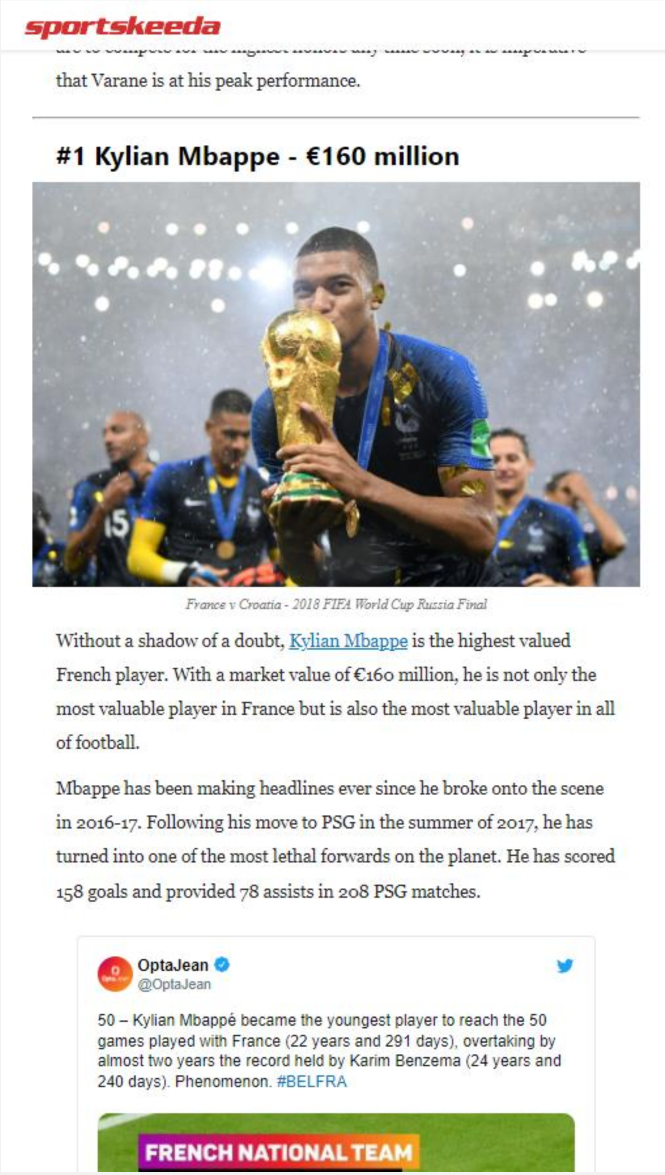
A Comparative Study of Multiple Linear Regression and Generalized Additive Model in Predicting Market Value of Players in FIFA 22

Christopher Salim

Background

Market value =/= salary


sportskeeda

that Varane is at his peak performance.

#1 Kylian Mbappe - €160 million

France v Croatia - 2018 FIFA World Cup Russia Final

Without a shadow of a doubt, Kylian Mbappe is the highest valued French player. With a market value of €160 million, he is not only the most valuable player in France but is also the most valuable player in all of football.

Mbappe has been making headlines ever since he broke onto the scene in 2016-17. Following his move to PSG in the summer of 2017, he has turned into one of the most lethal forwards on the planet. He has scored 158 goals and provided 78 assists in 208 PSG matches.

OptaJean
@OptaJean

50 – Kylian Mbappé became the youngest player to reach the 50 games played with France (22 years and 291 days), overtaking by almost two years the record held by Karim Benzema (24 years and 240 days). Phenomenon. #BELFRA

~ USD 4.86 Billions in 2021 were spent on transfer fees (Global Transfer Report, FIFA, 2021).

Need to be careful when assessing the market value of a player.

Some topics about market value analysis in academia include:

- Economic and technical approach to explain the market value of each player (Poli et al., 2021), (Ezzeddine, 2020)
- Importance of player’s age to market value (Metelski, 2021)
- Player’s reputation impact to market value (Valentini, 2020)

Can we apply the same concept to the soccer player data from the video game?
Can soccer video game enthusiasts apply the same market value analysis?

