Identifying what factors affect the wages of FIFA game players

Weiliang Hu

November 27, 2021

Summary

In the project, I use the proportional odds and multinomial logistic regression to model the odds of a FIFA game player earning a certain degree of salary based on a range of characteristics. The goal of this project is to identify important characteristics that are associated with wages of professional FIFA players and quantify the relationships. From the results, I find some variables in the dataset to possess an impact on earning a high-level wage. Specifically, variables such as the player's BMI, the player's overall rating, the player being in a strong club or not, have noticeable effects. However, there are a number of caveats which concern with the validity of the inferences.

Introduction

Soccer has been a game that is beloved around the world. According to the Big Count survey conducted by Fédération Internationale de Football Association (FIFA) in 2021, there are more than 265 million players actively involved in soccer around the world. When the last FIFA World Cup was held in 2018, FIFA grossed an annual revenue over 4.64 billion dollars. As a fan of soccer games, in this project I am interested in providing a reliable way of identifying the wage of a FIFA player. Particularly, I want to explore the following questions:

- Does the position of the player in the team affect his weekly wage?
- Every player has his preferred foot when playing soccer, does it affect the player's weekly wage?
- How does the club the player joins affect the player's weekly wage?
- Are there other interesting associations with weekly wage that are worth mentioning?

Data

1.Data Preparation: The dataset is obtained from Kaggle.com. It is provided by FIFA official to EA Sports in order to make the public game FIFA 2020. It contains 18278 variables on 74 attributes. Upon initial inspection, even though the dataset contains no missing value, it has several issues that need to be addressed in order to proceed to further analysis.

The first issue with the dataset is that there are too many variables. However, we can drop many of them because they are not scientifically meaningful when fitting our model: For example, we definitely do not need variables such as the player's website, first and last name, or the player's personal tags to fit our response variable; The columns after attacking_crossing are used to control the character representing the player in the video game so we do not need these columns either; Some columns essentially represent the same information, such as age and date of birth, or the player position, team position and national position, so we only keep one of them.

The second issue is that our numeric response variable, the weekly wage in Euro, does not following a normal distribution, which is not surprising because super stars such as Messi and CR can earn a lot while a very ordinary player just earn a base salary. The difference in wage varies a lot between every single player. This is an indicator that linear regression would not work well in here and I should start thinking about multinomial logistic/proportional odds model. As a result, I transform the response variable into a categorical variable *wage_brackets* with six levels:

- Base salary: For players earning between 1000 and 2000 Euro weekly
- Junior salary: For players earning between 2000 and 4000 Euro weekly
- Advanced salary: For players earning between 4000 and 10000 Euro weekly
- Senior salary: For players earning between 10000 to 20000 Euro weekly
- Top salary: For players earning between 20000 to 50000 Euro weekly
- Star salary: For players earning more than 50000 Euro weekly

The third issue is that some variables have high cardinality, which means that these variables have many unique levels. *Club*, for example, has 185 categories. This is problematic for three reasons. First, too many levels tend to result in fewer observations in each level. In our case, many clubs have less than 20 observations. Too few observations may affect the validity of model inference. Second, it would become very difficult to interpret any outcome related to this variable. Third, detailed breakdown of each variable could lead to higher probability of a category only exclusively contain a certain degree of salary, resulting in overfitting. For instance, all the players from Odense BK earn a junior salary (This may not be a big problem in this particular analysis but it is worth pointing out).

The fourth issue is that some predictors tend to be correlated with one another. For example, the overall rating of a player is calculated based on the pace, shooting, passing, dribbling, defending, and physic ratings of the player, indicating that they are highly correlated. Since we are considering a multinomial logistic regression model, we want to avoid high correlation that may lead to inaccurate standard errors and unreliable model inferences.

To deal with the two issues mentioned above, I implement the following solutions: First, for variables with high cardinality, I collapse and combine certain categories based on exploratory data analysis and external research. Take the *player_positions* variable for example. It indicates the player's position in the field, and can take one, two, or three of the total 11 positions in soccer. This variable is collapsed based on scientific knowledge: First only the position shown at the beginning is the player's major position, so we only keep that in our future analysis. Next we make the new variable *player_positions_new* contain only 4 position categories instead of the detailed 11 position: *Defender(DF), Forward(FW), Midfielder(MF) and Goalkeeper(GK)*. The variable then indicates one of the four position categories a player responsible for. We apply similar idea to the *club* variable to create a new variable *player_in_strong_club*, with levels 1 and 0. The 1 level contains the top 20 FIFA clubs in the year of 2020, while the 0 level contains the rest of the clubs. The re-leveled variable now indicates if a players plays in a top club or not. After re-leveling, highly correlated variables are also dropped based on exploratory data analysis and scientific importance. For example, *pace, shooting, passing, dribbling, defending, and physic* are dropped in favor of *overall*, since FIFA already calculated this variable and it is highly correlated with the preceding variables.

