Predicting Financial Distress across the Football Industry

Pedro de Almeida Conde

Dissertation written under the supervision of Professor Ricardo Reis

Dissertation submitted in partial fulfilment of requirements for the MSc in Finance, at the Universidade Católica Portuguesa, 4th January 2023.

Author: Pedro de Almeida Conde (152421018)

Topic: Predicting Financial Distress across the Football Industry

Abstract

Accurately forecasting financial distress within the football industry holds significant importance for various stakeholders, including creditors, investors, shareholders and local communities.

This research employs machine learning algorithms to forecast financial distress within the football industry over a 5-year period and by analyzing clubs' financial ratios. Two machine learning models are performed: a logistic regression and a neural network model. The primary objectives of this study are to test the effectiveness of these models, evaluate the financial performance of football clubs, provide an overview of the industry as a whole and examine the influence of the Covid-19 pandemic on financial distress within the sector. Despite the high levels of debt, unprofitability, irrationality and mismanagement that are prevalent in football clubs, bankruptcies are not such an ordinary event, being relatively rare. The machine learning models implemented in this study yielded interesting and favorable results, with the neural network model demonstrating a slightly higher level of predictive accuracy. However, the significant impact of Covid-19 on the overall industry partially impaired the predictive capabilities of the models, raising questions about their practical applicability. This study suggests that the unique status of football clubs, which shields them from being treated as ordinary businesses, may be the only factor that enables their survival.

Keywords: Financial distress; Football industry; Logistic regression; Neural network; Non-distressed clubs; Distressed clubs; Financial Ratios

Autor: Pedro de Almeida Conde (152421018)

Tópico: Previsão de potenciais dificuldades financeiras na indústria futebolística

Resumo

A previsão de potenciais dificuldades financeiras na indústria futebolística contém uma grande importância para todos os participantes no negócio, incluindo credores, investidores, acionistas e comunidades locais.

Nesta dissertação foram implementados algoritmos de *machine learning* para que se efetuasse a previsão de dificuldades financeiras para um período de 5 anos, através do uso de rácios financeiros. Os dois modelos elaborados foram uma regressão logística e uma rede neuronal. Os principais objetivos deste estudo são testar o desempenho destes modelos, avaliar a performance financeira dos clubes de futebol, efetuar uma visão geral da indústria futebolística e examinar o impacto da Covid-19 no setor. Apesar dos elevados níveis de endividamento, prejuízos, irracionalidade e má gestão, a verdade é que o número de falências entres clubes de futebol é reduzida. Os modelos de *machine learning* aplicados neste estudo apresentaram resultados interessantes e positivos. Contudo, o impacto da pandemia na indústria afetou a capacidade de previsão dos modelos, levantando questões acerca da sua potencial aplicação no mundo real. Este estudo sugere ainda que o estatuto dos clubes de futebol, que os diferencia de um negócio normal, pode ser o único fator que promove a sua sobrevivência.

Palavras-chave: Dificuldades financeiras; Indústria futebolística; Regressão logística; Rede neuronal; Clubes financeiramente saudáveis; Clubes com dificuldades financeiras; Rácios financeiros

Acknowledgements

Firstly, I would like to express my gratitude to my family for their constant support and motivation, which have been crucial in shaping who I am today. Without their inspiration, many of my achievements would not have been possible. I am lucky to have them as the foundation of all my successes.

I also want to express my appreciation to my friends, especially Rodrigo, Mariana, and Gonçalo, for their enduring friendship and for their constant encouragement to be a better person every day. As the saying goes, “friends are the family we choose”, and I am fortunate to have them in my life.

I am deeply grateful to my colleagues and supervisors at Millennium bcp for giving me the opportunity to work with them during this semester. Their advice and insights have been invaluable, and their support in allowing me to complete my master's degree has been indispensable.

Lastly, I would like to thank Catolica Lisbon School of Business & Economics and ISCTE – Business School for their contributions to my studies and for helping me to grow during my academic journey. A special thank you to my supervisor, Ricardo Reis, for his valuable guidance and expertise.