FIFA 22 Dataset

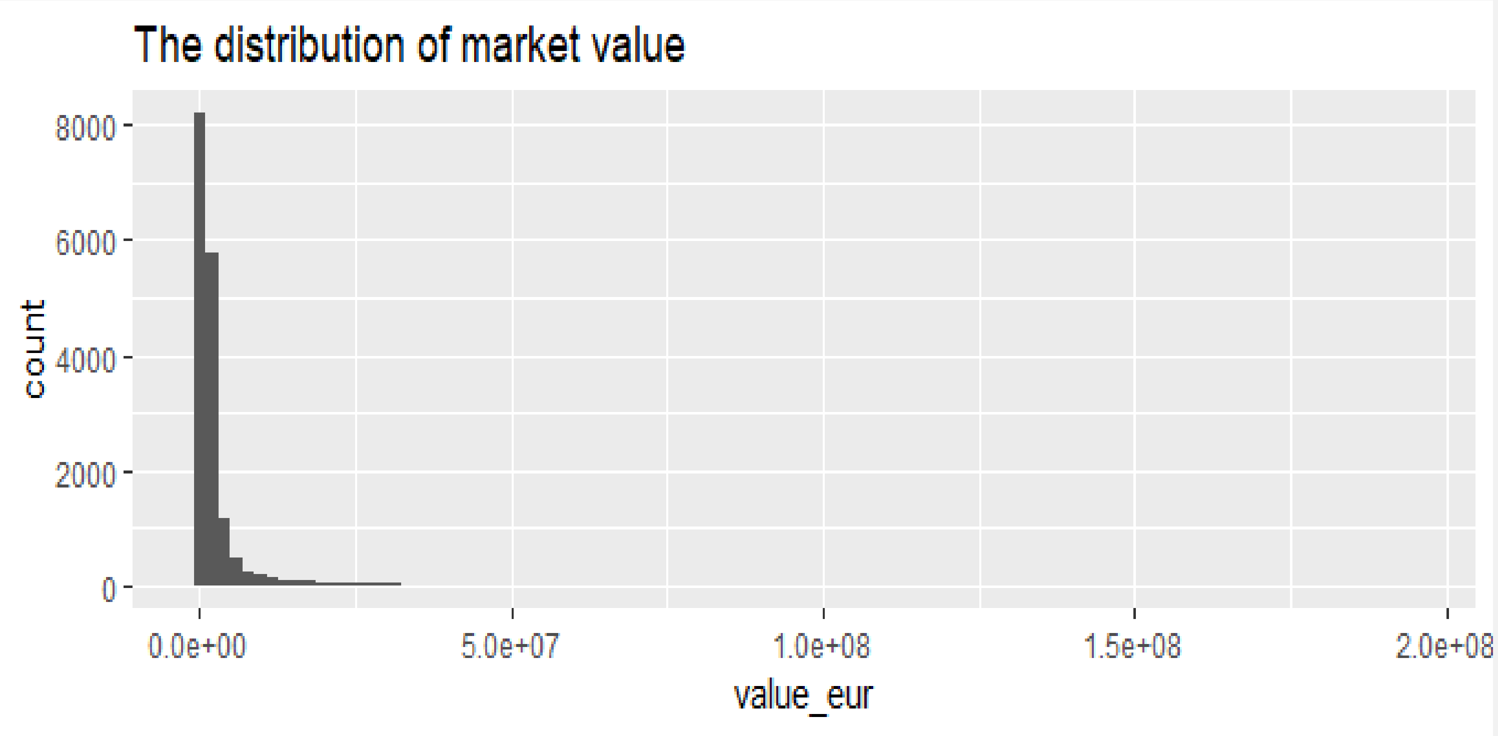
Response variable

Features

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.

Data source:
https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset
19,239 observations

Dataset at a glance



	long_name	value_eur
1	Kylian Mbappé Lottin	194000000
2	Erling Braut Haaland	137500000
3	Harry Kane	129500000
4	Neymar da Silva Santos Júnior	129000000
5	Kevin De Bruyne	125500000
6	Robert Lewandowski	119500000
7	Gianluigi Donnarumma	119500000
8	Frenkie de Jong	119500000
9	Jadon Sancho	116500000
10	Trent Alexander-Arnold	114000000

Left: histogram of market value distribution

Right: overview of some of the players with the highest market values.

Objectives

- Study and compare results between multiple linear regression (MLR) and generalized additive model (GAM) to model the market value variable vs 10 other features
- Choose a model that helps FIFA 22 video game players plan their transfer market strategies better.

Methods

- All were done in R

Data preprocessing and visualization → Parametric and nonparametric model fitting → Metrics calculation and added-variable plots for model selection

Data preprocessing

- Columns deletion e.g., jersey number, body type, URLs to the image, club name
- Combining similar features, for example attacking score = (finishing + heading + short passing + crossing + volley kicks)/5
- Removing missing values (2328 observations, mostly goalkeepers whose other aspects in soccer were usually not recorded)

Multiple Linear Regression

- Let *n* independent variables, $X = (X_1, X_2, \dots, X_n)$, with each variable having *p* observations such that

$X_i = \begin{pmatrix} X_{i1} \\ \dots \\ X_{ip} \end{pmatrix}$, then the predicted value would be in the form of $Y = \begin{pmatrix} Y_{i1} \\ \dots \\ Y_{ip} \end{pmatrix}$.

$$Y \approx f(X) + \varepsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

(James et al., 2021)

estimate $\beta_0, \beta_1, \dots, \beta_n$ coefficients to model the data

Generalized Additive Model

- Instead of estimating coefficients and assuming the model is linear, GAM estimates the function f(X) itself.

$$E(Y \mid X_1, X_2, \ldots, X_n) = \alpha + f_1(X_1) + f_2(X_2) + \cdots + f_n(X_n)$$

In this case, f_j's are unspecified smooth functions for each independent variable. If using link function g:

$$g(\mu(X)) = \alpha + f_1(X_1) + f_2(X_2) + \cdots + f_n(X_n)$$ (Hastie et al., 2017).

Backfitting Algorithm for GAM

- To find fj's, backfitting algorithm was used by first estimating alpha and assuming fj hat was 0, then:

1. Initialize: $\hat{\alpha} = \frac{1}{N}\sum_1^N y_i,\ \hat{f}_j \equiv 0, \forall i, j.$
2. Cycle: $j = 1, 2, \ldots, p, \ldots, 1, 2, \ldots, p, \ldots,$

$$\hat{f}_j \leftarrow \mathcal{S}_j\left[\{y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ik})\}_1^N\right],$$

$$\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{N}\sum_{i=1}^N \hat{f}_j(x_{ij}).$$

until the functions $\hat{f}_j$ change less than a prespecified threshold.

Smooth function at each iteration

Find new fj hat by subtracting the old fj hat with the mean centering.

