JBP040 Introduction to Machine Learning

# Predicting association football player value using open source Data

R. Haverkort, R. Van Der Heijden, T. Jongenelen, E. Mans and J. Sanderink

December 11, 2019

Jheronimus Academy of Data Science, ‘s-Hertogenbosch

## Introduction

Sports are everywhere in our modern society. Parents encourage their children to join a sports club at a young age and often try to work out themselves as well. Not only is it healthy for our physical selves, participating in a (team) sport can also be seen as a manner of maintaining one's social life. Alongside its individual physical and social benefits, professional sports have an increasingly large commercial side as well. Fans buy merchandise, watch matches that are transmitted after negotiations over broadcasting rights, and awe at the large sums of prize money and player transfer values.

Association football is currently the most popular sport in the world (Sawe, 2018). Billions of fans support their favorite clubs, and its best players reach worldwide-celebrity status. These players can transfer from one club to another two times per season. These so-called 'transfer periods' take place in summer - the 'main' transfer period - and winter. Individual player value can go as high as many millions of euros, and billions change hands overall each year (McVeigh, Christenson & Blight, 2019). Interestingly enough, the method of determining the value of a player is nearly always subjective and untransparent, although it is believed to directly reflect player quality in most of the cases (Foster, 2016; Müller, Simons & Weinmann, 2017). A model that can predict a player's market value could shine light on this black box, and guide clubs in their training of players that are young and promising.

So far, sports statistics are used for a variety of sports, but they have yet to make a solid entrance in others. The most analytics-driven sports are baseball and basketball, two well-known sports that are especially popular in the United States. Researchers and analysts started to apply their knowledge to try and uncover the underlying patterns of sports games as early as the nineties. The main intention behind this was to make monetary profit out of the newly gained insights. These models have also been the stars of several highly popular movies that cover the subject of sports

statistics. Moneyball, based on the novel by Michael Lewis (2003) is an especially well-known example of this.

It is striking that the application of machine learning and statistics is a lot less developed for football. Even with its more than 250 million transferable players of the sport globally, which is a lot more than for basketball and baseball, there have not been many inquiries in the inner workings of the sport and/or the trade (Vroonen et al., 2017). One of the first pieces of literature that considered the use of statistics in football is *The Numbers Game* by Chris Anderson and David Sally. This book appeared only in 2013. The authors wrote their book with the intention to gain more insight into football matches. The objective was to constitute a strategy for how to win a game (Anderson & Sally, 2013).

An explanation for why there have been fewer successful investigations into football might be the unpredictability of the sport. Ironically, precisely this unpredictability is one of the factors that is supposed to contribute to the enormous popularity of football (Creditor, 2014). Most investigations into basketball and baseball focus on on-ball statistics. This does not work for football, because the quality of a player's game is mainly determined by off-ball actions (Bornn, Cervone & Fernandez, Significance Magazine 2018). This phenomenon was neatly summarized in a quote attributed to (no formal source exists) football all-star Johan Cruyff: "When you play a match, it is statistically proven that players actually have the ball three minutes on average … So, the most important thing is: what do you do during those 87 minutes when you do not have the ball? That is what determines whether you're a good player or not."

Traditionally, player projections have been evaluated by human scouts, who are subjective and may suffer from biases (Vroonen et al., 2017). Over the past years, scientific inquiries have tried to formalize the process of scouting by using the power of statistical modeling and machine learning. This report aims to create a model that can accurately predict a player's market value based on their current qualities. Because there is a lack of large-scale football datasets (Cotta, de Melo, Benevenuto & Loureiro, 2016), we will use the database from EA sports' FIFA 19 video game. According to the

chief of the team responsible for data maintenance to the FIFA franchise, this dataset is updated twice a week with the help of 25 producers, 400 contributors, and 8.000 coaches (WIRED Brand Lab, 2019). Crowd evaluations for football player performance has proven to be quite accurate in the past (Herm, Callsen-Bracker & Kreis, 2014), and FIFA video game data has already been used in multiple scientific studies (Cotta et al., 2016; Dey, 2017; Vroonen et al., 2017). Similar to these studies, our study is restricted to field players, discarding the data about goalkeepers. Our research question therefore is: How accurately can we predict a football player's value using data from the FIFA video game franchise?

## Methods

Several well-known machine learning approaches were used in order to get an answer to our research question. This section will describe each of them and give reasons as to why we included it in our study. The models are divided into categories, based on the class of machine learning approach they belong to. First, regression models will be discussed. Second, decision tree strategies are covered. Third, algorithms that do not necessarily belong to a certain family will be discussed in the 'other models' section.

Next to the development of different models, a second analysis will be performed on the veracity of the dataset itself. Even though it might have been used by other studies in the past, and it is the best data there currently is available on association football, does not mean the 'value' in the FIFA video game dataset is close to the sums being paid in the transfer seasons. We will also compare the player value reported in the FIFA video game data with all actual transfer sums of the 2019 season. Exception criteria are: players without transfer sums, players that have been loaned to another team, and players whose transfer sum is not publicly known.

## Regression models

### OLS regression

Not only is it renowned for being one of the first models statistics students learn, Ordinary Least Squares (OLS) regression is one of the most intuitive models of the regression family (Graybill, 1961; Hutcheson, 1999). With its low complexity, OLS regression is often used as the first model to run on a dataset, to get familiar with its structure and to set an early benchmark. In this research project a comparable approach is used; OLS regression was run both before the development of other models, and during intermediate iterations to receive a feel of the effect a change made. Also, near the end of the study, its performance was compared with the performance of other models.

Several methods were applied for feature selection. Automatic selection via stepwise regression was used, while considerations resulting from manual inspection - with a focus on p-value and size of coefficient - were also taken into account. The best results were obtained by using a combination of automatic and manual selection, iteratively building the model until further adjustments yielded negligible improvements.

### Multilevel regression

Data is often encountered in some kind of grouped or hierarchical structure - e.g. municipal studies that investigate villages within a county, or educational studies that consider pupils within schools (Hox, 1998). Our FIFA dataset is no different. It is very likely that *position* is a grouping variable. *Position* is categorical with three values: the first contains all forward positions, the second value contains the midfielders and the third consists of all positions in the defensive part of the field. To be able to investigate whether the prediction of player values is different for defensive positions in comparison to center and forward positions, a multilevel regression model was applied to the dataset.

Before creating the actual model, all continuous features have to be scaled, while the categorical features can stay untouched. Also, it is essential to make a sensible selection of the predictor features. By means of trial and error, one can select and deselect which features to include in

the model, with the objective to include only those features that have a meaningful impact on the dependent variable. These decisions will be made by taking the p-values in account, where $p < 0.05$ is chosen to be the threshold of significance. Using Python this model is run iteratively until the best fitting model has been reached.

