Compre Report for Study Project on 'Sports Analytics'

Ahmed Thahir[1]

[1] BITS Pilani Dubai Campus

Abstract

Sports analytics is a field of study involved in understanding and improving sports performance of team(s) and player(s), using relevant data. It has been found that current research in Sports Analytics primarily focuses on prediction using statistical correlation, rather than making decisions using causation. My research proposes using causal inference to identify parameters that can improve a team/players' performance. Moreover, the idea of using economics principles for financial decisions such as player transfers has also been introduced here. Furthermore, a basic use-case of Granger Causality has also been implemented.

Keywords: Sports Analytics, Causal Analysis, Machine Learning, Python, Statistics, Economics, Regression, Granger Causality, Time Series Data, Cross-Sectional Data

Acknowledgements

Firstly, I would like to thank Almighty God for giving me the strength, knowledge, ability and opportunity to take this project.

Secondly, I would like to thank my family and friends, who have always supported me in this journey of finding my true passion.

Thirdly, I would like to mention people who have inspired me. Mohammed Azharudin is the person who started it all - he sparked my interest in Computer Engineering. Dr. Sartaj Rasool, my Economics professor, opened my eyes to the beautiful world of economics. A big thank you to all the teachers who have helped me get to the stage I am at right now.

Moreover, I would like to express my heartfelt gratitude to the Director of BITS Pilani, Dubai Campus, Prof. Srinivasan Madapusi, who has ushered a new light on our college.

Lastly, I would like to thank my project supervisor, Dr. Raja Muthalagu, for providing me with the opportunity of performing this project. His guidance and encouragement over the past few months has helped me learn and to be innovative.

Contents

List of Figures

List of Tables

Chapter 1: Introduction

Authorization

This report for 'Study Project on Sports Analytics' has been authorized by Dr. Raja Muthalagu, HOD of Computer Science at BITS Pilani Dubai Campus, on 4^{th} April 2022.

Historical Background

Sports Analytics reached a global market size of $\$2.5B$ in 2021. It has become widely-used by teams all over the world, with an expected market size of $\$8.4B$ by 2026. [1] Due to increasing competitiveness and limited resources, it has become essential to use data for optimization.

Objectives

- Identify parameters/factors which produce effective outcomes in sports.
- The goal is not prediction, but to find causes.

Scope

- The report covers an outline of the research performed over the past few months for 'Study Project on Sports Analytics'
- The sport under this study is European Football.
- Details not relevant to the topic in hand have been omitted

Limitations

- As this is a Study Project, the report only focuses on the learned concepts
- The report does not go into depth into all details, as the field is quite vast
- My proposed methods may be improvised, once I learn all the required concepts
- Implementation of my proposed methods will be worked in the upcoming months

Methods and Sources of Data Collection

The main sources for research were online videos, articles, periodicals. Datasets were taken from open-access databases.

Report Review

Including the introduction, the report is divided into five chapters. Chapter-2 contains the **literature review** which highlights the existing research in this field. Chapter-3 is the **discussion**. Chapter-4 contains **implementation** of the existing research, to get an idea of statistical computing. Chapter-5 contains the **conclusion**, which sums up the discussions, insights and outlines the major issues faced.

Chapter 2: Theory

This chapter highlights the major theoretical concepts that were come across during this study project. The references for this section are [2]–[4].

Definitions

This section highlights the major keywords and definitions relevant to this study.

Data Analytics. Data analytics is a field of study, which aims at obtaining useful insights using relevant data. Upon taking actions, further data is collected to verify the initial insights.

Sports Analytics. Sports analytics is a form of data analytics, involved in understanding and improving sports performance of team(s) and player(s).

Pythagorean Expectation. Pythagorean expectation is a sports analytics formula, which tries to quantify a team's performance. It was devised by Bill James.

$$\text{Expected Win\%} \propto \frac{x^2}{x^2 + y^2}$$

where

- x = parameter scored
- y = parameter conceded

Coefficient of Determination. It shows how well data fits within the regression. It is represented as R^2. It has a range of $[0, 1]$. Higher the better.

Inflation. It is defined as the rate at which prices of commodities increase.

CPI. "The Consumer Price Index is a measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services".[5] It helps include the effect of inflation during analyses.