Results

MLR

```
Call:
lm(formula = value_eur ~ ., data = data_selected_regression)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.23857 -0.40593 -0.03761  0.37217  2.92487 

Coefficients:
                            Estimate Std. Error t value
(Intercept)                6.9789752  0.0579596 120.411
age                       -0.0888630  0.0013002 -68.348
league_level2              0.0225301  0.0141217   1.595
league_level3             -0.2839382  0.0199479 -14.234
league_level4             -0.4857380  0.0269590 -18.018
league_level5             -0.6103990  0.1238485  -4.929
international_reputation2  0.8119474  0.0220319  36.853
international_reputation3  1.3283648  0.0399937  33.214
international_reputation4  1.8236804  0.0852494  21.392
international_reputation5  1.9804704  0.2480731   7.983
physic                     0.0473845  0.0011149  42.500
attacking_avg              0.0673678  0.0013929  48.364
contract_remaining_yr1     0.1393991  0.0162961   8.554
contract_remaining_yr2     0.2031533  0.0175982  11.544
contract_remaining_yr3     0.3438615  0.0189072  18.187
contract_remaining_yr4     0.3397093  0.0227383  14.940
contract_remaining_yr5     0.4566378  0.0340772  13.400
contract_remaining_yr6     0.7620085  0.1632213   4.669
contract_remaining_yr7     0.6451409  0.6062956   1.064
contract_remaining_yr10    1.2705769  0.6059179   2.097
skill_avg                  0.0108885  0.0013050   8.344
movement_avg               0.0301776  0.0007273  41.491
power_avg                 -0.0187991  0.0014860 -12.651
mentality_avg              0.0040178  0.0015433   2.603
defending_avg              0.0135565  0.0004492  30.178
```

```
                          Pr(>|t|)    
(Intercept)                < 2e-16 ***
age                        < 2e-16 ***
league_level2              0.11064    
league_level3              < 2e-16 ***
league_level4              < 2e-16 ***
league_level5             8.36e-07 ***
international_reputation2  < 2e-16 ***
international_reputation3  < 2e-16 ***
international_reputation4  < 2e-16 ***
international_reputation5 1.51e-15 ***
physic                     < 2e-16 ***
attacking_avg              < 2e-16 ***
contract_remaining_yr1     < 2e-16 ***
contract_remaining_yr2     < 2e-16 ***
contract_remaining_yr3     < 2e-16 ***
contract_remaining_yr4     < 2e-16 ***
contract_remaining_yr5     < 2e-16 ***
contract_remaining_yr6    3.06e-06 ***
contract_remaining_yr7     0.28731    
contract_remaining_yr10    0.03601 *  
skill_avg                  < 2e-16 ***
movement_avg               < 2e-16 ***
power_avg                  < 2e-16 ***
mentality_avg              0.00924 ** 
defending_avg              < 2e-16 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6056 on 17016 degrees of freedom
Multiple R-squared:  0.745,	Adjusted R-squared:  0.7446 
F-statistic:  2071 on 24 and 17016 DF,  p-value: < 2.2e-16
```