**Decision tree models**

Decision tree learning has become one of the most used machine learning techniques over recent years (Brid, 2018). A decision tree tries to estimate the dependent variable using a flowchart-like structure. The model learns by comparing the output to the actual data by making estimations on how 'wrong' the model is (the difference between predicted data and the target variable) (Drakos, 2019). As a measure of how well our model predicts the player value, we use the mean squared error (MSE).

**Random forest**

A random forest consists of multiple decision trees, all grown to a predefined maximal depth and constructed with a unique bootstrapped sample of the data. In addition to this, random forests construct the decision trees slightly different than in 'regular' decision tree learning: in regular trees, each node is split using the best split among all features, whereas in a random forest, each node is split using the best among a subset of predictors randomly chosen at that node (Wiener, 2003). The eventual prediction is made by aggregating the predictions of all the trees (Svetnik et al., 2003). The FIFA 2019 dataset consists of both numeric and categorical features, the random forest model only uses the numeric features in order to predict the - also numeric - player value. In order to secure decent model performance, but prevent overfitting, the maximum depth the trees can grow to is set at five. The quality of a split was measured with the mean squared error.

## XGBoost

Extreme Gradient Boosting (XGB) is an efficient and scalable implementation of gradient boosting framework that was proposed by Friedman (2001). Each new tree is fit on a modified version of the original dataset, so new trees are built upon older versions of the tree. Gradient boosting identifies the shortcomings of the old trees while applying gradients to these shortcomings. As a measure indicating how good a model's coefficients are at fitting the data, the loss function is used (Singh, 2018). Both XGBoost follows the same principle as 'regular' gradient boosting, but there are minor differences, XGBoost uses a more regularized model to control over-fitting, and this gives better performance. The name XGBoost actually refers to the engineering goal to push the limit of computations resources for boosted tree algorithms, which is the reason why it sees much use (Chen, 2015).

A grid search was performed to find the best performing set of parameters. Several values for maximum depth, learning rate and number of trees were used to train models, which were then evaluated on the test set on percentage of correct predictions within a margin of 10% around the actual player value in FIFA.

## Other models

### Support Vector Machine

A Support Vector Machine (SVM) can be used when a categorical target value can be split by an object such as a line or a plane, depending on the dimension of the problem. The model will attempt to find a way to separate different target clusters based on their predictors, and it can be done in multiple dimensions, making it applicable to many problems. However the target variable for an SVM has to be categorical, which is not the case in this study. In order to be able to build an SVM, the player value was transformed into a categorical variable by sorting the data on player value - in ascending order - and then creating ten groups with an equal amount of players (n/10) in them. By splitting the data in this way, class imbalance problems become less severe.

The most straightforward way to run an SVM in Python is installing and using the 'sklearn' package. The data will be divided using a 20/80 split in a test and training set before running the SVM so that the performance of the algorithm can be checked. There are many possible settings when running an SVM, but the parameter of most interest and main experimentation was the choice of kernel.

**CART**

The Classification And Regression Tree (CART) method is a tree-building technique which is unlike traditional data analysis methods. It explores the structure of a set of data and develops easy decision rules for a continuous variable. The regression tree does not create classes of dependent variables (as in classification). The output of a regression tree algorithm is a value to each of the new observations for the dependent variable (Lewis, Ph, & Street, 2000). In our case this dependent variable is the value of a player.

CART methodology consists of three parts: construction of the maximum tree, choice of the right tree size, and classification of new data using constructed tree (Timofeev, 1995). The maximum tree is set up by splitting, by using the *squared residuals minimization algorithm*.

$$i(t) = 1 - \sum_{k=1}^{K} p^2(k|t) \qquad (1)$$

The choice of the right tree size is done by trial and error, it was chosen to be set at a maximum depth of five. This was assumed because the maximum tree may turn out being too complex, resulting in overfitting, therefore the nodes or subtrees of lesser significance are to be cut off. Cross-validation is a good way to do this, this procedure is based on optimal proportion between the complexity of the tree and the misclassification error. The primary task is to find the optimal proportion between the tree complexity and misclassification error (Timofeev, 1995).

CART has some advantages over other classification methods, one of which is that CART is non-parametric. This means that there are no assumptions made regarding the underlying distribution of values of the predictor variables (Lewis et al., 2000).

## Results

### OLS regression

With an adjusted $R^2$ of 0.795 with 39 degrees of freedom, the final model seems to fit the data decently. It has a mean squared error of 8.00e12, and the regression results of all included predictor variables can be found in Appendix A. Even though nearly four-fifths of variance is explained by this model, the resulting performance in predictive accuracy is rather low. Important to note is that the regression coefficients, are displayed as value difference in thousands in order to increase the legibility. This performance, expressed in a predicted value that's within 10% of the actual value, was correct in 6.81% of observations of the test set (with an 80/20 train test split).

Curiously, several predictors have a negative coefficient, where domain knowledge would suggest a positive coefficient. Examples include short passing, shot power, penalty taking ability and ability to make standing tackles. Features where higher scores would entail better players, yet the model claims that increasing these features would result in a decrease in player value. This could be explained by the presence of multicollinearity in the dataset. Unfortunately, the structure of the dataset does not allow the complete removal of multicollinearity.

Looking at how the predicted values differ from the actual values, Figure 1 indicates how the predicted values differ from the actual values, and this gives some interesting results. A suggestion is raised that the residuals are not randomly distributed. Plotting residuals versus the predicted player values does indeed confirm this suspicion (see Figure 2). Further investigation, shown in Figure 3, where the errors should closely match the line, rejects normality of the target variable.

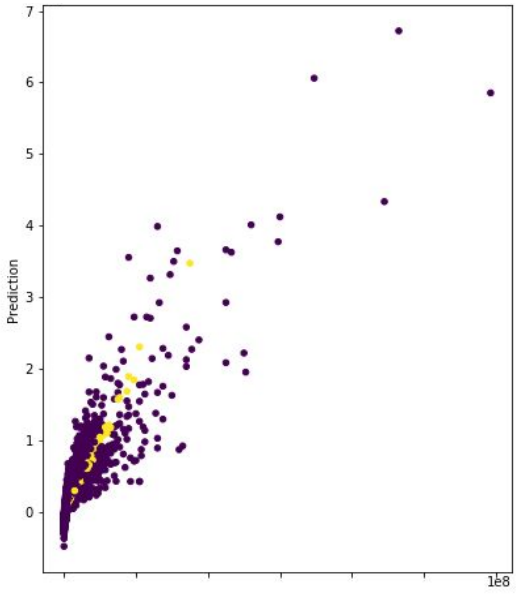


**Figure 1:** Predicted values vs actual values. Purple dots are predicted values within the 10% margin of the actual value. Only 6.81% of predictions was correct in the best OLS model, using 39 predictors. Yellow: predictions that fall within the 10% margin. Purple: Predictions outside of the margin

