Evaluating the Predictive Power of Expected Goals (xG) in European Football

Kevin F. Hegarty

Western Governors University

Table of Contents

A. Proposal Overview

A.1 Research Question or Organizational Need

The research question this project aims to solve is: “Is Expected Goals (xG) a better predictor of match outcomes (Win, Draw, Loss) in European football compared to traditional metrics like possession and shots on target?”.

A.2 Context and Background

Understanding the predictive power of performance metrics in football can significantly enhance strategic decisions made by coaches, analysts, and managers. Traditional metrics like possession and shots on target have long been used to gauge team performance, but the xG metric, which quantifies the quality of goal-scoring opportunities, offers a potentially more nuanced insight. By comparing the predictive accuracy of xG with that of traditional metrics, this study aims to provide empirical evidence on the most reliable indicators of match outcomes. Such findings could improve match preparation, player evaluation, and overall tactical planning, ultimately contributing to better performance and results.

A.3 and A3A Summary of Published Works and Their Relation to the Project

Review of Work 1

The first published work reviewed is "Expected Goals in Soccer: An Analysis of Predictive and Descriptive Power" by José M. García-García, Antonio Pérez-Rubio, and Francisco J. Molina-Carmona (2021). This article, published in *PLOS ONE*, provides a thorough examination of the Expected Goals (xG) metric and its effectiveness in predicting football match outcomes. The authors delve into the methodology behind xG, analyzing its predictive power compared to traditional performance indicators such as possession and shots on target. They discuss the strengths and limitations of xG, emphasizing its ability to quantify scoring opportunities more accurately than traditional metrics (García-García, Pérez-Rubio, & Molina-Carmona, 2021).

This study is directly relevant to the research question as it establishes a foundational understanding of how xG can be utilized to predict match outcomes. By highlighting the effectiveness of xG over

traditional metrics, this published work supports the hypothesis that xG is a better predictor of match outcomes. The methodological insights and empirical evidence provided in this paper will inform the data analytics approach used in the project, ensuring that the analysis is grounded in proven techniques.

Review of Work 2

The second work reviewed is "An Examination of Expected Goals and Shot Efficiency in Soccer" by A. Rathke (2017), published in the *Journal of Human Sport and Exercise*. This article explores the correlation between expected goals and actual match results, focusing on how shot efficiency impacts game outcomes. The author provides a detailed statistical analysis, comparing xG with other performance metrics and demonstrating the strong predictive relationship between xG and match results. The study also discusses the implications of shot quality versus shot quantity, providing insights into how teams can optimize their strategies (Rathke, 2017).

This article is important for the project as it provides empirical evidence on the relationship between xG and match outcomes. The author's findings reinforce the hypothesis that xG is a superior predictor.

Review of Work 3

The third work, "A Goal Scoring Probability Model for Shots Based on Synchronized Positional and Event Data in Football (Soccer)" by J. Gudmundsson and M. Horton (2020), published in *Frontiers in Sports and Active Living*, examines different xG models and their applications in football analytics. The authors discuss various methodological approaches for calculating xG and demonstrate how these models can be used to predict match outcomes. The study emphasizes the importance of integrating positional data and event data to enhance the accuracy of xG models. (Gudmundsson & & Horton, 2020).

This work informs the project by explaining advanced xG modeling techniques and their practical applications in predicting football outcomes. The findings from this article support the project's hypothesis and provide a solid framework for evaluating the predictive power of xG compared to traditional metrics.

A.4 Summary of Data Analytics Solution

The proposed data analytics solution is designed to address the research question of whether Expected Goals (xG) is a better predictor of match outcomes (Win, Draw, Loss) in European football compared to traditional metrics such as possession and shots on target. This solution can be realistically implemented, and logically addresses the research question, by utilizing a combination of logistic regression analysis (both binary and multinomial) and correlation analysis. By leveraging these methods, the project will provide empirical evidence on the predictive power of xG compared to traditional metrics. The findings will be detailed in a comprehensive report, complete with statistical summaries and visualizations, ensuring that the insights are accessible and actionable for football analysts and decision-makers.