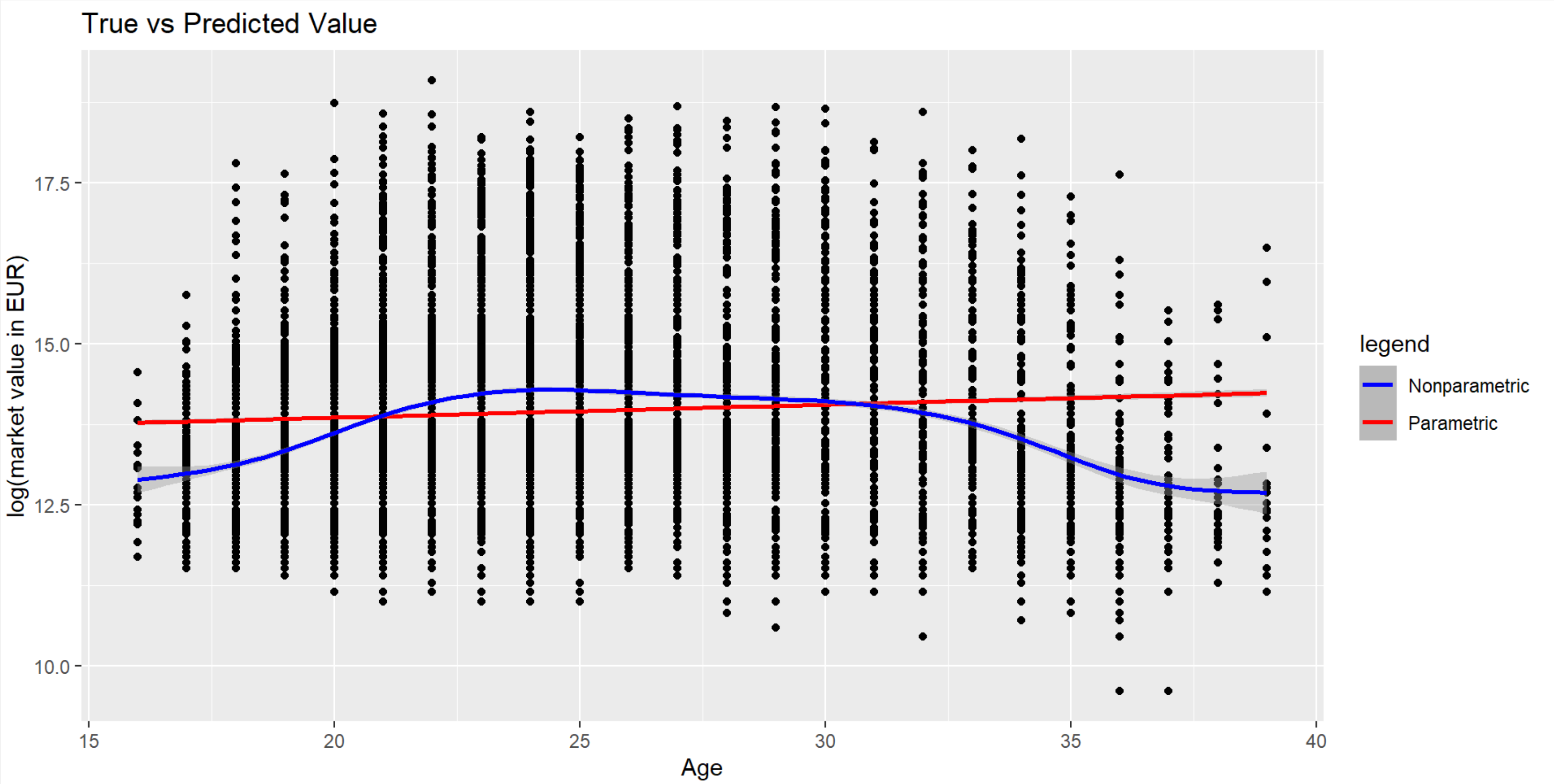
Results

GAM, in both models categorical variables were not significant, therefore they were not used for analyses

```
Family: gaussian
Link function: identity

Formula:
value_eur ~ (s(age) + s(attacking_avg) + league_level + international_reputation +
    s(physic) + contract_remaining_yr + s(skill_avg) + s(movement_avg) +
    s(power_avg) + s(mentality_avg) + s(defending_avg))

Parametric coefficients:
                          Estimate Std. Error  t value
(Intercept)               13.89035    0.01020 1361.150
league_level2              0.04341    0.01007    4.312
league_level3             -0.13076    0.01425   -9.173
league_level4             -0.31699    0.01928  -16.442
league_level5             -0.38826    0.08775   -4.425
international_reputation2  0.24546    0.01780   13.787
international_reputation3  0.34709    0.03450   10.060
international_reputation4  0.53672    0.07461    7.194
international_reputation5  0.33318    0.23006    1.448
contract_remaining_yr1     0.03345    0.01161    2.880
contract_remaining_yr2     0.06573    0.01257    5.231
contract_remaining_yr3     0.10097    0.01356    7.448
contract_remaining_yr4     0.15432    0.01623    9.509
contract_remaining_yr5     0.21511    0.02424    8.874
contract_remaining_yr6     0.32231    0.11589    2.781
contract_remaining_yr7     0.49588    0.43250    1.147
contract_remaining_yr10    0.74250    0.42896    1.731
```

```
                          Pr(>|t|)
(Intercept)                < 2e-16 ***
league_level2             1.63e-05 ***
league_level3              < 2e-16 ***
league_level4              < 2e-16 ***
league_level5             9.72e-06 ***
international_reputation2  < 2e-16 ***
international_reputation3  < 2e-16 ***
international_reputation4 6.58e-13 ***
international_reputation5  0.14756
contract_remaining_yr1     0.00398 **
contract_remaining_yr2    1.71e-07 ***
contract_remaining_yr3    9.95e-14 ***
contract_remaining_yr4     < 2e-16 ***
contract_remaining_yr5     < 2e-16 ***
contract_remaining_yr6     0.00542 **
contract_remaining_yr7     0.25158
contract_remaining_yr10    0.08349 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                    edf Ref.df       F  p-value
s(age)            8.916  8.997 1465.46  <2e-16 ***
s(attacking_avg)  8.294  8.821  465.53  <2e-16 ***
s(physic)         5.547  6.586  205.53  <2e-16 ***
s(skill_avg)      8.513  8.903  202.41  <2e-16 ***
s(movement_avg)   8.311  8.833  426.06  <2e-16 ***
s(power_avg)      7.554  8.370   41.91  <2e-16 ***
s(mentality_avg)  7.984  8.673  219.62  <2e-16 ***
s(defending_avg)  8.547  8.929 1174.09  <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.872   Deviance explained = 87.3%
GCV = 0.18427  Scale est. = 0.1834    n = 17041
```

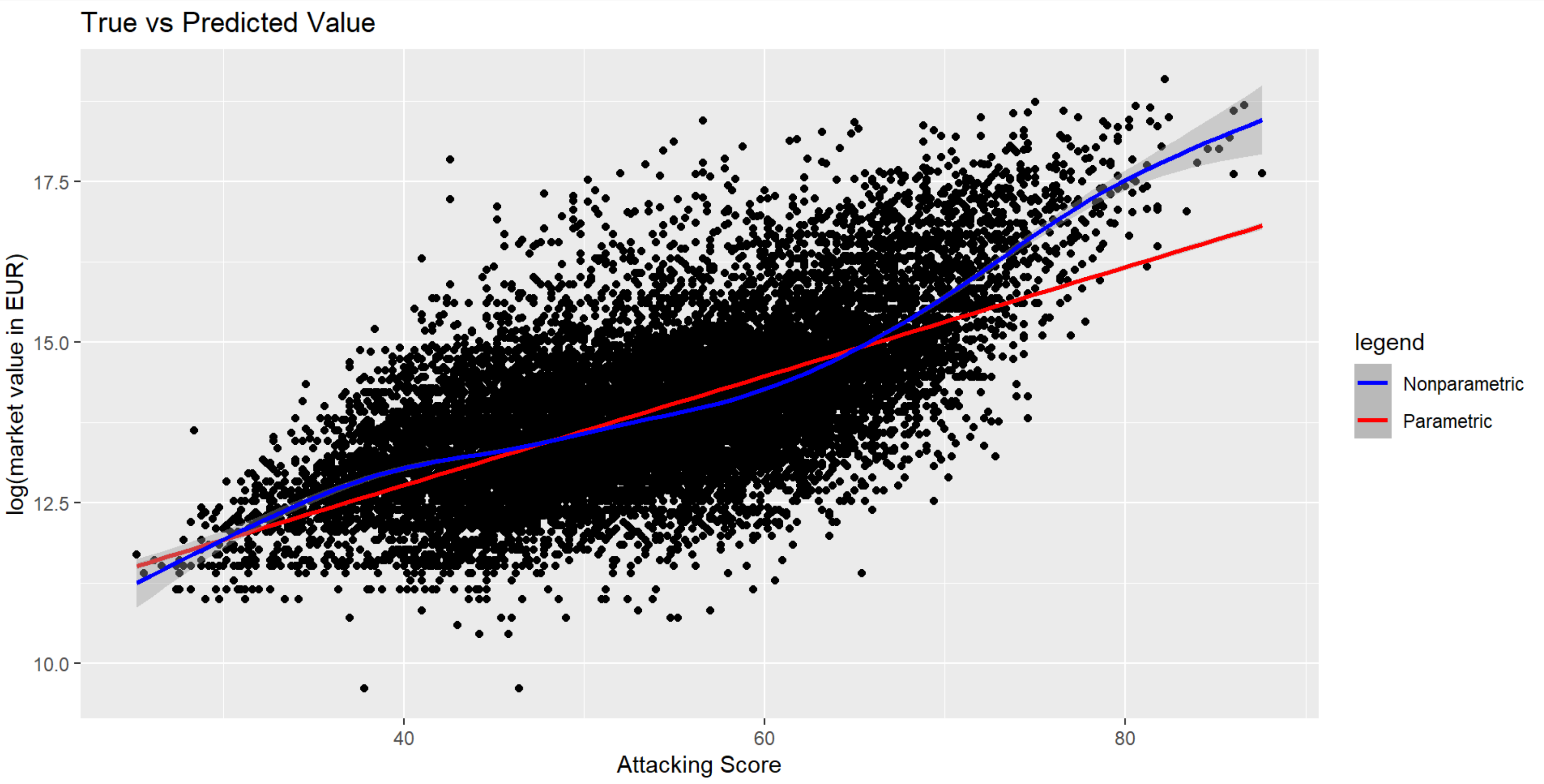
Results (Value vs age)



Interesting trend from GAM not found in MLR:

Young players usually retain market values really well and their values decline as they get older.

