

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 are 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 scope of the project, I will be looking at home wins.

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 do feature selection. Pandas profiling is used for EDA report generation.

STEPS TAKEN:

1. **Pre-Processing**
   * Frist I will generate the average odds as a singular dependent variable. I plan to take an average of all the odds provided by the betting companies.
     1. First issue to come up: Some betting companies have a very high level of missing values
     2. Text

        Description automatically generated
   * Any variable with more than 1% “missing values” are simply dropped.
     1. 14585 observations prior to dropping columns with missing values
     2. 14503 observations. after dropping columns with missing values.
     3. Data loss of less than 1%
     4. Below is comparison of the missing values % between each betting company

|  |  |
| --- | --- |
| Attribute Name | Missing Values |
| 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. Once missing data is removed, an “avgOdds” (average odds) variable was created by taking the average of between all odds provided by betting companies for each match.
  + Output Variable/dependent variable (avgOdds) is a continuous, numerical piece of data. As such may be hard to build models against. It is converted into a set of categories.

Graphical user interface

Description automatically generated with low confidence

* + 1. Outlier removal and detection was done. Anything outside of 3 z scores are all removed as they can be considered outliers.
       - Ignorable amount of data loss here (dependent observations went from 85122 to 83140, roughly a 3% loss of observations, not a huge deal).
       - Likewise for categorical data, it reduced from 85122 observations to 83140 observations (< 3% data loss)
    2. The continuous numerical data are converted into 4 categories/

From the EDA report, we can see this output data is not evenly distributed. As such the categories should be divided into groups that are representative. (e.g. simply taking the ranging and dividing by 4 would result in a biased grouping where one is weighed much heavier than the rest).

To resolve this issue, the four groups are created based on IQR provided by the EDA. IQR is chosen as it’s generated based on median rather than averages.

* + - 1. Almost Guaranteed Win (0 – 1.682)
      2. Likely Win (1.683 – 2.100)
      3. Unlikely loss(2.101 – 2.684)
      4. Likely Loss (2.684 +)

This stratification works as the resulting dataset has a fairly even distribution:

Chart, bar chart

Description automatically generated

* + Next is looking at the independent/input variables.
    1. From the database provided, I will only look at the table marked as “team Attributes” as this is relevant to how the team performs. Individual details are ignored.
    2. From the initial EDA, I am able to tell the database consists of numerical and categorial independent variables.
    3. Screenshots below show categorial vars on the left and numerical vars on the left.

Text

Description automatically generated Text

Description automatically generated

* + 1. For Numerical Data:

Graphical user interface, text, application, email

Description automatically generated

The above is the warnings for the EDA report on the numerical variables. 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. The numbers are on a standardized scale of 0 – 100. Nothing exceeds or goes below this range.

Below are random samples of distributions for the input numerical variables. We can see its mostly normal. With no warnings in EDA report, I can assume normality amongst the inputs.

Chart, bar chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

* + 1. For Categorical Data:

Table

Description automatically generated

The above is the warnings for the EDA report on the categorical variables. (Note the report included other items such as date, names, IDs, as well).

Again, no outliers are detected.

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 randomly picked examples below, normality is not a given:

A picture containing graphical user interface

Description automatically generated

This categorical trend 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, an amateur, hard coded encoding was done to covert the strings into a workable set of ordinal data. Each class was converted into a scale of 1 – 3 (depending on the number of categories) with the lower end (1) always representing a slower, restrictive, defensive style of play and the higher end (3) representing a fast-flowing, positive attacking style of play.

As 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 relevant inputs themselves.

Chart

Description automatically generated

Chart, waterfall chart

Description automatically generated

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

Graphical user interface, application

Description automatically generated

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

*Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated*

* Now that an understanding of the independent variables is completed. I will attempt to combine the two into one dataframe.
* The categorical data is crudely encoding using the following way:
  + Each category is looked at and converted into a scale of 1 – 3 or 1 – 2 depending on the number of categories.
  + 1 and lower numbers are associated with a restrictive, slow, defensive style of play. 3 and higher number are associated with a faster, free-flowing attacking style of play.
* All values were put onto one dataframe and a new EDA report is generated. A couple of key things to note as shown by the warnings below:

Table

Description automatically generated

* Categorical data has a good chunk of missing data throughout.