Finally, we want a variable to indicate the body condition of the players, so we calculate BMI using the player's height and weight. In addition, we mean center the continuous predictors to reduce multicollinearity (The names of mean centered variables are followed by a "c"). The reduced dataset now has 14 variables.

2. Exploratory Data Analysis: We do not have even distribution of players across the levels in our response variable *wage_brackets*. We have 7943 layers earning base salary, 3128 players earning junior salary, 3130 players earning advanced salary, 1723 players earning senior salary, 1540 players earning top salary, and 574 players earning star salary. This characteristic of data passes on when we examine the factor variables. For example, there are 2022 defense, 1309 forward, 1738 goalkeeper, and 2874 midfielders earning base salary, and only 390 defense, 343 forward, 209 Goalkeeper, and 598 midfielder earning top salary. Looking at the conditional probability between response and factor variables *preferred foot, player_in_strong_club, player_position_new and international_reputation*, we can see difference across each salary level and between each group. The corresponding chi-square test of independence are all significant, also suggesting that there are connections between wage and these variables. However, whether these variables are statistically significant needs to be further examined.

When looking at continuous variables, there are also interesting information. For example, we can see that people over 25 are more likely to earn advanced or higher level salary, as shown in figure 1 below. In addition, though not so apparent, players who earn star level salary do seem to have slightly higher BMI than rest of the groups, and people with higher-level overall rating earn higher level salary. When exploring interactions, by looking at the trends across different variables, we are not able to observe many potentially significant interactions. For instance, by looking at figure 2, one may argue there is interaction between *age* and *player_in_strong_club*, however the interaction may not be significant. After all, we should not make conclusions about the variables and interactions until further exploration and assessments.

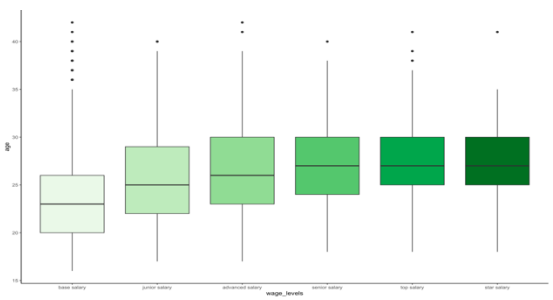

Fig 1. wage_brackets vs. age Plot

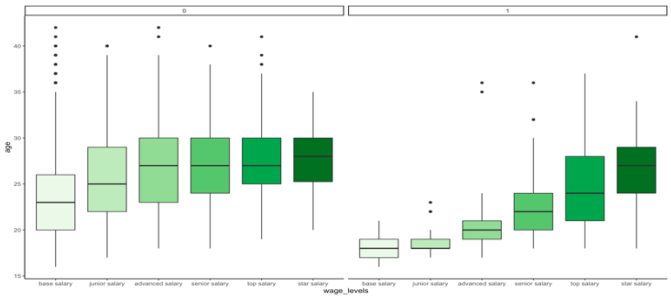

Fig 2. wage_brackets vs. age *by* player_in_strong_club Plot

Model

Our final model is a proportional odds model:

$$\log\left(\frac{\Pr(y_i \le j|x_i)}{\Pr(y_i > j|x_i)}\right) = \beta_{0j} + \beta_1 x_{i\ bmic} + \beta_2 x_{i\ Age} + \beta_3 x_{i\ Agec}{}^2 + \beta_4 x_{i\ Agec}{}^3 + \beta_5 x_{iAgec^4} + \beta_6 x_{iAgec^5} + \beta_7 x_{ioverallc} + \beta_8 x_{ioverallc^2} + \beta_9 x_{ioverallc^3} + \beta_{10} x_{ioverallc^4} + \beta_{11} x_{i\ player\_in\_strong\_club} + \beta_{12} x_{iplayer\_position\_new} + \beta_{13} x_{i\ international\_reputation}, \quad j = base, junior, advanced, senior, top$$

To begin with, we start by fitting a multinomial logistic regression model, using all the main effects (*agec, bmic, overallc, player_in_strong_club, preferred_foot, player_position_new, international_reputation*). After generating p-values for the covariates, we can see that *preferred_foot* does not seem to be a significant predictor. To validate this idea, we perform a chi-square test on *preferred_foot*, resulting in a p-value of 0.13, indicating that it indeed isn't significant. The results of the chi-square test on other covariates have p-values close to zero, so we keep them and move on to model assessment.