Table of Contents

1. Introduction 8
2. Literature Review 10
   2.1. Predictive models of financial distress 10
   2.2. Football Industry 11
      2.2.1. Football against common businesses 11
      2.2.2. Ownership structure 12
   2.3. Bankruptcy across Football Clubs 13
   2.4. Financial Fair Play (FFP) 14
   2.5. Covid-19 Impact 15
3. Data 17
   3.1. Original Sample 17
   3.2. Outliers Treatment 18
   3.3. Labelled Data 19
   3.4. Descriptive Statistics 20
4. Methodology 22
   4.1. Financial Ratios Selection 22
      4.1.1. Variance Inflator Factor 22
      4.1.2. Mutual Information Classification 23
      4.1.3. Statistical Tests 24
   4.2. Logistic Regression Model 24
   4.3. Neural Network 25
      4.3.1. Neural Network Model 25
      4.3.2. GridSearch 27
   4.4. Predictive Ability 27
5. Results 29
   5.1. Financial Ratios Analysis 29
   5.2. Logistic Regression Analysis 30
   5.3. Robustness Test 34
   5.4. Neural Network Analysis 36
6. Discussion of Results 38
7. Limitations and Future Research 40
   7.1. Limitations 40
   7.2. Future Research 40
8. Conclusion 42

9. References 43
10. Appendixes 46

List of Tables

Table 1 – Financial Ratios in the original sample 18
Table 2 – Means from non-distressed and distressed clubs between t-5 and t-1 20
Table 3 – Percentage of clubs with negative performance between t-5 (2017) to t-1 (2021) . 21
Table 4 - Financial ratios remained in the study after VIF and MI 29
Table 5 – Results from Levene's tests and T-tests 30
Table 6 – Logistic regression results for t-1 30
Table 7 – Logistic regression results for t-2 31
Table 8 – Logistic regression results for t-3 32
Table 9 – Logistic regression results for t-4 33
Table 10 – Logistic regression results for t-5 33
Table 11 – Summary of logistic regression results 34
Table 12 – Logistic regression's robustness test results 36
Table 13 – Neural network model best scenario 37

List of Figures

Figure 1 – Example of the architecture of a Multilayer Perceptron (MLP) 26
Figure 2 – Framework of the potential outcomes from classification 27

List of Appendixes

Appendix 1 – Standard Deviations from non-distressed and distressed clubs from t-1 to t-5 46
Appendix 2 – Hyperparameters manually defined for the neural network 46
Appendix 3 - Confusion matrices for the logistic regression from t-1 until t-5 47
Appendix 4 – Confusion matrices for the neural network model from t-1 to t-5 48
Appendix 5 – Python code for cleaning and treating the data 49
Appendix 6 – Python code for the implementation of the logistic regression 52
Appendix 7 – Python code for the implementation of the neural network 56

1. Introduction

Financial distress is a condition that occurs when an entity lacks the financial resources or income to meet its own obligations, and as a result, is at risk of bankruptcy.

The prediction of financial distress has gained awareness since the early 1960s, when Beaver (1966) and Altman (1968) began their work on the subject. Despite all the developments and improvements in the topic, the financial crisis of 2008 had exposed several weaknesses and susceptibilities not only from the financial system but also from the overall corporations and entities worldwide. This historic event revealed that several entities were facing financial difficulties without anticipating them. The crisis also showed a dramatic increase in corporate failures. In the aftermath of this event, there has been a sustained focus on creating stronger predictive models and frameworks to prevent such failures in the future (Jones et al. 2017).

As the financial system before the crisis, the football industry appears to be an insurmountable business, due to its powerful economic, cultural and social status. The number of bankruptcies among football clubs are considerable low, despite their tendency to operate with unsustainable debts and incur continuous losses on a significant scale (Storm and Nielsen 2012). Clubs are well-known for acting close to financial collapse (Storm and Nielsen 2012). Their financial performance depends on various factors, including their country of origin, governance structure, FFP regulations, and sportive goals/ambitions. For these reasons and many more, the football industry embraces several peculiar characteristics (Neale 1964) which turn it in something very complex to understand, like an uncharacteristic "black box". Additionally, Covid-19 pandemic brought several challenges to the industry which have been receptive to cut costs, deteriorating even more clubs' financials.