Table

Description automatically generated with medium confidence

* Categorical data is very highly correlated with a lot of the numerical data. This is extremely evident in the correlations matrix produced by the EDA (shown below). From here it seems only 3 of the categorical variables are not correlated:
  + buildUpPlayDribblingClass
  + chanceCreationPositionClass
  + defenceDefenderLineClass
* Due to the high correlations between most of the categorical data and the numerical data, the whole of the categorical data frame will be simply dropped.
  + Upon a further look at what’s actually included in categorical data – it looks like they (the database owners) simply applied a classification on existing numerical data. The categorial data and numerical data (that are correlated) are simply different versions of themselves.
  + Due to this fact, the majority of the categorical dataframe can be considered redundant.
* The 33.8% missing data is also too significant to simply take an average or ignored. The 3 non-correlated variables will also be dropped due to their significantly high missing values %.

Chart, timeline

Description automatically generated

After all the considerations above, the dataframe that is left over is simply just the numerical data.

1. **Feature Selection:**

Given:

Output/Independent – Categorical, even distribution.

Input/Dependent – Numerical, normal.

* Filter Based – Information Gain:

The first feature selection method applied is using information gain to filter out the best.

Using a rank and sort filter, below shows the numerical variables with the highest information gain:

Text

Description automatically generated

From the details above, we can see first and foremost, none of the variables are super strong objectively as they all rank 0.05 or lower. This tells me the overall task to trying to predict victory in a football match is hard, it’s not going to be the most accurate prediction. It helps tell me as a baseline fact that football is simply unpredictable, and hence any factors for prediction of results should be taken with a grain of salt.

Relatively speaking, defencePressure, chanceCreatePassing, and buildUpPlayPassing seem to contribute the most to an accurate prediction of victory.

* Wrapper Based – Backwards Selection

The first thing that must be done is to change the categorical output variable into a number. I’ve tried multiple times on different machines to run a backwards selection in R using the string variable “catOdds” as the independent variable. All of which produces an error saying “Na/NaN/Inf in’y’”. I ran checks to see if there are any NaN or Inf or Na in the column but there are none. Based on this, I’m assuming lm() simply just doesn’t take strings as the independent variable.

As such, I will convert the strings into numbers. I will do this extremely simply by converting the strings into a ranked set of numbers:

1 = Almost Guaranteed Win

2 = Likely Win

3 = Unlikely Win

4 = Likely Loss

Using this, I will treat the numbers like variables. The model will simply tell about how each variable can help predict 1, 2, 3 or 4.

The following are the results generated in R:

Text

Description automatically generated

Text

Description automatically generated

Looking at the AIC generated (given low AIC = lower amounts of information loss from removal the respective variable), the amount of information loss from each variable is fairly similar.

Generally, speaking, the lower the AIC the better the model. This model has a pretty high one, as such it’s not the best.

This also failed to eliminate any variables. This feature selection will be taken with a grain of salt.

The Adjusted R-squared value also shows a very low amount of relationship.

The backwards selection here has told me the victory results are not easily predicted using the input variables. This does make sense, as if it’s easy to predict, sports betting would be a poor business.

* Embedded Method – Decision Tree

Final method deployed is the embedded method where feature selection is a part of the model building.

A decision tree was created on the input dataset, and generated the following abridged tree:

Diagram

Description automatically generated

From the tree, we can see the highest Gini value is 0.75 at the top of tree. Telling me the most influential classifier/variable is the chanceCreationShooting variable, followed very closedly by defencePressure with a Gini value of 0.748. This simple, shortened decision tree shows me the most influential variables. The key takeaway here is that only 1 of which is related to defence.

* Bonus method – SelectKBest

Using the recommended package, I will use this function to try and do feature selection as well.

I will be using K select with respect to the Chi Squared test, as this is an appropriate test to use when the output variable is categorical.

The k will set as 3 as I want to simply see the top 3 best predictors.

From the results of my k select, I am able to see the following are the top 3 predictors:

1. chanceCreationPassing
2. chanceCreationShooting
3. defencePressure

Interestingly, only 1 of 3 top features includes a defensive one.

Extending the “k”, I can see the following being top influencers as well. More defensive stats are included:

1. buildUpPlayPassing
2. defenceTeamWidth
3. defenceAggression
4. **Model Building and Analysis:**

Here, I propose to build 3 different classification models to predict high chances of winning. Each model will include a baseline model that includes all the input variables, and a defense model that includes only the defense variables, and a comparison model that includes all variables except defensive ones. In total there will be 9 models.

