Fifa Player Attributes: Analysis

Morgan Petersen, Garrett Voigt

STATS 401: Applied Statistical Methods II

December 12, 2023

Background

The FIFA data set used for this analysis gives data on all different types of attributes from professional soccer players from the FIFA 22 game. To limit this incredibly large data set, this report will only analyze players from League level 1.

This model has the explanatory variables are passing, shooting, dribbling, preferred foot, and player potential. The response variable is salary. . Passing, shooting, and dribbling are quantitative variables that are measured on a scale of 1 to 100. These ratings are determined by FIFA to score a player's abilities in each of these areas. Prefered foot is a categorical variable where right-footedness is numerically assigned as 1 and left-footedness is assigned 0 by R studio. The potential variable is also included for an interaction term, which is a way of adjusting the slope of a model based on a categorical variable. Similar to the other predictors, this is rated on a scale out of 100 by FIFA, but is related to how the player is developing and the future that they could have in their career. All of these variables are created using historical data from games that each player participates in. While these variables may be assigned somewhat subjective, they are all assigned by the same organization, so it is fair to assume there is no bias between ratings, and they are assigned independent of one another.

Analysis

Original Model

The original model for this data set was :

$$Wage = \widehat{\beta_0} + \widehat{\beta_1} x_{passing} + \widehat{\beta_2} x_{shooting} + \widehat{\beta_3} x_{defending} + \widehat{\beta_4} I_{Rightfoot}.$$

The idea behind this is that passing, shooting, defending, and footedness are all fundamental abilities of a player that will impact their ability to garner a higher wage.. In addition, the categorical variable is included as another method of predicting. However, through further

analysis, it was found that this model only predicted about 20% of the variance in the wage variable. Obviously, this is not adequate when attempting to create an accurate model to predict players wages, and changes were necessary.

Model Alteration

The first problem with the initial model is that it did not account for the fact that the passing term warranted a quadratic transformation. After further analysis of the data, there was evidence that the passing variable was not linearly associated with the response, and was instead quadratically related. Making this change increased the accuracy of the model, explaining roughly 10% more of the variance in wage. This is seen in the first iteration summary table.(Table 2) This initial change leads to the following CMF::

$$Wage = \widehat{\beta_0} + \widehat{\beta_1} x^2_{Passing} + \widehat{\beta_2} x_{Shooting} + \widehat{\beta_3} x_{Defending} + \widehat{\beta_4} I_{Rightfoot}.$$

Furthermore, there appears to be no significant relationship between the defending variable and wage, as seen in the original scatterplot matrix.(Figure 5) As a result, this is not a good predictor to include as it adds complexity to the model without adding any significant accuracy. To adjust for the insignificance of the defending variable, it was removed from the model.. Another fundamental aspect of many players is their dribbling ability, so it was selected to replace the defending term. This adjusts the conditional mean function to:

$$Wage = \widehat{\beta_0} + \widehat{\beta_1} x_{Passing} + \widehat{\beta_2} x_{Shooting} + \widehat{\beta_3} x_{Dribbling} + \widehat{\beta_4} I_{Rightfoot}.$$

. While these previous changes did improve the model, it still did not explain a large amount of the variance in wage. Therefore, there was evidence that an interaction term would be beneficial. This interaction would be between the potential variable, and footedness. This would increase the accuracy of the model, and further improve the R squared value.

With these changes implemented, the conditional mean function is defined to be:

$$Wage = \widehat{\beta_0} + \widehat{\beta_1} x^2_{Passing} + \widehat{\beta_2} x_{Shooting} + \widehat{\beta_3} x_{Dribbling} + \widehat{\beta_4} I_{Rightfoot} + \widehat{\beta_5} x_{Potential} I_{Rightfoot}.$$

Although this is more accurate and explains the most variance of any of the previous models, it still does not explain a majority of the variance in the response variable.To investigate possible reasons for this phenomenon, the histograms and QQ plots were analyzed. The histogram showed clear evidence that the wage variable as a whole is strongly right skewed, suggesting there is potential for a nonlinear trend associated with wage.(Figure 1) Furthermore, the QQ plot shows clear evidence of a non-linear trend with the large deviation above the line on the right side of the plot.(Figure 8) This provides evidence that a log transformation is more accurately suited for this model instead of a linear transformation.

Final Model

With this final transformation in place, the CMF for the final model is as follows:

$$Log(Wage) = \widehat{\beta_0} + \widehat{\beta_1} x^2_{Passing} + \widehat{\beta_2} x_{Shooting} + \widehat{\beta_3} x_{Dribbling} + \widehat{\beta_4} I_{Rightfoot} + \widehat{\beta_5} x_{Potential} I_{Rightfoot}.$$

Following this change, the QQ plot became much more linear, which suggests that the function now fulfills the linear assumption for analysis. After the stated changes, the model explained the most variance of any previous model. This can be seen in the summary table, where it has an R-squared value of 0.4856. The adjusted R squared also has a value of 0.4853. Considering this value has an upper limit of the regular R squared, an adjusted R-squared that is so close to the original R-squared shows that the model is not overly complicated and all of the information used to estimate the response is valuable to keep in the model.

Conclusion

The final model does the best job of predicting the response variable, but it still cannot account for a majority of the variance in weekly wages. This is likely because there are many

more variables that go into a player's weekly wage. Given the context of wage, it can be impacted by many outside factors that are difficult to quantify such as the political ideas at the time, the popularity of the player, or things like wage caps and team budgets. While none of these variables are available for analysis, they could all certainly have a hand in explaining the wage of a given player. That is not to say that this model isn't somewhat helpful. At the end of the day, roughly half of the variance in a pliers wage is explained by just 6 variables. All things considered, this model in its current state is not a reliable predictor of weekly wage for a professional soccer player because the final model only accounts for about 48% of the variance There are many more factors at play and a more complex model would be needed to better explain this response.