Regarding the prediction of financial distress, machine learning models have been highly controversial throughout time since their relative effectiveness compared to classical models has been studied intensively. Despite this uncertainty, machine learning models have been widely used in literature and have consistently demonstrated promising results (Barboza et al. 2017). Some of these models have been gaining prominence such as logistic regressions and neural networks. Logistic regressions have their roots on statical modelling being considered a classical model, although, it has been extensively used as a machine learning approach due to interpretability and success in predicting binary outcomes (Jones et al. 2015). In the real-world, not only this technique is used to predict financial distress but also has several practical applications in the field of healthcare, consumer behavior, sports and social sciences. On the other hand, neural networks are less interpretable than logistic regression due to their level of complexity, but they are more prone to produce better classification estimations (Zhang et al.

1999). By curiosity, neural networks are inspired by the structure of the human brain once both are encompassed by various interconnected neurons which work together to analyze information. This machine learning model is designed to capture hidden patterns and non-linear relationships that may not be immediately apparent. It is extensively used in the real world, for instance, on robotics, healthcare, stock market prediction and even on weather forecasting.

Therefore, this research proposes to predict financial distress across the football industry through the employment of a logistic regression technique and a neural network model. It pretends to predict it, using financial ratios as the sole tool, in order to isolate them from other potential factors and understand how severe and dangerous clubs' financial accounts look like. There is relatively limited literature on this theme and the existing one shows some signs of low transparency as it is not very descriptive about the methods conducted. Fortunately, the event of financial distress and bankruptcy has been broadly studied over other industries which can be used as an example. This research also uses a longer period of time (5 years) for analysis compared to other studies, in order to better scrutinize the effects of the pandemic.

The main goal of this research intends to add new insights into the available literature, by answering the following research questions: 1) How poor is the financial performance of football clubs? 2) Is it possible to establish patterns and predict financial distress across such a disruptive industry like football? 3) Do these machine learning algorithms fit well in the football industry? 4) How did the pandemic affect clubs' financial performance? The primary aim of these models is to assess the management and performance of football clubs, identify whether clubs need financial intervention, and understand the overall behavior of the industry. They do not directly intend to predict bankruptcy since it is a rare event in the industry.

This research thesis is mainly structured into seven sections. Section 2 highlights the literature on predictive models of financial distress and also discusses the unique characteristics and constraints of the football industry that can impact clubs' financial performance. Section 3 exposes the underlying data used in the research, its properties and how it was cleaned and treated. Section 4 evidences the conducted methodology which scrutinizes the used features to identify the most informative financial ratios of financial distress and describes the application of the logistic regression and neural network models. Section 5 and 6 present the results from the analysis as well as its discussion and interpretation. Section 7 discusses the limitations of this study and potentialities for future research. Lastly, Section 8 concludes the study making a summary of the research.

2. Literature Review

This literature review was designed to provide valuable insights about predictive models of financial distress and the behavior of the football industry. There is evidenced some of the most notable models developed over time and identified the unique characteristics and factors that make the football industry a particularly interesting area for research.

2.1. Predictive models of financial distress

The study of predictive models of financial distress has a long history dating back to the 1960s, when Beaver (1966) first attempted to identify financial ratios that could distinguish between bankrupt and non-bankrupt firms by analyzing data from a five-year period prior to bankruptcy. While this study was considered a good foundation, it used a univariate analysis, which only considered financial ratios individually and did not take into account their potential interconnections.

Few years later, Altman (1968) introduced a revolutionary bankruptcy model called Z-score. This model was created using a multivariate discriminant analysis (MDA) technique which allows to simultaneously analyze multiple financial ratios and their relationships. It was applied to sixty-six public manufacturing corporations and identified five financial ratios that have been widely used in literature (Altman 1968). The model demonstrated a 95% accuracy rate for predicting bankruptcy 1-year prior to the event and 72% for 2-years prior to it. Altman has developed his own model over time, adapting it for use with private firms and non-manufacturing firms (Altman 2000). Subsequently, several other studies have built upon the Z-score such as Ohlson (1980), Zmijewski (1984) and Shumway (2001).

