**A Comparative Study of Multiple Linear Regression and Generalized Additive Model in Predicting Market Value of Players in FIFA 22**

Christopher Salim

Department of Mathematics and Statistics, Faculty of Mathematics and Science, Brock University

E-mail: christopher.salim@brocku.ca

**Abstract**

Transfer activity in soccer can be an exorbitant investment, therefore it is beneficial for managers and club owners to determine the factors that affect the market value before investing money into players. A comparative study between multiple linear regression (MLR) and generalized additive model (GAM) for FIFA 22 dataset has been done by using R programming language. The purpose of this paper is to determine whether GAM can capture the nonlinearity of the variables surrounding market value better by using 11 other features related to the quantitative abilities of the players. After training and fitting the models, the AIC of MLR and GAM were 31292.04 and 19539.74, respectively. The BIC of the GAM was also smaller, 20172.11 compared to 31493.37 obtained from MLR. In addition, choosing GAM over MLR also lowered the MSE from 0.366 to 0.183. GAM also successfully explained several factors that could not be explained by the MLR. One of the key facts learned from GAM was improving attacking, defending, and skill scores to above 60 would increase the market value significantly compared to below 60. Overall, GAM was superior to MLR in terms of performance and this model could be meaningful for FIFA 22 players who wanted to plan their transfer activities better before spending money in the game.

Keywords: FIFA 22, generalized additive model, market value, nonparametric regression, soccer

## Introduction

* 1. **Transfer Market in Soccer**

In soccer, market value means the price that an engaging club pays for the player in the releasing club in order for them to sever their remaining contract and play with the engaging club, after having mutual agreement from both parties (Poli et al., 2021). Even after the financial disruption in the wake of the COVID-19 pandemic, player transfer is still a multibillion-dollar investment. According to the Global Transfer Report released by (FIFA, 2021), in 2021, about USD 4.86 billions were spent on transfer fees worldwide.

Due to its expensive nature, managers and club owners have to be mindful when analyzing why certain players worth more than others before making informed decisions. Research about the player’s market value has gained traction in the academia, because there can be many factors influencing it. Some of the research papers used econometric approach, such as inflations or deflations, economic level of the recruiting club, and other technical variables to explain the market value of each player (Poli et al., 2021), (Ezzeddine, 2020). A study from (Metelski, 2021) emphasized the importance of player’s age to their market value. To complement, (Valentini, 2020) also added that the player’s brand, or reputation, is also as important in determining the market value. One research also mentioned that number of Google searches, number of years on contract left, how a player is looked up to by the community, number of goals and assists, and player’s ethnicity impact the market value for a player. Furthermore, with the correctly selected independent variables, the transfer fee for the most expensive soccer player ever is not random and can be explained (Barbuscak, 2018).

While those findings are valid and helpful for general cases, it can be complicated to bring those concepts to people that only want to simulate real-life soccer using video games. In video games, it can be hard to quantify some parameters, such as the human interaction factors or the intricacies of financial situation of each individual club. Instead, the variables existing in the game can potentially work just as well to explain the market value, but do not exist in the previous studies, such as the skill of the players in each aspect of the soccer e.g., attacking, defending, and mentality.

* 1. **Data Description and Objective**

FIFA video game is a series of soccer video games created and published once per year by Electronic Arts, with the latest version being FIFA 22. The reason FIFA was chosen is FIFA is one of the best-selling sports video game franchises in the world, proven by the fact that it performed well in 17 out of 19 European countries (Gibson, 2022). One of the key factors that contributed to their success is because they consistently update their player rosters and clubs’ database so that the game reflects correctly what happens in the real-life soccer. The gameplay revolves around a player controlling soccer athletes in a team to play against another team and try to win by outscoring the opponent. Career Mode is also available in FIFA, where the player can act as a manager of a club. The player is then responsible for all activities in the club, one of them is managing player transfers during the transfer window. This is where each player’s market value comes into consideration. With such a large community in the game, it is in the best interest of a lot of players to make optimal plays, including studying what really affects their players’ market value and predicting the market value of each player.

To analyze the factors influencing the FIFA 22 soccer player’ market value and make a predictive model, the data obtained from sofifa.com is used (at March 23rd 2022). Sofifa is an online database storing daily-updated player information in all FIFA video games, including FIFA 22. By web scraping (https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset), the data was exported to the .csv format with 19,239 observations. In the beginning, the data contained 117 variables, but after data cleaning, only 12 variables are used in this research, as shown in Table 1. In this data, our variable of interest that we want to predict is value\_eur, or the market value. The other 11 variables will act as predictors.

Table 1. Description of variables used in the data

|  |  |
| --- | --- |
| **Name of variable (data type)** | **Description** |
| value\_eur (numerical) | The market value of each player in EUR. |
| age (numerical) | Age of player. Typical soccer players start their careers at around 18 and retire after 35. |
| league\_level (categorical, 1-5) | Some leagues are more well-known than the others i.e., English Premier League, Bundesliga, Ligue 1, and 1 indicates that the player is playing at the highest league level, 2 is the second highest, and so on. |
| international\_reputation (categorical, 1-5) | Indicates how well-received a player is in the international community, with 5 being the most well-received. |
| contract\_remaining\_yr (categorical, 0-10) | Calculated from 2021, for example if a player’s contract at that club ends in 2022, then this column will be equal to 1. |
| attacking\_avg (numerical 0-100) | The average score of how well a player can do crossing, finishing, heading, short passing, and volley kicks. |
| physic (numerical 0-100) | The score of how fit a player is physically. |
| skill\_avg (numerical 0-100) | The average score of how well a player can do dribbling, curve shooting, long passing, ball control, and free kicks. |
| movement\_avg (numerical 0-100) | The average score of a player’s quality in acceleration, sprint speed, agility, reaction, and balance. |
| power\_avg (numerical 0-100) | The average score of a player’s quality in shot power, jumping, stamina, strength, and long shots. |
| mentality\_avg (numerical 0-100) | The average score of how well a player can manage their mentality during penalties, field visions, composure, positioning, aggression, and interception. |
| defending\_avg (numerical 0-100) | The average score of how well a player can do marking awareness, standing tackle, and sliding tackle. |

Upon closer inspection, the market value is not normally distributed, as shown in Figure 1, 2, and 3. This violates the v: linearity, independence, normality, and homogeneity of variance (Cross & Daniel, 2010), therefore linear regression cannot be used, unless the data can be transformed to attain linearity. Another approach is to use nonparametric regression, which does not assume anything about the distribution of the data. These two possibilities are the main motivations of this paper. Similar comparisons have been done in the past, including predicting Bitcoin’s price (Gyamerah, 2022) and aviation weather forecasting (Vislocky & Fritsch, 1995).

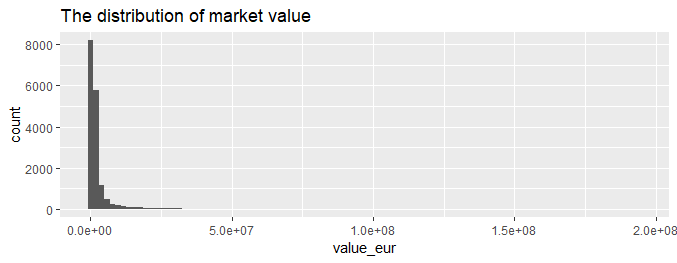


Figure 1. The distribution of market value is very left-skewed, meaning only selected top players receive big salaries.

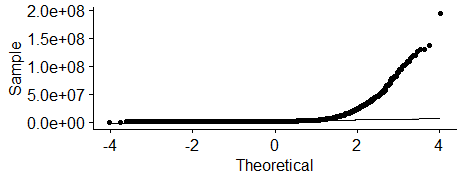


Figure 2. QQ plot is also used to test the normality, but the data points failed to fall under the straight line.

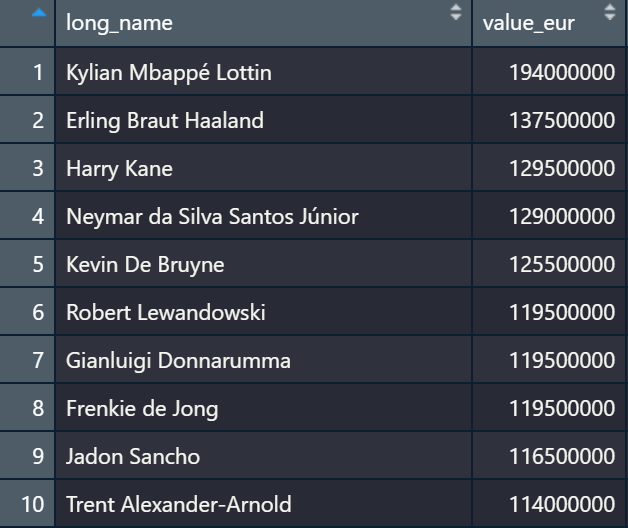


Figure 3. Ten highest market value worldwide, with all of them worth over 100 million euro. However, at the bottom, we can see players that worth exponentially less than the top players.

