Surrogate Modeling of FIFA 23 Ratings Using Linear Regression, Random Forest, and Polynomial Regression: A Comparative Analysis

Emre Guvenilir and **Chadi Bsila**
{emguvenilir,chbsila}@davidson.edu
Davidson College
Davidson, NC 28035
U.S.A.

1 Introduction

Each year, soccer players receive performance-based ratings in the video game "FIFA." The exact method for determining these ratings is not publicly disclosed. However, it is widely accepted that factors such as a player's international profile—specifically, the league and team they play for—along with detailed statistical analysis of their performance over the season are key considerations (Murphy 2019). Given this background, our paper seeks to develop a surrogate model that predicts these FIFA ratings for players who play in the 'Big 5 Leagues,' using regression techniques. The 'Big 5 Leagues' are commonly known as the Premier League in England, Bundesliga in Germany, La Liga in Spain, Serie A in Italy, and Ligue 1 in France.

Due to the fact that FIFA ratings are not publically disclosed, this paper aims to construct a surrogate model that mimicks the FIFA ratings for an arbitrary player profile.

To construct this surrogate model, we draw on key statistics from fbref.com, a reputable soccer statistics website. The process involved significant data cleaning and preprocessing to ensure the data was in a usable format for our models. We used three regression techniques: linear regression, polynomial regression, and random forest regression.

Our approach includes both categorical variables—such as "international recognition" factors and nationality and a range of performance metrics to determine the key drivers of FIFA ratings. By incorporating and excluding these categorical variables, we aim to identify potential biases in the FIFA rating system. Additionally, we created specific linear regression models for goalkeepers, given their unique role in soccer as the only players permitted to use both hands and feet during game play.

Through extensive testing and hyperparameter tuning, we optimized our models for accuracy and efficiency. The results were validated against a test dataset to ensure robustness. Beyond technical considerations, we also address the ethical implications of our surrogate model, examining the impact of potential biases on player ratings and discussing how these factors could influence perceptions within the soccer community.

2 Data Processing

In this section, we discuss the extensive data processing that was used to get our data to a more desirable format. Originally, we found a dataset on Kaggle that contained the information we were looking for, but we realized this dataset only contained data for about half of the season for some players, and had full season data for other players, which rendered it useless. However, the dataset linked the website where the data was taken from, fbref.com, so we decided to get our data directly from there and do more processing. In the past,one was able to directly download a csv file from the website, but that functionality was removed since the data provider no longer allowed it. [2]

There was a workaround as we found a Github repository which created methods to extract the data using R. Using the provided functions, we extracted the data from each category and merged them by player name, and then wrote the combined data back to a CSV to do further processing in Python. We found that this combined dataset contained many duplicate rows, so we filtered out these duplicate rows, which brought our dataset to a length of 3,317. Then we turned our focus to the columns, as there were duplicates and columns that were not useful as features, so those columns were dropped. Part of this processing was not written into Python as individual column headers may not have had the same name, but measured the same statistic, and thus were removed. It is important to note that the original Kaggle dataset contained significantly less columns of data than we included, and also did not include any goalkeeper specific metrics.

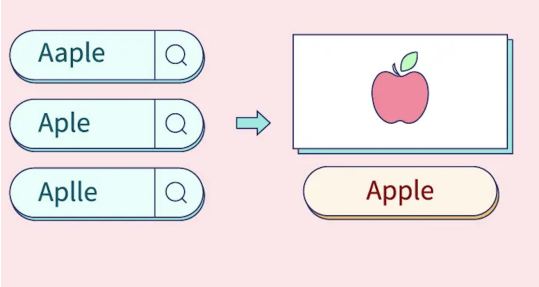

Figure 1: The Fuzzy Match to {"Aaple", "Aple", "Aplle"} is "Apple".

We used a second dataset from Kaggle which contained the FIFA ratings for each player. We filtered that dataset

down to players only in our target year, FIFA 23, and further filtered it to players only in the Big 5 leagues. We then faced our biggest challenge of matching the players from these two datasets so we could add the rating to our season statistics Dataframe. We faced difficulties with directly match player names since the FIFA 23 dataset did not include diacritic and accent marks, while the data retrieved from football reference did. To match the names, we used a best match process from a library called 'thefuzz.' This process calculated the percentage of how close the match, which worked for a majority of the players. The details for the matching process are beyond the scope of this paper.

