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

Kevin F. Hegarty

Western Governors University

Table of Contents

A. Project Highlights

Research Question or Organizational Need

The research question addressed by this capstone project 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?"

Scope of Project

The scope of this project included collecting and analyzing data from the top five European football leagues to compare the predictive power of xG with traditional metrics. The project involved data preprocessing, logistic regression analysis (binary and multinomial), correlation analysis, and visualization of results to determine the most reliable predictor of match outcomes.

Overview of Solution

The solution involved using Python for data collection, cleaning, and analysis. The key methodologies included logistic regression models to assess the impact of xG, possession, and shots on target on match outcomes, and correlation analysis to explore the relationships between these variables. The tools used were Python libraries such as pandas, statsmodels, seaborn, and matplotlib.content.

B. Project Execution

Project Plan

The project followed the initial plan outlined in the Project Proposal, with minor adjustments to accommodate additional data cleaning steps and model tuning based on initial findings.

Project Planning Methodology

The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology was followed, ensuring a structured approach to business understanding, data preparation, modeling, evaluation, and deployment.

Project Timeline and Milestones

The project timeline ultimately got accelerated due to early data collection and cleaning and prompt approval of the Project Proposal. All major milestones such as data collection, modeling, and reporting were completed within the following updated timeline:

Milestone or Deliverable	Duration (days)	Start Date	End Date
Data Collection	0.5	05/20/2024	05/20/2024
Data Cleaning	0.5	05/20/2024	05/20/2024
Modeling and Analysis	2	05/21/2024	05/23/2024
Evaluation	0.5	05/23/2024	05/23/2024
Reporting and Presentation	1	05/23/2024	05/24/2024

C. Data Collection Process

Data Selection and Collection

The data selection and collection process remained consistent with the plan in the Project Proposal. Data was sourced from FBRef using their export functionality. The dataset included match statistics and outcomes necessary for the analysis.

Handling Obstacles

A minor obstacle was using FBRef's export functionality. Using their tool, it is limited to export only the records shown on the screen, i.e. about 200 rows of data. This added to time anticipated for data collection due to the need to export 19 data files to get the entire dataset, and then include a cleaning step to combine all the files for analysis.

Data Governance Issues

No significant unplanned data governance issues were encountered. All data handling complied with the relevant guidelines and policies. FBRef data is publically available and can be used for research without restriction as long as they are credited with ownership.

C.1 Advantages and Limitations of Data Set

An advantage of using this dataset was its completeness. Every match over the 2022-2023 season for the top 5 European Leagues had an individual row, and all rows had all the information needed to complete the analysis.

A disadvantage of the dataset was that the result of each match were recorded as string that included both the score and the outcome. The outcome had to be extracted and mapped to a numerical value for meaningful analysis.

D. Data Extraction and Preparation

Data was extracted directly from FBRef using their export functionality, ensuring efficiency and accuracy. Preparation involved minor cleaning, formatting, and standardizing the data for consistency, as well as the mapping exercise of results as outlined in Part C.1. Python libraries such as pandas were used for these processes, which were appropriate given the structured nature of the data and the need for manipulation and analysis.

E. Data Analysis Process

E.1 Data Analysis Methods

- **Logistic Regression (Binary and Multinomial):** Used to model the probability of match outcomes (win, draw, loss) based on predictors (xG, possession, shots on target).
- **Correlation Analysis:** Used to examine the linear relationships between predictors and match outcomes.

Both methods were appropriate for understanding the predictive power of xG compared to traditional metrics.

E.2 Advantages and Limitations of Tools and Techniques

Advantages:

- Logistic regression provides a clear understanding of predictor significance.
- Correlation analysis helps identify relationships and multicollinearity.

Limitations:

- Logistic regression assumes a linear relationship between predictors and log-odds, which may not always hold.
- Correlation analysis only captures linear relationships.