Most classical bankruptcy models rely on a set of fixed ratios that were chosen, at a specific point in time, as having general predictive value for bankruptcy. However, it has been shown that the sensitivity of bankruptcy prediction is highly dependent on industry characteristics (Sayari and Mugan 2017) which means that using the same financial ratios for firms in different industries can diminish the predicting ability of these models. It is common to find models that perform well in the industries for which they were originally developed, but do not work as well when applied to a different one. According to Sayari and Mugan (2017), this may be due to the models' inability to capture industry characteristics, leading to difficulties in differentiating distressed from non-distressed firms.

Various other techniques have been explored to predict bankruptcy and address the issues mentioned above, such as logistic regressions, hazard models, neural networks, decisions trees and newer machine learning algorithms referred to as "new age" classifiers. Literature has also

been thorough and critical in evaluating different models. For instance, Bellovary et al. (2007) made a review of bankruptcy prediction studies since 1930 and they concluded that MDA and neural networks are the most accurate and promising methods. They also found that a greater number of factors does not ensure a higher model accuracy. In addition, Jones et al. (2015) showed that "new age" models (e.g. AdaBoost and random forest) tend to outperform the more classic ones, mainly because they need less data preparation, being resistant to some statistical issues, such as multicollinearity and heteroscedasticity. However, simpler model structures are often considered a viable option if statistical inference and interpretability are required.

As mentioned before, this study relies on a logistic regression and on a neural network. Both methods have been ordinarily used across literature and have demonstrated significant and reliable levels of accuracy and specificity (Sayari and Mugan 2017; Jones et al. 2017; Alaminos and Fernández 2019).

2.2. Football Industry

Overtime, researchers have conducted several studies on the football industry due to its uncommon characteristics which makes the field very interesting in testing diverse economic theories. Football clubs are considered to be part of a "peculiar" industry since their operations are guided according to financial and sportive objectives (Neale 1964) which are undoubtedly linked (Guzmán and Morrow 2007) but whose correlation is uncertain and inaccurate (signal terms).

2.2.1. Football against common businesses

A determinant aspect that distinguishes football clubs from normal companies rely on the way they compete. Generally, a firm operates to be the most profitable as possible to eliminate its competitors and, in a perfect scenario, to achieve a monopoly. Nevertheless, sports and football are more profitable when the competition is tougher and balanced since a single club does not have any advantage in competing alone (Neale 1964; Dobson and Goddard 2011). Indeed, what makes football so attractive for its followers derives from the "uncertainty of the outcome" (Dobson and Goddard 2011) which postulates that the chances of winning are almost the same for each club. Theoretically, higher levels of uncertainty can lead to higher attendance and broadcasting revenues, therefore, from a commercial perspective, football clubs cannot be perceived as rivals.

While in common businesses, viability is set by financial profitability, some authors argue that, in football, clubs' spending power will determine their competitive position and long-term success (Franck 2010). Clubs gain a competitive advantage when they successfully apply their funds into football rather than paying their stakeholders.

Football clubs tend to have high survival rates despite operating with unsustainable deficits, growing debts and a profound imbalance between revenue and expenditures (Storm and Nielsen 2012). All these unhealthy practices not only depend on soft budget constraints adopted by clubs but also on financial regulations provided by football leagues. However, the reality is that clubs which operate with lower debts and deficits tend to obtain poorer sportive results than clubs of equivalent size without budget restrictions. For instance, the German and French leagues have heavier financial regulations than the English, Spanish and Italian, although their sportive performance is worse (Drut and Raballand 2012). Indeed, clubs with lighter league regulations and budget constraints have there a competitive advantage once they can be more flexible financially. These clubs tend to have better sportive results because they are allowed to hire more quality players and overspend on wages.

In recent years, the question "Does money buy success?" has becoming a topic of discussion within the football industry. Over time, overinvestment have become highly worthy since the clubs which spend the most, are usually in the top positions. Franck (2010) found a positive correlation between talent investment and the likelihood of sportive success. However, when clubs invest too heavily in talent to achieve predefined goals such as a promotion, avoiding a relegation and winning the league, they often face uncontrollable financial problems (Evans et al. 2022).

2.2.2. Ownership structure

The ownership structure of a football club can significantly impact its financial performance. Like common businesses, clubs can operate as private or as public listed corporation. Many authors suggest that being privately owned embraces several advantages in terms of decision-making and investment policy (Franck 2010; Wilson et al. 2013).