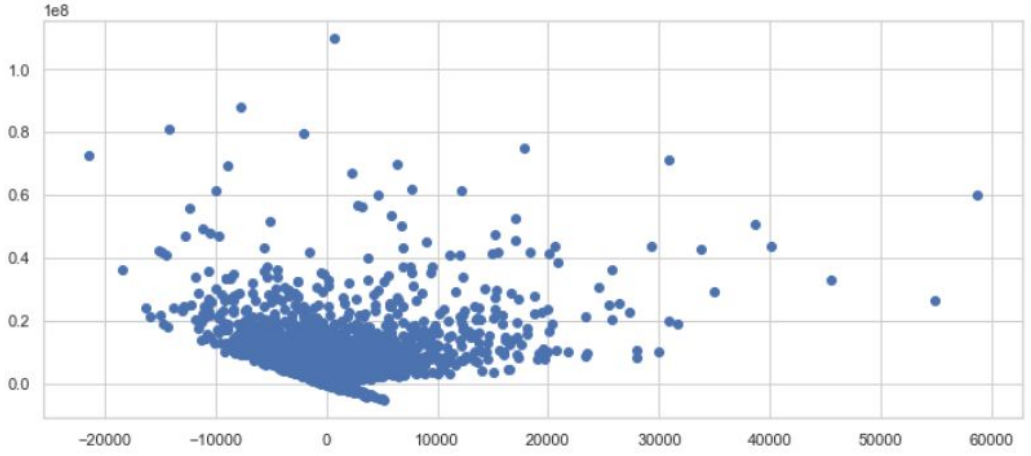


**Figure 2:** Residual plot.

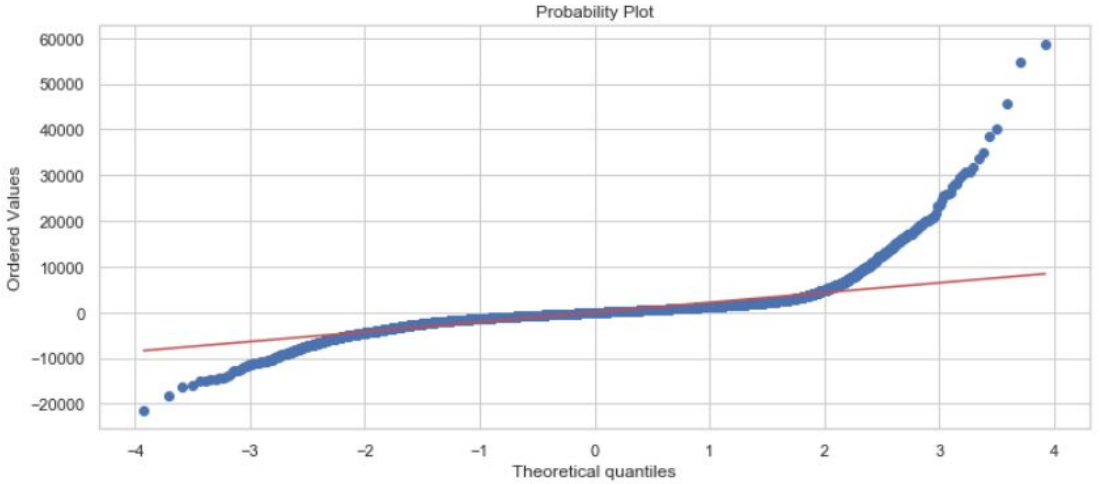


**Figure 3:** Prediction error

**Multilevel regression**

The fitted model resulting from the multilevel regression is shown in Appendix B. For each feature a coefficient has been calculated, from which it can be derived how much influence the variable has on the prediction of the target variable. It can be seen that *Group Var*, which represents the three (forward, center, defender) different positions of the players, is very large. This indicates that for each position, a different model can be constructed that predicts a player's value.

As for the predictor variables, there is a group of variables that changes a player's value positively, and another group that has a negative influence on player value. Looking at the positive predictors, it is evident that *Wage* has the biggest effect on *Value*. Logically, this makes sense, since it is obvious that a player who is worth more, is also given a higher salary. Next, *Basic Player, Position Stat* and *Jersey Number* have the strongest influence on the dependent variable, followed by a bigger group of predictors, which consists of all the features that indicate skills on the field (*Trick Skills, Free Kick Accuracy, Weak Foot,* etc). This implies that the more a player is skilled on the field, the higher his value, which seems fair. Within this group, *Reactions* and *Finishing* have the highest coefficients, while *Weak Foot* has the faintest positive effect on *Value*, followed by *Volleys*.

Looking at the negative predictors, it can be seen that *Age* has a fairly strong negative effect on *Value*, which can be regarded as logical. The older the player, the lower his value. The features that are negatively associated with *Value* most strongly are *Striker* and *Wingback*. These features describe a player's proficiency on the positions in the front and in the back of the field, respectively. This relationship is followed by *Agility* and *Penalties*, which have close to equal coefficients. Lastly, *Aggression* exerts a negative influence on *Value*, with a coefficient which is slightly lower than those of the previous two.

After testing the model on the test data, a few performance measures can be calculated such as the adjusted $R^2$ and the Mean Squared Error (MSE). As the model is imbalanced with way more

data points for low value players, we decided to also check the performance of the model separately for players with values below one million and above one million. These can all be found in Table 1.

Table 1

*Model performance measure for players valued under and over 1 million euros*

| Metric | Total dataset | Players Under Million | Players Over Million |
|---|---|---|---|
| Mean Absolute Error (E+03) | 140.16 | 91.994 | 1986.9 |
| Mean Squared Error (E+09) | 7751.0 | 14.150 | 9090.7 |
| Root of Mean Squared Error (E+06) | 2.7841 | 0.11895 | 3.0151 |
| Median of the absolute error (E+03) | 839.00 | 74.496 | 1408.5 |
| R-Squared | 0.78 | 0.79 | 0.81 |
| Adjusted R-Squared | 0.78 | 0.79 | 0.81 |
| Explained variance score | 0.78 | 0.79 | 0.81 |

In Figure 4, it is displayed per player value-group (under and over €1M) how many of its predictions fall in the 10% margin of correctness. The graphs indicate that for player values under €1M, there are many negative predicted values. Compared to the original model that includes all players however, it does perform better, with 30% of the predictions falling within the 10% margin, as opposed to 8% on the full dataset. The same goes for training and testing above €1M, which shows 14% of the predictions falling within the margin. When combining these findings with the errors mentioned in Table 23, it is safe to say that this model works best when it is trained and tested on a dataset containing only players with a value smaller than €1M.