Supported by the fact that the data are adequate and match to the parameters used in the previous studies, this paper will study and compare results between multiple linear regression (MLR) and generalized additive model (GAM) and test whether GAM performs better than the MLR for FIFA 22 dataset. Data processing and model analysis will all be performed in R programming language. From the result, it can also be determined which model captures the nonlinearity of the market value distribution better. This result can be useful for people who play FIFA 22 so they can plan their transfer strategies better when playing as coaches in the Career Mode.

## Methods

* 1. **Parametric and Nonparametric Regression Theory**

Generally speaking, the motivation of regression in general is to predict the output variable, or sometimes called the response or dependent variable, by using predictor, independent variables, features, or simply named variables. Compared to classification algorithms that return the category in which the response belongs to, regression methods aim to predict the numerical values of the response. Let *n* independent variables, , with each variable having *p* observations such that , then the predicted value would be in the form of .

The first approach is to use MLR. According to (James et al., 2021), the shape of the function is assumed to be linear in *X*:

(1)

Equation 1 is called a linear model. This model simplifies the true function in a sense that one only needs to estimate coefficients to model the data, instead of estimating the function itself. Note that is the error of the prediction after the model is trained/fitted. The problem is, in real life, it is rare to encounter a very simple linear problem. In addition, if the data does not fulfill four basic assumptions of linear regression as mentioned in Section 1, the model will produce a poor estimate.

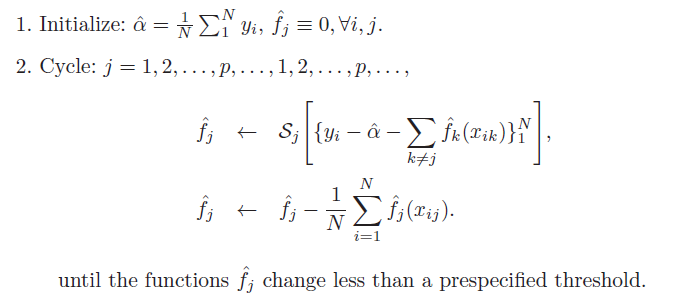
To solve this problem, nonparametric regression is preferred when linear regression is not suitable or when the data does not follow the assumptions needed for performing linear regression (James et al., 2021). Instead of assuming the linearity of the model, the method tries to estimate . The nonparametric model that will be used in this paper is Generalized Additive Models (GAM). GAM has the form:

(2)

In this case, ’s are unspecified smooth functions for each independent variable. The general function can also be expressed by a link function *g*.

(3)

To fit the GAM, the scatterplot smoother for flexible fitting of nonlinear effects is applied to the data. After that, backfitting algorithm as shown in Figure 4 is implemented to estimate each smooth function ’s (Hastie et al., 2017).



* 1. **Data Preprocessing**

Based on the data description in Section 1.2, initially a lot of columns were present in the dataset. Some columns were deleted because they do not contribute to the model, such as jersey number, body type, URLs to the image, club name, and player id in Sofifa. To avoid high dimensional problems in regression, some similar features were combined. For example, five columns (finishing, heading, short passing near the goals, crossing, and volley kicks) basically quantify the same idea: how well a player can contribute to a goal for their team, be it by scoring a goal or assisting. Therefore, those five similar columns were averaged into a single variable named “attacking”. This process is repeated for other qualities that exist in FIFA 22 soccer players e.g., defending, mentality, skill, movement, and power.

In addition to that, the heatmap showing the Pearson correlation between each variable is generated to function as a quick data analysis before performing the regression. To eliminate the skewness of the data, the value\_eur variable is also log-transformed. After selecting only necessary features and removing the missing values, the data are now ready to be passed to the model to train it.

* 1. **Parametric Regression Workflow and Model Evaluation**

MLR was done in R by using built-in lm command. To evaluate the model, several metrics were calculated, such as mean square error (MSE), Akaike information criterion (AIC), and Bayesian information criterion (BIC). MSE measures the deviation of predicted values from the true values, and can be calculated by using the following formula (James et al., 2021):

(3)

AIC is widely used for model selection and the model with the smaller AIC value is better at fitting new data. The formula is given by (Hastie et al., 2017):

(4)

BIC is similar to AIC, but this metric uses Bayesian approach. The formula is:

(5)

Where RSS is residual sum of square, n is the number of observation, d is the total number of parameters, and is the estimated variance.

* 1. **Generalized Additive Model Workflow and Model Evaluation**

GAM function was provided by the mcgv library. The same metrics were also calculated for this section to compare with the result obtained from the parametric regression.

## Results and Discussion

Based on the methods presented in Section 2, the resulting correlation heatmap plot can be observed in Figure 4. This heatmap gives a simple visualization and allows a quick representation on how each variable correlate to each other. The heatmap also serves as a guideline for the linear regression model to see whether it matches with the correlation heatmap. According to Figure 4, variables that positively correlate the most with the market value are mentality, attacking, and skill. This result is not surprising from a domain knowledge perspective, because players that possess good mentality scores tend to perform well under the highest level of competition. Attacking and skill variables also determine how well a player can contribute to a goal for the team, therefore those are the desirable qualities that a team naturally would seek for. Surprisingly, age is neutrally correlated with the market value. In a vacuum, a player’s ability on the field, such as stamina, should decline as they grow older. However, in the history, a lot of world-class players tend to gain hefty price tags from a young age and retain their values even when they start to hit 30s, take Lionel Messi and Cristiano Ronaldo as examples. Therefore, it makes sense for the age to correlate neutrally with the market value. Another thing that stands out is defending skill seems not to be as important in determining a player’s market value as skills associated with attacking, at least in a linear correlation.

