

An Immovable Object Meets an Unstoppable Force:

Does Defense Produce Success in Football?

Kyle Tan (500564538)

Supervised by Tamer Abdou

2/xxx/2023

Table of Contents

[Revised Abstract 4](#_Toc128570856)

[Foreword: 4](#_Toc128570857)

[Abstract: 4](#_Toc128570858)

[Literature Review 5](#_Toc128570859)

[Defining Research Questions: 5](#_Toc128570860)

[Proposed Methodology 6](#_Toc128570861)

[Bibliography 17](#_Toc128570862)

[Proposed Roadmap – Extremely General 18](#_Toc128570863)

# Revised Abstract

## Foreword:

This project a live work in progress, as such, things change. New information may come to light, old assumptions may be put into question. As more research, the initial exploratory analysis, and a more critical look at the information provided is completed, so changes some of the details previously discussed in the abstract. Below is a revised project abstract, with respect to exploratory analysis, overall report structure, and critical reading on existing works.

## Abstract:

“Attack wins you games, defense wins you titles”, these are the words of one of the most recognized football managers, Sir Alex Ferguson (Smith, 2017). This theory is an often-said rhetoric that spans all sorts of sports (Davis & Suryawanshi, 2023). However, with the scope of this project, this theory will be analyzed through the scope of international football (soccer). It needs to be recognized that the psychology, challenges, and difficulties in each different competition varies quite a bit. League football, for example, takes place over months, with different levels of support, control, time, and accountabilities than cup football. Time has passed since Sir Alex said the famous quote. Since then there has been major developments and changes in the dogma of both club and international football, from the rise of possession seen in Vincent Del Bosques’ successful Spain side (MARCA, 2012) to the rise in popularity and success in positional play seen in Pep Guardiola’s Manchester City (Breaking the Lines, 2022) all of which leads away from a traditionally “defensive” mindset. Despite the success and popularity of these developments, we saw in the most recent world cup, teams with fantastic defense records like Argentina and Morocco saw success and trophies (FOX Sports, 2023). The question this project aims to answer is: does the popular rhetoric still apply to today’s football? Using data provided from the European Soccer Database from Kaggle, this project will aim to use predictive analytics techniques to aim and see if a defensive team means success/higher chances of winning.

# Literature Review

## Defining Research Questions:

This project aims to answer one simple question: Does defense win titles in the context of international football? With more research done into previously completed projects and works, the question needs be further fine tuned and expanded.

LITERATURE REVIEW AS PER COURSE MODULE:

* What do you already know about the topic?

Personal knowledge of this topic comes from previous attempts at this project and purely rhetoric. From my previous attempts at the project and the extremely simple and poor results, chances are success is created from a balance of both attack and defence not a sole focus on one or the other.

Defensive performance is not a easily definable and measurable item for football. A major challenge would be to create something useable to measure defensive performance. As defensive statistics for soccer is an incredibly complex and understudied area. There has been attempts to develop a sound view to look at defensive by previous experts (Winterburn, 2017) (Brownell, 2013), they all agree on one fact: there is much works to be done, and there’s much more than meets the eye.

* What do you have to say critically about what is already known?

Research on defensive performance and machine learning in the field of football is fairly limited, most work is done on the attacking side of the game (Merhej, Beal, Ramchurn , & Matthews, 2021). However, there are researchers that have tackled the topic.

One proposed analysis of looking at defenders and rating defense is through using a GNN (graph convolutional neural network) to sort unstructured data to model certain defensive behaviours (Stöckl, 2021). The researchers trained a GNN model based on real time data using certain predictors they have created. They are able to create some good graphical analysis of defence. However with my limited time and skillset, and the lack of source code on their project, I am unable to create something similar.

Merhej et al.’s research has provided a solid foundation on using deep learning to value defensive actions. By predicting what is going to be stopped they are able to model what they call the DAxT measure (defensive action expected threat).

Critically, what I have to say about the overall topic is that research done is extremely specific and niche. It’s also a relatively new subject where there doesn’t seem to be a consensus on what is the one true measure for defense.

* Has anyone else ever done anything exactly the same?

There are research into looking at defence in football. But not yet any documents I’m aware of for the usage of seeing if defense truly wins through a data perspective.

* Has anyone else done anything that is related?

Yes, there are a lot of machine learning projects done on the topic of defensive and football. A lot of the work done is specific to other areas of the game like the offensive sides, injuries (Rossi, et al., 2018).

