Using Data to Find the Predict FIFA Wonderkids

Zack Eisman

STATS 401: Applied Statistical Methods II

April 15, 2024

Background

The dataset used for this project is from the EA sports video game FIFA 2022. Specifically, this data contains information on all the individual players play in Europe's top five leagues (English Premier League, Spanish La Liga, Italian Serie A, German Bundesliga, and French Ligue 1). Each row includes personal and professional information as well as ability related overalls.

This project explores how certain quantitative and categorical factors can predict a player's potential overall. This value generally refers to how good a player can become in the future. There are many potential quantitative predictor variables including the player's current overall which is a rating between 45 and 99 that references how good a player is, age in years, player transfer value (in euros), and wage (in euros). Additionally, the league that they play in can be used as a categorical predictor. Overall, the goal of this project is to predict a player's potential based on various personal and ability related factors. Initially, this analysis will measure the impact of overall, age, and transfer value on potential and determine if the league also has any impact on these relationships.

Analysis

Initial Model Evaluation

Initially, a linear model was created to assess the relationship between player potential and overall, age, value, and league, as shown in Figure 1. However, this model has significant flaws as it does not meet all the required assumptions needed to be properly analyzed. The residual vs fitted values plot in Figure 1 shows that assumption of constant variation is not met which means that the model will need some form of transformations in the future. However, the model does meet the assumption of linearity and the assumption of normality because the sample size is sufficiently large despite the QQ plot deviating from the identity line (Figure 1).

Log Transformations and Quadratic Fits

By analyzing the residuals vs fitted values graph in Figure 1, it can be determined that a log transformation will be needed for at least one of the variables because the model does not pass the assumption of constant variance due to the "fanning inward" trend from left to right. To determine which variable(s) require a log transformation, histograms of each variable can be analyzed (Figure 2). Clearly, the age and transfer value predictors have very heavy right skews indicating that they would benefit from a log transformation. By adding these two transformations, the overall model can now pass the assumption of constant variance, as shown in Figure 3 as all residual vs fitted values are evenly spaced across the graph.

Additionally, a scatter plot matrix can be used to confirm that the model does meet assumption of linearity as there are no strong non-linear relationship between player potential and the transformed predictors (Figure 4). As a result, there is no need to add a quadratic fit to the model.

By adding the two log transformations, a new model can be created that appropriately passes all the required assumptions. This model has a root mean squared error value of 2.061 and R^2 value of 0.8403, which are both strong values in context (Figure 5).

Interactions

One way to potentially further improve the model is by adding an interaction between the league and overall, which explores whether the relationship between potential and overall is dependent on the league. These relationships are plotted in Figure 6, where the different slopes and intercepts can be seen side-by-side. These differences are confirmed in Figure 7, which shows that the R^2-adjusted value increases from 0.8399 to 0.8434 when this interaction term is included in the model indicating that this addition improves the overall model.

Swapping Predictors

Another way to improve the model is by looking at altering which predictors are used in the model. Specifically, the transfer value predictor appears to have a high p-value in the improved model (Figure 7), which could mean that it is an unnecessary predictor. So, it is worth exploring whether an alternate model with a different predictor such as wage would still be equally effective. Like transfer value, wage has a heavy right skew so a log transformation must be performed to maintain the assumption of constant variance (Figure 8). This result of this swap is explored in the output of Figure 8 which shows that the resulting model would have a larger R^2-adjusted value of 0.8486 with the inclusion of the log(wage) predictor, indicating that it is better to swap the predictors in the model.

Final Model

Assumptions/Diagnostic Plots

After completing all the above steps, a final model can be created that best predicts potential overall based on the combination of predictors used. To verify that this final model can be appropriately used, the diagnostic plots must be analyzed to confirm the model passes all the assumptions. As seen in the residuals vs fitted values graph in Figure 9, the values appear to be evenly distributed and follow a linear relationship indicating that the assumptions of linearity and constant variance can both be reasonably met. Additionally, the QQ plot shows that there is a slight deviation from the identity line, but this the assumption of normality can still be met because the sample size of 2,878 rows is sufficiently large.

The assumption of independence is also very important to the legitimacy of a linear model so the data collection method for this FIFA dataset is important to explore. It is likely that most of the data in this dataset was taken directly from the game, where all of the player's data is inherently

independent of one another. However, there is a small chance that the independence assumption could have been violated if the FIFA ratings were created in comparison to other players of the same league, rather than across all the world's leagues.

Model Fit

Once the final model has been created, it is important to determine how well it performs by assessing the model fit. The first measure of model fit is root mean squared error (RMSE) which is 2.004. This is the average magnitude of an error in the model and in context, it appears to be a strong value. Additionally, the R^2 value of 0.8492 means that 84.92% of the variation in potentially can be explained by the predictors used in the final model. This is another strong value in context indicating that this model does a good job of predicting potential based on the data.

Multicollinearity and Overfitting

It is important to check if the model shows evidence of multicollinearity to ensure the proper contextualization of individual predictor results. From the VIF calculations in Figure 9, the overall and log(wage) predictors have a VIF value over 5, indicating there is evidence of multicollinearity. Moving forward, it is important to keep this in mind as the individual results for these variables may be counterintuitive and unable to be properly contextualized.

Overfitting is another important consideration of the model as the model must predict the true relationship between potential and the predictors, and not be too specific to this dataset. However, this model does not overfit the data as the model does not include any high-order polynomial fits, has a sample size greater than 10 times the number of predictors, and there are at least 10 observations for each league.

Predictor Results

Based on the summary output in Figure 9, all the predictor terms appear to be statistically significant as referenced by their p-values below 0.01. However, only a select few terms are practically significant because the evidence of multicollinearity may make the coefficients for terms including overall or log(wage) to be counterintuitive. For example, log(wage) should not have a negative relationship with potential as players with high wages would be expected to have greater potential. One example of a term that is practically significant is the log(age) term. Here, a 10% increase in a player's age would result in a 2.19 decrease in potential overall. This result makes sense as typically, the FIFA players with the highest potential are younger players.

Conclusion

This project created and enhanced a model that uses various FIFA and player specific statistics such as overall, age, transfer value, wage, and league to predict their potential overall in the FIFA 2022 video game. A simple linear model was not able to be used for analysis because it did not pass the assumption of constant variance, so log transformations were added to certain predictors. To further improve the model, an interaction between overall and league was added as well as the swap of wage for transfer value. These actions ultimately led to the completion of a strong finalized model that accurately predicts player potential.

Moving forward, this model could be further improved by analyzing different variables not included in the initial dataset such as years in the FIFA franchise or last season's overall to see if they would positively impact the effectiveness of the model.