Firstly, residual rights of control and residual claims in private clubs are all assigned to the club's owner. Thus, the owner of a private club will have more freedom to invest and allocate funds into the club since he does not have to consider the interests of stockholders. Secondly, private clubs are less likely to maximize profit, as their main goal centers on making clubs' supporters accomplished with sportive results. Lastly, many owners of private clubs make large

and unprofitable investments in the belief their related businesses and brands aside from football will gain popularity and recognition, compensating clubs' losses (Szymanski 2017).
On the other hand, publicly owned clubs must be profitable to attract potential stockholders and to retain the existing ones. Stockholders only invest in a club because of its profitability and return since most of them do not have the power to directly influence club's business. Therefore, excessive investment by public clubs will be seen as a red flag to stockholders who may withdraw their money and invest in a better business. According to a study conducted by Wilson et al. (2013) on the English league, publicly owned clubs are more stable financially than privately owned clubs and closer to accomplish with the Financial Fair Play regulations. The study also found that private clubs owned by foreign investors tend to reach better sportive results than those owned by domestic investors.

2.3. Bankruptcy across Football Clubs

Despite financial instability affecting many clubs, they are perceived as being "too big to fail". This concept postulates that the likelihood of a club becoming insolvent or filing for bankruptcy is reduced due to its powerful social, cultural and economic status. Nevertheless, whether we go back to the inception of 2008 financial crisis, financial institutions were also considered insurmountable but indeed they have collapsed against all the odds. Therefore, if financial sector crashed despite its unbreakable status, the football industry is surely not immune to failure.
As previously stated, football clubs are most of the time on the edge of insolvency. They operate in a "hypercompetitive environment" (Szymanski 2017) where sportive goals tend to lead them to a deterioration in their financial performance, well-evidenced by their balance sheets and net losses. Furthermore, Szymanski (2017) concludes that negative shocks in demand and productivity are the main causes for clubs to fall into financial distress and, consequently, bankruptcy. These negative shocks aggregate implications at the wage-performance, performance-revenue and investment-performance interconnectedness. Strategies to deal with those shocks are also very tricky and painful to introduce since a wage or an investment cut will have an indirect impact in clubs' league performance, affecting their revenue.
Among the European top-divisions, bankruptcies are not very usual once the participants are very protected against these events. Typically, when elite clubs experience financial distress, they are bailed out by international investors, funds, public entities, banks and fans through bulky investments or debt forgiveness. In contrast, financial distress and bankruptcy are more

common in the lower divisions since these clubs are less prone to raise money and generate significant revenue. Several researchers have conducted studies on insolvencies inside the football industry in various countries, including England (Szymanski 2017), Spain (Barajas and Rodríguez 2010) and Germany (Szymanski and Weimar 2019). Overall, these papers corroborate the fact that insolvencies are more common among lower divisions clubs, especially in those that have previously competed in top-divisions. These findings reinforce the impact of a relegation into clubs' financial stability. According to Szymanski and Weimar (2019), a promotion and relegation system can be financially risky for clubs because they may feel obligated to stretch their finances to the limit. This makes them extremely vulnerable to negative shocks. To mitigate these impacts some leagues provide funds to relegated clubs in order to keep them up.

2.4. Financial Fair Play (FFP)

Financial Fair Play (FFP) is a mechanism created in 2009 by the Union of European Football Associations (UEFA), but only fully introduced in 2011. It aims to prevent clubs from overspending in relation to their revenues, focusing on a stringent regulation with the goal of limiting unsustainable expenses and significant net losses among football clubs.

This project was created during a period of significant growth in revenue and external financing in the football industry. This was particular noteworthy given that the world was still recovering from the global financial crisis (Storm and Nielsen 2012). At that time, various concerns were increasing about the level of wages, operational costs and players transfers, which were promoting discrepancies in the competitiveness between the more and less prosperous clubs (Plumley et al. 2019). These concerns were also reinforced by the fact that 655 clubs had recorded consecutive net losses during 2009.