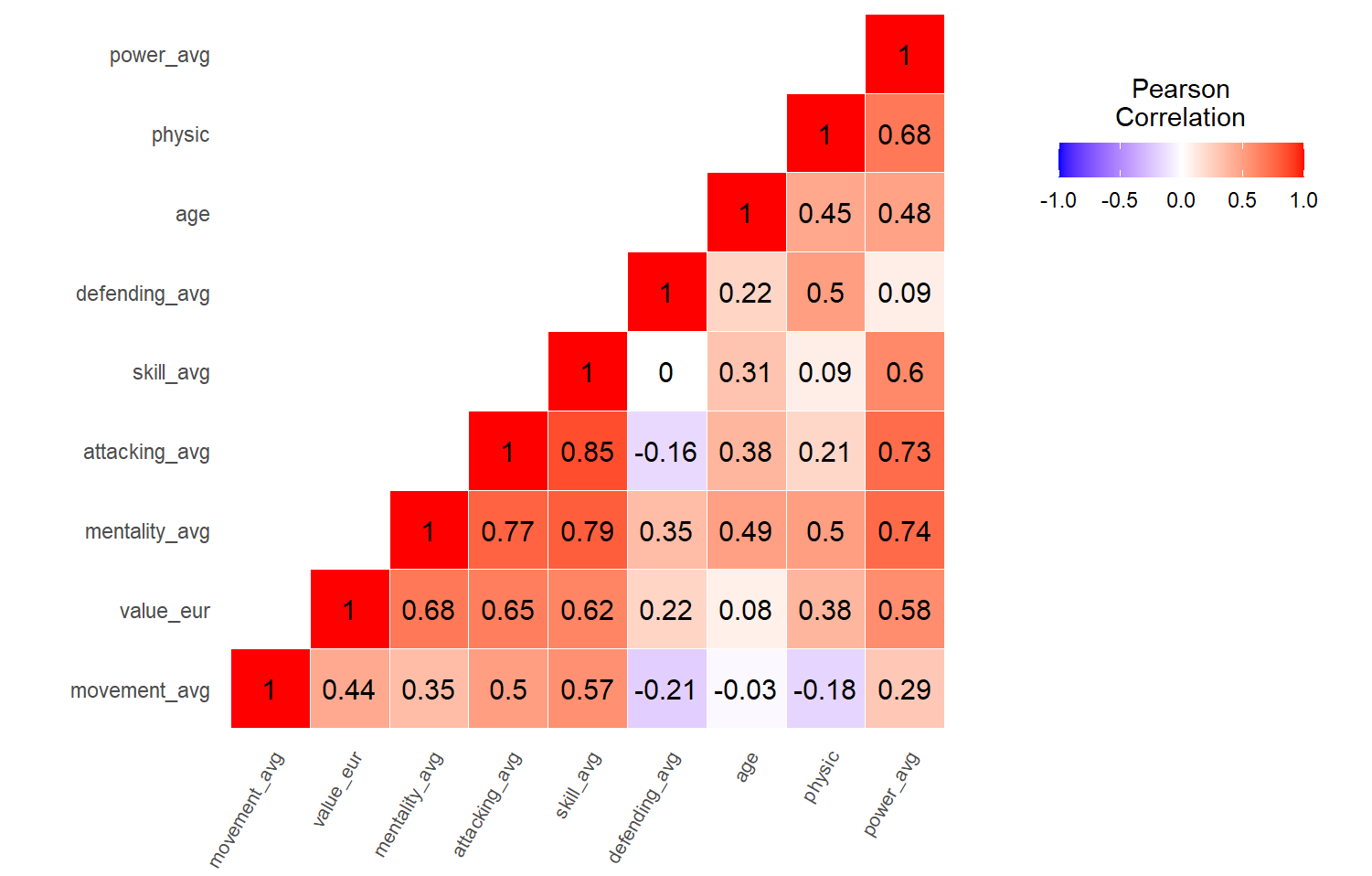
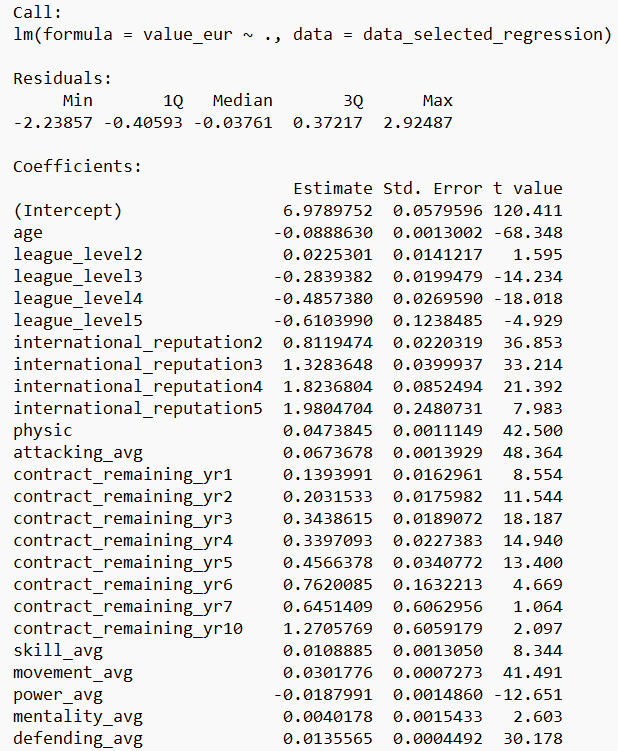
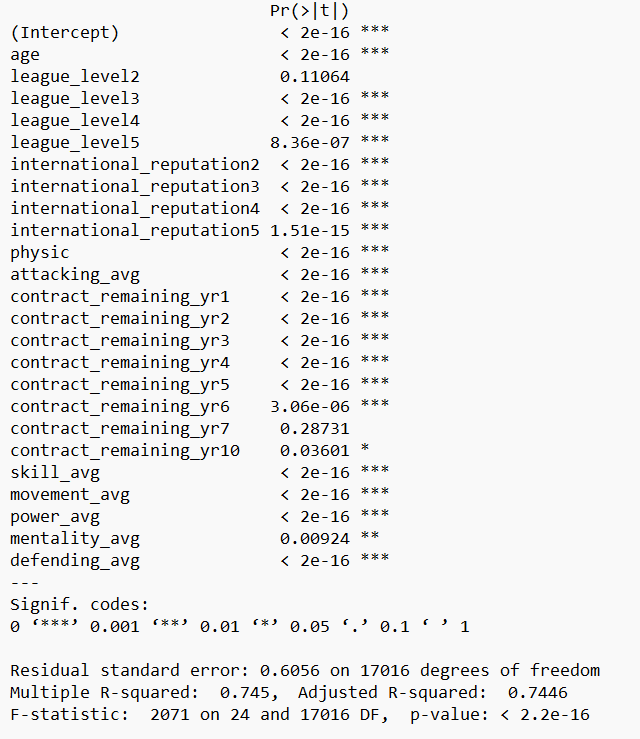


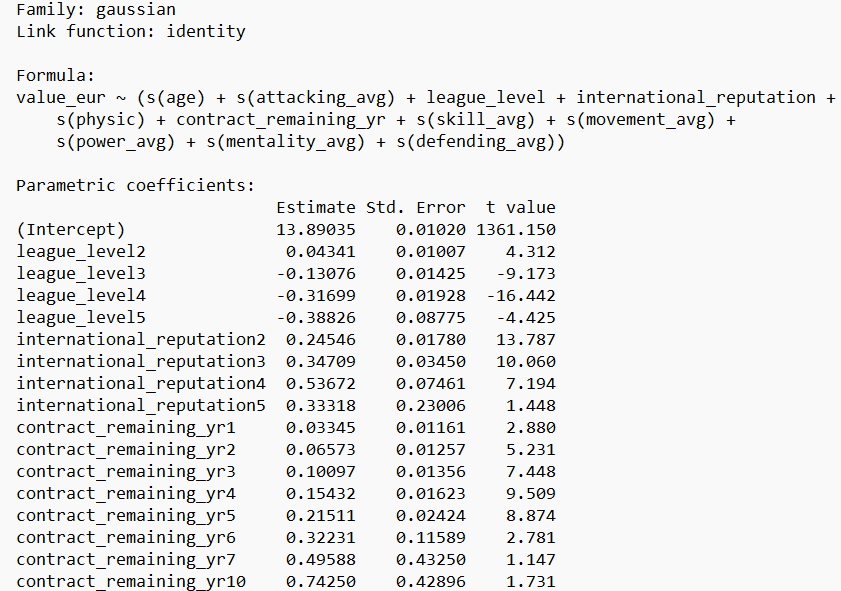
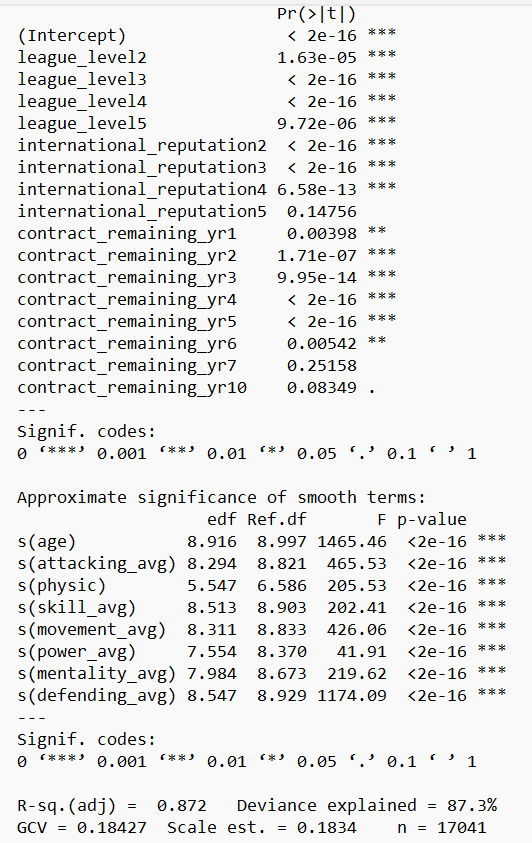
Figure 4. Correlation heatmap plot for the FIFA 22 players data.

The main reason a log-transformation was applied to the response variable (market value) is to comply with the assumptions for performing a linear regression, one of them is linearity. Figure 5a and 5b illustrate the parametric MLR result, and Figure 6a and 6b show the nonparametric GAM output. In MLR model, almost all of the parameters are significant, indicated by the \*\*\* next to their P-values. Only a very small portion of the players that have a very long contract could be the reason contract\_remaining\_yr\_10 was not significant in both models. Their R-squared values are not that different, 0.745 to 0.773. It might sound tempting to stop here and conclude that GAM improves from the MLR by a significant margin. However, R-squared values sometimes do not reflect the model’s ability to avoid overfitting and adjust to nonlinearity, therefore other metrics such as AIC, BIC, and MSE are used instead and will be discussed later. Overall, both models are significant and good enough to be used for further analyses.



(a) (b)

Figure 5a. The coefficients of each parameter and the intercept from MLR, 5b. The significance of each parameter in MLR model.

(a) (b)

Figure 6a. The coefficients of each parameter and the intercept from GAM, 6b. The significance of each parameter in GAM model

After deciding that both models can be used to explain the relationship of the response (market value) with the features, the models were plotted by using a library named visreg. Visreg is useful for producing added-variable plots that show the regression model for each independent variable. Figure 7 to 9 described the relationship between market value and duration of contract remaining (in years), international reputation rating, and league level in which the players are in, with the grey dots representing the data points generated by the models and blue line is the fit line. Because the smooth function could not be applied to categorical variables, those variables were left as they are, instead of being wrapped in s() function like other numerical variables that will be discussed later. Those three figures did not produce enough insight since the data were heavily imbalanced and therefore it was difficult to make a meaningful interpretation from skewed categorical variables. Also, looking from their p-values in both models, in some levels they were not significant, therefore it could be challenging to analyze and come to a correct conclusion.

