FIFA PLAYER PERFORMANCE ANALYSIS AND GOAL PREDICTION SYSTEM

Prudhvi Krishna Gangavarapu

Master's in information technology

University of New Hampshire-Manchester, gangavarapuprudhvi@gmail.com

1 ABSTRACT

Football is all about scoring goals, and this project dives deep into the numbers behind those crucial moments. The focus is on player stats and performance to understand the dynamics of goal-scoring. Additionally, a sophisticated prediction model has been developed, considering factors like time, position, and assist method etc. to forecast whether a particular situation will result in a goal. Accompanied by a user-friendly interface, this research enables anyone to input their own data and explore the model's predictions.

CCS CONCEPTS • **Computing methodologies~ Machine learning~ Machine learning approaches ~Classification and regression trees**

Keywords

Gradient Boosting Classifier, Logistic Regression, Soft Voting, **Python, Flask, HTML, CSS**

2 INTRODUCTION

Football is known as a beautiful game and is loved for its simplicity. However, behind the scenes, there are many complex statistics that can be confusing for fans. This project aims to simplify one key statistic: the goal. A goal not only excites fans but also determines the winner of the match. The team that scores the most goals wins, and the player who scores the most is often celebrated as the best.

The purpose of this project is to understand, analyze, and predict goal-scoring using a dataset from football events available on Kaggle. The entire project is developed in three phases. The first phase involves data analysis. The second phase is dedicated to building a machine learning model for prediction. Lastly, the third phase focuses on developing a user interface to showcase visualizations and the goal predictor.

I have cleaned the dataset and used it to analyze goal-related statistics and create visualizations of these analyses. For the machine learning aspect, the dataset is split into training and testing sets. The training set helps to train the model, while the testing set is used to validate the model's accuracy. Finally, I've developed a user interface (UI) that has been deployed as a website. This allows users to input their own values and test whether a given situation results in a goal.

3 APPROACH

The primary goal of this project is to leverage and understand different technologies to analyze football events data comprehensively. Utilizing methodologies such as Exploratory Data Analysis (EDA), Machine Learning, and Web Development, this project aims to extract meaningful insights from the Football Events dataset.

One of the key objectives is to utilize EDA techniques to explore and visualize the dataset, gaining insights into the patterns and trends of football events. This will involve generating summary statistics, creating visualizations, and identifying significant factors influencing game outcomes.

Furthermore, the project aims to employ Machine Learning to make models that can predict goal outcomes. I have used information like the time, side, assist method, location, fast break, body part, and situation from the dataset. I've tried using various other classification algorithms, but I have gone with the model that gave the best accuracy.

Additionally, the project intends to integrate the findings and models developed into a web application using Flask application. This web application will serve as an interactive platform for users to explore and analyze football event data, providing

insights into goal-scoring dynamics and facilitating strategic decision-making in professional football.

4 DATA ANALYSIS

In this section of my project, I look into the Football Events dataset using Exploratory Data Analysis (EDA). This method helps me find patterns and insights into goal-scoring trends and player performances. I checked if the dataset was labeled correctly, and fortunately, it was just right for my analysis. It's a sizable dataset, 22 MB in size, with lots of columns. However, before I could dive in, I needed to clean it up. There were plenty of N/A values that could mess up my analysis, so I used Excel to filter them out. Once the data was clean, I used pandas to read it and selected the columns I needed for my analysis.

I perform my analysis using Python and Jupyter Notebook, making use of libraries like Seaborn, NumPy, and Pandas. This process involves creating various visualizations to help me understand the data better. EDA is a crucial step in my research. [1] It involves a thorough examination of the data to spot trends, find outliers, and test hypotheses. By using summary statistics and graphs, I aim to improve my understanding of the dataset and draw valuable conclusions for the football domain.

4.1 Leading Scorers and Goals Per Match (GPM)

In this section, I investigate the top scorers across various football leagues and competitions. Utilizing the Football Events dataset and visualization techniques, I identify players with the highest number of goals, providing insights into their contributions to their respective teams' success. Additionally, I analyze the goals per match (GPM) ratio for these top players, offering a deeper understanding of their scoring efficiency and consistency. By comparing the GPM ratios of different players, I aim to uncover patterns and trends that that provide insight into the variables affecting a player's ability to score goals on an individual basis.