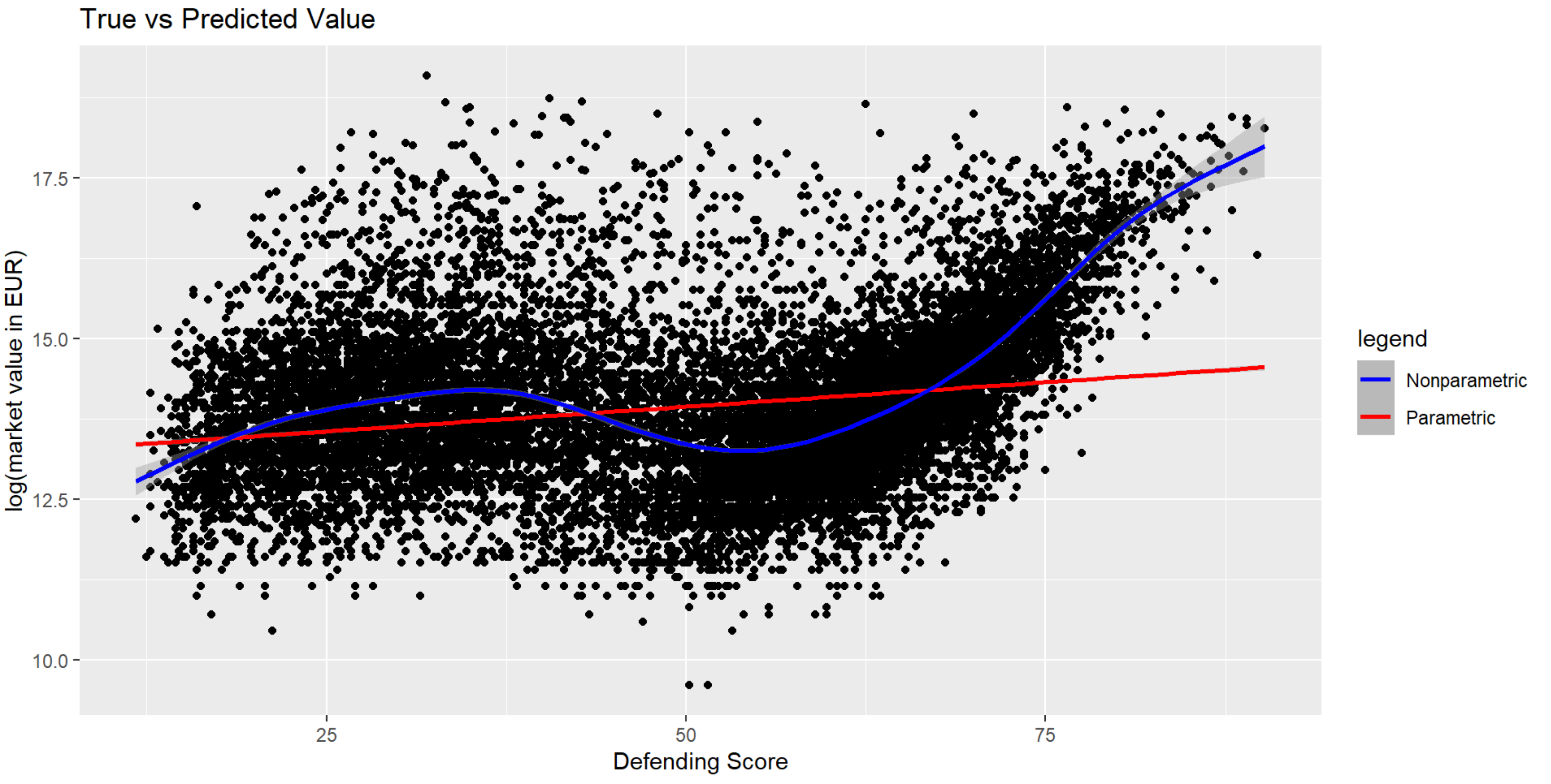
Results (Value vs Attacking)



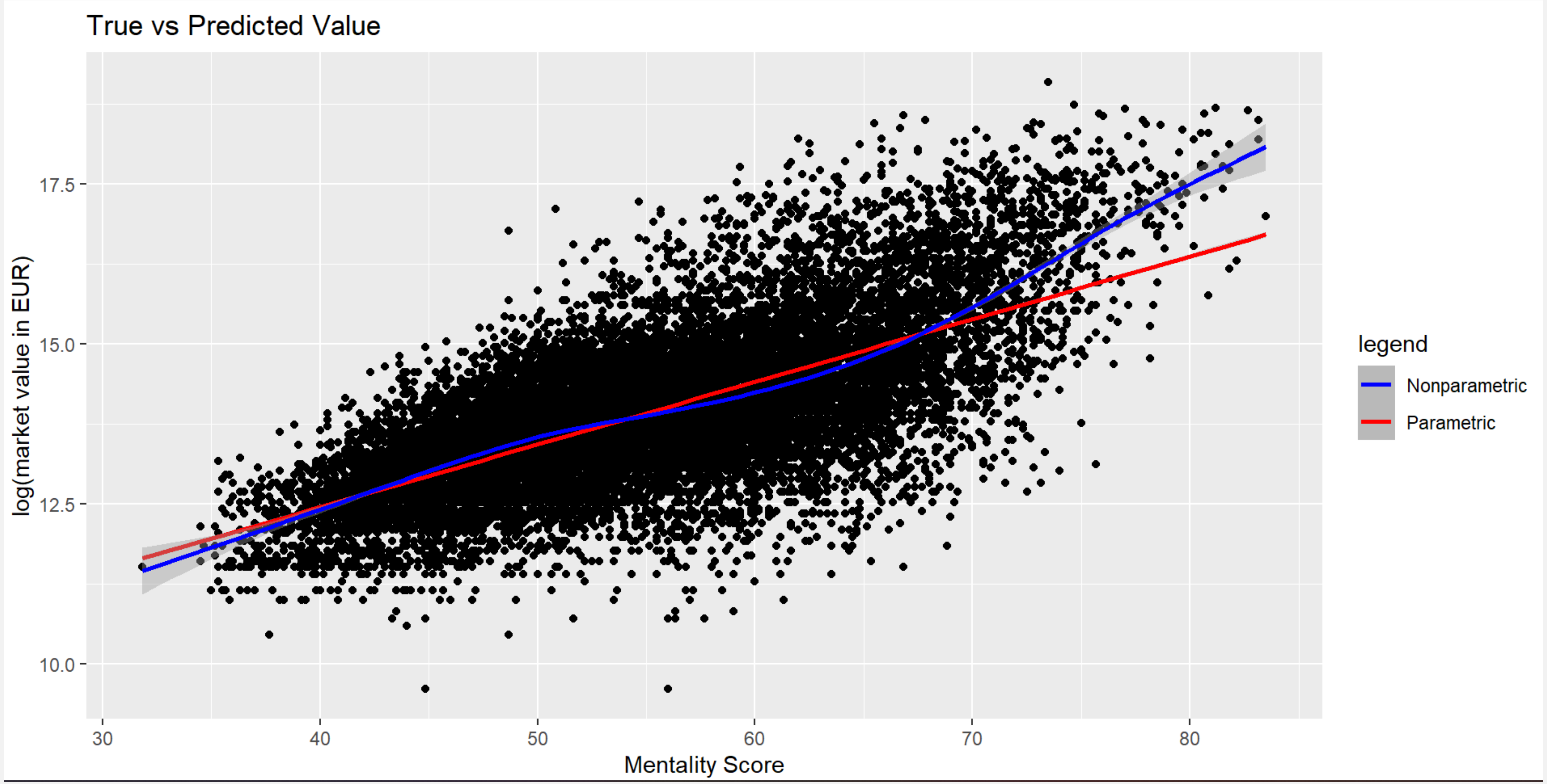
Interesting trend from GAM not found in MLR:

The market value increased at a greater rate as the attacking ability went from 60 upwards

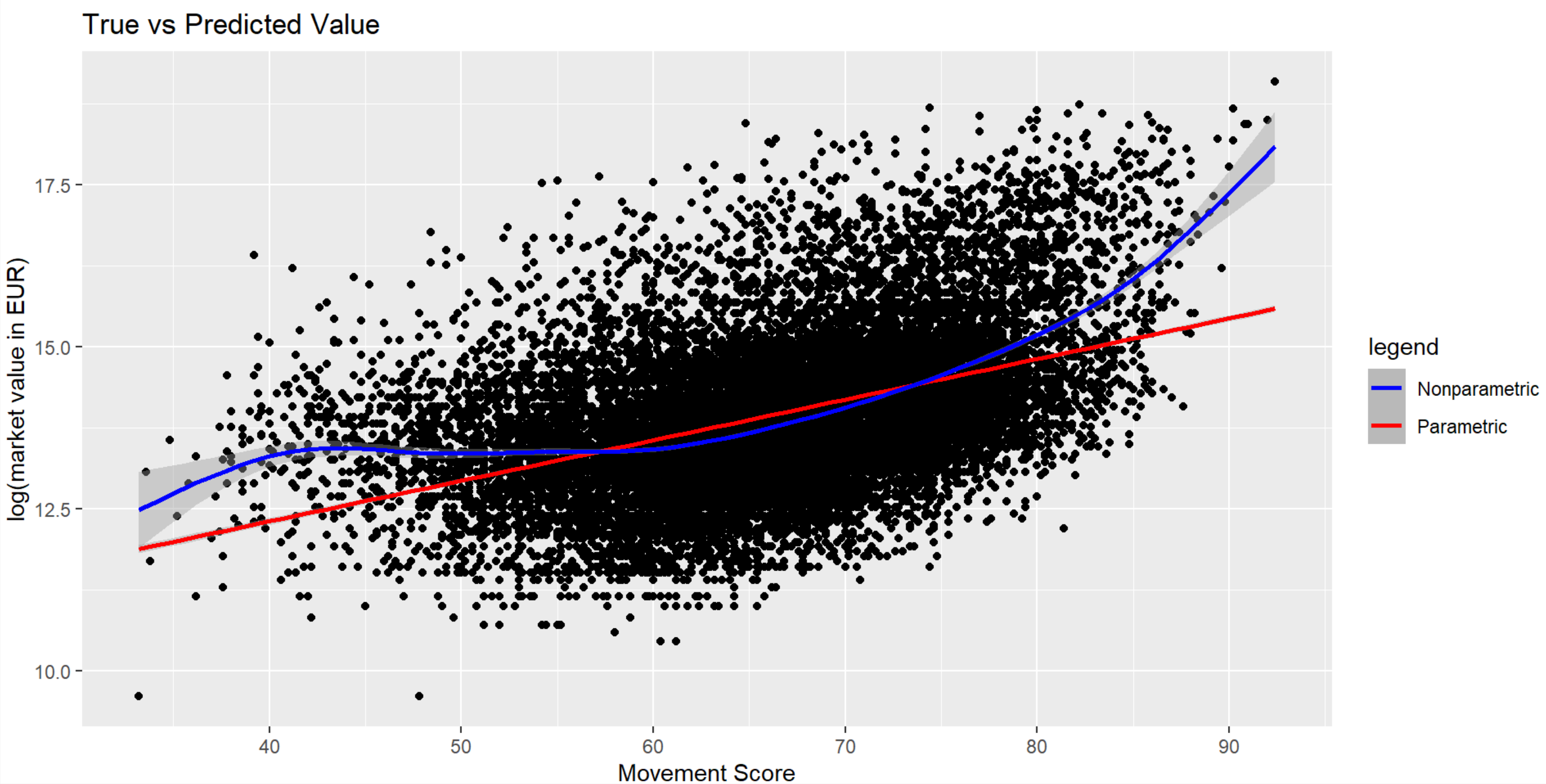
Results (Value vs Defending)