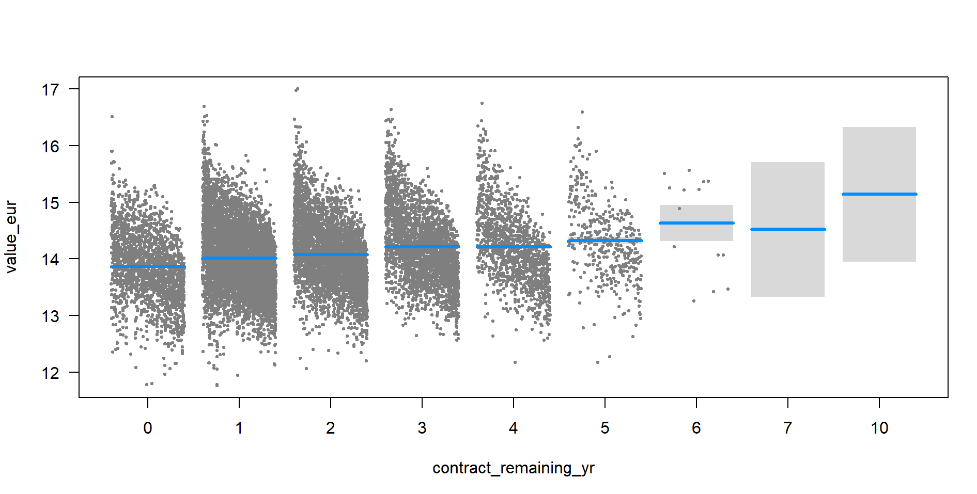


Figure 7. The relationship between value\_eur and each contract\_remaining\_yr category.

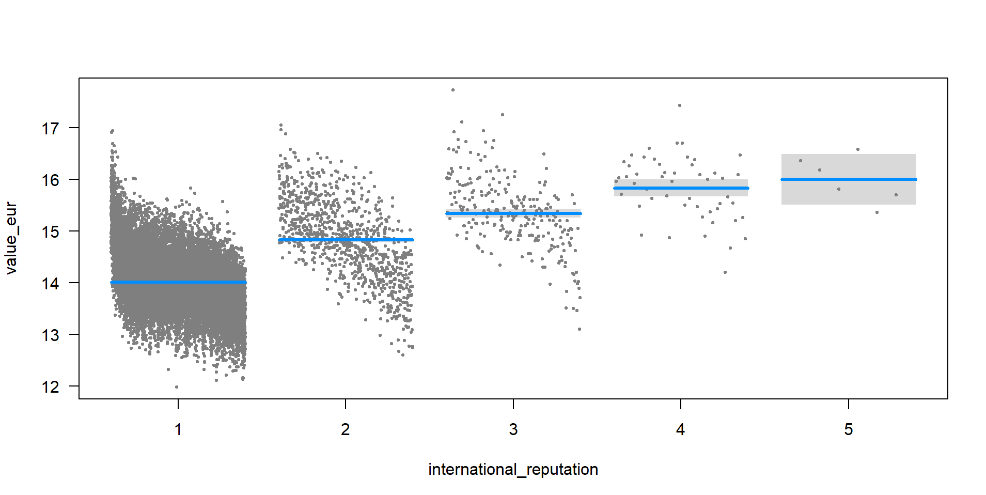


Figure 8. The relationship between value\_eur and each international\_reputation category.

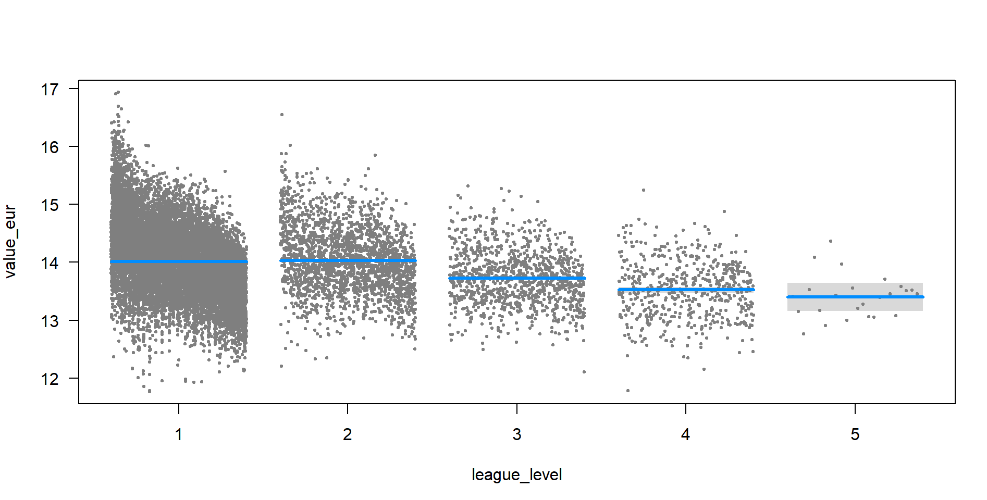


Figure 9. The relationship between value\_eur and each league\_level category.

Figure 10 displays the regression model for market value against age. While the MLR model showed an almost horizontal trend, GAM captured a more interesting trend. According to the result obtained from GAM, the decline in market value happened slightly slower from age 20 to early-30s, then started dropping faster as a player approaches 30s. However, there is a chance for the player to retain its market value, and this is usually true for more popular players that are close to retirement.

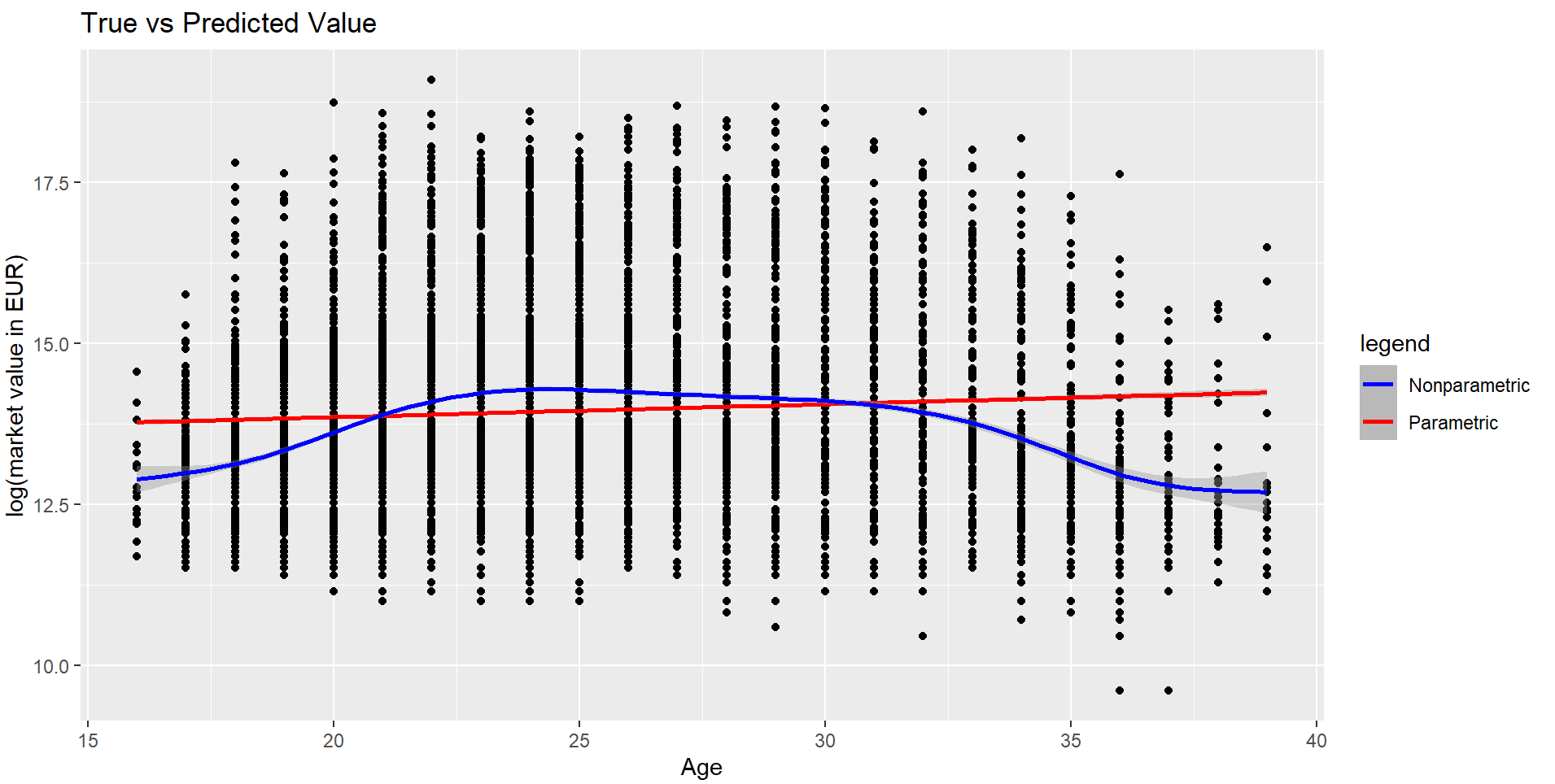


Figure 10. The comparison between value\_eur vs age from MLR and from GAM.