Opportunity Cost. It is defined as the real cost associated with any action. It is what you are sacrificing by doing a particular action, rather than t.

Types of Data

Cross-Sectional. Data that spans over various features, for the same time interval of interest.

Year	Company	Revenue
2015	A	1K
2015	B	2K
2015	C	3K

Table 1
Cross-Sectional Data

Time Series. Data that spans over various time intervals, for the same feature of interest.

Year	Company	Revenue
2015	A	1K
2016	A	2K
2017	A	3K

Table 2

Time Series Data

Panel. A combination of cross-sectional and time series data, spanning over various features and time intervals.

Year	Company	Revenue
2015	A	1K
2015	B	3K
2016	A	2K
2016	B	3K
2016	C	1K
2017	B	3K
2017	C	3K

Table 3

Panel Data

Statistics

Contrary to common understanding, Correlation $\neq$ Causation. That is, just because 2 variables are correlated does not necessarily mean that one causes the other.

There are 3 levels of statistical computing

Level	Meaning	Purpose
Correlation	Statistical relationship between two variables	Prediction
Causal Effect	Relationship between a cause and its resulting effect	Making decisions on tested sa
Causal Mechanism	Understanding the reason for causal effect	Making decisions on interested s

Table 4

Levels of Statistical Computing

Prediction Methods

Type of Prediction	Range Interval
Classification	$[0, 1]$
Value	Continuous

Table 5

Types of Prediction Models

Consider $h(x)$ to be hypothesis function (prediction). The following are the most commonly-used prediction models.

Linear Regression. A value prediction algorithm, which finds a best-fit straight line for the data.

$$h(x) = a_0 + a_1 x$$

Polynomial Regression. A value prediction algorithm, which finds a best-fit curve for the data.

$$h(x) = a_0 + a_1 x + \cdots + a_n x^n$$

Logistic Regression. A classification algorithm, which finds a best-fit line to separate 2 categories of data. It incorporates a *Sigmoid Function* $g(z)$ that maps the output range as $[0, 1]$.

$$h(x) = g(a_0 + a_1 x)$$
$$g(z) = \frac{1}{1 + e^{-z}}$$

Decision Tree. A classification and value prediction algorithm that contains a tree-like structure of nodes, with conditions in each of them. It is very useful for conditional problems.

Random Forests. A classification and value prediction algorithm that contains a collection of decision trees. The final output is an averaged output of the trees.

Supported Vector Machines. A classification algorithm that tries to find a *hyperplane*, that tries to maximize the margin between different categories of data.

Rubin's Causal Model

A model for causal inference, that depends on randomized testing to derive causality. It is widely used for deriving the effectiveness of treatments and drugs. Using a correct randomized experiment, the correlation obtained is equal to causal effect of the input.

'Treatment' refers to applying the input (1), and 'no treatment' refers to not applying the input (0). Outcomes are the output with/without the treatment.

The average treatment effect - the true causal effect of an input - is the difference between the average outcome with and without the treatment. Averaging improves accuracy, as the input is a random variable.

However, as it depends on experimentation and the input can only be binary - 0/1, it is not always feasible. The preferred causal inference model is Judea-Pearl model which derives causality from observational data.

Gini Coeffiecent

Quantifies disparities/inequality in a distribution. It is mainly used for quantifying income disparities of various locations, communities, etc.

$$G = \frac{2}{n^2 \bar{x}} \sum_{i=1}^{n} i \left(x_i - \bar{x}\right)$$
$$= \frac{\sum_{i=1}^{n} (2i - n - 1) x_i}{n \sum_{i=1}^{n} x_i}$$

Hypothesis Testing

A statistical tool involving a null hypothesis and an alternate hypothesis to determine validity of a hypothesis.

p-Value. It is the probability of the null hypothesis being true.

Statistical Significance. A statistical test is said to be significant, if it has a p-Value of $p \leq 0.05$.

Granger Causality

A model to determine if one time series affects another. It helps determine direction of causality, when there is a 2-way correlation. The past values of x are tested to determine if they have a statistically significant effect on the current value of y, taking past values of x as regressors. However, it may **not always** give true causality.

$$y_t = a + by_{t-1} + cx_{t-1}$$

- if $c \neq 0$, then x **granger causes** y
- t and $t - 1$ are not necessarily years; it just denotes a time period

Implementation

The following are few of the major tools that I will be using for my research.