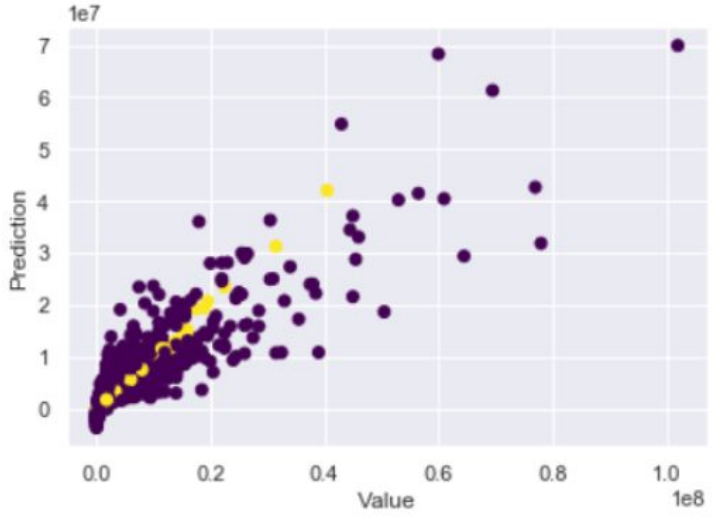


**Figure 4a:** Margin of correctness for multilevel regression, full dataset
Yellow: predictions that fall within the 10% margin. Purple: Predictions outside of the margin

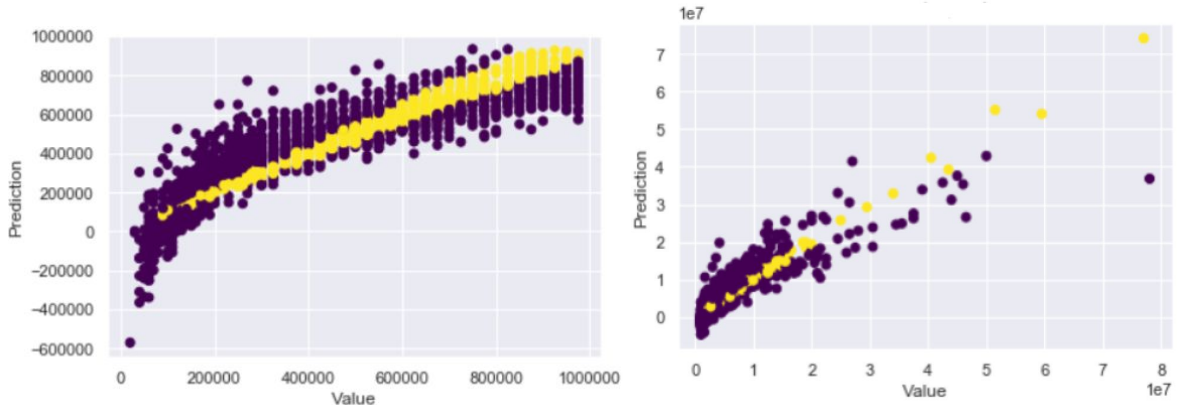


**Figure 4b:** Margin of correctness for multilevel regression, player values < 1M (left) and > 1M (right)
Yellow: predictions that fall within the 10% margin. Purple: Predictions outside of the margin

**Random Forest**

The results of the final Random Forest model are shown in Figures 7a and b. What immediately stands out is that the prediction model for players with a value of over €1M is reasonable; 448 out of 1213 (i.e. 36.9%) of the predicted values fall within a 10% margin. The model for players with a value of under €1M shows different results: there are 411 out of 1912 data points which were predicted within a 10% correctness-margin, which equals a percentage of only 21.5%.

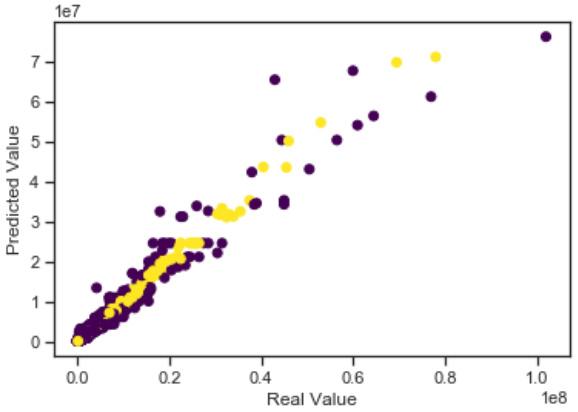


**Figure 7a**: Predictions for the entire dataset
Yellow: predictions that fall within the 10% margin. Purple: Predictions outside of the margin

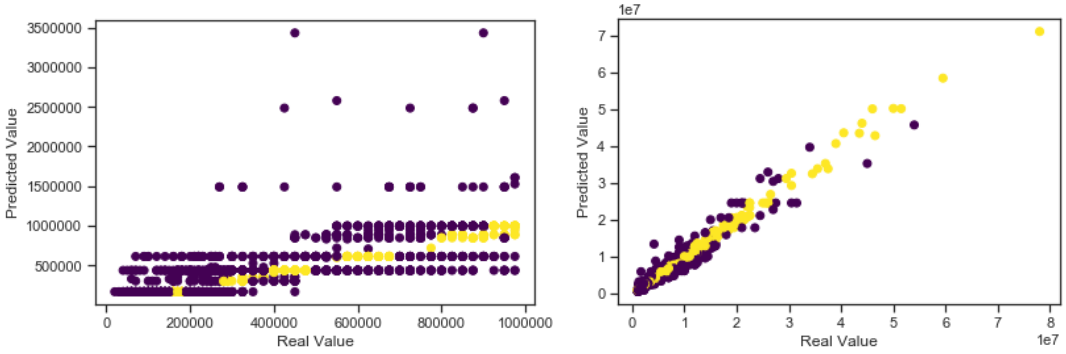


**Figure 7b**: Predictions for player values under 1M (left) and over 1M (right)
Yellow: predictions that fall within the 10% margin. Purple: Predictions outside of the margin

Furthermore, some performance measures regarding the found model are in Table 2.

Table 2

*Metrics for Random Forest model*

| Metric | Total dataset | Players Under million | Players over million |
|---|---|---|---|
| Mean Absolute Error *(in thousands)* | 444 | 145 | 778 |
| Mean Squared Error *(in billions)* | 1366 | 55.3 | 1594 |
| Root of Mean Squared Error *(in thousands)* | 1169 | 235 | 1262 |
| Log of Mean Squared Error | 0.152 | 0.225 | 0.0541 |
| Median of the absolute error *(in thousands)* | 163 | 97 | 447 |
| R-Squared | 0.962 | 0.189 | 0.969 |
| Explained variance score | 0.962 | 0.252 | 0.969 |

**XGBoost**

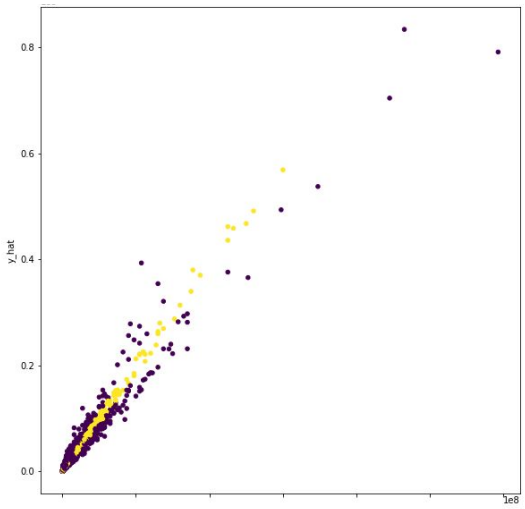


**Figure 5:** Predicted player value versus actual in-game player value.
Yellow: predictions that fall within the 10% margin. Purple: Predictions outside of the margin

A grid search was used to find the best performing set of parameters. Several values for maximum depth, learning rate and number of trees were used to train models, which were then evaluated on the test set on percentage of correct predictions within a margin of 10% around the actual player value. The best results were obtained using a max depth of 10, a learning rate of 0.07 and 250 estimators, which yielded an accuracy for the performance metric mentioned above of 39.48%, with an $R^2$ of 0.9499 and an MSE of 1.70e12 (see Figure 5). The top 10 best performing features for each category - gain or weight as importance metric - can be seen in Figure 6. A comparison of these top 10 best performing parameters can be found in Appendix B.

