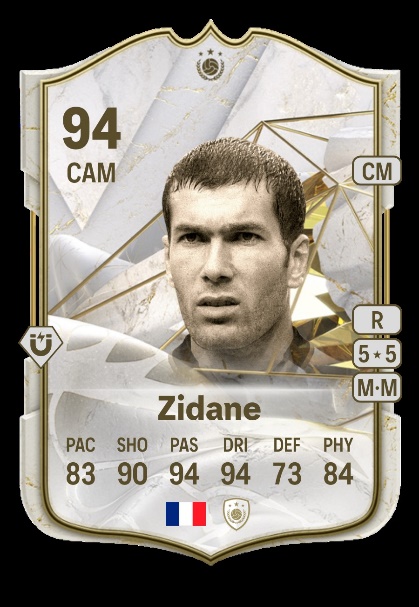
Pick 5-a-side: Machine Learning Project for Selection of a Competitive Soccer Team

Hamza Khattak  
*Dept. of Computer Science  
Eastern Kentucky University*Richmond, USA   
muhammad\_khattak@mymail.eku.edu

*Abstract*—This project aims to present a side-by-side comparison of the five most elite soccer players in Electronic Arts’ Fédération Internationale de Football Association computer game with a user selected legendary team in the same positions to determine if their ratings surpass the current best team. The project utilizes datasets from FIFA 2022 official data via Kaggle, The FIFA FUT 2017 to 2024, and EA FC 24 players. The project involves preprocessing, modeling, and evaluation using regression techniques. The outcome is visualized through radar charts, providing insights into the overall strength of the teams given the relevant attributes. This project addresses a real-world problem related to sports team ranking and its financial implications for the sport of soccer/football which is loved by over billions of people across the world.

**Keywords—FIFA, Soccer, 5-a-side, modeling, regression**

# Introduction

Electronic Arts’ FIFA series computer games are one of the most widely recognized and played games in the entire video gaming industry[1]. The game launched in 1994 with its first version titled FIFA Soccer 95[2]. Since its release FIFA released 29 iterations of the game with the latest being FIFA 24. Since the dawn of the EA FIFA empire, countless players have been represented in the game with various features. The goal of this project is to use the FIFA 2022 official dataset from Kaggle to create the most prolific 5-man team and compare it to a user inputted 5-man team of legendary players featured in any of the EA FIFA iterations. The project intends to tackle the practical issue concerning the ranking of sports teams, therein ranking players on their skills, abilities, strengths, and weaknesses.

**Fig.1: FIFA FC 24 icon card for French Midfielder Zidane with relevant features**

# Literature review

The collection of scholarly articles published on the content of EA’s FIFA computer games is limited. However, a substantial number of scholarly content has been published that encapsulates the spirit of this project in a broader sense. For example, Lijuan Mao’s ‘Identifying keys to win in the Chinese professional soccer league’ describes the correlations between soccer match statistics and match results. In this study, Mao’s team employed generalized linear modeling to analyze 21 performance-related factors and their impact on match outcomes in the Chinese Football Association Super League's 2014 and 2015 seasons[3]. It was discovered that several factors, including Shot on Target, Shot Accuracy, Cross Accuracy, Tackle, and Yellow Card, consistently exerted influence on match results, regardless of the teams' varying strengths and the quality of the opposition [4].

Another study by Jassim AlMulla and his team focused primarily on using deep learning models predicting the outcomes of football matches in the Qatar Stars League (QSL)[5]. The Team used Long Short-Term Memory (LSTM), which is a recurrent Neural Network, and the Multilayer Perceptron (MLP), which is a feedforward neural network that can handle complex patterns in data[6]. The goal of their study was to understand the features and their influence in determining the match outcomes to provide valuable insights to the coaching staff and team management.

These two projects are a great resource and can be used as a reference point for the ‘Pick 5-a-side’ project. Although it is apparent that the 'Pick 5-a-side' project shares a close alignment with Mao's study, in contrast to Jassim's project. A closer look at 'Pick 5-a-side' would enable the reader to see that the project employs a notably distinct approach in pursuing its objectives of creating a robust model that is able to predict scores of individual players rather than teams as a whole. This may appear a small deviation in the beginning, but the dynamics of a team leverages a number of metrics that are averaged by the players in the team and use that as an entity. In ‘Pick 5-a-side’ the driving entity are the player attributes rather than how they fit to be a part of a broader structure such as a team or an organization.