A.5 Benefits and Support of Decision-Making Process

The proposed data analytics solution significantly enhances the decision-making process for football clubs by improving predictive accuracy and providing actionable strategic insights. Demonstrating that Expected Goals (xG) is a superior predictor of match outcomes compared to traditional metrics enables clubs to rely on more accurate data for strategy formulation. This leads to better preparation for matches, improving on-field performance. Additionally, the project supports enhanced player evaluation, as identifying xG as a key predictor helps clubs make informed decisions regarding player acquisitions, transfers, and development. Players who create higher xG will, in turn, increase the odds of a positive match outcome.

The visualizations generated, including correlation matrices and heatmaps, make complex data accessible and understandable, ensuring findings are effectively communicated to stakeholders. Overall, leveraging the predictive power of xG offers substantial benefits in terms of accuracy, strategic insights, player evaluation, tactical adjustments, and long-term planning, contributing to enhanced performance and success on the field.

B. Data Analytics Project Plan

B.1 Goals, Objectives, and Deliverables

- Goal 1: Determine the most reliable predictor of match outcomes in European football.
    - Objective 1.1: Collect and preprocess data from the top five European leagues.
        - Deliverable 1.1.1: Clean and formatted dataset. This dataset will include comprehensive match-level data such as Expected Goals (xG), possession percentages, shots on target, and match outcomes (win, draw, loss). Ensuring the data is clean and properly formatted is crucial for accurate analysis and subsequent modeling.
    - Objective 1.2: Perform logistic regression analysis.
        - Deliverable 1.2.1: Logistic regression models and summaries. This involves creating binary and multinomial logistic regression models to analyze the relationship between match outcomes and the predictors (xG, possession, shots on target). The regression summaries will provide insights into the significance and impact of each predictor on match outcomes.
    - Objective 1.3: Conduct correlation analysis.
        - Deliverable 1.3.1: Correlation matrices and visualizations. The correlation analysis will examine the linear relationships between xG, possession, shots on target, and match outcomes. Visualizations, such as heatmaps, will help in identifying the strength and direction of these relationships, making it easier to understand the data patterns and dependencies.

B.2 Scope of Project

B.2.A Included in Project Scope

- Data collection from the top five European leagues
- Data preprocessing and cleaning

- Logistic regression analysis (binary and multinomial)
- Correlation analysis
- Reporting and visualization of findings

B.2.B Not included in Project Scope

- Analysis of data from leagues outside the top five European leagues
- Consideration of metrics beyond xG, possession, and shots on target

B.3 Standard Methodology

The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology will be used to organize and implement the project. This includes:

- Business Understanding: Define objectives and requirements.
- Data Understanding: Collect and explore data.
- Data Preparation: Clean and format data.
- Modeling: Apply statistical techniques.
- Evaluation: Assess model performance.
- Deployment: Present findings.

B.4 Timeline and Milestones

Milestone or Deliverable	Duration (days)	Projected Start Date	Anticipated End Date
Data Collection	1	05/27/2024	05/28/2024
Data Cleaning	1	05/28/2024	05/29/2024
Modeling and Analysis	7	06/03/2024	06/10/2024
Evaluation	5	06/10/2024	06/15/2024
Reporting and Presentation	7	06/17/2024	06/24/2024

B.5 Resources and Costs

- Hardware: Standard computer (No cost)
- Software: Python, Jupyter Notebook, pandas, statsmodels, seaborn, matplotlib (No Cost)
- Data: Stathead FBRef data (No Cost)

- Additional: Primarily work hours at No Cost, no additional financial costs anticipated

B.6 Criteria for Success

- **Accuracy of Models:** Statistical significance of predictors
- **Completion of Deliverables:** Timely and comprehensive report
- **Actionable Insights:** Practical recommendations for football clubs

C. Design of Data Analytics Solution

C.1 Hypothesis

Expected Goals (xG) is a better predictor of match outcomes in European football than traditional metrics such as possession and shots on target.

C.2 and C.2.A Analytical Method

C.2 Analytical Method

To test the hypothesis that Expected Goals (xG) is a better predictor of match outcomes in European football than traditional metrics such as possession and shots on target, two primary analytical methods will be used: logistic regression and correlation analysis.