A striking result is the lack of overlap between the top performing features for the two separate metrics. Though these metrics do follow a different train of thought, it was expected that their results would be more comparable than they turned out. Out of all features, only 3 were present in both top 10 lists. There features are *Wage* (7th and 1st, for gain and weight respectively), *CentralMidfielder* (8th and 5th) and *Striker* (10th and 7).

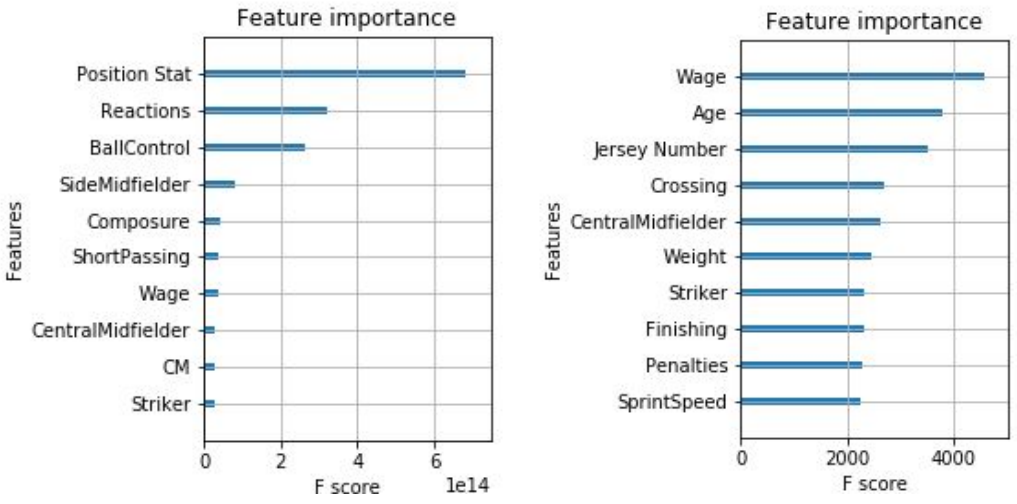


**Figure 6**: Feature importance plot of top 10 features using gain (left) and weight (right) as importance metric.

**Support Vector Machine**

After training the Support Vector Machine on the training data, it was checked to what extend the player values predicted by the model correspond to the values in the test set. The experimentation with different kernels eventually showed that, for this problem, a linear kernel performed best, outperforming polynomial and radial basis function kernels. With this linear kernel, the model reached an accuracy of 64%. In order to check for overfitting, the accuracy was also calculated on the training set, which resulted in a percentage of 65%. As the accuracies of both sets are almost equal, it is safe to say that the model is not overfitted. The more in-depth classification report that can be found in Table 3 below indicates that the model predicts more correctly for players in either the group with the lowest values or the highest values in comparison to the players that fall in between, as the former two groups have higher precision and recall.

Table 3

*Classification report for Support Vector Machine*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.85 | 0.79 | 0.82 | 315 |
| 1.0 | 0.60 | 0.57 | 0.59 | 311 |
| 2.0 | 0.57 | 0.58 | 0.58 | 351 |
| 3.0 | 0.48 | 0.48 | 0.48 | 313 |
| 4.0 | 0.47 | 0.44 | 0.46 | 312 |
| 5.0 | 0.57 | 0.65 | 0.61 | 353 |
| 6.0 | 0.56 | 0.50 | 0.53 | 269 |
| 7.0 | 0.67 | 0.68 | 0.67 | 329 |
| 8.0 | 0.74 | 0.76 | 0.75 | 329 |
| 9.0 | 0.84 | 0.91 | 0.87 | 302 |
| accuracy | | | 0.64 | 3184 |
| macro avg | 0.64 | 0.64 | 0.64 | 3184 |
| weighted avg | 0.64 | 0.64 | 0.64 | 3184 |

Furthermore, the plot in Figure 8 compares the FIFA player values from the dataset with player values that were predicted using the SVM model. The plot shows a positive linear relation with a reasonable correlation, implying that the model's predictions are quite decent.

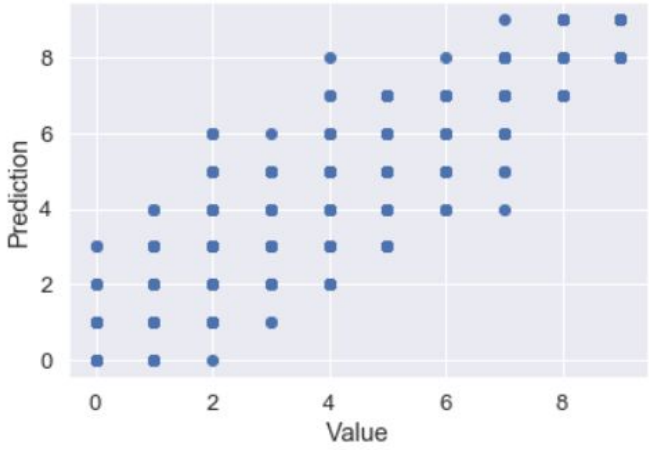


**Figure 8:** Visualization of the comparison between predicted player value and FIFA player value.

**CART**

Figure 9 shows that the predictive power of this model is reasonable. From the 3184 data points in the test set, 815 (25.65%) are predicted within the 10% margin of the real value. Because of the imbalance in our dataset, it is not clearly visible whether the trend still holds in the dense cloud of data points in the bottom left corner. This can be seen in Figure 10, which shows the players with an actual value lower than one million.