Overall, FFP aims to improve the financial stability of European football clubs through a number of key drivers, including transparency, protection of creditors, financial discipline, revenue-based operations, and long-term sustainability. These measures are intended to reduce the risk of default and promote the ongoing viability of clubs within the football industry (UEFA 2018). Most of FFP rules rely on a breakeven requirement where clubs' expenditures are not allowed to exceed income. UEFA is very restricted on the definition of income, the European football regulator only considers on this caption the "relevant income" which sets aside the sources of external financing (not related to clubs' operations) also known by "financial doping" (UEFA 2018). Moreover, these regulations are also quite narrowed

regarding overdue payables between clubs. FFP regulations prioritize the elimination of the systematic risk that occurs when a football club becomes insolvent, the "snowball effect". Therefore, overdue payables between clubs must be minimized.

One practical limitation of these regulations is the fact they only cover clubs that intend or have resources to compete in UEFA competitions which are less than one third of the total clubs in European top-divisions. Clubs which do not fulfill these regulations are penalized and forbidden from participating in such competitions. Another limitation of FFP regulations is that they do not directly constraint clubs' budgets (Francois et al. 2022). Instead, they only ensure that clubs do not spend more than their means allow.

Previous literature on the effectiveness of FFP has been critical, with multiple studies pointed out that FFP has an adverse effect on clubs, specially, in terms of competitiveness. Under the breakeven principle, these regulations favor the wealthiest clubs by limiting the spending power of the smallest (Peeters and Szymanski 2014), making almost impossible for them to compete without external funding. In theory, the lack of competitive intensity can also result in a relevant fall in income (Neale 1964).

Additionally, various studies have examined the effect of FFP on clubs' financials through a comparison of both pre-FFP and post-FFP periods (Ahtiainen and Jarva 2022; Francois et al. 2022). According to Ahtiainen and Jarva (2022), who studied the FFP effect on the top-five European divisions, its impact is not homogeneous across countries. For instance, they concluded that its effect is significant in Spain, weakly significant in England and Germany and insignificant in France and Italy. Despite these results, they argue that is not accurate to attribute the improvement in clubs' financial results solely to FFP, as some clubs which do not frequently participate in European competitions have improved their financial performance as well. In fact, it may be believed that football clubs are less prone to report a loss in the post-FFP period (Ahtiainen and Jarva 2022; Francois et al. 2022).

2.5. Covid-19 Impact

Since the end of 2019 the entire world has been facing a public health crisis which promoted a profound restructure at a structural, governmental, economic, cultural, and social level during the last two endless years. Although football is often considered a unique economy/industry (Neale 1964), it was also significantly affected by the Covid-19 outbreak reaching a point where all activity was completely suspended, and the uncertainty dominated it.

In response to the escalation of the pandemic, the season 2019/2020 was postponed or, in some extreme cases, cancelled by several leagues, following the guidelines established by

governments and football regulators. Financial difficulties worsened during these heartless times once the revenues which were feeding up most of them suddenly shortened. This event caused a unique decrease in revenue among the European top-division clubs never stated afore, shrinking 10,4%, from €23bn in 2019 to €20,6bn in 2020 (UEFA 2022). Not only did the pandemic reduce revenue but it also changed its structure. For instance, gate receipts and matchday revenues became less significant in clubs' accounts while TV broadcast and UEFA prizes and solidarity payments gained importance. Football competitions only returned in the financial year of 2021 with various restrictions regarding stadium attendances. According to UEFA, the number of spectators in 2021 was about 94% lower than in 2019.

Following this revenue collapse, several concerns emerged regarding clubs' ability to pay their tremendous wages as their affordability became unbearable. However, sports have a different cost structure making this industry resistant when it comes to reduce costs. During Covid-19, clubs' main costs were players and staff wages as well as players transfers, which made up 91% of the total revenue in 2021 compared to 66% pre-pandemic. Despite the economic instability, clubs have reported higher wages in 2021 than in 2019, with a total of €11,9bn compared to €11,3bn, not consistent with the world economic panorama (UEFA 2022). The increase in wages indicates an improving player empowerment throughout time. Over this period, clubs have also increased their long-term liabilities by €750 million to restructure their financing and almost 25% of them evidence a negative equity.