Figure 11 showed a clear distinction between MLR model and GAM. While the result from MLR is generally enough for the public to understand, GAM explained it better. According to MLR, clubs valued players that could score more goals more than those who were less capable in offensive department, and the relationship was linear. From GAM, it could be concluded that the market value increased at a greater rate as the attacking ability went from 60 upwards. Therefore, the importance of attacking ability to the market value was more visible in the GAM model result than in the MLR result. It could also be observed that GAM captured more data points where the attacking score and market value were both high.

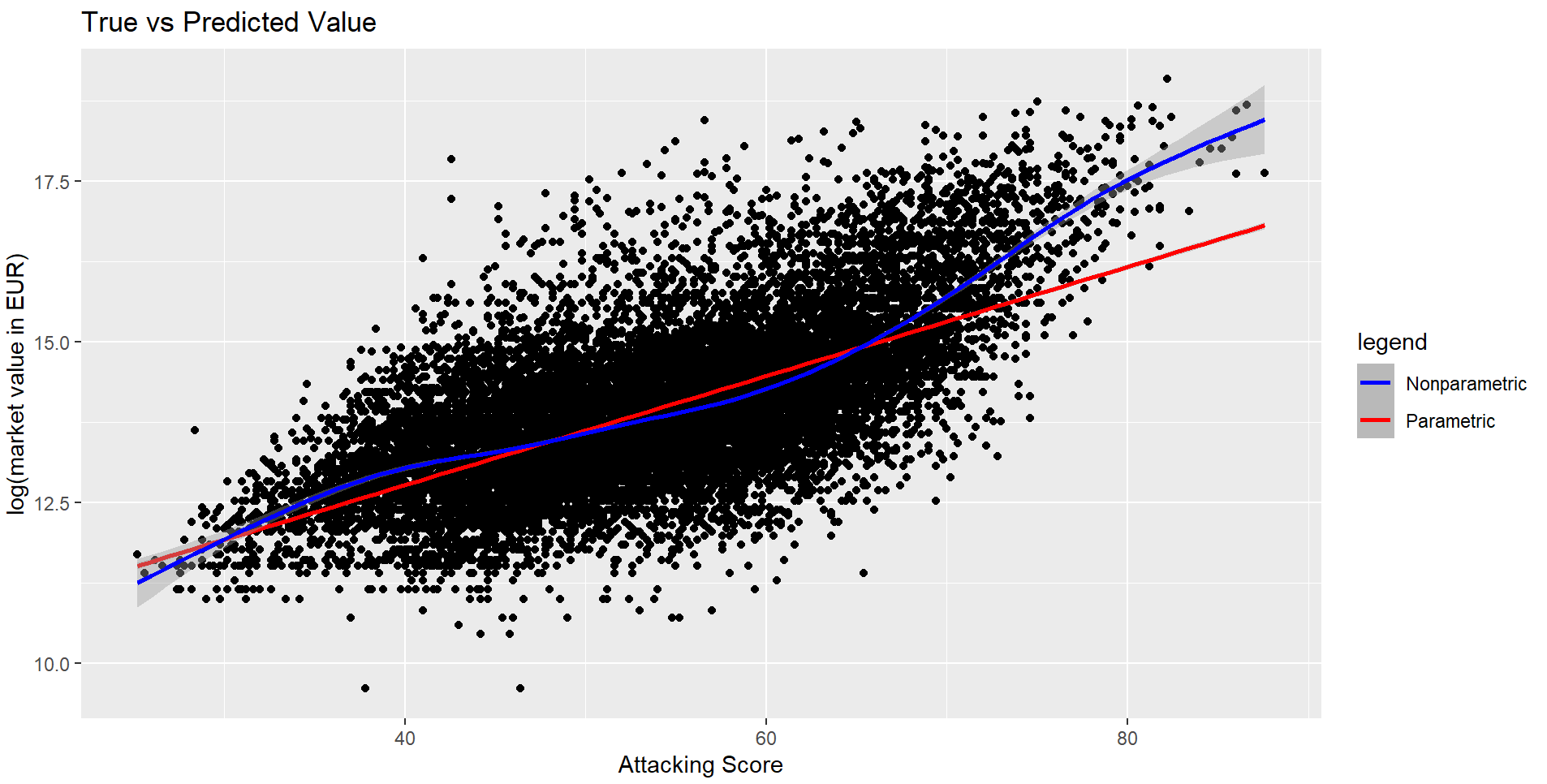
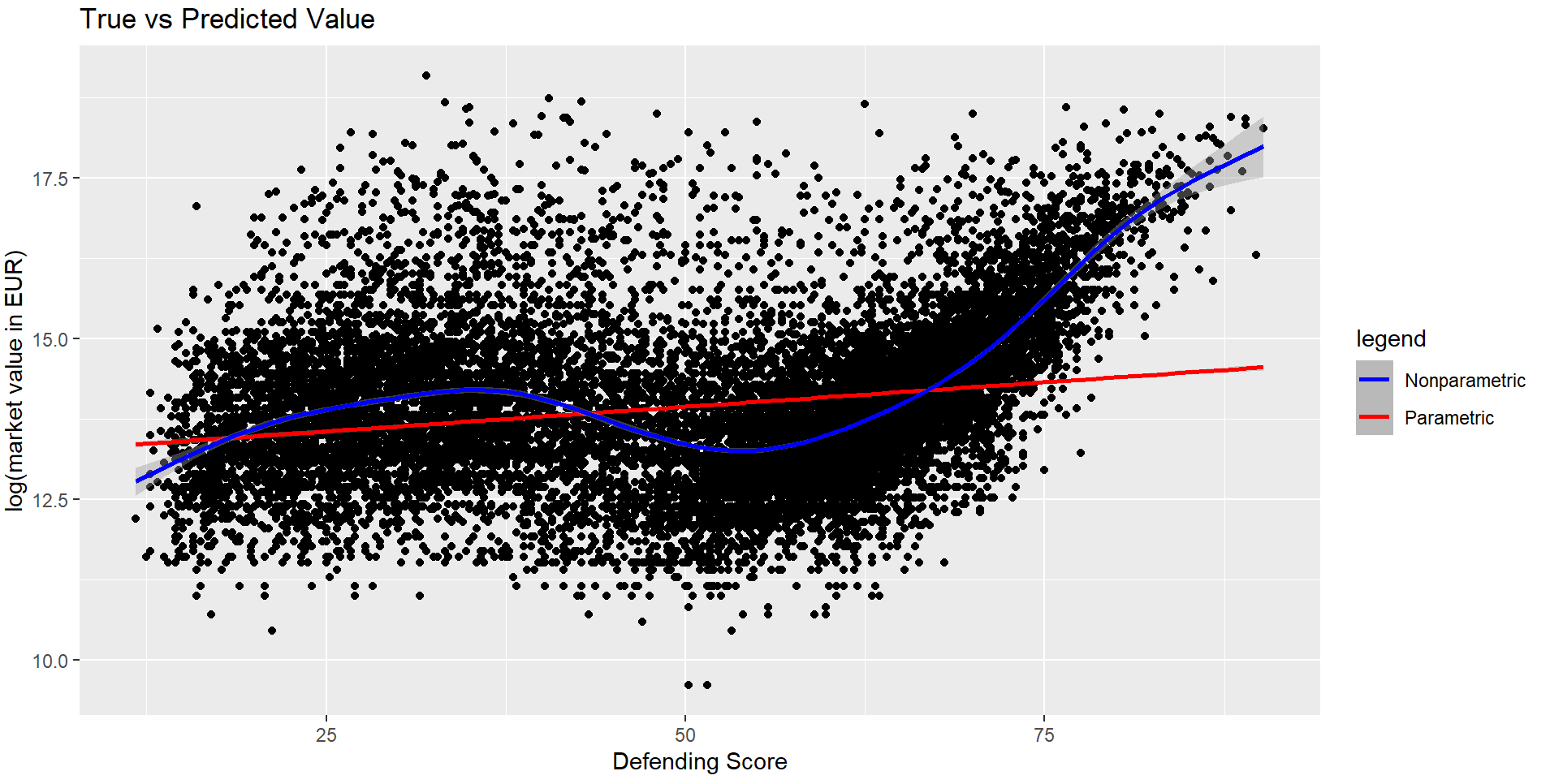


Figure 11. The comparison between value\_eur vs attacking score from MLR and from GAM.