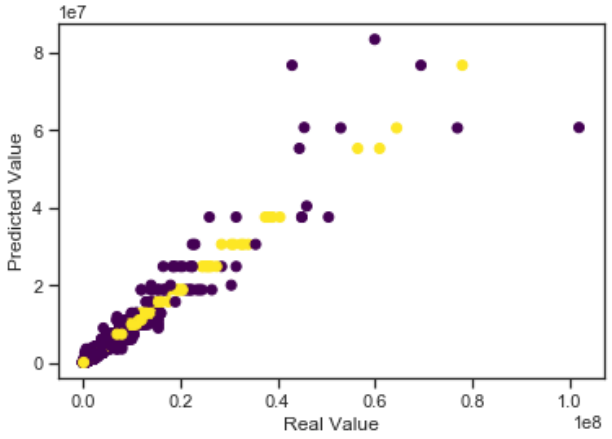


**Figure 9**: Predictions of Cart model.
Yellow: predictions that fall within the 10% margin. Purple: predictions outside of the margin

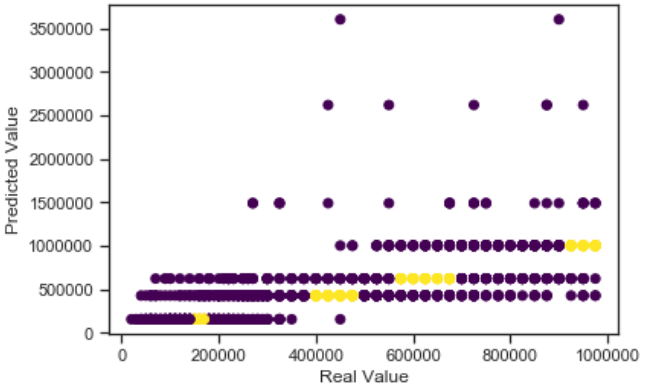


**Figure 10:** Predictions of Cart model for players with value < 1M
Yellow: predictions that fall within the 10% margin. Purple: Predictions outside of the margin

Table 4 shows that there are some significant differences in the model regarding the total dataset and the low valued players. Especially the $R^2$ score stands out. It explains to what extent the variance of the model explains the variance of the predicted variable (value in this case). Since there are so many low valued players, the $r^2$ value in the entire dataset is high. However, the variance around this segment of cheaper players is a fraction of the error of a higher cost segment of players.

Table 4

*Metrics for Cart model*

| Metric | Total dataset | Players Under million | Players Over Million |
|---|---|---|---|
| Mean Absolute Error *(in thousands)* | 502 | 151 | 917 |
| Mean Squared Error *(in billions)* | 2260 | 60.3 | 2290 |
| Root of Mean Squared Error *(in thousands)* | 1500 | 246 | 1520 |
| Log of Mean Squared Error | 0.160 | 0.229 | 0.0643 |
| Median of the absolute error *(in thousands)* | 178 | 103 | 519 |
| R-Squared | 0.937 | 0.115 | 0.955 |
| Explained variance score | 0.937 | 0.177 | 0.955 |

## Discussion

First and foremost, it could be argued that the dataset lacked several potentially impactful features. One could think of several seemingly relevant real-world factors that were not taken into account, such as marketability, mental fortitude, locker room presence, off-pitch behaviour and relation between current and possible new club, all of which could significantly impact transfer costs. Most of these features are quite vague and are very subjective concepts, making it unlikely that they will appear in any future version of the FIFA video game dataset.

With regard to the dataset as it was used, several issues could be pointed out as well. The distribution of association football player value follows an exponential curve, but it was considered to be linear in this study. Hereby, predictions estimated by the models that hold this assumption are likely to be biased, with a potential of increased performance when taking the target variable into account.

Next to this, we encountered multicollinearity with many features in the dataset. Conceptually, this can be readily explained; in general better players are rated better because of individual attributes. After all, a good striker will most likely rank higher in all striker attributes than a mediocre one, making it hard to distinguish exactly what striker quality is over- or undervalued.

Moreover, performance of the model is very dependent on the goal of its expected use. Though generalization is taken as the main goal in this research, creating models on partial data would result in preferable local results. In several models an insight is offered between two data segments (in addition to the overall data): players worth less than one million and players worth more. Significant changes in performance are found, therefore final models can be developed with an increased focus on the partial data that matches the transfer budget of the club that will eventually deploy the model.

Lastly, a large limitation of the model as a whole is its temporal bias; the data we have is very specific to the current age. Transfer values have exploded in the last decade, meaning that (especially the best) players now net a much higher transfer price than they would have several years ago. For reference: in 2009 the world watched in awe as Cristiano Ronaldo was bought by Real Madrid for €94 million, while in 2017 the (still-standing) record transfer of Neymar was paired with €222 million. It is likely that this trend holds for the future, meaning that our proposed models would underestimate players' values for future seasons.. Because the values are exponential, a simple multiplicative corrective factor would not sufficiently correct for this over the entire data set, though in our current approach with player value regarded as linear, this could help substantially to remedy this issue.

For future studies, it could be a reasonable  advice to continue working with our found models, and to compare the models' predictions with real life transfer sums of existing players. When looking at the difference between player values in the FIFA dataset and the transfer values of the same players however, an average error of 1.5 times the actual transfer value is found. This means that there is substantial divergence between the two metrics. A plausible reason for this is that transfer values are more often than not subject to human judgement and human subjectiveness. A big challenge lies

in finding an appropriate method to compare the player value, as defined in the video game, with real-life transfer sums.

The following part of this discussion will consider each model on its own.

**OLS regression**

A better model fit could be attained by focussing on minimizing one of the standard (and widely used) metrics, such as mean squared error, AIC or BIC, but more importantly by transforming the target data into linearity. Using a log-function would achieve this effect. Since results from that transformation and corresponding regression run would not be tangible, and due to man hour constraints, it was not realistic to sufficiently delve into the new realm of possibilities. For reference though: where the best performing model without transformed target variable reached an $R^2$ of 0.785, transforming the target variable gave an $R^2$ of over 0.95 in the first model that was tried, a model non-specialized for the situation. It is very likely that transformation of the dependent variable will lead to better-performing regression models.

**Multilevel regression**

Regarding the mixed model or multilevel regression, there are some issues that need extra attention. In the model as is, feature selection has been conducted mainly by looking at p-values, cutting the features that have p-values over 0.05. In future studies, progress can be made by investigating features more extensively before they are definitively excluded from the model. Also, another aspect that was limited in this research due to the scope, was the presence (or absence) of interaction variables and dummy variables.

**Random Forest**

Due to the nature of the method itself, it is hard to discern which features contribute to the eventual prediction. Also, in order to modify the model, several assumptions have to be made. There

is, however, not one single best fitting solution. The creation of models has conflict leading to compromise. Finally, and more specific to our eventual model, is that the predictive power is higher for the few players with a value above 1.000.000. This is peculiar, because given the imbalance in the dataset, one would expect predictions for the lower-priced players to be better.

**XGBoost**

It is important to note that grid search was done on the manually crafted evaluation metric of highest percentage of the test set predicted within a 10% margin of the actual (FIFA) value. This reduction into a binary classification could make judging a 'correct' or 'incorrect' prediction very interpretable and increased comparability between models, but at the expense of reducing each classification into a binary value. The difference between being of 0.0% and 9.8% of the actual target variable would be non-existent in this method, while the difference between being 9.8% and 10.1% would be 1 vs 0.

