**Title:** Professional Football Player Market Value and Skill Level.

**Introduction**

The purpose of my research is to find how does players physical traits and soccer skill effect the market value of the professional football player in the forward position. The importance of the research is that this research could help the teams see what kind of relationship players ability has with the value of the player without accounting for the market value based on the popularity. My interest in this area is from general interest in the question of how the values of the professional football players are evaluated.

From my background research on this topic, I learned that sprinting and power are most important ability of a striker so it will have a relationship with the market value. I should also note that the strikers position players are less effected by the big events compared to defenders so using strikers will be more accurate representation of players’ market value. Additionally, since there are super star players who bring revenue to the club, their market value may be significantly higher compared to the player ability.

**Method**

I first randomly split the dataset to make training and validation set. I would first need to create a model from a training dataset. To make an inference from the model, I must satisfy 4 assumptions: linearity/Mean zero errors, uncorrelated, common error variance, normality of errors.

I performed an EDA on the training set. From each set I needed to delete rows with NA values and impossible values if I found it from the summary. I first look at the univariate graph of each predictor and scatterplots plot of predictors to response. From these graphs, I would see if there were distribution that could cause normality or linearity to be violated. To fix the violation, I must look at residual plot of the model. There were 3 main residuals plot I used: residual versus predictor plots, residual versus fitted values plots, and normal QQ plot.

For residual plots to be valid, it needs to pass 2 conditions: conditional mean response is a single function of a linear combination of the predictors, and conditional mean of each predictor is a linear function with another predictor. I checked 1st condition by looking at the plot response against the fitted values. If the points are randomly scattered around the identity function, the 1st condition is satisfied. The 2nd condition is satisfied looking at a scatterplot of all the pairs of variables and determining if there is non-linear relationship. I could fix the conditions by transforming my variable or response.

Once the 2 conditions are satisfied, I look at residual plots to see if 4 assumptions are satisfied. I checked for linearity by looking if there is systematic pattern in the residual, uncorrelated errors by looking for large clusters of residuals, and constant variance by looking some form of pattern line fanning. I could check for normality by looking if the QQplot points follow the line. If the none of the assumption passed, then I transform the variable with problematic residual plots or response. After transformation, I check for all of conditions and assumption. I repeat until 4 assumptions are satisfied.

I check for F-test to see is there is a significant linear relationship exists overall. I then to t-test on each individual coefficient to test whether each predictor has a linear relationship in the presence of the other predictors. If predictor or predictors that does not pass, we perform a partial-f test and background research to see if removing the predictor is the right decision.

After two test passes, I create multiple models by removing some predictors in the model. I check VIF for multicollinearity and consider value above 5 as extreme collinearity. I also check for all the influential points. I also use adjusted R squared, AIC, and BIC values to choose the best model. After I find the best model, I apply the same transformation and model to the validation dataset after cleaning the validation dataset. To compare the training model and validation model, I look for the differences in the estimated coefficients, if the same predictors appear significant, no new model violations, and similar adjusted R squared.

**Results**

**Diagram, schematic, box and whisker chart

Description automatically generated**

Graph 1. Univariate Graphs of each variable from training dataset.

Based on Graph 1, there were skewness in some of the graphs which might violate the assumption of normality and linearity.

Before I looked at residual plot and normal QQplot to check for 4 assumptions, my model must satisfy 2 conditions.

Chart, scatter chart

Description automatically generated

Graph 2. The fitted market value versus market value plot before transformation

From Graph 2, I could see in fitted market value versus market value plot there is a curve pattern. The first condition is not satisfied so I must transform some variables. After going through multiple transformation options, I decided that the best option is just to log the response. With transformed dataset, I created a new model.

Chart, scatter chart

Description automatically generated

Based on Graph 3, the fitted market value versus market value plot showed datapoints are randomly scatter around the line and all the pairwise graphs from Graph A(Appendix) were not significantly nonlinear. The 2 conditions are now satisfied.