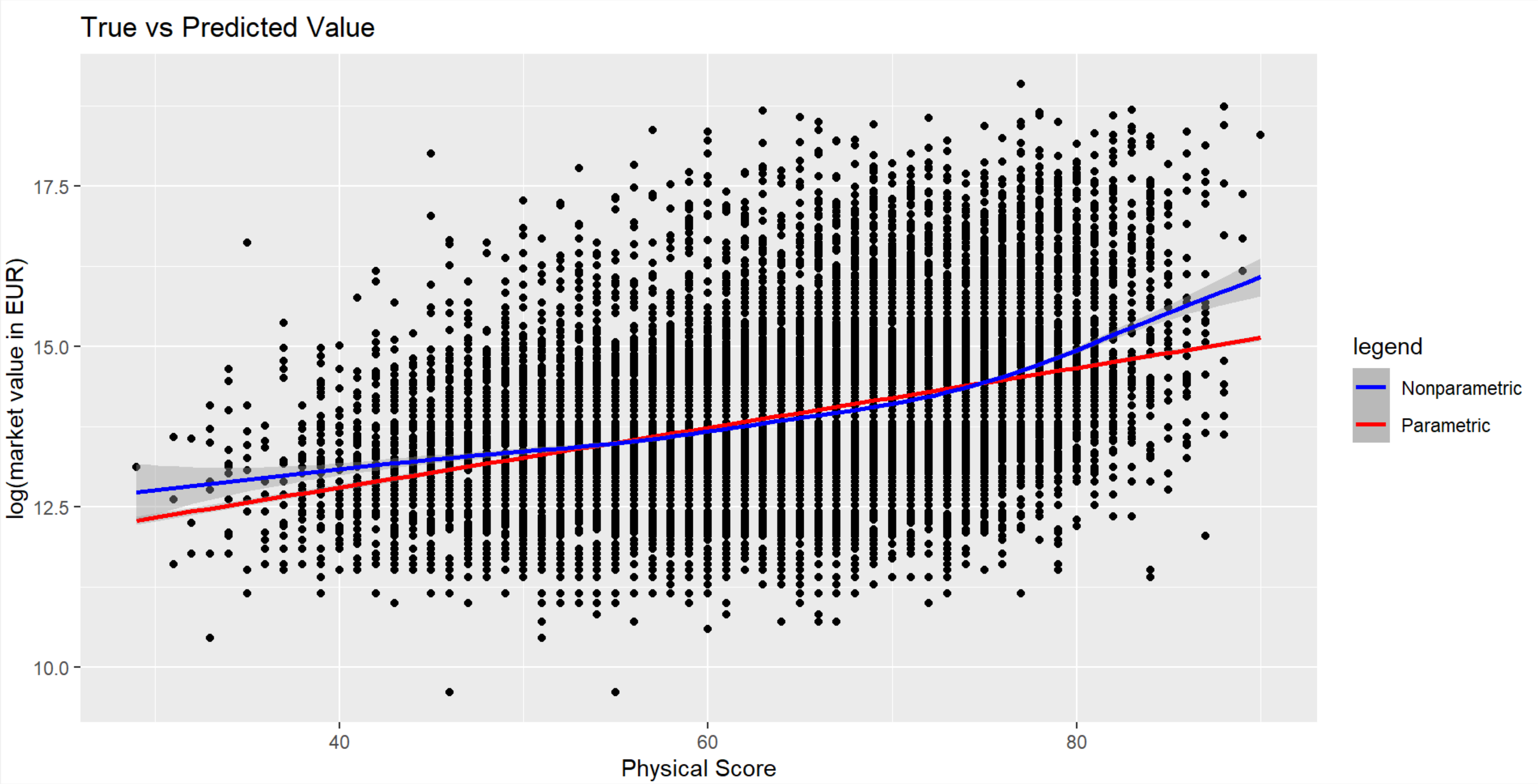
Results (Value vs Mentality)



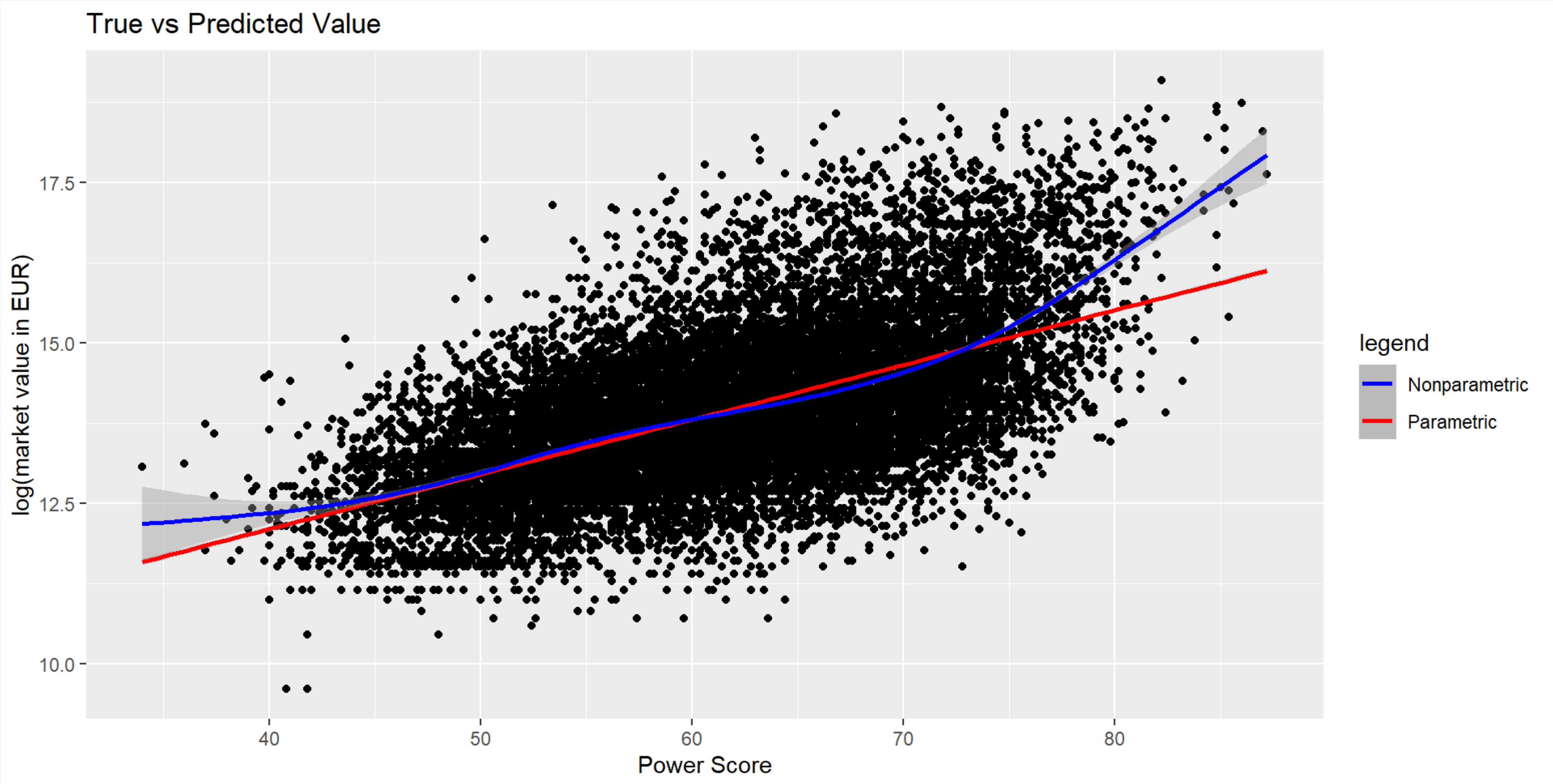
Results (Value vs Movement)