# Dataset

The EA FIFA 2022 official dataset includes 19,239 rows and 110 features. Almost half of the features in the dataset can be excluded from the analysis because they are insignificant in determining the target variable of ‘overall rating’. The overall rating feature is bound in values from 0 to 100, a numerical score to quantify the general abilities, skills, and performance of players in the real-world sport of soccer[7]. For illustration, an example of an icon card for the French midfielder and Real Madrid football club legend, Zinedine Zidane is given in figure 1.

The original dataset was reduced to 62 relevant columns and 16,710 records for pre-processing. The excluded data suffered from data quality and relevance issues. For instance, columns such as player tags, international team positions, and date of birth etc. Furthermore, data with units or characters mismatch according to the numeric standards were removed or substituted in excel before importing the dataset to Jupyter Notebook to be processed in Python. Once the dataset was imported into Jupyter Notebook, missing values were counted and replaced. Which led to 2,312 records being removed from the dataset. After further analysis, more columns were removed to reduce extraneous dimensionality, leaving the final dataset with 41 features and 14,398 records. Some of the columns were also renamed from the Kaggle dataset, most notably the target variable was renamed to ‘overall’ instead of ‘overall rating’. However, for the purpose of this project, both words are used interchangeably to convey the same meaning,

# Main Approach

## Logistic Regression

**Fig.2: Scatter plot of actual values in the dataset compared to the model predictions for the target value of y-train (‘Overall’)**

The first machine learning algorithm applied in the project is Logistic Regression. Although Logistic Regression comes from a family of linear regression packaged in the Scikit-Learn library used for machine learning in Python, it is a model used for classification[8]. The project used this model to see the data fit for a multinomial logistic regression where K=100. Using ‘lbfgs’ as the solver the goal is to employ ℓ2­ regularization, with intent to converge faster than the model usually would[9]. The idea behind using ‘lbfgs’ was to encapsulate the data faster than other solvers because with 41 features the input may very well be a highly dimensional dataset. The performance of this model was unsatisfactory with an accuracy score of only 21.53%.

## Ridge Regression

The second model used in the project is the Scikit’s Ridge Regression. Using the regularization term alpha (α), the goal of using Ridge Regression is to minimize the influence of highly correlated features by applying the ℓ2­ regularization. The amount of co-efficient shrinkage is controlled by tuning the hyperparameter alpha. The solver used for Ridge Regression was ‘cholesky’ and the alpha used is 1.0, which was selected after manually trying a set of alpha values. Using r-squared score metric for this model returned a value of 0.98, which is a decent score given the maximum score is a 1.0.

## Elastic Net for Ridge and Lasso Regression

The Elastic Net model from the Scikit library is a popular linear regression model that combines ℓ1 and ℓ2­ regularization the minimize the cost function. In the code, the alpha hyperparameter is used to regularize the coefficients and the l1\_ratio determines the mix of Ridge and Lasso regression, with l1\_ratio of 0 is pure Ridge Regression and l1\_ratio of 1 is pure Lasso Regression[10]. The project uses an alpha for of 0.1 and l1\_ratio of 0.5 to ensure the development of a stable model. For accuracy score, the metric used is r-squared score. The Elastic Net scored 0.98 as well for the accuracy score. As seen in the scatter plot in figure 2 of actual values on x-axis and predicted of y-axis, it is evident that the model is doing an adequate job in terms of prediction.

A graph of blue lines

Description automatically generated with medium confidence

|  |  |
| --- | --- |
| **Top 10 features from the elastic net** | |
| **Features** | **Co-efficient Value** |
| Best Overall Rating | 0.960667636 |
| Crossing | 0.032092043 |
| Potential | -0.031360803 |
| Reactions | 0.025910489 |
| Height(in cm) | -0.014846897 |
| GKDiving | 0.012499715 |
| Composure | 0.01183138 |
| LongPassing | -0.011275322 |
| Vision | -0.011016444 |
| GKKicking | 0.009783812 |

Furthermore, getting the co-efficient values would allow one to get a better idea of what features matter the most in terms of the predictions. Table 1 contains the top 10 most influential features in the model with its respective co-efficient value.

**Table 1: Top 10 influential features in the model with its co-efficient scores**

In purely mathematical sense, given below would be the model equation using intercept value of 2.145.

# Evaluation Metrics

A screenshot of a computer

Description automatically generatedAt this point it would be appropriate to see if the model suffered from multicollinearity. Using a correlation matrix and plotting it via the matplotlib package in Python renders the image in figure 3. The threshold for this correlation matrix is 0.7 since the threshold at 0.9 did not reduce any features from the model. However, at the threshold value of 0.7, the remaining features are not significantly influential for predicting the target variable. Thus, for testing purposes two parallel models were created. The original Ridge Regression and another Ridge Regression model used the dataset without highly correlated data. The second Ridge Regression model was also standardized using StandardScaler from Scikit library after treating multi-collinearity. The accuracy score for this metric returned an r-squared score of 0.97, which is a satisfactory score, but it still falls short of improving the previous high score of 0.98 from the Elastic Net model.