Chart, diagram

Description automatically generated

Graph 3. Some of the residual plots of model of each predictor and normal QQplot after transformation

From the Graph 3 and Graph B(Appendix), the residual plots are spread out and not have pattern or clusters. Also, the normal QQplot seem to follow the line. Based on these graphs, I concluded 4 assumptions are satisfied.

The model has F test p-value less than 2.2e-16 so there exist a linear relationship overall. For every individual T test, all the predictors are significant meaning all predictors have a linear relationship in the presence of other predictors.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | VIF | | | | | | | | |
|  | Finishing | Heading Accuracy | Sprint Speed | Ball Control | Dribbling | Shot Power | Acceleration | Age | Height |
| Mod2 | 2.465449 | 2.743497 | 4.554269 | 4.126995 | 4.610161 | 1.747640 | 5.949531 | 1.413929 | 1.952983 |
| Mod4 |  | 2.093624 | 4.539508 | 3.925059 | 4.604202 | 1.584074 | 5.932725 | 1.413071 | 1.952173 |
| Mod5 | 1.881439 |  | 4.534761 | 4.125271 | 4.470064 | 1.727881 | 5.837003 | 1.364633 | 1.794015 |
| Mod6 | 2.457458 | 2.731746 |  | 4.063634 | 4.570155 | 1.726986 | 2.163266 | 1.402249 | 1.924716 |
| Mod7 | 2.344813 | 2.742351 | 4.484348 |  | 1.885214 | 1.731586 | 5.946730 | 1.382837 | 1.949736 |
| Mod8 | 2.462262 | 2.660126 | 4.514748 | 1.687635 |  | 1.736866 | 5.820597 | 1.404819 | 1.946205 |
| Mod9 | 2.234701 | 2.712479 | 4.500446 | 4.089084 | 4.581740 |  | 5.903705 | 1.383051 | 1.937970 |
| Mod10 | 2.458485 | 2.691608 | 1.655945 | 4.125052 | 4.510253 | 1.734179 |  | 1.408911 | 1.767451 |
| Mod11 | 2.463951 | 2.647845 | 4.516647 | 4.036241 | 4.580456 | 1.709474 | 5.928414 |  | 1.829574 |
| Mod12 | 2.464426 | 2.520183 | 4.488352 | 4.120133 | 4.594162 | 1.734205 | 5.384329 | 1.324583 |  |
| Mod13 | 2.339048 | 2.690771 | 1.565082 |  | 1.806628 | 1.717473 |  | 1.377259 | 1.765180 |
| Mod14 | 2.453655 | 2.585446 | 1.449492 | 1.652332 |  | 1.726493 |  | 1.397452 | 1.745741 |
| Mod15 | 2.453654 | 2.212834 | 1.337861 | 1.590529 |  | 1.703800 |  | 1.318423 |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | | |
|  | Adjusted R Squared | AIC | BIC |
| Mod2 | 0.8461811 | 850.3546 | 903.3025 |
| Mod4 | 0.8018807 | 1079.6800 | 1127.8145 |
| Mod5 | 0.8385253 | 893.5664 | 941.7009 |
| Mod6 | 0.8448649 | 857.1186 | 905.2530 |
| Mod7 | 0.7978680 | 1097.9273 | 1146.0617 |
| Mod8 | 0.8393973 | 888.6386 | 936.7730 |
| Mod9 | 0.8229459 | 977.3835 | 1025.5180 |
| Mod10 | 0.8451761 | 855.2912 | 903.4256 |
| Mod11 | 0.8318268 | 930.5538 | 978.6882 |
| Mod12 | 0.8452136 | 855.0708 | 903.2053 |
| Mod13 | 0.7972217 | 1099.841 | 1143.162 |
| Mod14 | 0.8373615 | 899.1106 | 942.4316 |
| Mod15 | 0.8373309 | 898.2905 | 936.7981 |

Table 1. The VIF, AIC, BIC, and adjusted R squared of some of the possible the model