However, there were outliers such as a player named 'Marquinhos,' whose name in the FIFA 23 dataframe was his full name, 'Marcos Aoas Correa.' This made it nearly impossible to find matches for these types of players. Thus, for every best match, we confirmed the result by comparing the year they were born, and then adding the rating to our Dataframe if both identifiers matched. If both were not achieved, we dropped the row to not have our data possibly contain incorrect ratings for a given set of player statistics.

We then replaced all blank values with 0, as empty values were being read as NaN or Not a number. We also created binary indicators for the positions a player was assigned to, as we did not want this aspect of the players position to be a categorical variable.

3 Methodology

Our research tried to emulate the FIFA 23 rating system by training models to predict a player's FIFA rating based on a variety of features (categorical, numerical, etc). We used a variety of models, including linear regression, polynomial regression, and random forest regression, to explore which method would yield the most accurate or generalizable predictions (to FIFA's rating).

Linear and Polynomial Models

Linear regression, as a model, imposes a linear relationship between the features (x) and the output (FIFA rating, y). This model was a self-evident baseline for predicting FIFA ratings. For linear regression, we used various player attributes such as age, height, nation, position, and others to predict the FIFA rating.

When examining the relationships between different features, such as age and overall player rating, we noticed that the distributions appeared non-linear. To explore these complex patterns, we opted to train polynomial regression models, which incorporate polynomial terms to account for non-linear relationships. We experimented with various degrees of polynomial expansion to find the best fit for our dataset, but ultimately, the quadratic model provided the ideal balance of accuracy and computational efficiency.

The principles underlying linear and polynomial regression are relatively simple, so we will skip a detailed explanation to focus on our final type of ensemble-method model.

Random Forest Regressor

To capture non-linearities and interactions among features, we employed a random forest regressor. A random forest is

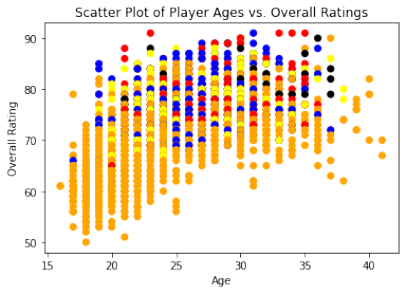

Figure 2: Player Ages vs FIFA Overall Ratings

an ensemble learning technique that consists of a collection of decision trees. The primary idea behind random forests is that by aggregating the predictions of multiple decision trees, the overall model becomes more robust and less prone to overfitting.

Each decision tree in the random forest is trained on a bootstrapped subset of the dataset, creating a different combination of training examples for each tree. Additionally, as the term "random" might indicate, when creating a new split in a tree, only a random subset of features is considered, which reduces the risk of overfitting.

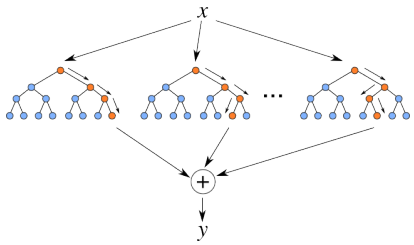

Figure 3: "The output of the random forest regressor is obtained by averaging the predictions of all individual trees."

The output of the random forest regressor is obtained by averaging the predictions of all individual trees. This aggregation helps reduce variance and generally provides better predictive performance compared to individual decision trees.

Here is a precise description of an algorithm for a random forest regressor:

- **Bootstrap Sampling**: The algorithm starts by creating multiple bootstrap samples from the training data. This means that for each of the B trees, a random sample with replacement is drawn from the original dataset.
- **Tree Growth**: Each bootstrap sample is used to grow a separate decision tree. The process of growing a tree involves recursively splitting the nodes, with these key steps:
  1. **Random Variable Selection**: At each node, a subset of m variables is randomly selected from the total p variables in the dataset.

2. **Best Split Point**: Among these m variables, the algorithm identifies the best variable and split point to divide the data into two child nodes.
3. **Recursion**: The node-splitting process continues until the node size reaches a specified minimum, creating the tree structure.