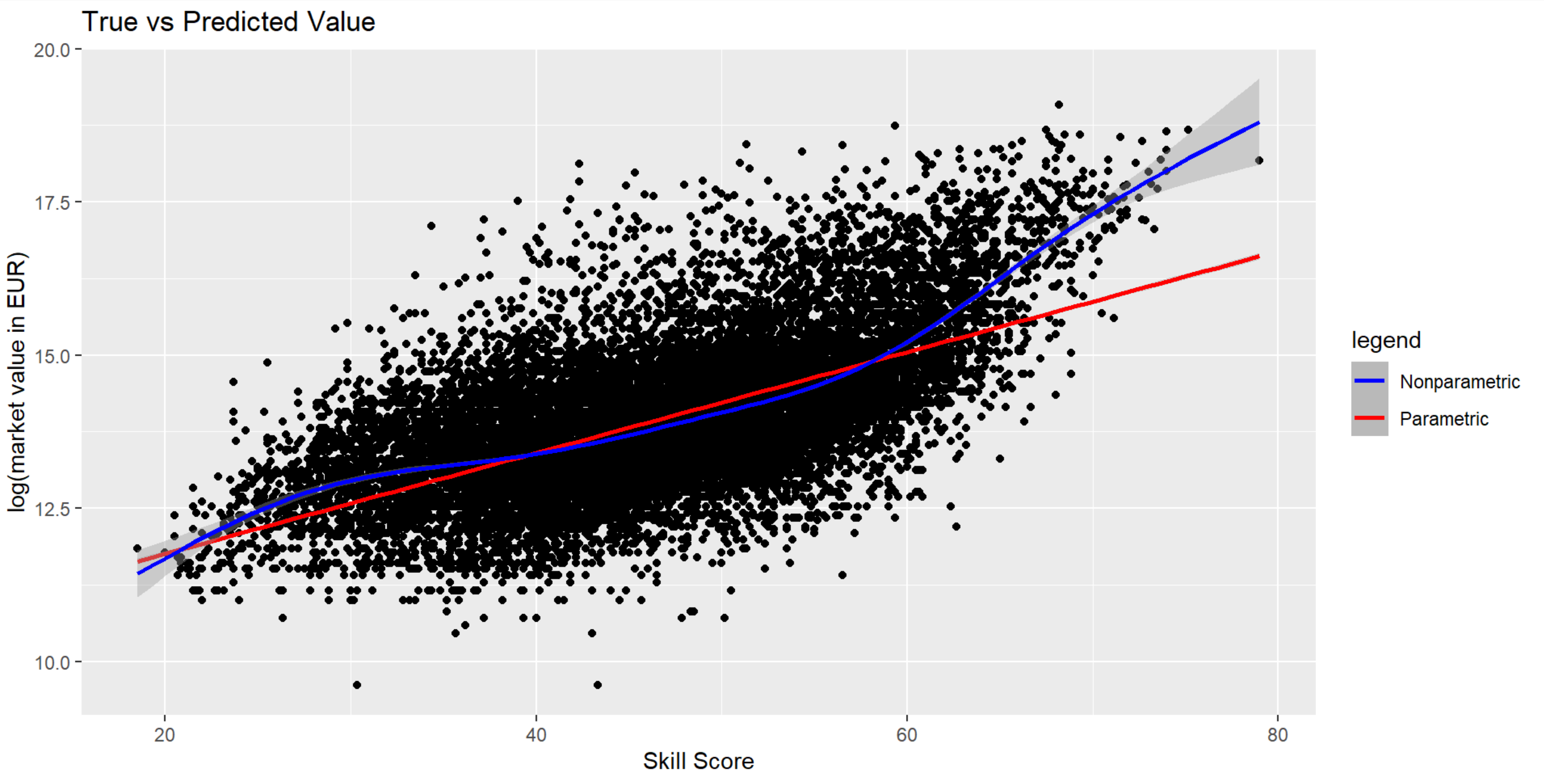
Results (Value vs Physical)



Results (Value vs Power)



Results (Value vs Skill)



Model Selection

Formulas:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

$$AIC = \frac{1}{n}(RSS + 2d\widehat{\sigma^2})$$

$$BIC = \frac{1}{n}\left(RSS + \log(n)d\widehat{\sigma^2}\right)$$

Where RSS is residual sum of square, n is the number of observation, d is the total number of parameters, and $\widehat{\sigma^2}$ is the estimated variance.

	MLR	GAM
AIC	31292.04	19539.74 ↓
BIC	31493.37	20172.11 ↓
MSE	0.366	0.183 ↓

Conclusion

- GAM successfully explained trends that MLR failed to do.
- GAM decreased the MSE, AIC, and BIC by about 50%, 37%, and 36% respectively compared to the MLR model
- GAM performed superior to MLR

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*.

THANK YOU, ANY QUESTION?

Christopher Salim

Email: christopher.salim@brocku.ca

LinkedIn: https://www.linkedin.com/in/christophersalim/