Defending was not seen as a more important factor in the linear regression model, as depicted by MLR by the slightly increasing trend. While it is true that defending ability affects market value positively, this model failed to capture the second trend from defending\_avg with the score of 50 and above. GAM explained that as the player got better at keeping their team from conceding more goals, the clubs started to reward them more in a nonlinear trend. In soccer, emphasis has been put to attacking capabilities of a team and player, and there is even a great adage applied to many sports and strategic games, “The best offense is a good defense”. It basically means that constantly attacking the opposing team will put them into defensive, leaving them less chance to strike back. However, the converse is also true: a good defense can win games because there is even less chance of the opposing team to outscore on a crackback if the team can defend well.

 Figure 12. The comparison between value\_eur vs defending score from MLR and from GAM.

Similar to attacking and defending, according to GAM as shown in Figure 13, mentality also increased faster as the score increased above 60. The MLR result showed a straight increasing line. While generally the result from MLR was enough to explain the impact of mentality score versus market value, GAM elaborated further that high mentality score could make a difference between winning and losing, such as in penalty shootout or extra time. There were not many players with high mentality score, therefore most clubs were willing to spend more money to acquire such players.

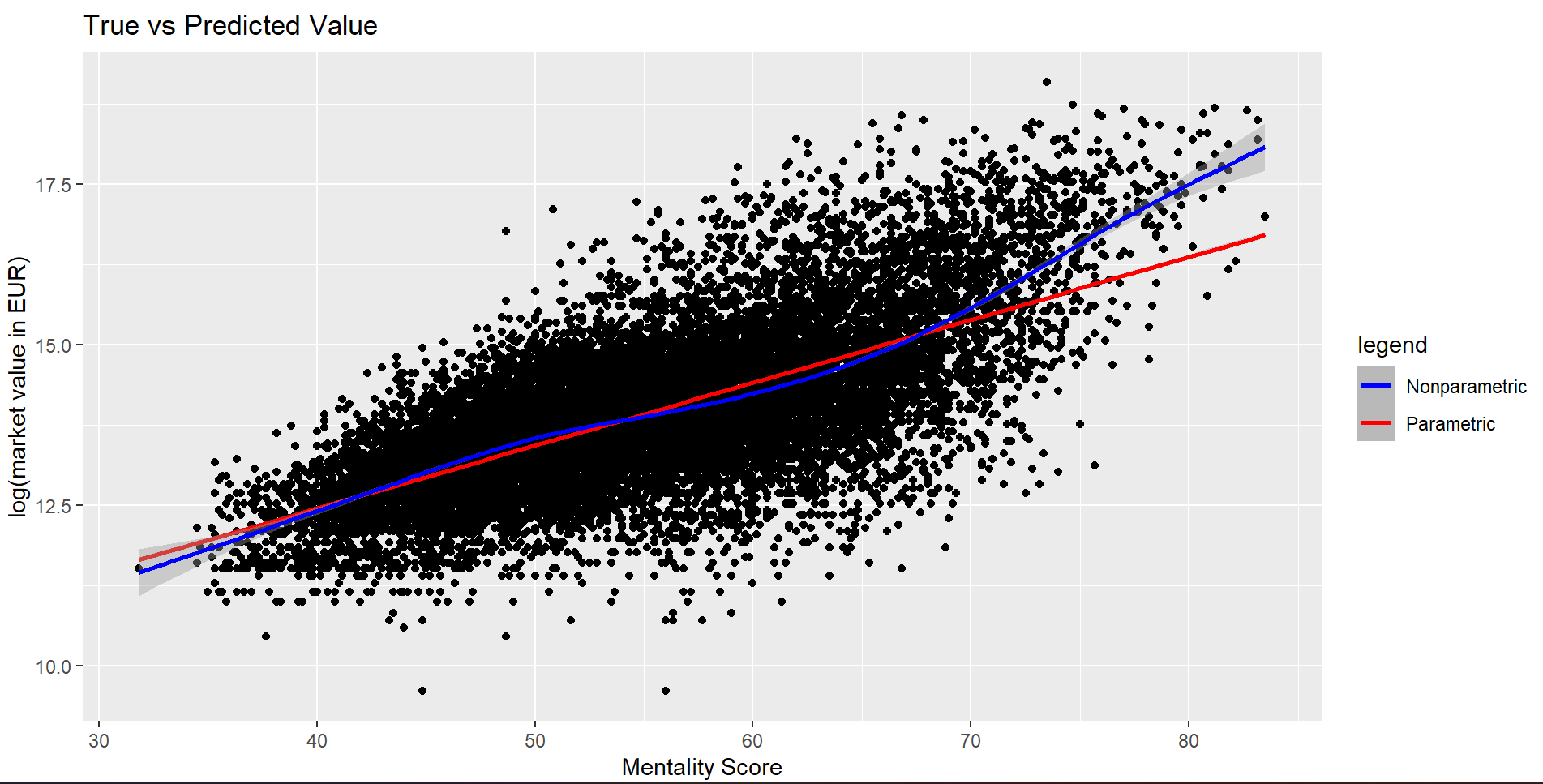


Figure 13. The comparison between value\_eur vs mentality score from MLR and from GAM.

Both models agreed that market value increased as the players had better movement capabilities, as shown in Figure 14. However, GAM performed better visually in a sense that it successfully captured the higher movement average score trend, just like the previous 3 features. Players with finesse and could navigate the field better could open up more opportunities to score goals, and therefore obtain more wins. Therefore, when the movement average score was above 60, the trend started to climb faster in a nonlinear fashion. Also, fitness is an important aspect for every soccer player. From both models in Figure 15, the market value was proportional with the physical score of the player. Although the GAM model indicated there was a bit nonlinearity in the relationship, the difference was not much.

