Predictive Analytics in Football - Predicting Top Goal Scorers -

Chanyoung Park

Hussain Alsharif

Oluwasegun Adegoke

OR-568, Applied Predictive Analytics

George Mason University Dr.

Vadim Sokolov

ABSTRACT

This study explores predictive modeling to identify the top goal scorer in football using mid-season data. Utilizing a dataset covering three seasons, we employed models like Boosting Trees, Linear Regression, Neural Networks, and Random Forests. Key findings highlighted 'goal_per_game' and 'shots' as significant predictors across models. The study balanced predictive accuracy with model interpretability, revealing strengths and limitations within each approach. Our research contributes to sports analytics by demonstrating the applicability of diverse predictive methods in football and suggests future directions for more comprehensive analyses.

Keywords ~ Sports Analytics, Predictive Models, Football, Goal Scoring, Machine Learning

---

INTRODUCTION

In the ever-evolving world of sports analytics, the ability to accurately predict outcomes based on statistical data is invaluable. This study, spearheaded by a dedicated team comprising Hussain Alsharif, Oluwasegun Adegoke, and Chanyoung Park, delves into the realm of football analytics with a specific objective: to predict the season's top goal scorer using data available at the mid-season mark.

This ambitious project is not just about forecasting a single player's success; it's about understanding the intricate dynamics of football performance and leveraging statistical models to make informed predictions. By examining various aspects of players' performances in the first half of the season, we aim to develop a predictive model that can accurately project who will emerge as the top goal scorer by the season's end.

Through this endeavor, we not only contribute to the field of sports analytics but also provide valuable insights that could be of interest to football clubs, managers, analysts, and fans alike. The blend of statistical rigor and football knowledge forms the cornerstone of this project, making it a unique and exciting challenge in the world of sports data analysis.

METHODS

Data Collection and Preparation: Our journey begins with the meticulous assembly of a dataset that encompasses the 2019/2020, 2020/2021, and 2021/2022 football seasons. This dataset is not just a collection of numbers; it is a comprehensive record capturing 1442 observations across 26 distinct features. These features include but are not limited to, players' goals, assists, minutes played, shots on target, and other vital statistics that could influence their goal-scoring capabilities.

The data was sourced with careful consideration to ensure relevance and accuracy. By focusing on the first half of each season, we aimed to capture a snapshot of players' performances, which we believe are indicative of their potential to be the top goal scorers.

Data Cleaning and Preprocessing: The raw data, though rich in information, required significant preprocessing to be useful for our analysis. This involved several critical steps:

- Renaming columns for better clarity and consistency.
- Cleaning the data to remove any inconsistencies or errors.
- Excluding irrelevant columns to focus on the most impactful features for our predictions.
- Converting data types where necessary to ensure compatibility with our analytical models.

This preprocessing stage was crucial in setting a strong foundation for our predictive analysis, ensuring that the data fed into our models was of the highest quality and relevance.

Analytical Approach: Armed with this cleaned and preprocessed data, we embarked on a journey to apply various predictive models. Each model was chosen for its unique ability to analyze and interpret different aspects of the data:

- Decision Trees and Random Forests for their robustness in handling non-linear relationships.
- Linear Regression for its effectiveness in identifying trends and relationships among variables.
- Deep Learning models for their capability to capture complex patterns and interactions within the data.
- Gradient Boosting Machines (GBM) for their strength in optimizing prediction accuracy.

Each model was meticulously tuned and evaluated to ensure that our predictions were not only accurate but also meaningful in the context of football analytics.

PREDICTIVE MODELING

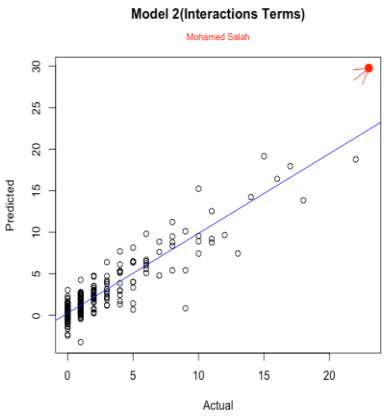
Linear Models (Linear Regression)