**Fig.3: Correlation matrix after applying the 0.7 threshold for correlated features.**

# RESULTS & Analysis

## Selecting the best performaning model

The model with the best accuracy for the project is the Elastic Net model with 98% or 0.98 r-squared score. As previously discussed, the elastic net model used a 0.1 alpha and a 0.5 l1\_ratio. The model was trained on 80% of the data as training and the performance was measured against the target value using 20% of the testing data or 2880 records. Since the goal of the project is to select the top players in the entire dataset, after fine tuning, the model was applied to the entire dataset to add an additional column titled ‘Predicted Score’.

## Selecting the best team from FIFA 22

Using Predicted Score as the driving force, the project selects the top five players based on the predicted scores. The top five players comprise of one goalkeeper, one defender, two midfielders and one attacker. In the code this is performed by creating a list of positions for each of the domains and iterating through these positions to select players for each category. The selected players are then assigned to a list variable which is used later for render an html output.

The goalkeeper selected by the model is the Slovenia national team and Atletico Madrid player, Jan Oblak with a predicted score of 91.03 and an actual score of 91. For the defensive position, the model predicted French national and Chelsea football club defender, N’Golo Kante as the best rated player with a predicted score of 88.67 and actual score of 90. For midfield positions, one of the midfielders is the Belgian player Kevin De Bruyne who plays for an English football club with a predicted score of 90.15 and actual score of 91. The other midfielder selected was Son Heung-min, who belongs to South Korea and plays for the English club of Tottenham. Son has a predicted score of 88.14 and actual score of 89. For the attack position, the model decided Argentinian player and global soccer icon, Lionel Messi who plays for the French club Paris Saint-Germain. Messi has an overall rating of 93 while the model scored him a total of 91.62.

## User input

The next step in the project is to compare the team selected by the Elastic Net model and provide an alternative team to see if the model is able to predict scores for the alternative team and compare relevant features of this team with the current team. For this purpose, five legendary players were selected for each position. For the goalkeeping position, Spaniard Iker Casillas who played in Real Madrid football club during his prime was provided. For defensive position, Italian international Paolo Maldini’s values were inputted. For midfield, English international and long time Liverpool football club Steven George Gerrard and French Legend Zinedine Zidane were presented. Finally for attack, Portuguese international Cristiano Ronaldo was inputted to complete the five-man team.

The model predicted high scores for the legends team due to their indisputable skills and abilities. The predicted score for Casillas was 92.97, for Maldini the score was 91.13, for Zidane and Gerrard the predicted scores were 95.48 and 88.64 respectively. Finally for Cristiano Ronaldo the predicted score was 90.69, which falls in the generally accepted scores for his FIFA ratings[11].

Although there are various perspectives to evaluate the superiority of one team over another, a straightforward spider web chart, which compares and contrasts a selection of their most pertinent abilities, can be regarded as a fundamental and standard analytical tool. Figure 4 shows one such instance.

## Pros and Cons

The pros of using the defaults, such as the choice of the using an Elastic Net instead of pure Ridge Regression or using an alpha value of 0.1 in this project, are that the model results in a satisfactory performance. The time used by the model to train is negligible and the model appears to be stable. Furthermore, the model predicting scores the user input is in the ballpark region of what is available online, which means the model can be considered to be scalable. Overall, the model does a good job in dealing with the issues in the dataset and producing a model that can be employed for future work.

The cons for this project are that the model is a pretty simple model. While there is some fine tuning in shape of hyperparameter application for feature regularization, one would expect the official FIFA data scientist team to employ a much more stable and complicated model. This is primarily because the model important feature in this model is the ‘best overall rating’, which is a legacy feature carried over from previous iterations of the game. In terms of result analysis, it must be noted that these features are picked arbitrarily and are all not part of the important features described above in table 1. If a reader decides to select a set of other features or expand the list of features selected for comparison the result may end up being different than the one furnished above.