Purpose	**Tool**
Programming Language	Python
Dataframes Library	Pandas
Plotting Library	Matplotlib
Prediction Library	SciKit-Learn
Math Library	Numpy
Granger Causality Library	Statsmodel
Integrated Development Environment	Jupyter Notebook

Table 6
Tools used

Python is the chosen programming language, as it is

1. free
2. open source

3. easy to learn
4. widely-used
5. easy to collaborate

Chapter 3: Literature Review

Current Research

The main reference paper for my research is . It provided a comprehensive explanation of Sports Analytics.

Sports data can be in different forms - qualitative/quantitative or structured/unstructured. This data could include player actions, spatio-temporal data, biographical data, scouting reports, etc. Sports analytics involves collecting and analyzing this data in order to gain insights. [6]

[7] predicted the winner of a game, using a Poisson Distribution to predict the goals scored by each team. [8] created a model to predict match outcome, using logistic regression. [9] used numerous machine learning algorithms on n-grams (a specific sequence of letters/words) to predict match outcomes. [10] concluded that that shot efficiency is more important than the number of shots; number of passes and ball contacts are also important factors; the distance covered was not found to be important. [11] analyzed the 2006 World Cup and found that winning teams enter the opponent's penalty area more often. [11] analyzed the 2006 World Cup and found that winning teams hardly allowed opponents enter their penalty area.

[12] found that there are more second half goals than in the first half; moreover, the probability of scoring increases with time. [13] found that top teams are more efficient and score more goals from within the penalty area, inferred to be due to tactical superiority; another differentiator between top teams and other teams is the number of key passes (passes that lead to a goal scoring opportunity). [14] compared one-sided games with competitive ones; they found that one-sided winners had much better possession %, "one-v-one" duels, number of passes, shots, shots on target, and shooting accuracy; but these were inapplicable close range games. [15] analyzed two different FIFA World Cups and found that were more goals scored from passing sequences that were longer rather than shorter; however, when the number of passes gets too large, more attempts were required to score a goal.

[2] created a binary classification model to predict how each team of various European Leagues would perform, using numerous parameters such as previous year position, wins, draws, losses, net transfer expenditure, etc. Moreover, it focused on player performance prediction, with a focus on central defenders. Various features such as passing, heading, aggression were obtained from the game *Football Manager*. These were used to predict match rating of players. The best predictors were found to be interceptions, clearances, jumping reach and strength.

Finally, [16] took a different approach. It performed a *Granger causality test* to find that "the influence of lagged revenue on current performance is greater than the influence of lagged performance on current revenue". However, despite this research highlighting some form of causality, trying to improve performance through increasing revenue isn't feasible, as it is not an easily controllable parameter.

Gaps

1. Current research mainly focuses on prediction
2. Over-reliance on correlation

   Correlation is useful for making predictions, but is insufficient for making effective decisions
3. Lack of causal inference

   Decisions without casual understanding may prove to be costly, as they may **not produce** efficient outcomes
4. Financial values are in absolute terms, rather than relative terms

   This is not accurate, as £50M was much more valuable 10 years ago compared to present-day, due to inflation.

Chapter 4: My Research

The following are the insights I obtained when doing this study project, that were worth mentioning.

Causal Inference

Correlation between Win% of a team and a team's number of passes does **not** necessarily imply that passing more during games will lead to higher Win%. The same goes to other parameters such as possession percentage, passing accuracy, etc.

The best approach to identify effective parameters would be to use causality tests, and hence obtain true causes of improved performance for a team/player. These effective parameters may vary from team to team, as each team differs in its ambitions and style of play.

Economics

When analyzing the cost of buying a player, economic principles such as correcting values for inflation and opportunity cost analysis may be used. This will give a more holistic understanding of whether or not a player is worth buying.

A point to note is that the inflation within the football transfer market has been much higher than the consumer sector. Players that would've cost around £30M few years ago now cost around £80M. This is due to large transfers made by big clubs for players, such as *Neymar Jr.* in 2017 - costing Paris Saint-Germain around £222M. Moreover, there has been a sudden influx of investments in the footballing world, thereby resulting in clubs demanding higher transfer fees to sell their players.