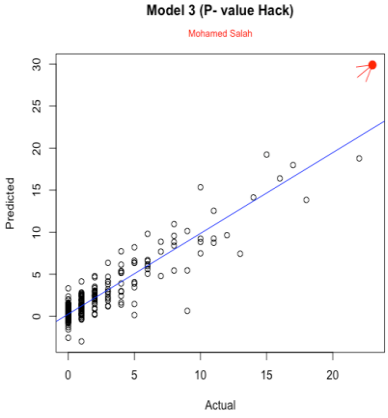
Implementation of Linear Regression:

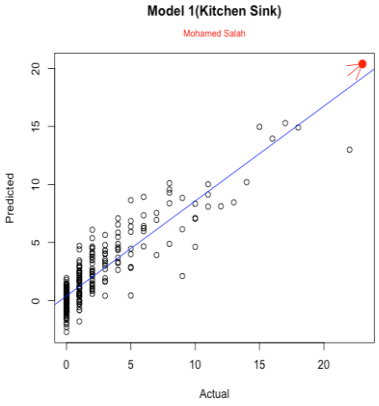
- Model Development: Linear regression models were implemented with a focus on understanding how different variables interact and influence the prediction of the total goals.
- Variable Selection: Initial models included all variables, followed by a refined approach using p-value-based selection to identify the most relevant predictors.
- Interaction Consideration: Interactions between variables were considered to capture the combined effect of different player attributes on the prediction.

Evaluation:

- Statistical Analysis: The model's coefficients, standard errors, t-values, and p-values were carefully analyzed to understand the significance of each predictor.
- Performance Metrics: The models showed RMSE values of 1.56 and 1.47. These figures suggest a moderate level of prediction accuracy.
- Results Interpretation: Analysis of the linear models highlighted the importance of not only individual player statistics but also their interactions in predicting goal-scoring performance







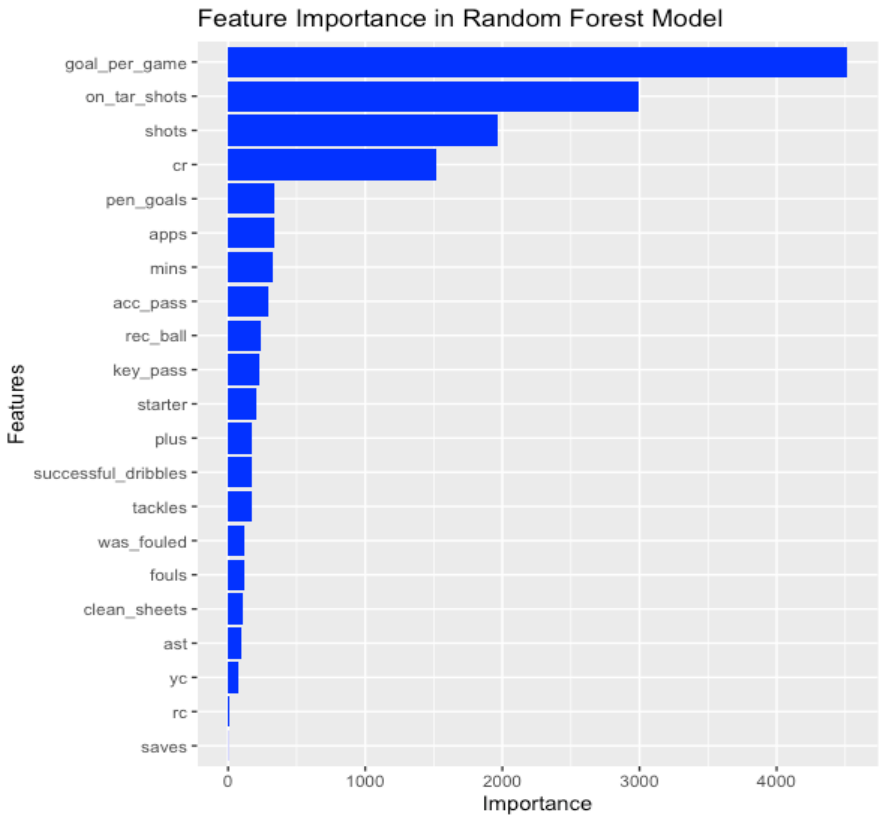
Random Forests

Implementation of Random Forest Model:

- Model Configuration: A random forest model was configured to predict total goals. The model used numerous decision trees to capture a broad spectrum of data patterns.
- Feature Consideration: Like the GBM model, the random forest model utilized various player statistics to predict goal-scoring performance.