1. Logistic Regression:
    - Binary Logistic Regression: This statistical method will model the probability of a binary outcome (win = 1, not win = 0). By analyzing the relationship between xG, possession, shots on target, and match outcomes, we can determine the significance and impact of each predictor on the likelihood of winning a match. The logistic regression model will be developed using Python and relevant statistical libraries such as statsmodels.
    - Multinomial Logistic Regression: This method extends the binary model to predict all three possible match outcomes (win, draw, loss). It allows for a comprehensive analysis by examining how each predictor influences different match results. The multinomial logistic regression model will also be implemented using Python and statsmodels.
2. Correlation Analysis:

- This method will examine the linear relationships between xG, possession, shots on target, and match outcomes. It helps in understanding the strength and direction of these relationships, providing additional context to the regression analysis. Correlation coefficients will be calculated using Python functions, and visualizations such as heatmaps will be created using seaborn and matplotlib.

C.2.A Justification of Analytical Method

Logistic regression is appropriate for this analysis because it models binary and multinomial outcomes, allowing us to assess the significance and impact of multiple predictors simultaneously. It quantifies the predictive power of each metric and provides coefficients that indicate the strength and direction of relationships between predictors and match outcomes. Correlation analysis complements logistic regression by revealing linear relationships and identifying potential multicollinearity issues, ensuring our regression models accurately reflect underlying data patterns. Together, these methods provide a robust and comprehensive approach to evaluating the predictive power of xG compared to traditional metrics.

C.3 Tools and Environments

Tools: Anaconda, Python, Jupyter Notebook, pandas, statsmodels, seaborn, matplotlib

Third-Party Code: None required.

C.4 and C.4.A Methods and Metrics to Evaluate Statistical Significance

C.4 Methods and Metrics to Evaluate Statistical Significance

To evaluate the statistical significance of our hypothesis that Expected Goals (xG) is a better predictor of match outcomes than traditional metrics, we will use logistic regression and correlation analysis.

1. **Logistic Regression:**
    - **Null Hypothesis:** The null hypothesis (H0) is that xG is not a better predictor of match outcomes compared to possession and shots on target.
    - **Planned Statistical Test**: We will perform both binary and multinomial logistic regression.

- **Binary Logistic Regression:** Models the probability of a binary outcome (win = 1, not win = 0).
- **Multinomial Logistic Regression:** Extends to predict all three match outcomes (win, draw, loss).

- **Metrics Generated:**
  - **Coefficients (β):** Indicates the strength and direction of the relationship between predictors and outcomes.
  - **p-values:** To test the significance of individual predictors.
  - **Log-Likelihood:** Measures model fit.
  - **Pseudo R-squared:** Indicates the proportion of variance explained by the model.
- **Alpha Value (α):** We will use an alpha value of 0.05. If the p-value is less than 0.05, we will reject the null hypothesis, indicating that xG is a significant predictor of match outcomes.

2. **Correlation Analysis:**
   - **Null Hypothesis:** The null hypothesis (H0) is that there is no significant linear relationship between xG, possession, shots on target, and match outcomes.
   - **Planned Statistical Test:** Pearson's correlation coefficient.
   - **Metrics Generated:**
     - **Correlation Coefficients (r):** Indicates the strength and direction of the linear relationship between variables.
     - **p-values:** To test the significance of the correlation coefficients.
   - **Alpha Value (α):** We will use an alpha value of 0.05. If the p-value is less than 0.05, we will reject the null hypothesis, indicating significant correlations between the metrics and match outcomes.

C.4A Justification of Methods and Metrics

Logistic regression is an appropriate choice for this analysis as it allows us to model binary and multinomial outcomes, providing insights into the significance and impact of multiple predictors on match outcomes. By using logistic regression, we can quantify the predictive power of xG compared to traditional metrics, ensuring comprehensive analysis. The p-values and coefficients generated from logistic regression offer clear evidence of the relationship between predictors and outcomes, supporting our hypothesis if xG shows significant predictive power.

Correlation analysis complements logistic regression by revealing the linear relationships between xG, possession, shots on target, and match outcomes. Pearson's correlation coefficient is used for assessing the strength and direction of linear relationships, making it an ideal choice for this analysis. The combination of logistic regression and correlation analysis provides a thorough examination of the data, ensuring that our findings are both statistically and practically significant.