Most of the work are predictive modelling. For example this following paper that looks to use a set of attributes to help predict transfers using Random Forest, Naïve Bayes,and AdaBoost algorithms (Ćwiklinski, Gielczyk, & Choras, 2021). I plan to take a similar approach, whereas their goal is to create a good predictive model. My goal is to look at which variables are most impactful.

The positive side is that there are a lot of machine learning work done, but just not for my specific topic.

* Where does your work fit in with what has gone before?

My work here aims to further explore the field of defense in the context of world football. It will aim further help understand defense and which ones are key to a team’s success.

* Why is your research worth doing in the light of what has already been done?

The main goal is to develop a further understanding of the game. No project has really looked at predicting success. Most research in relations to defence is based on creating a good measure/predictor for defence.

# Proposed Methodology

Previous attempts and methodology is removed and ignored as database was not large enough.

New proposed methodology using data from <https://www.kaggle.com/datasets/hugomathien/soccer?datasetId=63&sortBy=voteCount>

I will load in all match data from the top division in England, France, Germany, Italy, Spain.

The predictor I will use for success will be based on the available betting odds, as low return odds generally mean a favorable match and I will use this as my measure of “success”. This method is a very general relation based on predictions made by betting companies.

I hypothesize little link between most of the predictive stats and the chances of success. There are other experts that also have said football is one of the more unpredictive sports (Anderson & Sally, 2013).

A team will be defined as successful it they’re predicted to win their matches (defined by low odd returns) and are considered unsuccessful if their victory is unpredictable and less likely (defined by high odd returns).

For the sake of this project, defense is defined by the traditional sense that a team will look to sit back in their own half, allow the opposition to be on the ball. The main aim not to create and score, but rather not allow the opposition to score. This can be defined as teams with low build up speeds, higher defensive aggression, lower defensive team width, plays less expressively and plays conservatively.

As the original database is a SQLLite file, I will be taking inspiration from: <https://www.kaggle.com/code/yonilev/the-most-predictable-league/notebook> on how to read in and organize the data. I will not using their entropy analysis but simply how they read in the data

As recommended from project feedback I will be initially using filter-based, wrapper, and embedded/hybrid techniques to gain a greater understanding of the input variables and detect any patterns that may arise.

STEPS TAKEN:

1. Generate a output variable “home win odds” that will show the respective odds of winning of each home game in the top division of England, France, Germany, Italy, Spain (the big 5) from the years 2008 – 2016.
   1. Subset for home win odds.
   2. Take a look at EDA to see NA.
      1. Following chart summarizes NA
      2. To deal with any missing values companies with anything above 1% missing is removed form consideration.
   3. With the new subset, missing values are simply removed as they’re fairly insignificant(<1%)
      1. 14585  prior to dropping na
      2. 14503  after dropping na
      3. 99% of data is still captured.

|  |  |
| --- | --- |
| Attribute Name | NA |
| B365H | 0.1% |
| BWH | 0.2% |
| IWH | 0.3% |
| LBH | 0.01% |
| PSH | 50% |
| WHH | 0.01% |
| SJH | 24.1% |
| VCH | 0.2% |
| GBH | 37.7% |
| BSH | 37.7% |

1. Generate a new column “average betting odds” so that the output variable is univariate that captures a “consensus by betting experts on who’s likely to win”.
2. I will then take a look at the input variables which I will pick from team Attributes table of the database (as opposed to individual data). The goal here is to find which ones can help predict average betting odds best and see its relations to a defensive style of play.
3. From EDA: Data here is both categorial and numerical.
4. I will split it into two data frames so they’re not mixed up.
   * 1. One numeric attribute (buildUpPlayDribbling has a very high missing value % of 66.7%. This attribute will simply be dropped.
   1. Brief data description from EDA:

Graphical user interface, text, application, email

Description automatically generated

Above image shows EDA warnings on numData – none of the warnings are in relations to the predictive input stats. High correlation is showing the betting companies agree with each other the outcomes. Other warnings are mostly included on variables that are used for identification (IDs, dates, names) rather than as an input for prediction.

No outliers are detected.

Table

Description automatically generated

Again, no outliers are detected,

This longer list is the warnings of the EDA for the categorical data. All previous issues with the identification variables (IDs, etc.) remain true. However, this time there is a great imbalance in distribution for many of the variables. Showing two examples below, normality is not a given:

A picture containing graphical user interface

Description automatically generated

This does show one striking piece of information. There is a lot of agreement between different teams on they play. It seems most teams play one way and there’s not a of variation.