Evaluation:

- Performance Metrics: The model achieved an RMSE of 1.44 and 1.26 in different scenarios. Additionally, the mean absolute error was calculated to assess the average deviation of the model's predictions from the actual values.
- Results Analysis: The random forest model's predictions were subjected to a confusion matrix analysis to evaluate its classification accuracy in predicting the exact number of goals.



Boosting Trees (Gradient Boosting Machine)

Implementation of GBM Model:

- **Model Setup**: The Gradient Boosting Machine (GBM) model was set up with specific parameters: 950 trees (n.trees), an interaction depth of 5, and a shrinkage value of 0.01. These parameters were selected to optimize the model's ability to learn complex patterns in the data while preventing overfitting.

- **Feature Selection**: The model used a wide range of features, including player stats such as goals per game, shots, assists, and more, to predict the total goals.
- **Training Process**: The GBM model was trained on a subset of the dataset, with a 10-fold cross-validation to ensure the model's robustness.

Evaluation:

- **Performance Metrics**: The model's performance was evaluated using the Root Mean Squared Error (RMSE), which came out to be 1.29. This metric indicates the model's accuracy in predicting the total goals scored by players.
- **Results Interpretation**: An RMSE of 1.29 signifies that the model predictions are close to the actual data, although there is still room for improvement.
- **Key Feature Influence**: The model identified 'goal_per_game' and 'shots' as some of the most influential features in predicting a player's total goals.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  22  23
        0  112  14   1   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        1   20  23   8   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        2    0  15  11   5   1   2   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0
        3    0   2   7   5   4   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        4    0   0   5   2   4   1   0   1   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0
        5    0   1   0   1   1   0   2   0   1   1   0   0   0   0   0   0   0   0   0   0   0   0   0
        6    0   0   0   0   0   0   0   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        7    0   0   0   0   0   2   4   1   0   0   0   2   0   0   0   0   0   0   0   0   0   0   0
        8    0   0   0   0   0   0   0   1   1   1   0   1   0   0   0   0   0   0   0   0   0   0   0
        9    0   0   0   0   0   0   0   1   1   2   0   0   1   0   1   0   0   0   0   0   0   0   0
        10   0   0   0   0   0   0   0   0   0   0   1   0   1   1   0   0   0   0   0   0   0   0   0
        11   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0
        12   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   1   0   0   0   0   0   0   0
        13   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        14   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        15   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0
        16   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0
        17   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0
        18   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0
        19   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        20   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        22   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        23   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0

Overall Statistics

               Accuracy : 0.5575
                 95% CI : (0.4979, 0.6158)
    No Information Rate : 0.4599
    P-Value [Acc > NIR] : 0.0005751