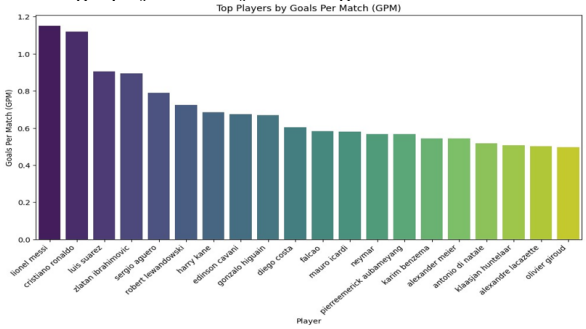

Figure 1: Goals Per Match

4.2 Effectiveness and League Distribution

For this section, I assess player effectiveness based on their conversion rates and goal-scoring performances. By analyzing the relationship between goals scored and attempts made, I identify players with the highest conversion rates, highlighting their efficiency in front of goal. Furthermore, I categorize top scorers by league, providing insights into the distribution of goal-scoring talent across different football leagues and competitions. This analysis not only showcases the diversity of talent but also offers valuable insights into the competitive landscape of professional football, helping teams identify potential transfer targets and strategic opportunities.

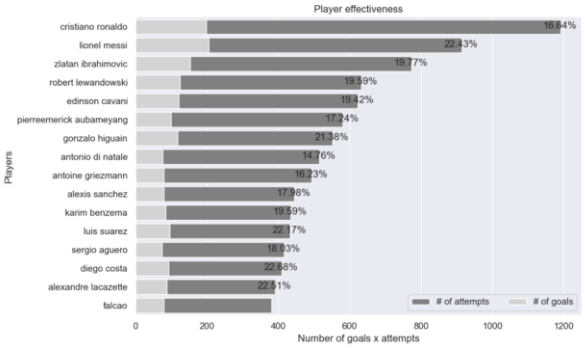

Figure 2: Player effectiveness

4.3 Offensive Teams and Goal-Scoring Techniques

In this part, I examine the offensive capabilities of football teams based on their goal-scoring percentages and techniques. By analyzing the distribution of goals scored using different body parts, situational contexts, and assist methods, I uncover patterns in goal-scoring techniques and strategies. Additionally, I identify the most offensive teams based on their goal-scoring percentages, providing insights into their attacking prowess and strategic approaches. This analysis not only enhances our understanding of goal-scoring dynamics but also offers practical insights for coaches, analysts, and players seeking to optimize their offensive strategies and performance on the field.

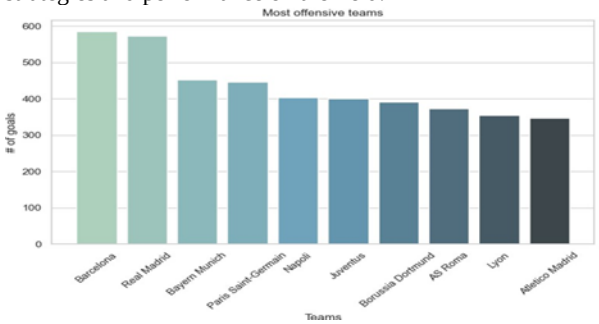

Figure 3: Most offensive teams

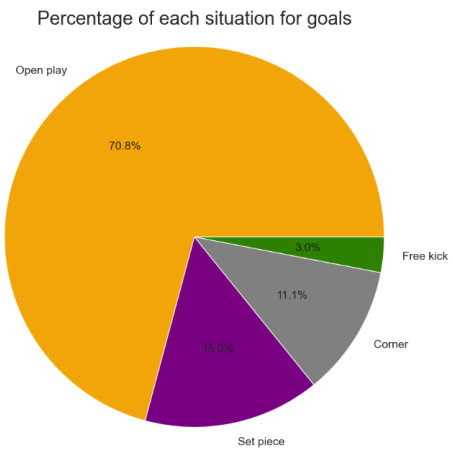

Figure 4: Percentage distribution for each situation of Goals.

5 MACHINE LEARNING MODELS:

The machine learning component of the project focuses on developing predictive models for football goal outcomes using the Football Events dataset. Leveraging Python alongside essential libraries such as NumPy, Pandas, and scikit-learn, various machine learning algorithms are trained and evaluated to predict the likelihood of a shot resulting in a goal during a football match.