The accuracies and various relevant statistics between the baseline and defence models in order to generate knowledge about just how important defence is to winning.

The 3 models that will be built are:

* Decision Tree
* Random Forest
* KNN Classifier

For the building if all models, we will always use a 70:30 training to testing ratio with a random state of 88 for consistency.

**Decision Tree:**

Baseline decision tree model was created during the embedded feature selection.

Below is the confusion matrix and confusion reportfor the baseline:

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

A decision tree that includes only the defensive variables are predictors is generated, and the following abridged tree is a quick show of the classification being done on it:

Diagram

Description automatically generated

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

The tree goes down many levels and gets very specific with the sorting. Notice how defencePressure is on the top level as well as the second level.

The following is the abridge diagram to the “non-defensive” model aka the model that has defensive variables removed:

Diagram

Description automatically generated

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

By generating the accuracy scores, I am able to see the following as a comparison:

|  |  |
| --- | --- |
| Model | Accuracy Score |
| Defensive | 0.1487 |
| Non-Defensive | 0.1457 |
| Baseline | 0.1451 |

First things first, all models produce terrible accuracies. Each of which produce a rough accuracy of 15%. This does make sense is a real world application as if I am able to create a model that can accurately predict betting odds, I would be a billionaire and invest all my money into sports betting. This just tells me football wins are not easily nor accurately predicted.

As mentioned in the previous sections, the high gini values do tell me this is not a bad model, just a model with bad accuracy.

Whatever difference exists is very minor.

Interestingly, the purely defensive model produced the relatively highest accuracy score. By removal defensive stats from the baseline model, the accuracy score improved, but the predictions aren’t as strong as a purely defensive model.

Looking at the confusion matrix, despite the terrible 15% accuracy, the model is ok at classifying matches that are considered almost guaranteed wins. Not so much with matches that are considered a likely loss.

**Random Forest:**

For Random Forest, a similar numerical variable is imposed on the dataset like we had done on backwards selection above. Rather than using the string variables as the dependent variable, I used their numerical counterpart:

1 = Almost Guaranteed Win

2 = Likely Win

3 = Unlikely Win

4 = Likely Loss

Below is the confusion matrix and classification report for the baseline model:

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

Below is the confusion matrix and classification report for the defensive model:

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

Below is the confusion matrix and classification report for the non-defensive model:

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

Below is a chart to summarize the respective accuracies of each model under a random forest classifier:

|  |  |
| --- | --- |
| Model | Accuracy Score |
| Defensive | 0.4510 |
| Non-Defensive | 0.4625 |
| Baseline | 0.4624 |

Overall, this model produced a much higher accuracy score. Making it slightly more reliable for making predictions. However, an average accuracy of roughly 46% is still not good. Not enough for me to quit my job and do sports betting full time using my algorithm. Using this algorithm, I’m still most likely to lose any bets I place.

Looking at the comparisons, once again the difference in accuracies is minute. The defensive model this time created the, relatively, lowest accuracy with 45%. The non-defensive model showed a tiny improvement (0.001%) from the baseline model. These scores suggest by removing defensive, the predictions will be better by 0.0001%. Honestly, an improvement of 0.0001% in this context is near negligible.

\*\*\*Will expand on precision and recall in next iteration of the project

**KNN Classifier**

Finally the KNN classifier models are created.

For these comparisons keep consistent a number of neighbours value “k” was set as 5 for all models.

Below are the respective confusion matrices and classification reports of the various KNN models:

Baseline:

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

Defensive:

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

Non-Defensive:

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

The following chart comparing accuracies is derived:

|  |  |
| --- | --- |
| Model | Accuracy Score |
| Defensive | 0.4078 |
| Non-Defensive | 0.4176 |
| Baseline | 0.4183 |

Similar to all models above, accuracy is not strong. This is still not favorable and reliable to make predictions on betting odds.

Similar to the Random Forest models, the defensive model’s accuracy predictions seem to be the relative lowest. Telling me these factors would have the less of a say in the overall prediction of a winning result.

Unlike Random Forest however, the baseline model here seem to perform stronger than the model that has no defensive stats in it.

# 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