Shapiro Wilkes performed on output statistics avgOdds. It’s safe to say this is not a normal distribution. As such non-parametric tests should be used.



Normalization is not done on the numerical data as the data is all already on a scale of 0 – 100. As for the Categorical data, a numerical ranked categorical data frame is also created to convert the string data into ranked numbers with a higher ranking representing an offensive style of play.

For multicollinearity:

Below are the two charts for the correlations between all variables on both the numerical dataset and the categorial. As previously mentioned. The only high correlation belong to the betting odds (which indicate agreement, to deal with this an average is taken), the irrelevant variables (id, teamname, date, etc.). The variables that are to be the input, do not show any hint of high levels of positive nor negative correlations amongst themselves. At least, not to the extend where the EDA reported needed to warn me.

Having looked at the various ‘interactions’ between input variables on the EDA reports, there also seems to show little to no correlation between the inputs.

Chart

Description automatically generated

Chart, waterfall chart

Description automatically generated

Below is the scatterplots created for numData data frame. From the quick observation below, it’s safe to say linearity cannot be assumed for the numData.

Graphical user interface, application

Description automatically generated

As catData is categorical, linearity is not checked. ANOVA will be done.

Next set of charts show the fitted values vs residuals and this can tell me about the assumption of homoscedasticity. The left chart shows catData and the spread would suggest heteroscedasticity. The right chart shows numData and the spread would also suggest heteroscedasticity.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

1. Merge
   1. Categorical data frame with average odds.
   2. Numerical data frame with average odds.
      1. Merge will be based on home team ID. This attribute will be rename homeTeamID for easy merging for all data frames.
2. Use filter based, wrapper, and embedded/hybrid techniques to choose good predictors.

Influences taken from links below:

* 1. <https://sebastianraschka.com/faq/docs/feature_sele_categories.html>
  2. <http://sigmaquality.pl/models/feature-selection-techniques/feature-selection-by-filter-methods_-categorical-input-categorical-output-260320201223/>
  3. <https://github.com/codingnest/FeatureSelection/blob/master/Data%20Science%20Lifecycle%20-%20Feature%20Selection%20(Filter%2C%20Wrapper%2C%20Embedded%20and%20Hybrid%20Methods).ipynb>
  4. <https://ademos.people.uic.edu/Chapter22.html>
  5. **Filter based**
     1. Correlation Coefficient – done on num data
        + Pearson

|  |  |
| --- | --- |
| Attribute name | Correlation to avgOdds |
| **buildUpPlaySpeed** | 0.003013 |
| **buildUpPlayPassing** | 0.112958 |
| **chanceCreationPassing** | -0.015613 |
| **chanceCreationCrossing** | 0.015509 |
| **chanceCreationShooting** | -0.064868 |
| **defencePressure** | -0.090507 |
| **defenceAggression** | -0.038118 |
| **defenceTeamWidth** | -0.030806 |

Kendall

|  |  |
| --- | --- |
| Attribute name | Correlation to avgOdds |
| **buildUpPlaySpeed** | 0.019835 |
| **buildUpPlayPassing** | 0.122729 |
| **chanceCreationPassing** | -0.026335 |
| **chanceCreationCrossing** | 0.006865 |
| **chanceCreationShooting** | -0.088601 |
| **defencePressure** | -0.119189 |
| **defenceAggression** | -0.040733 |
| **defenceTeamWidth** | -0.045314 |

Spearman\* non-parametric

|  |  |
| --- | --- |
| Attribute name | Correlation to avgOdds |
| **buildUpPlaySpeed** | 0.029654 |
| **buildUpPlayPassing** | 0.179454 |
| **chanceCreationPassing** | -0.038683 |
| **chanceCreationCrossing** | 0.010666 |
| **chanceCreationShooting** | -0.129550 |
| **defencePressure** | -0.173616 |
| **defenceAggression** | -0.059658 |
| **defenceTeamWidth** | -0.066553 |

* Despite any differences between the tests, it’s very interesting to see that there is consensus amongst them on which are the one to correlate with the outcome.

Two way ANOVA – done on categorical data

Text

Description automatically generated

Below is the Information Gain on ranked categorical data to show the features that has the most presence in making the correct prediction for the outcome. These are the ones that has the most say in who is most likely the winner.