Implementation

A basic implementation was performed to learn using Python for statistical computing. The codes are available on the GitHub repository for this project.

English Premier League Predictor. This implementation tries to predict the 2nd half win% of a team using data from the first half of the season. Using Pythagorean Expectation gave better results, as the correlation and R^2 value was higher. Referred to [4]

2nd Half Win% predicted using 1st Half	R^2	**Correlation**
Win%	0.572	0.757
Pythagorean Expectation	0.633	0.796

Table 7
Win % Prediction

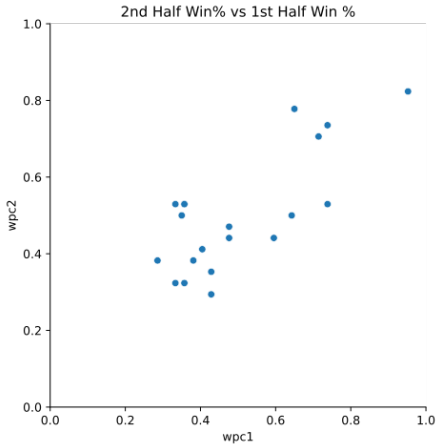

Figure 1. Win %

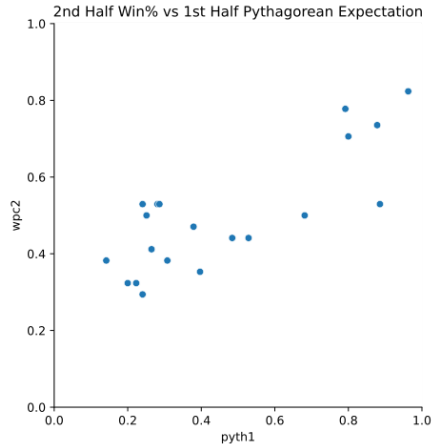

Figure 2. Pythagorean

FIFA21 Player Rating. This implementation aims at predicting players' FIFA21 (an official football game) rating using their market value. An attempt at including the effect of inflation using CPI failed. Referred to [2]

Method	Degree	R^2 Value
Uncorrected	1	0.5974
	2	0.7076
	3	0.7308
	4	0.6696
Corrected	1	0.6157
	2	0.7204
	3	0.7413
	4	0.7476

Table 8

Corrected using CPI and Mean Value

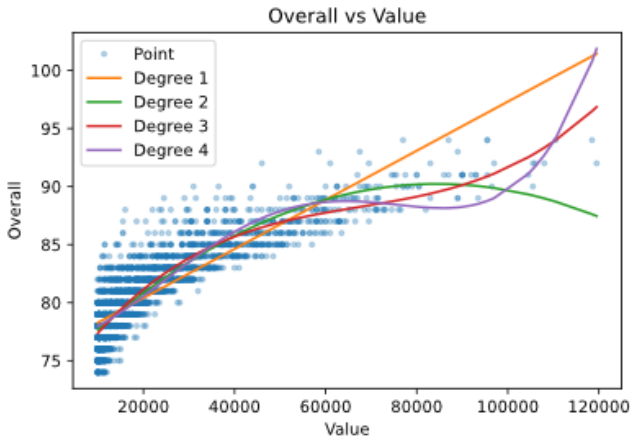

Figure 3. Rating vs Value

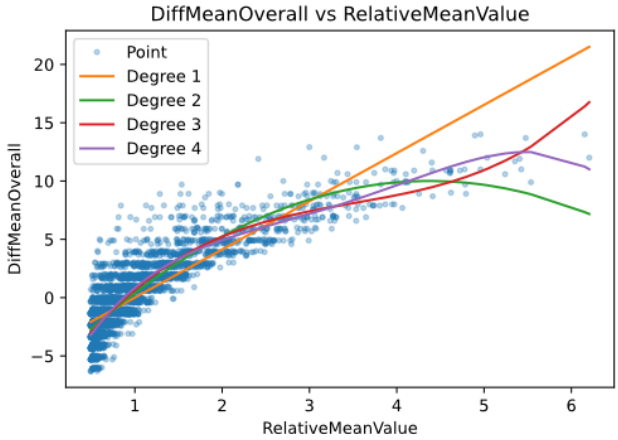

Figure 4. Corrected using CPI and Mean Value

Inferences.

