AFCON Match Outcome Prediction with Logistic Regression

Mini Presentation

Theophilus Dwamena Frimpong

Supervisor
Issa Karambal, Ph.D.

African Institute for Mathematical Sciences

Outline of Presentation

- Introduction
- Data Preparation
- Logistic Regression Model
- Model Evaluation
- Future Predictions
- Conclusion and Recommendations

Introduction

Background of Study

The African Cup of Nations (AFCON) stands as a premier football competition in Africa, captivating millions of fans worldwide. The diversity of participating teams and the unpredictability of match outcomes make AFCON a compelling field for sports analytics.

Objectives

1. Identify features that affects match outcomes
2. Predict future match outcomes using logistic regression

Data Preparation

The Dataset

- The dataset was extracted from 65 games and comprised 22 features along with 1 outcome variable.
- These features included pre-game and in-game statistics and their values were primarily obtained from FIFA's official website [3] and SofaScore [4].
- There were no missing values in the dataset.
- Python programming language (Version 3.9.2)

Feature Engineering

Features that could impact match results were identified, encoded, and checked for multicollinearity, leading to the removal of highly correlated features while observing causality. [1]

Data Preparation

Features

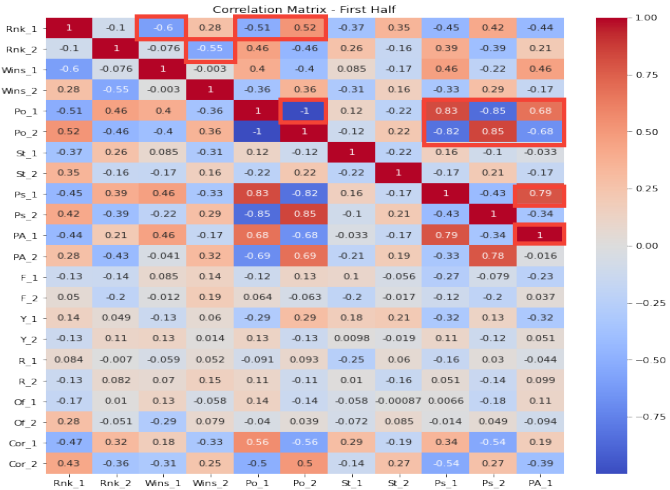
Labels	Description	Type
Rnk_1,2	FIFA ranking prior to the game	numeric
Wins_1,2	Wins in last 5 competitive matches before the game	numeric
Po_1,2	Possession	numeric
St_1,2	Shots on target	numeric
Ps_1,2	Passes	numeric
PA_1,2	Pass Accuracy	numeric
F_1,2	Fouls	numeric
Y_1,2	Yellow Cards	numeric
R_1,2	Red Cards	numeric
Of_1,2	Offside	numeric
Cor_1,2	Corners	numeric
Outcome	Win or Not win in 90 mins	category

Table: Selected Features That Affects Match Outcomes

Data Preparation

Multicollinearity Check

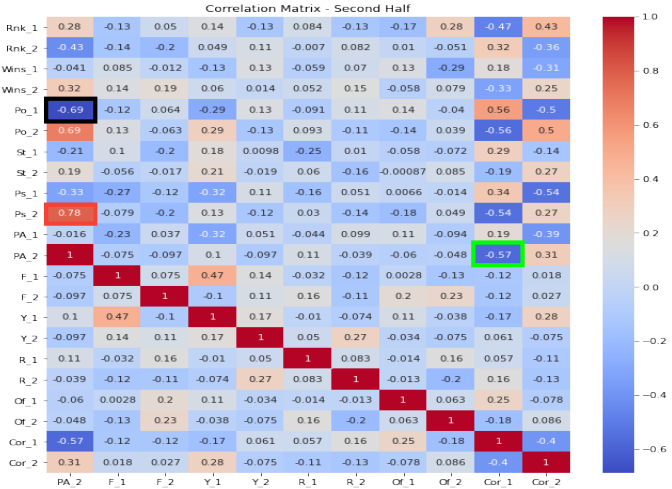
Correlation Heat Map 1



Data Preparation

Multicollinearity Check

Correlation is not causality.