When examining the binned residual plots, we can see that both *raw residual vs. agec* plots and *raw residual vs. overallc* plots have multiple salary levels containing points not randomly distributed (forming a trend) and scattered outside the lines (meaning that they lie outside the 95% confidence interval). Since different levels contains different trend, we apply polynomial transformations to the two predictors. After multiple attempts, it seems that a 3rd order polynomial transformation on the predictor *agec* and a 5th order polynomial transformation on the predictor *overallc* smooth the issues. An example of the effect of transformation is shown below, now no assumptions are clearly violated:

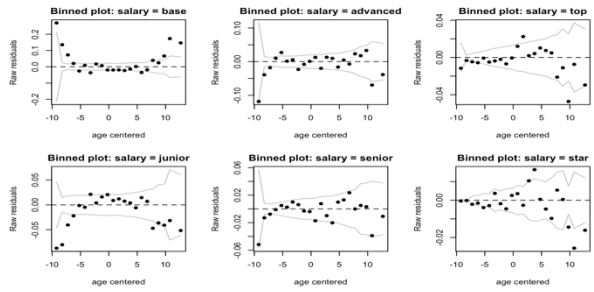

Fig 3. Raw residual vs. agec before transformation

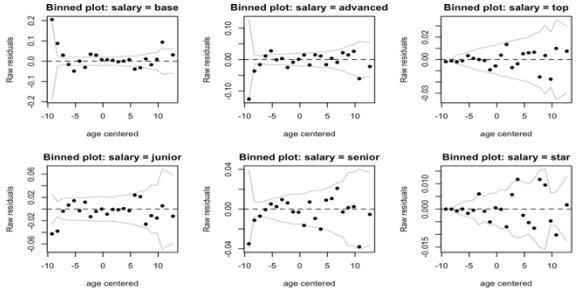

Fig 4. Raw residual vs. agec after transformation

It is possible that there are some interaction terms we should consider in our model. However, since the chi-square tests on all the possible interactions do not generate p-values small enough to indicate statistical significance, we would not include interaction terms in our model. When we finally have a model for multinomial logistic regression, we notice two facts that drive us to consider the proportional odds model:

- The output of multinomial logistic regression model is very hard to interpret given the predictors now possessing so many levels.

- The categories of our response variable are essentially ordinal, making them a better fit for proportional odds model compared to multinomial logistic regression model.

Since the proportional odds model is essentially a more parsimonious version of the multinomial logistic regression model, everything about the model we have so far can carry over to the proportional odds model when we fit it in R (using ordered response variable). After slight coordination on the order of the polynomial transformed terms during model assessment (in order to make most observations scattered randomly and within the 95% confidence interval), we have our final model shown at the beginning of this section.

Our final model has low VIF score (ranging between 1 and 3), meaning that we avoid potential multicollinearity issue. In addition, our model performs arguably well in terms of prediction. Under the best threshold, some salary levels have good prediction results such as base salary (0.88 sensitivity, 0.76 specificity) and star salary (0.78 sensitivity, 0.99 specificity), while levels such as junior and senior salary have very low sensitivity scores (0.15 and 0.30). The overall accuracy of the model is 0.62 and the AUC scores are shown below. Generally, the AUC scores are quite high.

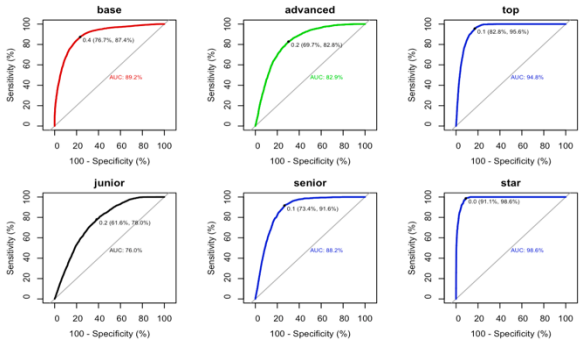

Fig 5. The ROC plots for the final proportional odds model

Results

	Value	Std. Error	t value
poly(agec, 5)1	4.67	2.92	1.60
poly(agec, 5)2	-45.55	2.85	-15.98
poly(agec, 5)3	10.24	2.76	3.70
poly(agec, 5)4	-17.71	2.65	-6.69
poly(agec, 5)5	26.37	2.51	10.49
bmic	-0.03	0.01	-2.28
poly(overallc, 4)1	358.67	4.48	80.14
poly(overallc, 4)2	68.02	4.09	16.62
poly(overallc, 4)3	-67.82	3.62	-18.72
poly(overallc, 4)4	-23.10	3.32	-6.95
player_in_strong_club1	2.85	0.11	26.38
player_position_newFW	0.35	0.05	7.48
player_position_newGK	-0.27	0.05	-5.43
player_position_newMF	0.07	0.04	1.71
international_reputation2	0.58	0.07	7.76
international_reputation3	0.60	0.18	3.41
international_reputation4	0.44	0.49	0.90
international_reputation5	4.12	1.64	2.51
base salary|junior salary	-0.82	0.04	-22.12
junior salary|advanced salary	0.68	0.04	17.95
advanced salary|senior salary	2.79	0.04	62.89
senior salary|top salary	4.60	0.06	83.10
top salary|star salary	7.94	0.10	79.78