- Only CPI is not enough to include the hyper inflation in football. CPI only shows an

inflation of around 5%; hyper-inflation is much higher than that. Hence, getting the deviation from the average transfer value of each year may provide a better result.
- Relationship is only satisfied for players with value $\geq 1M$. This **could** mean that there are many good under-valued players.
- Corrected Degree 3&4 Regression gave the best-fit. However, due to over-fitting, it has a downward slope at the end; hence, it is to be rejected.

Granger Causality test of Cristiano Ronaldo's Market Value, Performance, and Age. Hypothesis: Market Value is a function of previous year's performance and age.

$$
\begin{aligned}
V &\propto P \\
V &\propto \frac{1}{A} \\
V &= aV_{t-1} + bx_{t-1} \\
x &= \text{Goals} + \text{Assists} - \text{Age}
\end{aligned}
$$

The base/'zero' value of every player must be their initial value, as every player starts off from a different initial value. This is not just restricted to talent/skill, due to factors such as nationality. Moreover, it helps accommodate the market value of a player who started off with a high value won't really go $< 1M$.

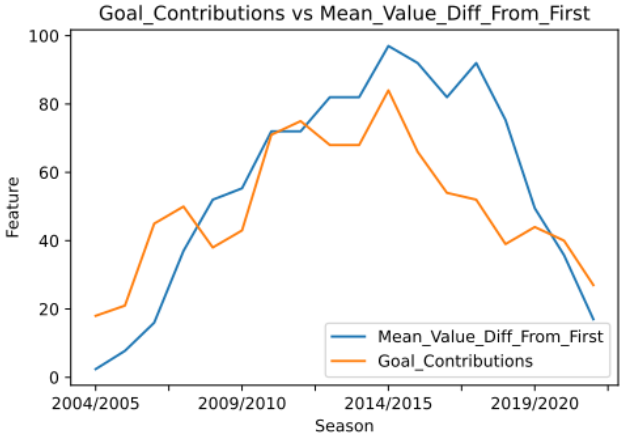

Figure 5. Granger Causality test of Goal Contributions as a function of Market Value

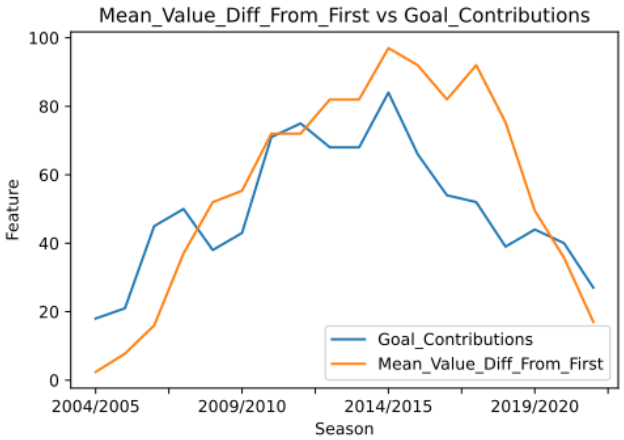

Figure 6. Granger Causality test of Market Value as a function of Goal Contributions

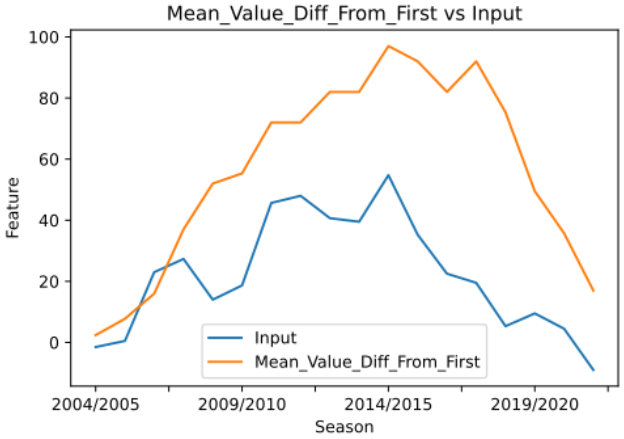

Figure 7. Granger Causality test of Market Value as a function of Goal Contributions and Age

Inferences.