After checking the VIF of the original model, I noticed significant multicollinearity on acceleration and sprint speed. I noticed that sprint speed and acceleration is correlated since its both related to speed of the player. Also, the predictors ball control and dribbling could have correlation since its both related to the players ability with ball movement. After creating model with one missing on each group, the VIF values are lowest when dribbling and acceleration is removed. Partial F test with reduced model was significant however from my knowledge there were other predictors that represented the Dribbling and Acceleration. So, I decided on removing them to lower VIF values. The adjusted R squared, AIC, and BIC values between original model and reduced model are not significantly different enough to consider different reduced model or to not use reduce model. The T test of reduced model reveals that predictor height was not significant, so I removed the Height predictor.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Intercept | Finishing | Heading Accuracy | Sprint Speed | Ball Control | Shot Power | Age |
| Train | -10.240609 | 0.051755 | 0.009293 | 0.017671 | 0.096119 | 0.031234 | -0.038331 |
| Val | -9.78790 | 0.04940 | 0.01014 | 0.01773 | 0.09259 | 0.03114 | -0.04269 |

Table 2. These are the coefficient of final model of train and validation data.

Based on the Table 2, I could see that there was no big difference in each coefficient between train and validation model. The same predictors in both models appeared as significant. The adjusted R squared of a model in validation dataset was 0.824163 which is close to the value of model with train dataset. Since there are no substantial changes, I could say my model has been validated.

**Discussion**

My final model is log (Value) = -10.240609 + 0.051755(Finishing) + 0.009293(Heading Accuracy) + 0.017671(Sprint Speed) + 0.096119(Ball Control) + 0.031234(Shot Power) – 0.038311(Age). The interpretation is if the Finishing rating increase by 1 unit, we expect log (Value) to increase by 0.051755 in the presence of the other predictors. It answers the question of football player physical and skill traits effecting the market value of the players. This answer is important since this shows there is some relationship to players market valuation based on their skill level.

The limitation is that there were a lot of missing players market value and 0 as the market value. This dataset does not include players with market value less than 1 million euro. Additionally, there are superstar players who have higher rating and high market value. The market value of these kind of players are bloated compared to their skill level since they bring large revenue to the club. The train model also included multiple influential points which could have influence the values of the coefficient. Based on background result, players height is one of the affect factors of market value. However, my model T test found height as non-significant, so I removed it. This may be caused by the difference in the dataset.

**Appendix**

**Graphical user interface, application

Description automatically generated with medium confidence**

Graph A. pairwise graphs of predictors of transformed model.

Chart, diagram, schematic, scatter chart

Description automatically generatedChart, diagram, schematic, scatter chart

Description automatically generated

Graph B. Rest of residual plots of predictors not in Graph 3.

**Reference**

Faude, O., Koch, T., & Meyer, T. (2012). Straight sprinting is the most frequent action in goal

situations in professional football. *Journal of Sports Sciences*, *30*(7), 625–631. <https://doi.org/10.1080/02640414.2012.665940>

Kiefer, S. (2014). The impact of the euro 2012 on popularity and market value of football

players. *International Journal of Sport Finance*, *9*(2), 95–110.

Singh, P., & Lamba, P. S. (2019). Influence of crowdsourcing, popularity and previous

year statistics in market value estimation of football players. *Journal of Discrete Mathematical Sciences and Cryptography*, *22*(2), 113–126. <https://doi.org/10.1080/09720529.2019.1576333>

Serna Rodríguez, M. (2021). Factor analysis of the market value of high-performance

players for three major European association football leagues. *Managing Sport and Leisure*, *26*(6), 484–507. <https://doi.org/10.1080/23750472.2020.1771197>

Kologlu, Y., Birinci, H., Kanalmaz, S. I., & Ozyilmaz, B. (2018). *A Multiple Linear Regression*

*Approach For Estimating the Market Value of Football Players in Forward Position*.