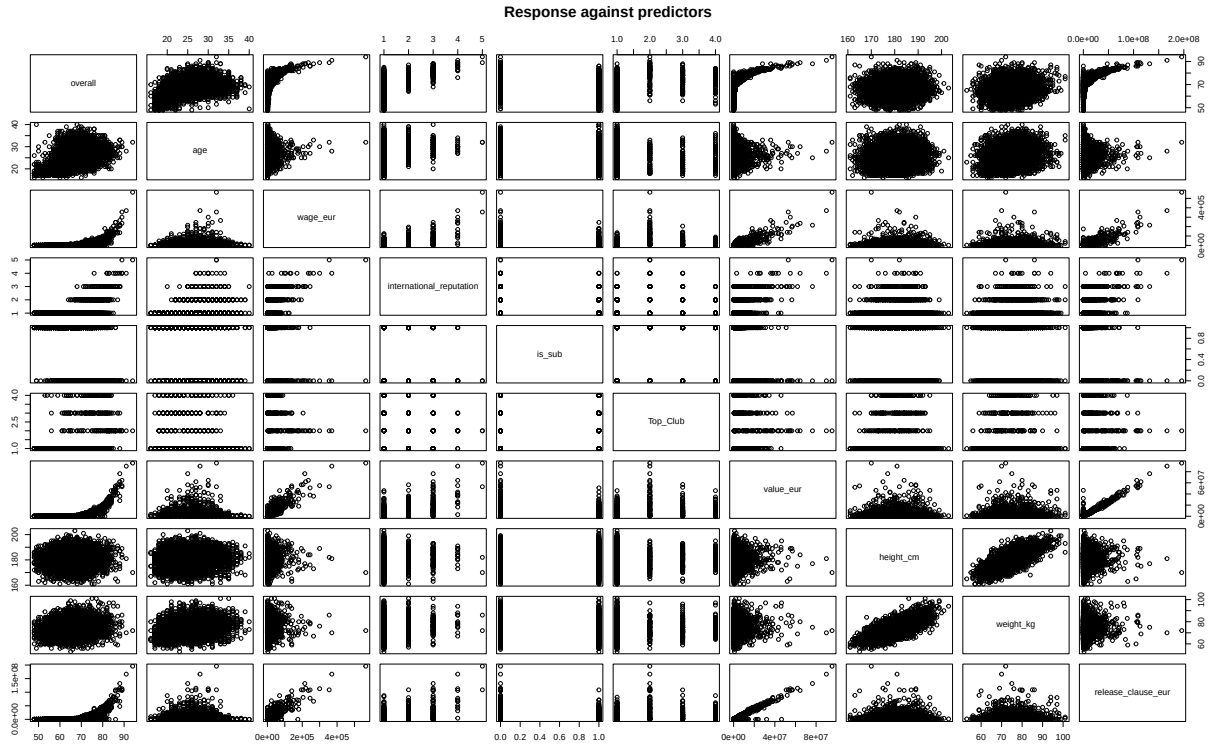

Figure 4: The response v.s. predictor and predictor v.s. predictor diagram. We can infer linearity from inspecting the relationship between different predictors and response.