E.3 Application of Analytical Methods

Logistic Regression:

- Steps: Model fitting using **statsmodels**, interpreting coefficients, p-values, and pseudo R-squared values.
- Requirements: Linear relationship assumption, binary/multinomial outcome.
- Verification: Checking model diagnostics and goodness-of-fit measures.

Correlation Analysis:

- Steps: Calculating correlation coefficients using **pandas**, visualizing with heatmaps.
- Requirements: Continuous variables.
- Verification: Ensuring significant p-values for correlations.

F Data Analysis Results

F.1 Statistical Significance

To evaluate the statistical significance of our analysis, we employed both binary and multinomial logistic regression models, as well as correlation analysis.

Binary Logistic Regression

- Null Hypothesis (H0): Expected Goals (xG) is not a better predictor of match outcomes compared to possession and shots on target.
- Statistical Test: Binary logistic regression.

- Metrics Generated:
    - Coefficients (β): Indicates the impact of each predictor on the log-odds of winning a match.
    - p-values: Assess the statistical significance of each predictor.
    - Log-Likelihood: Measures the model's goodness of fit.
    - Pseudo R-squared: Indicates the proportion of variance explained by the model.
- Alpha Value (α): 0.05.
- Results:
    - xG: Coefficient = 1.0000, p-value < 0.0001. This indicates a significant positive relationship between xG and the probability of winning.
    - Possession: Coefficient = -0.0244, p-value < 0.0001. This shows a slight negative impact on the probability of winning, though much less influential than xG.
    - Shots on Target: Coefficient = 0.2478, p-value < 0.0001. This suggests a significant positive relationship with the probability of winning.
- Conclusion: The p-values for xG, possession, and shots on target are all less than 0.05, allowing us to reject the null hypothesis. There is sufficient evidence to support the hypothesis that xG is a better predictor of match outcomes.

Multinomial Logistic Regression

- Null Hypothesis (H0): Expected Goals (xG) is not a better predictor of match outcomes compared to possession and shots on target.
- Statistical Test: Multinomial logistic regression.
- Metrics Generated:
    - Coefficients (β): Indicates the impact of each predictor on the log-odds of each match outcome (win, draw, loss).
    - p-values: Assess the statistical significance of each predictor for each outcome.
    - Log-Likelihood: Measures the model's goodness of fit.

  - Pseudo R-squared: Indicates the proportion of variance explained by the model.
- Alpha Value (α): 0.05.
- Results:
  - xG: Significant positive coefficients for predicting wins and draws, with p-values < 0.0001.
  - Possession: Varying impact, less significant than xG.
  - Shots on Target: Significant positive impact on winning, with p-values < 0.0001.
- Conclusion: The results reinforce the binary logistic regression findings, supporting the hypothesis that xG is a superior predictor of match outcomes. This model adds depth by demonstrating the predictive power of xG across multiple outcome categories (win, draw, loss).

Correlation Analysis

- Null Hypothesis (H0): There is no significant linear relationship between xG, possession, shots on target, and match outcomes.
- Statistical Test: Pearson's correlation coefficient.
- Metrics Generated:
  - Correlation Coefficients (r): Indicates the strength and direction of the linear relationship between variables.
  - p-values: Assess the statistical significance of the correlations.
- Alpha Value (α): 0.05.
- Results:
  - xG and Match Outcome: Moderate positive correlation (r = 0.42), p-value < 0.0001.
  - Possession and Match Outcome: Weak positive correlation (r = 0.066), p-value < 0.0001.
  - Shots on Target and Match Outcome: Moderate positive correlation (r = 0.39), p-value < 0.0001.
- Conclusion: Significant correlations between xG and match outcomes support the hypothesis, confirming that xG is a more reliable predictor.hypothesis.

The correlation analysis for the multinomial logistic regression also highlights the relationships between xG, possession, shots on target, and the different match outcomes (win, draw, loss). This analysis adds to the discussion by providing a broader perspective on how these predictors influence all possible match results, further supporting the overall findings of the logistic regression models.