- **Creating the Forest**: After generating B trees, the complete ensemble forms the Random Forest. This ensemble is then used to make predictions.
- **Making Predictions**: For regression tasks, predictions are made by averaging the outputs from all the trees in the forest. Each tree produces an estimate, and the final prediction is the mean of these estimates.

The regression in "random forest regressor" part comes in when we run a polynomial regression model on each internal node.

With and Without Categorical Variables

Given the relevance of categorical variables, such as nationality and international recognition, for the overall rating of a soccer player, we trained two sets of models for each model type (linear, polynomial, random forest regressor): one excluding categorical variables and one including them. When including categorical variables, we converted them into dummy variables (one-hot encoding). Dummy variables are binary columns (think yes or no decisions) that represent each category, enabling models to process categorical data in a numerical manner.

Goalkeeper-Specific Model

Due to the uniqueness of goalkeepers in soccer, we trained a specific linear regression model to account for these players. Goalkeepers were outliers compared to field players in their on-field role, potentially skewing predictions. Thus, we created a subset of features that are goalkeeper specific to accurately assess this model.

Cross-Validation with K-Fold

Given the relatively small dataset (when excluding non-goalkeepers), we used k-fold cross-validation to assess model performance for the last goalkeeper-specific model. In k-fold cross-validation, the dataset is split into k subsets (folds). The model is trained on $k-1$ folds and tested on the remaining fold, with the process repeated k times. This approach helps ensure that the models are generalizable and not overfitting to specific data subsets and works perfectly for small datasets as it increases the number of instances the model trains. In this paper, our k value is 10.

Model Evaluation Metrics

To evaluate the performance of our models (with and without categorical variables), we notably used R^2 (coefficient of determination) and mean squared error (MSE) among others. MSE measures the average squared differences between the predicted and actual values, offering a metric for model accuracy.

- **Mean Squared Error (MSE)**: The MSE measures the average of the squares of the errors (the differences between the predicted and actual values). It is calculated as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the total number of data points.

- **Coefficient of Determination (R^2)**: The R^2 score indicates the proportion of variance in the dependent variable that is predictable from the independent variables. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the actual values.

MSE is a measure of accuracy—the lower the MSE, the more accurate the predictions. R^2 indicates the proportion of variance in the output prediction (rating), providing a measure of the model's goodness-of-fit. Simply, it shows how well our models fit the datapoints. We use these metrics to evaluate the accuracy of our models and their ability to mimic the FIFA Rating.

In the next section, we present our results and the performance of each model type to determine the most effective approach for mimicking the FIFA 23 rating system.

4 Results

This section provides a detailed summary of our results using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2).

Table 1: Performance Metrics for Regression Models

Model	MAE	MSE	RMSE	R^2
Linear Regression (with categorical)	0.0308	0.001675	0.04093	0.6738
Linear Regression (without categorical)	0.0334	0.0019	0.04359	0.6301
Linear Regression (Goalkeepers)	0.0414	0.002846	0.05334	0.6129
Ridge Regression (K-Fold)	0.0427	0.002945	0.05340	0.5800
Random Forest (without categorical)	0.03292	0.001814	0.04259	0.6725
Random Forest (with categorical)	0.03064	0.001611	0.04013	0.7092
Random Forest (without categorical and After Hyperparameter Tuning)	0.03278	0.001772	0.04210	0.67997

5 Analysis

Our results indicate that the most effective model was the Random Forest regression when accounting for categorical variables, as mentioned in the introduction. Among all our

models, incorporating categorical variables improved the R^2 scores. Our best-performing model achieved a Mean Absolute Error (MAE) of 0.03064, a Mean Squared Error (MSE) of 0.001611, a Root Mean Square Deviation (RMSE) of 0.04013, and an R-squared (R^2) of 0.7092. The linear regression model with categorical variables showed slightly lower performance. These findings align with other machine learning studies in sports analytics, many of which use some form of Random Forest. A similar project by the International University of Malaya-Wales, focusing on FIFA 20, used a dataset approximately one-third the size of ours, but their study utilized a classification system, making direct comparisons challenging. They also examined different features, such as player wage and player value (M. 2023).