These methods and metrics are chosen because they offer a clear, quantifiable means of evaluating the predictive power of xG. They are well-supported by the literature and are appropriate for addressing the research question, providing a detailed and accurate assessment of how well xG predicts match outcomes compared to traditional metrics.

C.5 Practical Significance

The practical significance of the data analytics solution provided in this project will be shown through improvement in predictive accuracy, practical applicability of insights for decision-making in football strategy. In practice, this may be evaluated through feedback from football analysts, and comparative analysis of pre-and post-implementation outcomes.

C.6 Visual Communication

The project report will include graphical visualizations that effectively communicate the findings of the data analytics solution. These visualizations will help stakeholders easily understand the relationships between the variables and the results of the statistical analyses.

Planned Visualizations:

1. **Correlation Matrix Heatmap:**
    - Type of Graph: Heatmap
    - Purpose: To visualize the linear relationships between the predictors (xG, possession, shots on target) and match outcomes.
    - Details: The heatmap will display the correlation coefficients between the variables, with color gradients indicating the strength and direction of the correlations. This visualization helps in identifying significant relationships and potential multicollinearity issues.
    - Tools: The heatmap will be generated using Python libraries such as seaborn and matplotlib. Seaborn provides a high-level interface for attractive and informative graphics, while matplotlib allows for detailed customization of the plots.
2. **Predicted Probabilities Plot from Logistic Regression:**
    - Type of Graph: Line Plot
    - Purpose: To visualize the predicted probabilities of match outcomes based on xG, possession, and shots on target.
    - Details: This plot will show the predicted probabilities of winning, drawing, or losing a match as a function of xG values. It can help in understanding how changes in xG affect the likelihood of different match outcomes. The line plot will depict the probabilities over a range of xG values.
    - Tools: The plot will be created using Python libraries such as matplotlib and statsmodels. Matplotlib will be used for plotting the data, while statsmodels will provide the regression results and predicted probabilities.

These visualizations will be included in the project report to provide a clear and comprehensive representation of the analysis results. The correlation matrix heatmap will offer a quick overview of the relationships between the variables, while the predicted probabilities plot will provide deeper insights into

the logistic regression model's performance and the predictive power of xG. Together, these graphical representations will effectively communicate the key findings and support informed decision-making.

D. Description of Dataset

D.1 Source of Data

The dataset was sourced from the publicly available football statistics website FBRef.com by Sports Reference, which provides comprehensive match-level data for European leagues. The data includes match statistics and outcomes necessary for the analysis.

D.2 Appropriateness of Dataset

The data from FBRef contains comprehensive match-level data required for the analysis, including xG, possession, shots on target, and match outcomes.

D.3 Data Collection Methods

The data being used is from FBref.com and collected using their built-in export functionality. Raw data is exported as .csv(s) from https://stathead.com/tiny/iDceP .

D.4 Observations on Quality and Completeness of Data

The data is high quality and clean, with detailed and complete records for the required metrics.

D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances

Data Governance:

Data governance will be ensured by maintaining structured data management practices. This is essential for reliable analysis and results.

Privacy:

The dataset contains publicly available match-level statistics and no personal data.

Security:

The data is publicly available and not stored in encrypted storage, however, access to the analysis will be controlled to prevent unauthorized use, ensuring data integrity and appropriate handling.

Ethical, Legal, and Regulatory Compliance:

Compliance with all relevant legal and regulatory requirements, including copyright laws and data usage policies specified by FBRef.com will be followed. Their policy includes encouragement to use their data and for projects including academic research and projects. Ethical standards will be maintained by using data responsibly and reporting findings accurately.

References

García-García, J. M., Pérez-Rubio, A., & Molina-Carmona, F. J. (2021). Expected goals in soccer: An analysis of predictive and descriptive power. *PLOS ONE*. Retrieved from https://doi.org/10.1371/journal.pone.0282295

Gudmundsson, J., & & Horton, M. (2020). A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). *Frontiers in Sports and Active Living*. Retrieved from https://www.frontiersin.org/articles/10.3389/fspor.2021.624475/full

Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise*. Retrieved from https://rua.ua.es/dspace/handle/10045/68771?locale=en