5.1 Data Preparation

- The dataset is loaded into a Pandas Data Frame, filtering out non-shot events and irrelevant locations to focus solely on goal-scoring opportunities.
- Relevant features such as time, side, body part, location, situation, assist method, and fast break are selected for model training.
- Categorical variables are encoded using Label Encoder to transform them into numerical representations suitable for machine learning algorithms. [2]
- The dataset is split into training and testing sets using a stratified approach to ensure balanced representation of goal and non-goal instances in both sets.

5.2 Model Training and Evaluation

- I've opted for classification algorithms to predict football goal outcomes. Since the dataset used is labeled, supervised learning techniques are appropriate. Two powerful algorithms, Gradient Boosting Classifier (GBC) and Logistic Regression (LR), were chosen for training on the dataset.
- Gradient Boosting Classifier (GBC) is a boosting ensemble method that iteratively builds multiple decision trees to minimize errors.[3] It's a widely used algorithm known for its high predictive accuracy and robustness.
- Logistic Regression, despite its name suggesting regression, is commonly utilized as a classification algorithm. It models the probability of a binary outcome,[5] making it suitable for binary classification tasks like predicting whether a football shot will result in a goal or not.
- Both GBC and Logistic Regression algorithms are trained to predict the likelihood of a shot resulting in a goal based on the selected features from the dataset. Each algorithm learns patterns and relationships in the data to make accurate predictions.
- To further enhance predictive performance, a Soft Voting Classifier is implemented. This ensemble method combines the predictions of multiple base estimators, namely the GBC and Logistic Regression models, by averaging their probabilities.[10] By leveraging the collective accuracy of these models, the Soft Voting Classifier aims to provide more robust and accurate predictions.
- The Soft Voting Classifier undergoes training on the same training data used for individual model training. Subsequently, it is evaluated alongside the individual models to assess its performance in predicting football goal outcomes. This comprehensive evaluation allows for a comparison between the ensemble approach and the individual algorithms, providing insights into the effectiveness of the combined predictions.

5.3 Model Deployment

- The performance of each individual model (Gradient Boosting Classifier, Logistic Regression) and the Soft Voting Classifier is evaluated using accuracy score metrics on the testing set.
- Accuracy scores provide insights into the effectiveness of the models [4] in predicting goal outcomes.
- The trained Soft Voting Classifier model is serialized using joblib and saved to a file for future use.
- The saved model can be deployed in production environments or integrated into web applications for real-time predictions.

5.4 Model Validation

- Model validation is crucial to ensure the reliability and generalization capability of the trained machine learning

models. In this project, k-fold cross-validation technique is employed for robust model validation.

- K-fold cross-validation involves partitioning the dataset into k equal-sized folds, where each fold is used once as a validation set while the remaining k-1 folds are used for training. This process is repeated k times, with each fold being used as the validation set exactly once.
- By using k-fold cross-validation, we obtain multiple estimates of model performance across different subsets of the data, allowing us to assess the model's stability and generalization ability.
- The accuracy scores obtained from k-fold cross-validation provide a more reliable estimate of the model's performance compared to a single train-test split.
- The k-fold cross-validation process ensures that each data point is used for validation exactly once, leading to a more robust evaluation of the model's performance on unseen data.

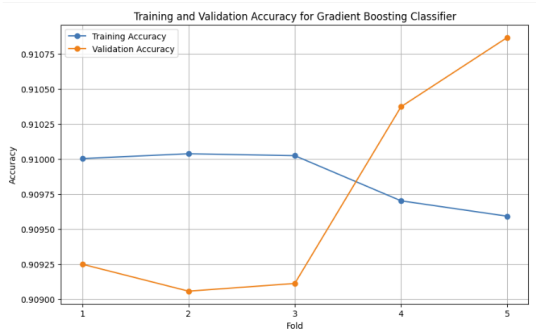

Figure 5: Comparison of Training and validation accuracy for GBC

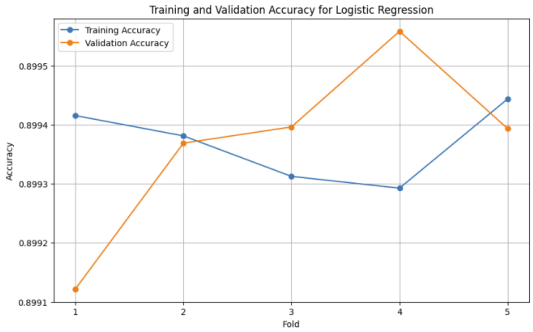

Figure 6: Comparison of Training and validation accuracy for LR

6 WEBSITE DEVELOPMENT:

The website development component of the project focuses on creating an interactive platform to showcase the insights derived from the data analysis and machine learning models. Using the Flask framework alongside HTML, CSS, and Bootstrap components, the goal is to design an intuitive and visually appealing web interface for users to explore football event data and predictions seamlessly.