Table 1. Proportional Odds Model Result

	2.5 %	97.5 %
poly(agec, 5)1	-1.02	10.43
poly(agec, 5)2	-51.20	-40.01
poly(agec, 5)3	5.36	15.73
poly(agec, 5)4	-22.19	-13.28
poly(agec, 5)5	21.88	30.75
bmic	-0.05	0.00
poly(overallc, 4)1	349.91	367.45
poly(overallc, 4)2	59.98	76.03
poly(overallc, 4)3	-74.91	-60.70
poly(overallc, 4)4	-29.58	-16.54
player_in_strong_club1	2.64	3.05
player_position_newFW	0.26	0.45
player_position_newGK	-0.37	-0.17
player_position_newMF	-0.01	0.15
international_reputation2	0.44	0.73
international_reputation3	0.26	0.95
international_reputation4	-0.38	1.46
international_reputation5	1.04	7.32

Table 2. Proportional Odds Model Confidence Interval

First, by looking at Table 2, we can see that most predictors are significant at 5% significance level. For example, the *overallc, bmic* are significant given that the confidence interval does not contain zero (upper bond of *bmi* is rounded to 0 but it's actually negative). *Player_in_strong_club* is significant given 1 compared to 0. For the *player_position_new*, all the positions are significant compared to defender except for midfielder, where the confidence interval contains 0. The similar idea applies to *agec* and *international_reputation*, most levels are significant except for one. This may affect how we interpret the result.

By looking at Table 1, a lot of interesting results can be explored. For conciseness we only discuss the most important ones. Note that the results are in log-odds scale so we have to exponentiate them for interpretation. For player in a strong club, the probability of him earning salary level in a star direction compared to him earning salary level in a base direction is 17.3 times more likely than a player not in a strong club, the probability of him earning salary in a star direction compared to him earning salary in a base direction. Now let's look at the player's position: for a forward, the probability of him earning a salary level in a star direction compared to him earning salary level in a base direction is 1.42 times more likely than the probability of a defender earning a salary level in a star direction compared to him earning salary level in a base direction. For a goalkeeper, the probability of him earning a salary level in a star direction compared to him earning salary level in a base direction is 0.76 times less likely than the probability of a defender earning a salary level in a star direction compared to him earning salary level in a base direction. For midfielder, there is no statistically significant difference between it's the salary level and the defender's salary level.

Looking at the intercepts through the last 5 rows of Table1, it would not be surprising to see that the odds of earning high levels of salaries are many times less compared to earning low levels of salaries. For instance, the odds of earning a base/junior/advanced salary are 16.44 times more compared to the odds of earning a senior/top/star level salary. The result justifies our common sense.

One interesting finding is the BMI of FIFA players, from the result it seems that the odds of earning salaries in a star direction compared to earning salaries in a base direction decrease by about 3% (after exponentiation), for each unit increase in the BMI score of the players.

Conclusion

In this project, I used multinomial logistic regression model and proportional odds model to identify the factors that affect a FIFA player's wage. From the regression results, it is find that the age of the player, the BMI score of the player, the overall rating of the player, the position of the player, the international reputation of the player, and whether the player plays in the top 20 clubs of FIFA, are significant predictors at 5% significance level, meaning that they all play a role in determining the wage of a FIFA player's wage. On the other hand, whether the player's preferred foot is left or right does not influence the amount of money he receives from his career.

There are a few limitations of this project. First, even though this is the largest public dataset of FIFA players I was able to find on the internet, it is still rather subjective and contains incorrect information. Algorithm for calculating the overall rating and classifying the international reputation was designed by EA sports which may not reflect the real status of some players. In the original dataset, there are 240 players with 0 wages, which does not make sense (and we dropped these observations in our analysis); 7902 players earn weekly salary of 1000 Euro, which is also far from being scientifically reasonable. Unlike NBA, the salary of players from many FIFA clubs are confidential and EA do not have access to the information. These problems suggest that the inferences using our model could be dubious. In addition, the overall prediction accuracy of our model is not very good. If we want to do prediction tasks, we should consider naïve-bayse model or some neural networks.

Source codes/plots/tables: https://github.com/wh153/FIFA_Player_Wage_Analysis

Data source: https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset?select=players_20.csv