F.2 Practical Significance

The practical significance of the data analytics solution is evident in its ability to provide actionable insights for football clubs. The results demonstrate that xG is a superior predictor of match outcomes, which can directly influence strategic decisions in football management.

- **Predictive Accuracy:** The high significance of xG in predicting match outcomes suggests that teams should focus on increasing their xG during matches. This can lead to better offensive strategies and ultimately more wins.
- **Application Example:** A football club can use the findings to enhance their training programs, emphasizing scenarios that increase xG, such as creating high-quality goal-scoring opportunities rather than merely increasing possession or the number of shots on target.

These practical applications underscore the value of integrating xG into performance metrics, leading to improved match preparation and strategic planning.

F.3 Overall Success

The project was successful in demonstrating that Expected Goals (xG) is a better predictor of match outcomes compared to traditional metrics like possession and shots on target. The logistic regression models and correlation analysis provided robust statistical evidence supporting this hypothesis. The practical significance of the findings was demonstrated through actionable insights for football clubs, confirming the overall effectiveness of the project. This success validates the hypothesis and provides a strong foundation for future applications of xG in football analytics.

G. Conclusion

G.1 Summary of Conclusions

The primary goal of this project was to determine if Expected Goals (xG) is a better predictor of match outcomes in European football compared to traditional metrics like possession and shots on target. Our analysis, which included both binary and multinomial logistic regression models as well as correlation analysis, demonstrated that xG is indeed a stronger predictor of match outcomes. The logistic regression results showed that xG had a significant and positive impact on the probability of winning a match. These findings align with our hypothesis and suggest that xG provides a more accurate measure of a team's performance than traditional metrics.

G.2 Effective Storytelling

To effectively communicate our findings, two main visualizations were utilized: a correlation matrix heatmap and a predicted probabilities plot from the logistic regression model. The correlation matrix heatmap illustrated the linear relationships between xG, possession, shots on target, and match outcomes, highlighting significant correlations. The predicted probabilities plot provided a clear visualization of how xG influences the likelihood of winning, demonstrating the logistic regression model's predictive power. These visualizations, created using Python libraries such as **seaborn** and **matplotlib**, were essential in making complex statistical results accessible and understandable to stakeholders.

G.3 Recommended Courses of Action

Based on the results of our analysis, we recommend the following actions:

1. **Integrate xG into Team Strategies:** Football clubs should incorporate xG into their performance analysis and decision-making processes. This will allow them to focus on creating high-quality scoring opportunities, which are more strongly correlated with positive match outcomes.
2. **Revise Training Programs:** Coaches should emphasize drills and strategies that increase xG during training sessions. By focusing on the quality of scoring opportunities rather than just possession or number of shots, teams can improve their chances of winning matches.

H Panopto Presentation

D195 - xG Analysis Capstone Panopto Presentation

References

No sources were cited.

Appendix A

Additional Files

Project Code

Included with this report is a copy of the project code, completed in a Jupyter Notebook, in an .html format as well as .ipynb format.

- D195 – xG Analysis.html
- D195 – xG Analysis.ipynb

Data Set

Included with this report is a copy of the data set in .csv format.

- team_match_stats_22-23_1.csv
- team_match_stats_22-23_2.csv
- team_match_stats_22-23_3.csv
- team_match_stats_22-23_4.csv
- team_match_stats_22-23_5.csv
- team_match_stats_22-23_6.csv
- team_match_stats_22-23_7.csv
- team_match_stats_22-23_8.csv
- team_match_stats_22-23_9.csv
- team_match_stats_22-23_10.csv
- team_match_stats_22-23_11.csv
- team_match_stats_22-23_12.csv
- team_match_stats_22-23_13.csv
- team_match_stats_22-23_14.csv
- team_match_stats_22-23_15.csv
- team_match_stats_22-23_16.csv
- team_match_stats_22-23_17.csv
- team_match_stats_22-23_18.csv
- team_match_stats_22-23_19.csv

Data Source

The source for the above Data Set files was a generated report from FBRef by Sports Reference at the following link: https://stathead.com/tiny/iDceP