6.1 Flask Framework:

- Flask, a micro web framework for Python, serves as the backbone of the website development process. Its lightweight and modular design make it ideal for building web applications with minimal overhead.
- Flask enables the creation of routes to handle HTTP requests and render dynamic content, facilitating seamless interaction between the backend and frontend components of the website.[6]

6.2 HTML, CSS, and Bootstrap:

- HTML (Hypertext Markup Language) provides the structure and content of the web pages, defining the layout and organization of elements.[8]
- CSS (Cascading Style Sheets) is used to enhance the visual presentation of the web pages, including styling, colors, fonts, and layout adjustments.[9]

6.3 Integration of Data Analysis and Machine Learning:

- The website integrates the insights derived from data analysis and machine learning models, presenting them in a user-friendly manner for easy exploration and interpretation.
- Visualizations generated during the data analysis phase, including top scorers, most effective players, and percentage distributions, are embedded within the website to provide interactive insights into goal-scoring dynamics.
- Predictions from the trained machine learning models, such as the likelihood of a shot resulting in a goal, are incorporated into the website, allowing users to make informed decisions and explore hypothetical scenarios.

6.4 User Interaction and Navigation:

- The website features intuitive navigation and user interaction elements, including menus, buttons, and search functionality, to enhance the user experience.
- Users can navigate through different sections of the website, explore visualizations, and interact with predictive models to gain insights into various aspects of football analytics.

- Interactive elements such as dropdown menus, sliders, and input forms enable users to customize their experience and delve deeper into specific areas of interest.

6.5 Responsive Design and Accessibility:

- The website is designed with responsiveness in mind, ensuring optimal viewing and usability across a range of devices, including desktops, laptops, tablets, and smartphones.
- Bootstrap, a popular front-end framework, offers pre-designed components and responsive layout utilities, streamlining the development process and ensuring compatibility across different devices and screen sizes.[7]
- Bootstrap components such as navigation bars, cards, buttons, and forms are utilized to design a modern and user-friendly interface for the website.

7 RESULTS AND DISCUSSION:

- Through a comprehensive exploration of football event data and predictive modeling, valuable insights into goal-scoring dynamics, player performance, and strategic decision-making in professional football are uncovered and discussed.

7.1 Data Analysis Insights:

- The data analysis phase reveals various trends and patterns in goal-scoring dynamics, including top scorers across different leagues, distribution of goals by body parts, and situational contexts leading to goals.
- Insights from the analysis reveal that Lionel Messi and Cristiano Ronaldo consistently top the goal-scoring charts across different leagues. However, when it comes to effectiveness, Messi outshines Ronaldo with a better conversion rate. Ronaldo, on the other hand, has more attempts at goal.
- In terms of leagues, the Spanish leagues dominated, with Barcelona and Real Madrid topping the charts, followed by Bayern Munich, PSG, and Napoli. Surprisingly, no English team made it to the top 10 scoring clubs.
- Additionally, the analysis of goal-scoring distribution shows that most goals are scored in open play situations. The most effective assist method is a cross, while the right foot is the most utilized body part for scoring.

7.2 Machine Learning Model Performance

The trained machine learning models, including Gradient Boosting Classifier, Logistic Regression, and Soft Voting Classifier, exhibit promising performance in predicting goal outcomes during football matches. Notably, out of all the models trained, Gradient Boosting Classifier and Logistic Regression achieved the highest accuracies of 91.7% and 90.8%, respectively. Therefore, both models were utilized as base learners, with the Soft Voting Classifier serving as the meta-learner to aggregate their predictions. These accuracy scores obtained from model evaluation underscore the effectiveness of the models in capturing underlying patterns and making accurate predictions based on selected features.

```
Accuracy: 0.9177001307106194
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.93      0.96    183236
           1       0.18      0.61      0.28      4966

    accuracy                           0.92    188202
   macro avg       0.59      0.77      0.62    188202
weighted avg       0.97      0.92      0.94    188202
```

Figure 7: Classification report for GBC

7.3 Website Functionality and User Experience

- The website developed using Flask, HTML, CSS, and Bootstrap components offers an intuitive and visually appealing platform for users to explore football event data and predictions.