Data Preparation

16 Selected Features

Labels	Description	Type
Wins_1,2	Wins in last 5 competitive matches before the game	numeric
St_1,2	Shots on target	numeric
PA_1,2	Pass Accuracy	numeric
F_1,2	Fouls	numeric
Y_1,2	Yellow Cards	numeric
R_1,2	Red Cards	numeric
Of_1,2	Offside	numeric
Cor_1,2	Corners	numeric
Outcome	Win or Not win	category

Table: Selected Features That Affects Match Outcomes

Data Preparation

Data Sampling

The outcome training set was highly imbalanced with 40 (61.5%) wins and 25 (38.5%) not wins so SMOTE was used to address the issue of class imbalance.

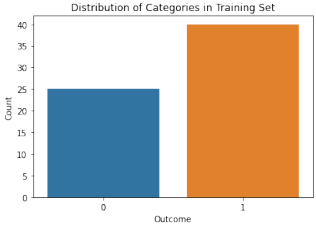


Figure: Unbalanced Training Set

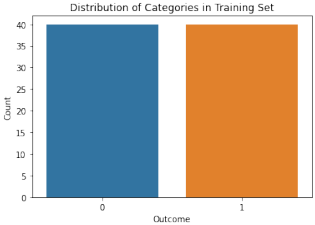


Figure: Balanced Training Set

Logistic Regression

Logistic Regression

Logistic Regression is a classification algorithm used when the dependent (Outcome) variables are categorical and binary in nature.[2]

The logistic Regression Equation

$$\ln\left[\frac{p(y)}{1-p(y)}\right] = \sum_{i=0}^{n} \beta_i x_i$$

Where;
β_i are the coefficients of the model
x_i are the predictor variables
y is the binary outcome variable

The model was regularized to reduce overfitting, and the dataset was randomly split into an 80% training set and a 20% validation set in a cross-validation process for model selection.

Model Evaluation

Confusion Matrix and Evaluation Metrics

The confusion matrix is a table that assesses a classification model's performance.

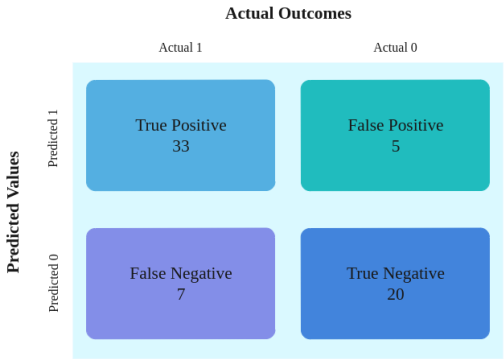


Figure: Confusion Matrix

- **Accuracy** $= \dfrac{33+20}{65} = 0.815$
- **Precision** $= \dfrac{33}{33+5} = 0.868$
- **Recall** $= \dfrac{33}{40} = 0.825$
- **F1 Score** $= \dfrac{2 * 0.868 * 0.825}{0.868 + 0.825} = 0.846$

Future Predictions

In-game Features Problem

In-game features may not be accessible before the game for future predictions. The average of all features across the tournament will ensure unbiased predictions for each team's in-game features.

Future Predictions

In-game Features Problem

AFCON Finals Predictions (Estimate)

Tm1	Tm2	Wins		St		PA		F		Y		R		Of		Cor		Outcome	
		1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	Pred	Act
NGA	RSA	4	3	3.2	5	77.8	79.6	17.8	15	1.6	1.2	0	0	3.2	2	3.4	4	0	0

Table: Model Predictions

AFCON Finals Predictions (Actual)

Tm1	Tm2	Wins		St		PA		F		Y		R		Of		Cor		Outcome	
		1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	Pred	Act
NGA	RSA	4	3	3	5	80	87	10	21	2	0	0	0	0	1	4	2	0	0

Table: Model Predictions

Conclusion and Recommendations

Conclusion

In conclusion, the model demonstrated promising results with high evaluation metrics. However, it is important to acknowledge the uncertainties inherent in football, limiting the model's ability to predict outcomes with absolute certainty.

Limitations

- The model does not capture draw or away team victories.
- The data collected may not cover all relevant factors influencing outcomes.

Recommendations

Consider employing and comparing various machine learning models to identify the most suitable approach for AFCON match outcome prediction.

References

- Maalouf, Maher. (2011). Logistic regression in data analysis: An overview. International Journal of Data Analysis Techniques and Strategies. 3. 281-299. 10.1504/IJDATS.2011.041335.
- Borucka, Anna. (2020). Logistic regression in modeling and assessment of transport services. Open Engineering. 10. 26-34. 10.1515/eng-2020-0029.
- FIFA, www.fifa.com.
- SofaScore, www.sofascore.com.

Thank you

“All models are wrong, but some are useful”
George E. P. Box