This effect could be circumvented by choosing to use mean squared error as primary evaluation (optimization) metric. However, because the mean squared error is very susceptible to points far removed from the prediction, in addition to being very counterintuitive in use for our target audience, the decision was made focus on our primary evaluation metric. Perhaps a modified error metric could be divized, taking into account the error in terms of relative rather than absolute quantities, and this pursuit could yield improved models.

An unforeseen and surprising effect was found when comparing the best performing features over the gain and weight metrics. When taking the top 10, only 3 features were present on both lists, with several high-performing features on one metric scoring below average on the other. Though a bit counter-intuitive, this does not necessarily spell out problems. A possible explanation would be that very informative splits are done first, resulting in just a single tally for the weight metric. On the other hand, lesser informative features, or informative features on a more granular scale, that are used at a higher depth in the tree, might be used several times but have lesser overall impact. It is hypothesized

that this effect is buried beneath these results, though time constraints prevented performing an analysis to the extent of the presence of this effect with different maximums of tree depths.

**Support Vector Machine**

As always, there are a  few discussion points about the use of support vector machines in this context. First of all, in this project, in order to make categories, the data is split in 10 groups of approximately equal size, to avoid class imbalance. Unfortunately, the original dataset is imbalanced, which means that even though the groups are equal sized and the model can predict them reasonably, high valued players will not be predicted accurately because the scope within this cluster is simply too big.

Support vector machines can run based on different kernels to change the dimensions of the input data, during this project there was some experimentation with which kernel would work best. Due to time related limitations, this has not been done as in depth as it could have been done with a longer project traject. For future research, more kernels could be tested to see if they would improve the predictions. The same goes for the additional parameters that come with each kernel. For this project, these parameters were kept at their default values as the focus was on trying multiple models instead of going extremely in depth on one model. The most obvious parameter to experiment with is the C value, which indicates how many mistakes the model is allowed to make.

**CART**

CART uses decision trees, which means it will automatically be more suitable for classification problems than our continuous estimation problem. We encountered this issue with several of the models we deployed during the study. Also, comparable to the random forest model, assumptions had to be made in order to build the model, but there is no clear-cut solution, and the model fits better for players with a value of more than 1.000.000, which remains peculiar. Finally, the

model itself is very sensitive to overfitting. Measures were taken in order to prevent this, but cart definitely struggles more with this compared to the other models.

**References**


Adler, D. (2019, September 30). MLB sees fan growth across the board in 2019. Retrieved November 14, 2019, from MLB website: Adler, D. (2019). MLB sees fan growth across the board in 2019. [online] MLB.com. Available at: https://www.mlb.com/news/mlb-increased-viewership-attendance-in-2019 [Accessed 30 Oct. 2019].

Anderson, C., & Sally, D. (2014). *The numbers game why everything you know about football is wrong*. London Penguin Books.

Anderson, P. M., Ian Stewart Blackshaw, Siekmann, R. C. R., Janwillem Soek, & Hill, D. (2012). *Sports betting : law and policy*. The Hague, The Netherlands: T.M.C. Asser Press.

Bornn, L., Cervone, D., & Fernandez, J. (2018). Soccer analytics: Unravelling the complexity of "the beautiful game." *Significance*, *15*(3), 26–29. https://doi.org/10.1111/j.1740-9713.2018.01146.x

Brid, R. S. (2018, October 26). Decision Trees - A simple way to visualize a decision. Retrieved November 2019, from Medium Greyatom website: Brid, R. S. (2018). Decision Trees. A simple way to visualize a decision. Retrieved December 10, 2019, from https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb

Chen, T. (2015, September 4). What is the difference between the R gbm (gradient boosting machine) and xgboost (extreme gradient boosting)? Retrieved November 2019, from Quora website:

https://www.quora.com/What-is-the-difference-between-the-R-gbm-gradient-boosting-machine-and-xgboost-extreme-gradient-boosting

Cotta, L., vaz de Melo, P. O. S., Benevenuto, F., & Loureiro, A. A. F. (2019). *Using FIFA Soccer video game data for soccer analytics*. Retrieved from https://homepages.dcc.ufmg.br/~fabricio/download/lssa_fifa_CR.pdf

Dey, S. (2017). *Pricing Football Players using Neural Networks*. Retrieved from https://arxiv.org/ftp/arxiv/papers/1711/1711.05865.pdf

Drakos, G. (2019, May 23). Decision Tree Regressor explained in depth. Retrieved December 11, 2019, from GDCoder website: https://gdcoder.com/decision-tree-regressor-explained-in-depth/

Foster, R. (2017, May 17). How football clubs calculate the cost of buying players in the transfer market. Retrieved from The Guardian website: https://www.theguardian.com/football/2016/apr/04/clubs-calculate-cost-transfer-market-leicester-southampton

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/doi.org/10.2307/2699986

Graybill, F. A. (1961). *An introduction to linear statistical models*. New York, N.Y.: Mcgraw-Hill.

Herm, S., Callsen-Bracker, H.-M., & Kreis, H. (2014). When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community. *Sport Management Review*, *17*(4), 484–492. https://doi.org/10.1016/j.smr.2013.12.006

Hox, J. (1998). Multilevel Modeling: When and Why. *Classification, Data Analysis, and Data Highways*. Springer.

Hutcheson, G. D. (1999). Ordinary Least-Squares Regression. In *The Multivariate Social Scientist*. https://doi.org/10.4135/9780857028075.d49

Lewis, R. J. (2000). An introduction to classification and regression tree (CART) analysis. *Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California*.

Liaw, A., & Wiener, M. (2002). *Classification and Regression by Random Forest* (pp. 18–22). Retrieved from R Project website: https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf

Müller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research*, *263*(2), 611–624. https://doi.org/10.1016/j.ejor.2017.05.005

Niall McVeigh, Christenson, M., & Blight, G. (2019, June 4). Summer transfer window recap – every deal from Europe's top five leagues. Retrieved November 2019, from the Guardian website: https://www.theguardian.com/football/ng-interactive/2019/jun/04/football-transfer-window-2019-every-summer-deal-from-europe-top-five-leagues

Pallotta, F. (2019, October 10). The NFL continues its ratings momentum one month into the 2019 season. Retrieved October 30, 2019, from CNN website: https://edition.cnn.com/2019/10/10/media/nfl-ratings-analysis/index.html

Sawe, B. E. (2018, April 5). The Most Popular Sports in the World. Retrieved December 3, 2019, from WorldAtlas website: https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html

Singh, H. (2018, November 3). Understanding Gradient Boosting Machines. Retrieved from Medium website: https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, *43*(6), 1947–1958. https://doi.org/10.1021/ci034160g

Temkin, N. R., Holubkov, R., Machamer, J. E., Winn, H. R., & Dikmen, S. S. (1995). Classification and regression trees (CART) for prediction of function at 1 year following head trauma. *Journal of Neurosurgery*, *82*(5), 764–771. https://doi.org/10.3171/jns.1995.82.5.0764

Vyas, H. (2018, December 21). Record audience watched "best World Cup ever" - FIFA. Retrieved October 30, 2019, from U.S. website: https://www.reuters.com/article/uk-soccer-worldcup-viewers/record-audience-watched-best-world-cup-ever-fifa-idUKKCN1OK19B

Vroonen, R, Decroos, T, Van Haaren, J, & Davis, J. (2017). Predicting the potential of professional soccer players. CEUR Workshop Proceedings, 1971, 1-10.