Due to this economic uncertainty, UEFA launched several guidelines and regulations to promote clubs' stability. One of those focuses on overdue payables. Basically, UEFA and some leagues recognised transfer payables among clubs as preferential which means that in case of financial distress this type of debt must be paid first to prevent defaults in a domino effect if payments are missed. To mitigate covid impacts, not only UEFA has been increasing its money prizes and solidarity payments to compensate clubs' losses, but several states and municipal authorities are providing subsidies to support them as well.

Since the covid-19 outbreak impacted the football industry and its underlying businesses, it is expected that it will be noticeable on clubs' future performance and on their financial statements. When testing financial distress and bankruptcy prediction, an event like this one engages a huge importance for the overall study once it will degrade clubs' financial conditions and the external environment by an indefinitely time horizon. According to UEFA (2022), many more clubs would not have survived if the pandemic had arrived 10 years ago.

3. Data

The dataset created for the current study was subject to various processes to ensure its quality and accurately represent the real-world. In this section, a comprehensive overview is provided of the steps involved in its construction.

3.1. Original Sample

The dataset covers clubs that have participated in the top European divisions from 2002 to 2021, with a particular emphasis on the English, Spanish, German, Italian, French and Portuguese leagues. The clubs were selected based on a 21-year period in order to create a more robust and representative sample as well as to capture clubs with heterogeneous characteristics operating in different contexts. This historic data was obtained from the Transfermarket database, a distinguished and reliable provider of football information. The original sample was formed by 267 clubs.

For each of the 267 clubs, their financial information was retrieved from the database Orbis and complemented with data from their published financial accounts. The data covers a 5-year period from 2017 to 2021. A significant problem in the football industry is the reduced financial transparency, as some football clubs do not regularly disclose their financial accounts unless they are compelled to. Therefore, clubs that did not provide financial information for at least 2 out of the 5 years were excluded from the analysis. For clubs with only 4 years of available financial information, an average of their financial captions was done for the remaining one, considering the premise that present financial performance is a mirror of past performance.

The financial ratios chosen to be part of this study not only were selected according to their relevancy across literature (Bellovary et al. 2007; Sayari and Mugan 2017; Alaminos and Fernández 2019), but also according to the ones that the database Orbis considered important. In Table 1, all financial ratios analyzed during this research are reported.

After application of exclusion criteria, the final sample of this study is composed by 222 football clubs and 43 financial ratios related to the past 5 years (2017-2021). These 5 years will be further considered as a proxy for t-5 (5 years prior to bankruptcy) to t-1 (1 year prior to bankruptcy), respectively, for predicting distress.

Table 1 – Financial Ratios in the original sample

Attribute	Code	Financial Ratio	Attribute	Code	Financial Ratio
Leverage Ratios	Lev_1	Current Liabilities/Fixed Assets	**Liquidity Ratios**	Liq_1	Accounts Receivables/Revenue
	Lev_2	Total Liabilities/Equity		Liq_2	Cash/Current Liabilities
	Lev_3	EBIT/Interest Paid		Liq_3	Cash/Total Assets
	Lev_4	Equity/Non-current Liabilities		Liq_4	Cash/Total Liabilities
	Lev_5	Equity/Total Assets		Liq_5	Current Assets/Equity
	Lev_6	Gearing Ratio		Liq_6	Current Assets/Revenue
	Lev_7	Long Term Debt/Total Assets		Liq_7	Current Assets/Total Assets
	Lev_8	Long Term Debt/Total Liabilities		Liq_8	Current Ratio
	Lev_9	Net Debt/EBITDA		Liq_9	Fixed Assets/Total Assets
	Lev_10	Payables Accounts/Revenue		Liq_10	Net Assets/Fixed Assets
	Lev_11	Total Debt/Fixed Assets		Liq_11	Operational Cash Flow/Revenue
	Lev_12	Total Debt/Total Assets		Liq_12	Quick Ratio
	Lev_13	Total Liabilities/Total Assets		Liq_13	Working Capital/Net Income
	Lev_14	Working Capital/Total Debt		Liq_14	Working Capital/Revenue
	Lev_15	Working Capital/Total Liabilities		Liq_15	Working Capital/Total Assets