Text

Description automatically generated

* 1. **Wrapper Methods**
     1. Recursive feature elimination
        + Numerical data – backwards selection

Text, table

Description automatically generated with medium confidence

* + - * Catergorical data = backwards selection

Text

Description automatically generated

* + 1. **Embeded Methods**
       - Regression tree for numerical
       - <https://data36.com/regression-tree-python-scikit-learn/>
         1. Respective r squared, MSE, and RMSE:

0.16776824013258596 2.354175235038137 1.534332178844639

Below chart is a summary of regression tree’s predictions.

Chart

Description automatically generated

* + - * Decision tree for categorical

Taking in instruction from but adding on training and testing: <https://www.w3schools.com/python/python_ml_decision_tree.asp>

* + - * 1. As per the set of instructions input variable are to be numerical categories representing string categories.
        2. All categorical data are converted to a scale of 1 – 3 with 1 representing defensive styles as outlined above, and 3 representing a more offensive, positive, and creative playstyle.
        3. From the abridged tree below we can see which input variables are the more impactful ones:

Diagram

Description automatically generated

1. Choose predictors to create a model (baseline vs defense) and compare the models to see any trends.

Based on feature selection, I will choose and look at the features that are standout (and create a baseline model based off this). And I propose to do a similar sort of modelling to the following project linked below albeit a much more general model, as I won’t be touching the models mentioned that I do not understand or has its assumptions violated (log regression).

I plan to do something similar by creating KNN, Random Forest, and Decision Tree on the num/catData sets.



<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9261335>

A set of the models with the “best predictive” inputs will be compared to a set of the models that hold “defensive” inputs. These comparisons will be made on accuracy and such details.

For the sake of the project, defensive inputs are the ones to describe what’s defined as the defensive playstyle. A limitation of this is that this is fairly subjective, not only to the researcher’s bias but also the team’s systems, and where and how a team truly defends is not easily measured.

A comparison of the accuracies between “best” and “defensive” inputs will be looked at. If there is a difference then defense is not the sole contributor to overall success of a team, and a conclusion can be made.

# Bibliography

Anderson, C., & Sally, D. (2013). *The Numbers Gae.* Penguin.

Breaking the Lines. (2022, July 28). *What is Juego de Posición?* Retrieved from Breaking the Lines: https://breakingthelines.com/tactical-analysis/what-is-juego-de-posicion/

Brownell, P. (2013, July 31). *Which Stats Are Most Important for Measuring Defenders?* Retrieved from Bleacher Report: https://bleacherreport.com/articles/1722602-which-stats-are-most-important-for-measuring-defenders

Ćwiklinski, B., Gielczyk, A., & Choras, M. (2021). Who Will Score? A Machine Learning Approach to Supporting Football Team Building and Transfers. *Entropy*, 90.

Davis , A., & Suryawanshi, N. (2023, 1 14). *Northwestern Sports Analytics Group*. Retrieved from Does Defense Win Championships?: https://sites.northwestern.edu/nusportsanalytics/2020/12/15/does-defense-win-championships/

FOX Sports. (2023). *FIFA World Cup 2022 Defensive Stats*. Retrieved from Fox Sports: https://www.foxsports.com/soccer/morocco-men-team-summary-stats?category=defensive&sort=t\_int&season=2022&sortOrder=desc&groupId=12

MARCA. (2012, 9 12). *Spain beats its possession record for the Del Bosque era*. Retrieved from MARCA: https://www.marca.com/2012/09/12/en/football/national\_teams/1347468397.html

Merhej, C., Beal, R., Ramchurn , S., & Matthews, T. (2021). What Happened Next? Using Deep Learning to Value Defensive Actions in Football Event-Data. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 3394-3403.

Rossi, A., Pappalardo, L., Cinitia, P., laia , F. M., Fernandez J,, & Medina, D. (2018). Effective injury forecasting in soccer with GPS training data and machine learning. *PloS on*.

Smith, A. (2017, 8 30). *Does attack or defence win titles?* Retrieved from Premier League: https://www.premierleague.com/news/464218

Stöckl, M. S. (2021). Making offensive play predictable-using a graph convolutional network to understand defensive performance in soccer. *In Proceedings of the 15th MIT Sloan Sports Analytics Conference*.

Winterburn, S. (2017, January 7th). *What can we learn from defensive statistics?* Retrieved from Football3665: https://www.football365.com/news/what-can-we-learn-from-defensive-statistics

# Proposed Roadmap – Extremely General

Diagram, letter

Description automatically generated