WIRED Brand Lab. (2018, May 7). Meet the Data Master Behind EA Sports' Popular FIFA Franchise. Retrieved October 30, 2019, from Data Makes Possible website: https://datamakespossible.westerndigital.com/meet-data-master-ea-sports-fifa/

## Appendix A: OLS regression model results

| Feature | Coefficient (E+07) | Standard error (E+05) | t-value | P>\|t\| | [0.025 (E+05) | 0.975] (E+05) |
|---|---|---|---|---|---|---|
| Constant | -1000 | 3,45 | -29,02 | 0,000 | -107,0 | -93,30 |
| Age | -205,0 | 65,8 | -31,10 | 0,000 | -218,0 | -1,920 |
| Wage | 0,176 | 0,01 | 142,1 | 0,000 | 17,31 | 1,779 |
| Weak Foot | 4,830 | 0,38 | 1,260 | 0,208 | -0,268 | 1,230 |
| Trick Skills | 11,20 | 0,58 | 1,935 | 0,053 | -14,51 | 2,250 |
| Finishing | 9,428 | 34,8 | 2,707 | 0,007 | 26,02 | 0,163 |
| ShortPassing | 1,320 | 48,9 | -2,700 | 0,007 | -0,023 | -36,13 |
| Volleys | 7,281 | 31,5 | 2,314 | 0,021 | 11,12 | 0,135 |
| FreeKickAccuracy | 9,933 | 24,8 | 4,012 | 0,000 | 50,80 | 0,148 |
| Acceleration | 6,091 | 33,3 | 1,830 | 0,067 | -4,339 | 0,126 |
| Agility | -1,270 | 32,0 | -3,986 | 0,000 | -0,190 | -64,77 |
| Reactions | 4,290 | 53,5 | 8,031 | 0,000 | 0,324 | 0,534 |
| ShotPower | -2,260 | 30,6 | -7375 | 0,000 | -0,286 | -0,166 |
| Stamina | 8,493 | 26,5 | 3,209 | 0,001 | 33,05 | 0,137 |
| Interceptions | -8,469 | 35,9 | -2,373 | 0,018 | -0,155 | -14,75 |
| Vision | 8,524 | 36,8 | 2,318 | 0,020 | 13,17 | 0,157 |
| Penalties | -6,939 | 31,5 | -2,202 | 0,028 | -0,131 | -7,623 |
| StandingTackle | -7,851 | 36,4 | -2,156 | 0,031 | -0,150 | -7,121 |
| Basic Player | 13,80 | 0,50 | 2,759 | 0,006 | 0,400 | 2,360 |
| Position Strat | 22,00 | 0,10 | 21,49 | 0,000 | 2,000 | 2,400 |
| CB | 17,50 | 1,37 | 1,273 | 0,203 | -0,942 | 4,430 |
| CDM | -2,240 | 1,28 | -0,175 | 0,861 | -2,730 | 2,280 |
| CM | -4,340 | 1,10 | -0,396 | 0692 | -2,580 | 1,720 |

| | | | | | | |
|---|---|---|---|---|---|---|
| LAM | -54,30 | 6,75 | -0,805 | 0,421 | -18,70 | 7,800 |
| LB | -40,40 | 1,23 | -3,291 | 0,001 | -6,450 | -1,630 |
| LCB | 49,90 | 1,67 | 2,985 | 0,003 | 1,710 | 8,260 |
| LCM | 50,80 | 1,63 | 3,113 | 0,002 | 1,880 | 8,270 |
| LF | 6,080 | 8,23 | 0,074 | 0,941 | -15,50 | 16,70 |
| LM | -24,10 | 1,18 | -2,033 | 00,042 | -4,730 | -86,04 |
| LS | 56,10 | 2,33 | 2,411 | 0,016 | 1,050 | 10,20 |
| LW | -43,70 | 1,70 | -2,57 | 0,010 | -7,710 | -1,040 |
| RAM | 3,760 | 5,85 | 0,064 | 0,949 | -11,10 | 11,80 |
| RB | -36,60 | 1,25 | -2,934 | 0,003 | -6,100 | -1,210 |
| RCB | 13,60 | 1,67 | 0,812 | 0,417 | -1,920 | 4,640 |
| RCM | 31,40 | 1,67 | 1,878 | 0,060 | -0,137 | 6,420 |
| RF | 80,80 | 6,76 | 1,197 | 0,231 | -5,160 | 21,30 |
| RM | -17,90 | 1,18 | -1,521 | 0,128 | -4,100 | 0,517 |
| RS | -16,80 | 2,26 | -0,742 | 0,458 | -6,110 | 2,750 |
| RW | -51,60 | 1,75 | -1,940 | 0,003 | -8,600 | -1,720 |
| ST | -51,60 | 1,25 | -4,129 | 0,000 | -7,610 | -2,710 |

**Appendix B: Comparison of top 10 best performing parameters found using XGBoost by two evaluation metrics**

| Feature | F score gain | Gain rank | F score weight | Weight rank |
|---|---|---|---|---|
| Position Stat | 679181883959628 | 1 | 1325 | 40 |
| Reactions | 323217529881602 | 2 | 1847 | 20 |
| BallControl | 264845704238414 | 3 | 1745 | 24 |
| SideMidfielder | 81754190689466 | 4 | 2024 | 14 |
| Composure | 41070994552130 | 5 | 1405 | 38 |
| ShortPassing | 38919418713515 | 6 | 1744 | 25 |
| Wage | 38198911619605 | 7 | 4551 | 1 |
| CentralMidfielder | 29880330396184 | 8 | 2616 | 5 |
| CM | 29054545559096 | 9 | 152 | 50 |
| Striker | 27498918424219 | 10 | 2320 | 7 |
| Age | 13393573117392 | 18 | 3795 | 2 |
| Jersey Number | 1936026247416 | 54 | 3510 | 3 |
| Crossing | 2612604716226 | 47 | 2687 | 4 |
| Weight | 1987929937116 | 53 | 2467 | 6 |
| Finishing | 11097791349977 | 20 | 2304 | 8 |
| Penalties | 2128267608256 | 50 | 2274 | 9 |
| SprintSpeed | 4029821968581 | 38 | 2239 | 10 |