We employed two tuning approaches for our models. The first involved grid search and randomized search, while the second used manual testing of various hyperparameters on our test data. However, we couldn't implement grid or random search for Random Forest regression models.

In linear regression, we started with a basic model without adjusting the regularization constant (α). By experimenting with different α values from 0.001 to 100, we observed minor improvements in MSE and R^2. To avoid underfitting, we decided to keep α at its default setting and used ordinary least squares for the full dataset. For our linear regression model focusing on goalkeepers, we employed grid search to find the optimal α value, which was 100. Given the smaller size of this subset, we applied K-Fold cross-validation, dividing the dataset into k subsets to ensure robust validation. While experimenting with different α values, we found that increasing α beyond 10 did not significantly improve the results.

In polynomial regression, we used a random search to tune our hyperparameters, usually finding the optimal α value to be around 10. However, even with tuning, our evaluation metrics, particularly R^2, consistently yielded negative results, indicating that our polynomial model with a degree of 2 performed worse than a horizontal line. This suggests that polynomial regression is not suitable for this data. Attempting to increase the polynomial degree caused Jupyter to crash, preventing further experimentation. Given the linear relationships observed during data exploration, we believe a higher-degree polynomial model would not have offered significant improvements.

For Random Forest regression, we initially evaluated the model without tuning hyperparameters, then used a grid search to optimize them. The most significant hyperparameter was $n_{\text{estimators}}$, representing the number of trees in the forest, which leveled off around 250. Another key hyperparameter, *max_depth*, was less significant unless it was set too low. To avoid overfitting, we left *max_depth* at its default (none). Using a minimum sample split (*min_sample_split*), which refers to the minimum number of rows needed to split a node, helped control overfitting. For a dataset with high dimensionality, having a low *min_sample_split* (default is 2) can easily cause overfitting. Our experiments showed that, without categorical variables, the best *min_sample_split* was 16, while with categorical variables, it was 8.

We addressed overfitting in various ways, including reducing the number of irrelevant features and maintaining a separate validation dataset. We used an 80/20 train-test split to evaluate model performance more accurately.

6 Ethics

Creating a surrogate model to replicate FIFA player ratings, derived from undisclosed criteria, raises significant ethical concerns. The models in this paper that included categorical variables like nationality or race consistently outperformed those that did not, indicating potential biases in the original FIFA rating system. This preference for models with categorical variables not only suggests a biased foundation but also questions the ethicality of using such models for future predictions or ratings. Additionally, employing a surrogate model with these characteristics risks perpetuating underlying prejudices, favoring some groups of soccer players while disadvantaging others.

Research by Francesco Principe and Jan C. van Ours (2021) underscores the presence of racial bias in professional football player ratings, specifically in Italian newspapers (Principe and van Ours 2022). Their study revealed that black players often received lower ratings compared to non-black players, especially at the lower end of the rating distribution, even when objective performance indicators were considered. Notably, the bias was not evident among top-tier black players, suggesting that the racial prejudice was more pronounced among average or lower-rated players.

Given this context, building surrogate models with potential biases linked to nationality or race poses a challenge. While these models can help identify and highlight existing biases, they also risk reinforcing them if not handled carefully. This requires a delicate balance: using the model to uncover and address systemic inequities without contributing to them through the modeling process. To maintain ethical integrity, machine learning researchers must approach this issue with caution, transparency, and a commitment to confronting unjust practices. Ultimately, the goal should be to use these models to drive positive change and fairness in the evaluation of professional soccer players.

References

M., A. 2023. awwalm/fifaprediction. https://github.com/awwalm/FIFAPrediction#user-content-fn-1-3d8370a45a252eb0a7b03bb6db14b38f. Accessed: 2024-05-08.

Murphy, R. 2019. Fifa player ratings explained: How are the card number stats decided? https://www.goal.com/en-us/news/fifa-player-ratings-explained-how-are-the-card-number--1hszd2fgr7wgf1n2b2yjdpgynu. Accessed: 2024-05-07.

Principe, F., and van Ours, J. C. 2022. Racial bias in newspaper ratings of professional football players. *European Economic Review* 141:103980.