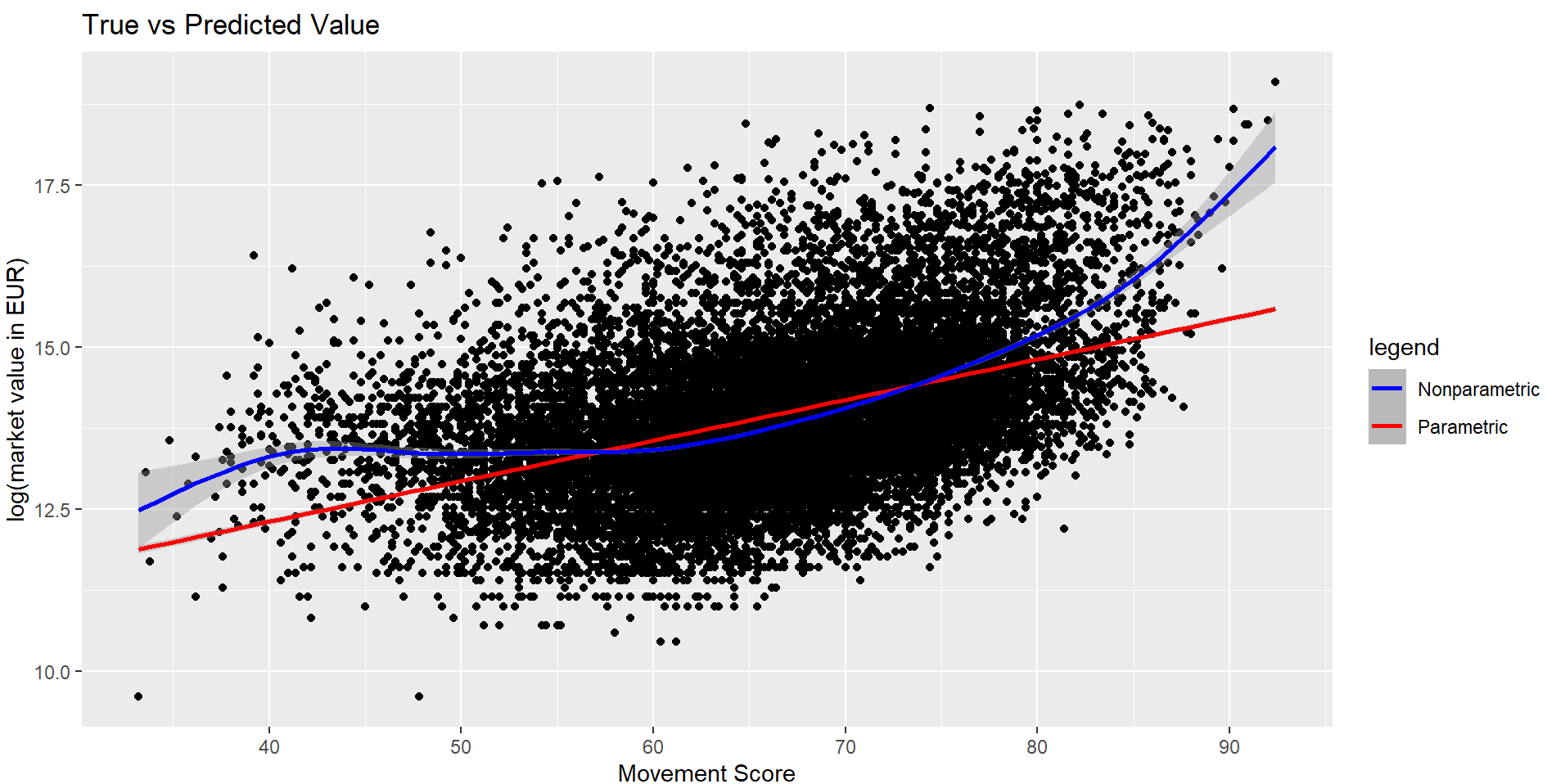
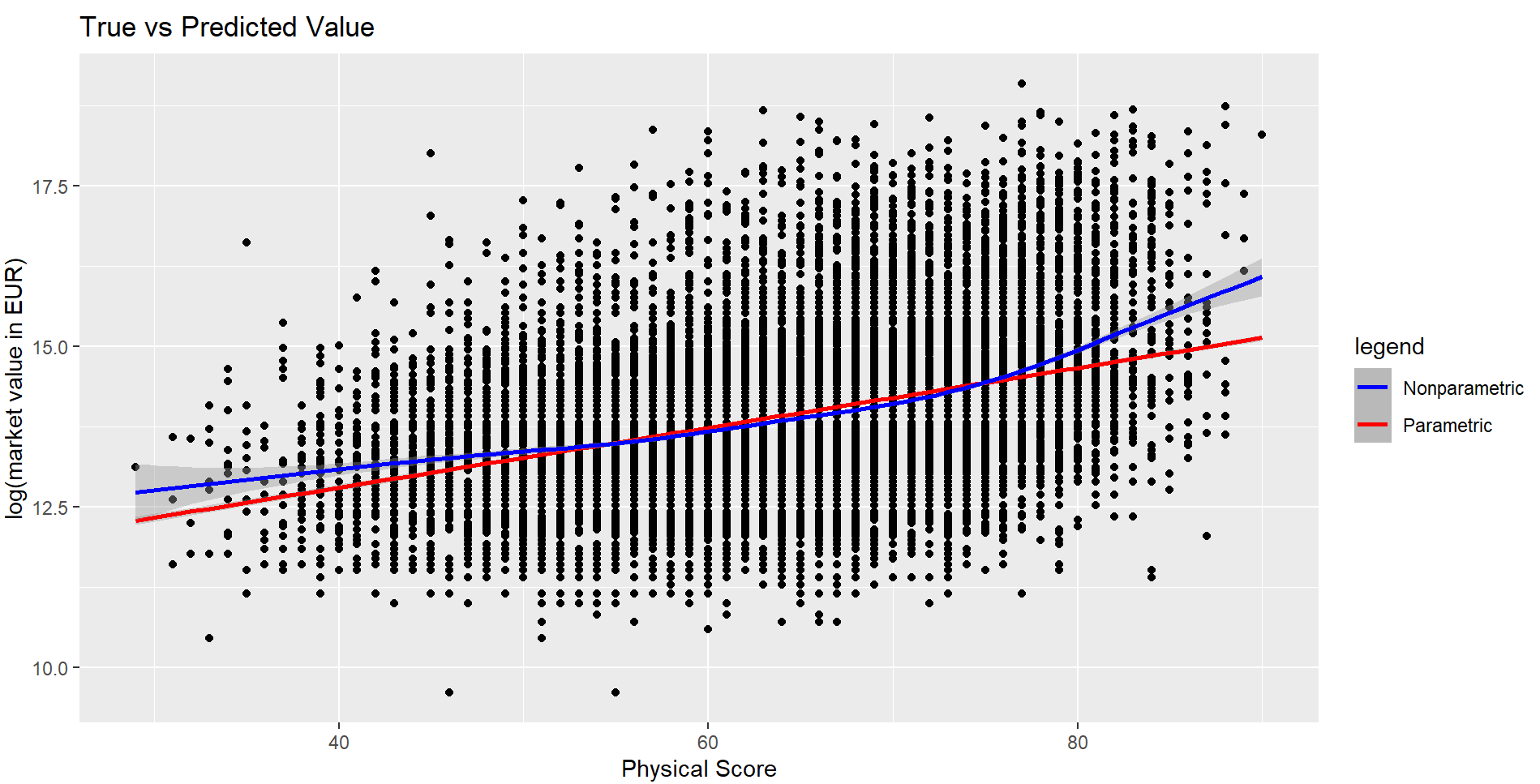
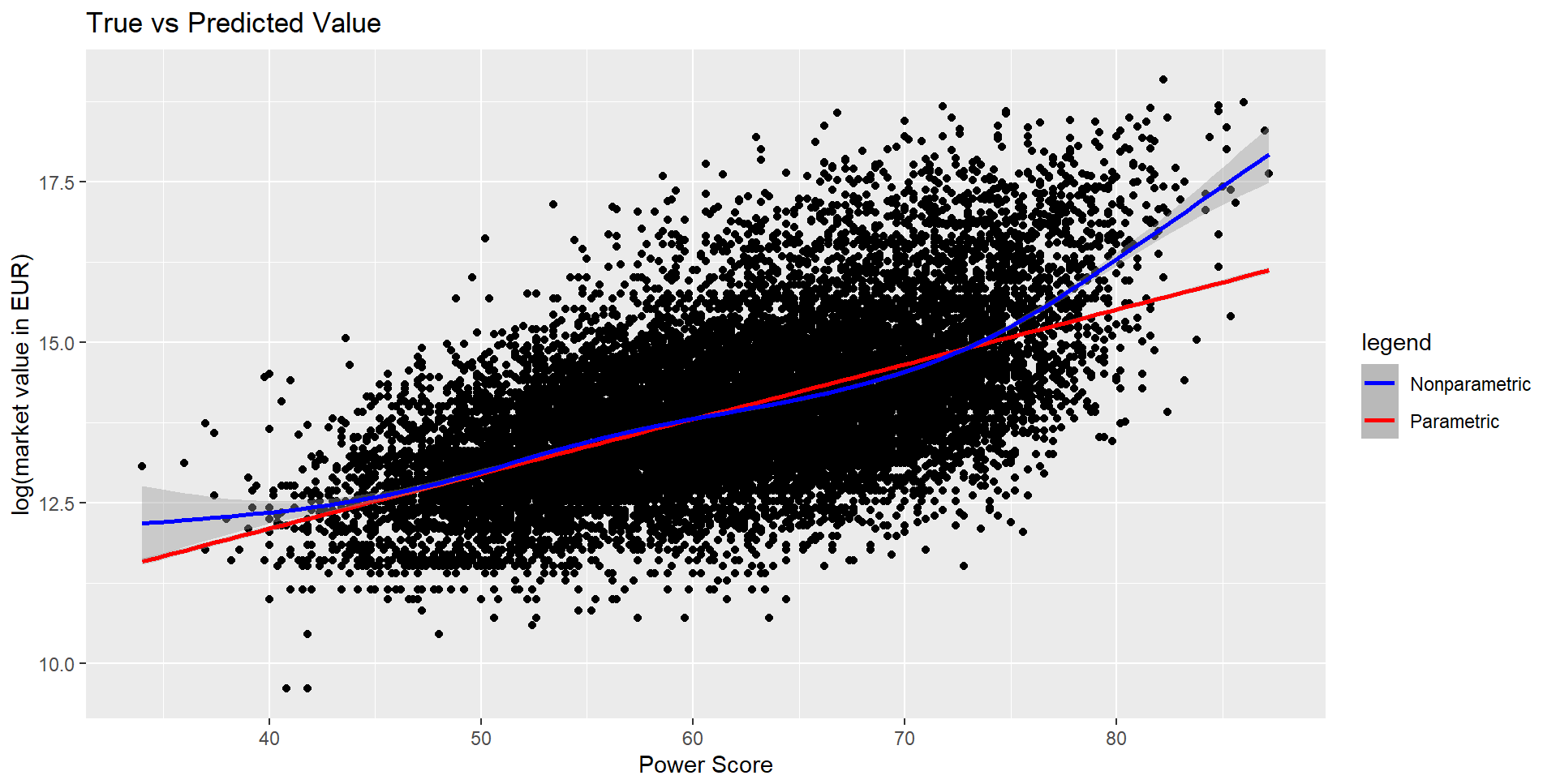


Figure 14. The comparison between value\_eur vs movement score from MLR and from GAM Figure 15. The comparison between value\_eur vs physical score from MLR and from GAM

The last two variables that will be discussed in this section are power and skill. From both models in Figure 16, the insight obtained was the market value started increasing as a player put their focus on power. GAM provided a clearer explanation that power would start being more rewarded after a certain threshold. When the power score started to increase above 60, the market value increased faster. It is true that power is not essential in winning soccer games; efficiency is. However, stamina also contributed to the power average, and players that can play longer games usually generate more return of investment for the club as the club gains more milage out of those players. Therefore, clubs are willing to pay more for players that have better stamina. Similar to other features, as shown in Figure 17 , skills such as dribbling, free kicks, and long passes were significantly more rewarded after the score passed a certain threshold. Some of the best players in the world, for example Andrea Pirlo, have been in demand because of their capabilities in converting free kicks into goals.