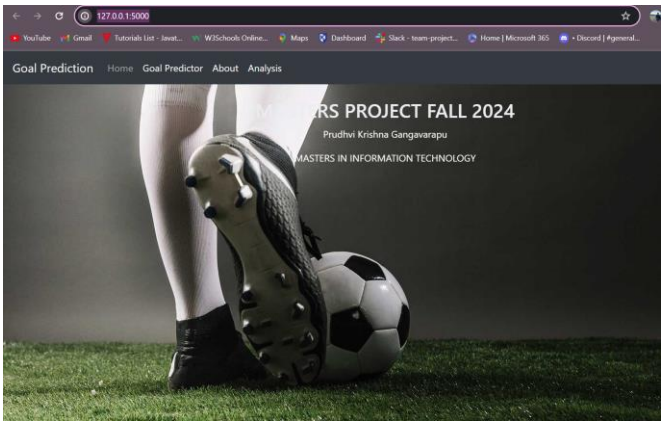

Figure 8: Web Home Page

- Interactive visualizations and predictive models are seamlessly integrated into the website, providing users with an immersive and informative experience.

Figure 9: Goal predictor page

8 LIMITATIONS AND CHALLENGES:

The primary challenge encountered during the project was the cleaning of the dataset, which comprised a substantial 900,000 events. Ensuring data integrity and preparing it for analysis posed a significant hurdle due to the sheer volume of data. Additionally, integrating the trained models into a user interface (UI) presented another notable challenge, requiring efficient implementation and optimization to ensure smooth functionality.

As for the limitations of the model, it is essential to acknowledge its inherent bias towards predicting 'NOGOAL' situations, as the dataset predominantly consists of instances where goals were not scored. Out of the 900,000 rows in the dataset, only approximately 40,000 represent goal situations, which accounts for a mere 1% of the total data. This imbalance may affect the model's performance and accuracy, particularly in scenarios where the goal occurrence is relatively rare.

9 CONCLUSION

In conclusion, this project has successfully leveraged machine learning techniques to develop predictive models for football goal outcomes using the Football Events dataset. Through exploratory data analysis (EDA), various insights into goal-scoring trends and player performances were uncovered, shedding light on the dynamics of goal-scoring in football.

The machine learning models, including Gradient Boosting Classifier, Logistic Regression, and Soft Voting Classifier, demonstrated promising performance in predicting goal outcomes, with the best accuracy achieved by Gradient Boosting Classifier and Logistic Regression at 91.7% and 90.8%, respectively. However, it's crucial to acknowledge the limitations of the model, primarily its bias towards predicting 'NOGOAL' situations due to the significant class imbalance in the dataset.

Despite the challenges faced during data cleaning and model integration into a user interface, the project has provided valuable insights and tools for analyzing and predicting goal outcomes in football matches. Moving forward, addressing the model's limitations, and exploring techniques to mitigate class imbalance will be essential for further enhancing its accuracy and applicability in real-world scenarios.

10 REFERENCES

[1] Prasad Patil. [n. d.]. What is Exploratory Data Analysis? https://towardsdatascience.com/exploratory-data-analysis- 8fc1cb20fd15.

[2] scikit-learn. [n. d.]. Labelencoder.https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

[3] scikit-learn. [n.d]. GBC. https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

[4] scikit-learn. [n. d]. Accuracy. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

[5] Vijay Kanade. [n. d.] What Is Logistic Regression? Equation, Assumptions, Types, and Best Practices https://www.spiceworks.com/tech/artificial intelligence/articles/what-is-logistic-regression

[6] Moraneus [n. d.]. Python Flask: A Comprehensive Guide from Basic to Advanced https://medium.com/@moraneus/python-flask-a-comprehensive-guide-from-basic-to-advanced-fbc6ec9aa5f7

[7] Jigar Shan [n. d.]. bootstrap-vs-angular. https://wpwebinfotech.com/blog/bootstrap-vs-angular/

[8] Mdn web docs. [n. d.]. HTML: HyperText Markup Language https://developer.mozilla.org/en-US/docs/Web/HTML

[9] Dev Mountain. [n. d.] What Is CSS and Why Should You Use It? https://devmountain.com/blog/what-is-css-and-why-use-it/

[10] Ilyas Ahmed [n. d.] What is Hard and Soft Voting in Machine Learning? https://ilyasbinsalih.medium.com/what-is-hard-and-soft-voting-in-machine-learning-2652676b6a32