                  Kappa : 0.4007
```

Neural Networks (Deep Learning)

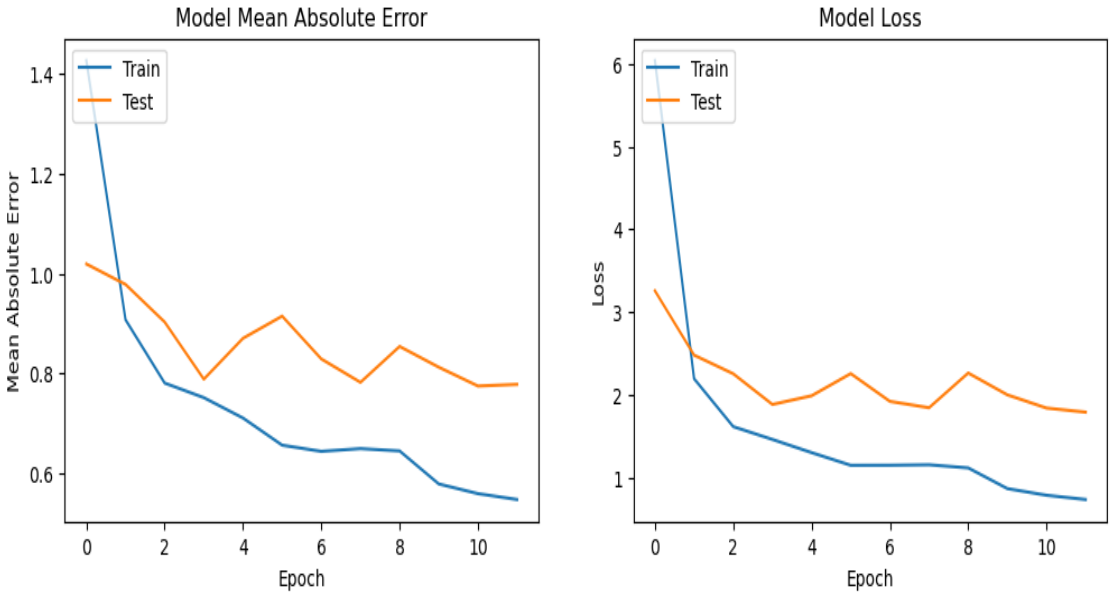
Implementation of Neural Network Model:

- **Model Architecture**: A deep learning model with multiple dense layers was created. It included layers with 256, 128, 64, and 32 neurons, respectively, all using the ReLU activation function. The output layer had a single neuron, as the task was regression (predicting total goals).

- **Training Process**: The model was trained over 15 epochs with a validation split to monitor performance on unseen data.

Evaluation:

- **Loss and Accuracy**: The model's training process was closely monitored, noting the loss and mean absolute error at each epoch. The final evaluation on the test set resulted in a loss of 1.5128 and a mean absolute error of 0.7182.
- **Model Insights**: The deep learning model's ability to capture nonlinear relationships and complex patterns in the data was evident from the training and validation metrics.



- **Comparative Analysis**: Each model brought its unique strengths and weaknesses to the study. While linear models provided a clear understanding of variable relationships, boosting trees and random forests captured complex nonlinear patterns more effectively. The neural network model excelled in identifying intricate interactions within the data.
- **Feature Importance Across Models**: A consistent observation across all models was the significance of 'goal_per_game' and 'shots' as predictors of goal-scoring performance.
- **Model Selection Considerations**: Choosing the right model depends on various factors, including the specific nuances of the dataset, the desired interpretability of the model, and the trade-off between complexity and performance.

CONCLUSION AND FUTURE RESEARCH

Concluding Insights

This study embarked on an ambitious journey to predict the season's top goal scorer in football using a diverse array of statistical models. The findings offer intriguing insights into the capabilities and limitations of different predictive techniques in sports analytics.

- **Boosting Trees and Random Forests** showcased their prowess in handling complex data, yielding high accuracy but at the cost of interpretability.
- **Linear Models** provided valuable insights with greater interpretability, though they fell short in capturing more complex relationships within the data.
- **Neural Networks**, with their ability to model intricate data interactions, pointed towards the potential of advanced machine learning techniques in sports analytics, despite requiring extensive data and computational resources.

A consistent theme across all models was the prominence of 'goal_per_game' and 'shots' as significant predictors, underscoring the critical role these metrics play in a player's scoring capability.

Model	RMSE	MAE	Accuracy
Linear Model	1.4713	0.9827	43.32%
Random Forest	1.4383	0.8627	48.61%
Boosting	1.2921	0.7924	55.75%
Deep Learning	1.4351	0.7455	49.13%

Future Research

The study opens several avenues for future research:

1. **Data Enrichment**: Incorporating additional variables such as in-depth player statistics, team strategies, or even psychological factors could provide a more holistic view and potentially enhance the predictive power of the models.
2. **Advanced Modeling Techniques**: Exploring more sophisticated machine learning algorithms, including ensemble methods or more complex neural network architectures, could yield improvements in both accuracy and interpretability.
3. **Real-World Application**: Applying these models in practical scenarios, such as player performance analysis, team strategy development, or even in betting industries, could demonstrate the real-world utility of such predictive analytics.

4. **Interdisciplinary Approaches**: Combining data analytics with insights from sports science, psychology, and tactical studies could lead to more comprehensive models that better capture the multifaceted nature of sports performance.

In conclusion, while this study has made significant strides in predicting the top goal scorer using mid-season data, it also highlights the complexity and unpredictability inherent in sports. The fusion of data science and sports expertise opens exciting prospects for further exploration, promising not only to enrich the field of sports analytics but also to enhance our understanding and enjoyment of the game.

REFERENCE

Kickest. (2022). Premier League Player Statistics 2021-2022. Retrieved from https://www.kickest.it/en/premier-league/stats/players/table/2021-2022?sortparam=101.