 Figure 16. The comparison between (a) value\_eur vs power\_avg from MLR and (b) f(power\_avg) vs power\_avg from GAM.

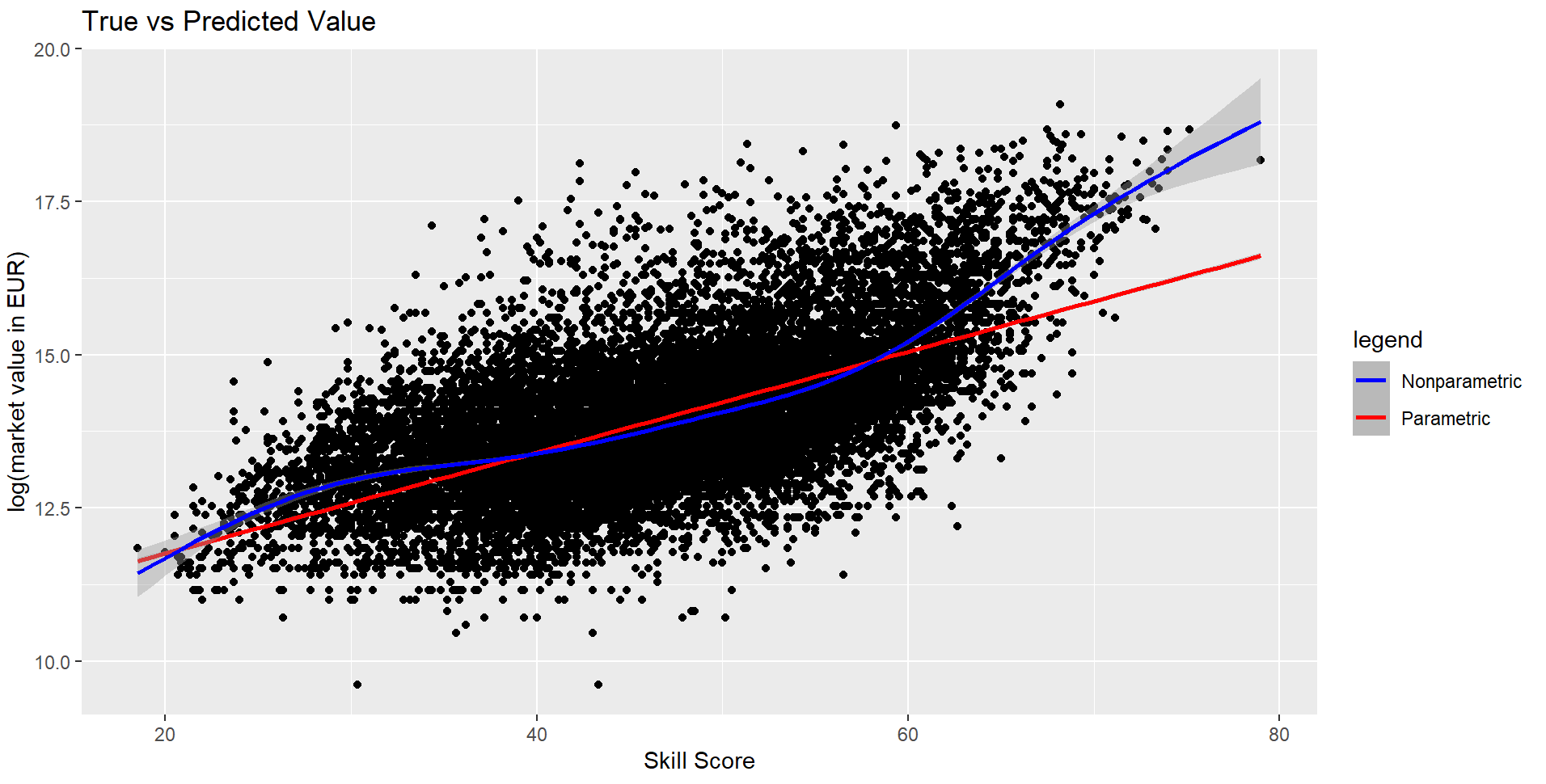


Figure 17. The comparison between (a) value\_eur vs skill\_avg from MLR and (b) f(skill\_avg) vs skill\_avg from GAM.

After analyzing the comparison of both models visually, the performance of both models could also be compared by using AIC, BIC, and MSE as demonstrated in Table 2.

Table 2. Metrics for model selection

|  |  |  |
| --- | --- | --- |
|  | MLR | GAM |
| AIC | 31292.04 | 19539.74 |
| BIC | 31493.37 | 20172.11 |
| MSE | 0.366 | 0.183 |

When performing model selection, the model with the lowest AIC, BIC, and MSE are preferred. GAM proved to be superior to MLR in predicting the market value of FIFA 22 players, even after performing log-transformation to improve the performance of MLR before the model training. GAM predicted better because the data contained nonlinearities that could not be described by parametric regression, but GAM could adjust to the data by using the smooth function. If, for some reason, the data was linear, it could still be possible that MLR performed better than GAM in terms of those three metrics. In the end, it all depends on the type of data that are being used in the study.

The only limitations in this research were the lack of numerical international reputation data from 0-100 and how the model handled categorical variables. International reputation would have explained the market value a lot because popularity dictates the market value of the player most of the times.

## Conclusion

After comparing MLR and GAM, GAM was superior to MLR both visually and quantitatively speaking. By using the GAM model to predict the market value of the players in FIFA 22 video game, the model successfully explained the nonlinear properties from each predictor that the MLR failed to do. For example, defending, attacking, power, movement, and skill scores were more rewarded once they passed a certain threshold, which was not explained in the MLR model. In addition to that, GAM decreased the MSE, AIC, and BIC by about 50%, 37%, 36% respectively compared to the MLR model. Overall, this model is better for FIFA 22 video game players so they have a better understanding of the factors affecting the market value and how to adjust their investment in the game accordingly.

## References

Barbuscak, L. (2018). *What Makes a Soccer Player Expensive? Analyzing the Transfer Activity of the Richest Soccer.* Augsburg Honors Review.

Cross, C. L., & Daniel, W. W. (2010). *Biostatistics: basic concepts and methodology for the health science.*

Ezzeddine, M. (2020). Pricing football transfers: determinants, inflation, sustainability, and market impact: finance, economics, and machine learning approaches. *Economics and Finance, Université Panthéon-Sorbonne*.

FIFA. (2021). *Global Transfer Report 2021.* FIFA.

Gibson, A. (2022, January 19). *FIFA 22 Is the Best-Selling Game in 17 of 19 EU Nations*. Retrieved from Twinfinite: https://twinfinite.net/2022/01/fifa-22-is-the-best-selling-game-in-17-of-19-eu-nations/

Gyamerah, S. A. (2022). On forecasting the intraday Bitcoin price using ensemble of variational mode decomposition and generalized additive model. *Journal of King Saud University - Computer and Information Sciences*.

Hastie et al., T. (2017). *The Elements of Statistical Learning.* Springer.

James et al., G. (2021). *An Introduction to Statistical Learning.* Springer.

Metelski, A. (2021). Factors affecting the value of football players in the transfer market. *Journal of Physical Education and Sport*.

Poli et al., R. (2021). Econometric Approach to Assessing the Transfer Fees and Values of Professional Football Players. *Economies*.

Valentini, K. (2020). *Transfer Pricing: An Analysis of The Impact of Player Brand Value on Transfer Fees in European Football.* Pennsylvania: Joseph Wharton Scholars.

Vislocky, R. L., & Fritsch, J. M. (1995). Generalized Additive Models versus Linear Regression in Generating Probabilistic MOS Forecasts of Aviation Weather Parameters. *Weather and Forecasting*.