# Future Work

An orange hexagon with white text

Description automatically generatedThe project boasts a remarkable score of 0.98 on the r- squared for the model but it must be noted that there are a set of other metrics that could be used to evaluate model performance. One such metric is the Root Mean Squared Error or the RMSE. The RMSE is the quality of the prediction of the model. The RMSE is calculated by measuring measures the average magnitude of the errors or residuals between predicted values and actual observed values[12]. Moreover, this project relies heavily on linear or generalized linear regression models. Some community projects benefit from introducing more advanced machine learning algorithms such as Decision Tree Regression and Random Forest Regression in Jean-Paul’s ‘FIFA Ultimate Team Machine Learning Project’ [13]. Jean’s project uses 35 features which makes it a relevant comparison for the ‘Pick 5-a-side’ project. The ‘Pick 5-a-side’ project can certainly leverage some of the steps taken in Jean’s project and enhance its algorithmic depth in terms of machine learning.

**Fig.4: Spider web chart to compare features of current top 5 to the user input top 5.**

The ‘Pick 5-a-side’ can also benefit from the introduction of the Gradient Boosting with XGBoost algorithm as demonstrated by Awwal M’s the in “FIFAPrediction” project on Github[14]. While neither Jean’s nor Awwal’s model has an r-squared metric score, it is necessary to consider that multi-collinearity in the dataset can influence how RMSE or r-squared values can be interpreted.

If time constraints were ignored, ideally the model should be using cross-validation to ensure a high quality of hyperparameter tuning. Especially for the Ridge Regression model, the full potential of the model will not be reached unless a rigorous cross-validation process is applied. This will enable the model to weed out any extraneous columns, which would reduce the dimensionality of the dataset and create a much more stabilized model. Furthermore, at that point an r-squared score of 0.98 can be considered impressive because r-squared on its own does not necessarily represent the performance of the model especially if the variables are not independent.

A blue hexagon with white text

Description automatically generated

1. Lerma, Alberto Zaragoza. “The Best-Selling FIFA in EA Sports History.” Meristation, September 26, 2023. https://en.as.com/meristation/news/the-best-selling-fifa-in-ea-sports-history-n/.
2. Contributors to FIFA Football Gaming wiki. “FIFA Soccer 95.” FIFA Football Gaming wiki. Accessed October 5, 2023. <https://easportsfc.fandom.com/wiki/FIFA_Soccer_95>.
3. Lijuan, Mao, and Zhaofang Peng. Identifying keys to win in the Chinese professional soccer league. Accessed October 6, 2023. https://www.tandfonline.com/doi/abs/10.1080/24748668.2016.11868940.
4. Lijuan, Mao, and Zhaofang Peng. Identifying keys to win in the Chinese professional soccer league. Accessed October 6, 2023. https://www.tandfonline.com/doi/abs/10.1080/24748668.2016.11868940.
5. AlMulla, Jassim, Mohammad Tariqul Islam, Hamada R H Al-Absi, and Tanvir Alam. “Soccernet: A Gated Recurrent Unit-Based Model to Predict Soccer Match Winners.” PloS one, August 1, 2023. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10393150/>.
6. Pettersson, Julius, and Petter Falkman. “Comparison of LSTM, Transformers, and MLP-Mixer Neural Networks for Gaze Based Human Intention Prediction.” Frontiers, May 5, 2023. <https://www.frontiersin.org/articles/10.3389/fnbot.2023.1157957/full>.
7. “Zinedine Zidane Icon 94 OVR.” FUT.GG. Accessed October 5, 2023. <https://www.fut.gg/players/1397-zinedine-zidane/24-1397/>.
8. “1.1. Linear Models.” scikit. Accessed October 5, 2023. <https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression>.
9. “1.1. Linear Models.” scikit. Accessed October 5, 2023. <https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression>.
10. Lerma, Alberto Zaragoza. “The Best-Selling FIFA in EA Sports History.” Meristation, September 26, 2023. <https://en.as.com/meristation/news/the-best-selling-fifa-in-ea-sports-history-n/>.
11. FUTWIZ. “Cristiano Ronaldo Ultimate Team History.” FUTWIZ. Accessed October 5, 2023. <https://www.futwiz.com/en/player/history/20801>.
12. C3 A1. “Root Mean Square Error (RMSE).” C3 AI, September 28, 2021. <https://c3.ai/glossary/data-science/root-mean-square-error-rmse/>.
13. Zambrano-Leon, Jean-Paul. “FIFA Ultimate Team Rating Prediction Machine Learning Project.” Medium, May 19, 2021. <https://towardsdatascience.com/fifa-ultimate-team-rating-prediction-machine-learning-project-3a02767fcb38>.
14. M., Awwal. “AWWALM/FIFAPREDICTION: FIFA Overall Player Ratings Prediction Using Regression Methods.” GitHub. Accessed October 5, 2023. https://github.com/awwalm/FIFAPrediction.