Prediction	**Predictor**	**p-Value**	Significant
Goal_Contributions	Mean_Value_Diff_From_First	0.8667	N
Mean_Value_Diff_From_First	Goal_Contributions	0.0491	Y
	Goals_Contributions - Age	0.0068	Y

Table 9

Granger Causality test of Cristiano Ronaldo's Market Value, Performance, and Age

Chapter 5: Conclusion

There is clearly a lot of scope for sports analytics. Current research only focuses on prediction using correlation. However, this report highlighted the dangers of relying on statistical correlation. Causal inference is crucial for taking decisions that improve efficiency and avoid costly errors.

Concepts to Learn

To progress with my research, the next steps would be to learn

- Judea Pearl Causality Model
- when to implement each prediction model
- handling panel data

Possible Future Work

- Forecasting performance of player in the next season
- Determining if a player would be a good purchase
- Determining if collection of players with large variety of experience would perform better than a collection with the same experience, using Gini-Coefficient.

Bibliography

[1] "Sports Analytics Market with COVID-19 Impact Analysis by Component, Application, Deployment Mode, Organization Size, Industry Vertical And Region - Global Forecast to 2026." https://www.reportlinker.com/p03825782/Sports-Analytics-Market-by-Type-by-Applications-by-Deployment-Type-by-Region-Global-Forecast-to.html?utm_source=GNW.

[2] V. Chazan - Pantzalis, "Sports Analytics Algorithms for Performance Prediction," Jun. 2020.

[3] "Modern Data Analysis for Economics," *Modern Data Analysis for Economics*. http://jiamingmao.github.io/data-analysis/.

[4] "Foundations of Sports Analytics: Data, Representation, and Models in Sports | Coursera." https://www.coursera.org/learn/foundations-sports-analytics.

[5] "Consumer Price Index (CPI)," *Investopedia*. https://www.investopedia.com/terms/c/consumerpriceinde

[6] R. Bhatnagar and M. Babbar, *A systematic review of sports analytics*. 2019.

[7] M. J. Dixon and S. G. Coles, "Modelling Association Football Scores and Inefficiencies in the Football Betting Market," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 46, no. 2, pp. 265–280, Jan. 1997, doi: 10.1111/1467-9876.00065.

[8] L. Mao, Z. Peng, H. Liu, and M.-A. Gómez, "Identifying keys to win in the Chinese professional soccer league," *International Journal of Performance Analysis in Sport*, vol. 16, no. 3, pp. 935–947, Dec. 2016, doi: 10.1080/24748668.2016.11868940.

[9] S. Kampakis and A. Adamides, "Using Twitter to predict football outcomes," p. 10.

[10] H. Broich, J. Mester, and F. Seifriz, "Statistical Analysis for the First Bundesliga in the Current Soccer Season," p. 8, 2014.

[11] C. Ruiz-Ruiz, L. Fradua, Á. Fernández-GarcÍa, and A. Zubillaga, "Analysis of entries into the penalty area as a performance indicator in soccer," *European Journal of Sport Science*, vol. 13, no. 3, pp. 241–248, May 2013, doi: 10.1080/17461391.2011.606834.

[12] V. Armatas, A. Yiannakos, S. Papadopoulou, and D. Skoufas, "EVALUATION OF GOALS SCORED IN TOP RANKING SOCCER MATCHES: GREEK " SUPER-LEAGUE " 2006-07," *Serbian Journal of Sports Sciences*, vol. 3, pp. 39–43, Feb. 2009.

[13] "DIFFERENCES IN OFFENSIVE ACTIONS BETWEEN TOP AND LAST TEAMS IN GREEK FIRST SOCCER DIVISION. A RETROSPECTIVE STUDY 1998-2008."

[14] B. Evangelos, G. Aristotelis, G. Ioannis, K. Stergios, and A. Foteini, "Winners and losers in top level soccer. How do they differ?" p. 8.

[15] M. Hughes and I. Franks, "Analysis of passing sequences, shots and goals in soccer," *Journal of Sports Sciences*, vol. 23, no. 5, pp. 509–514, May 2005, doi: 10.1080/02640410410001716779.

[16] S. M. Dobson and J. A. Goddard, "Performance and revenue in professional league football: Evidence from Granger causality tests," *Applied Economics*, vol. 30, no. 12, pp. 1641–1651, Dec. 1998, doi: 10.1080/000368498324715.