Attribute	Code	Financial Ratio
Profitability Ratios	Prof_1	EBIT/Revenue
	Prof_2	EBIT/Total Assets
	Prof_3	EBITDA Margin
	Prof_4	EBT Margin
	Prof_5	Net Income/Capital Employed
	Prof_6	Net Income/Equity
	Prof_7	Net Income/Revenue
	Prof_8	Net Income/Total Assets
	Prof_9	NOPLAT Margin
	Prof_10	Revenue/Current Assets
	Prof_11	Revenue/Equity
	Prof_12	Revenue/Fixed Assets
	Prof_13	Revenue/Total Assets

3.2. Outliers Treatment

Due to the variety of clubs in the sample in terms of size and value, outliers were expected to emerge. Since the outliers came from real-world data variability and not from a computational error or sampling problems, they were not eliminated but treated.

In this study, outliers' treatment contains a great importance in order to be correctly captured the differentiated patterns between non-distressed and distressed clubs. The main concern about outliers is that they can negatively affect the statistical analysis and the training process of the algorithms, impacting models' accuracy, stability and reliability (Hodge and Austin 2004).

Through a boxplot visualization was possible to scan the median, first quartile and third quartile and the outliers for each of the financial ratios in the sample. To address the observed outliers,

a technique defined as Interquartile Range (IQR), as analyzed by Rousseeuw and Hubert (2011), was used, which is a statistical measure of the dispersion of a dataset. Firstly, it was calculated for each financial ratio the first and third quartile and the corresponding IQR. Secondly, the lower bound and the upper bound of the boxplots were, respectively, computed for each financial ratio, according to the following formulas:

$$lower = Q1 - (1{,}5 * IQR); upper = Q3 + (1{,}5 * IQR) \qquad (1)$$

Lastly, the values that fell below the lower bound or above the upper bound were identified as outliers. To reduce their impact in the overall analysis, they were replaced with the median value (Rousseeuw and Hubert 2011). However, some of the outliers were not detected as they were situated too close to the bounds and therefore were included in the sample. Fortunately, the most accentuated and impactful outliers were scanned and treated without any constraints and without damaging the data dispersion.

There are other techniques to address outliers like log transformation. Although, the chosen method was the most suitable for the dataset and did not significantly transform the data.

3.3. Labelled Data

In the following sections of this research, both the logistic regression and the neural network will use a supervised learning approach which involves training a computer algorithm on a labelled input data to predict a particular output. This labelled data is designed to train the algorithm, allowing it to identify some implicit patterns and relationships in the data and then, using what it had learned to label new and unseen data. The main goal of training is to minimize estimation errors that may occur (Zhang et al. 1999).

As these models are intended to predict financial distress, the desired output for both models is whether a football club is in financial distress or not. To reflect this, the labelled data created was a binary variable with a value of 0 indicating that the club is healthy and do not show any signs of financial distress, and 1 if the club is in financial distress. Due to the limited information on football clubs that have faced financial distress and bankruptcy in the past few years, clubs considered to be in financial distress were those which have showed financial difficulties during the analyzed 5-year period. Based on prior literature, clubs experiencing financial difficulties are recognized as those which had a negative net income for 3 consecutive periods between 2017 and 2021 (Sayari and Mugan 2017). Consequently, the sample is composed à priori by 131 non-distressed clubs and 91 distressed clubs.

3.4. Descriptive Statistics

Table 2 – Means from non-distressed and distressed clubs between t-5 and t-1
The sample sizes for non-distressed and distressed clubs are 131 and 91, respectively. Some financial measures (means) are not directly comparable between the two groups due to the negative equity and negative results of some football clubs that underestimate the real values.

	T-1		T-2		T-3		T-4		T-5	
Variables	Non distressed	Distressed	Non distressed	Distressed	Non distressed	Distressed	Non distressed	Distressed	Non distressed	Distressed
Lev_1	2.398	3.056	1.695	3.116	2.486	3.094	2.279	3.571	2.084	2.104
Lev_2	1.105	1.243	1.274	-0.329	1.008	0.521	1.171	0.026	1.299	0.004
Lev_3	-6.572	-10.052	-2.295	-6.746	8.626	12.000	0.120	1.700	11.718	3.805
Lev_4	1.601	0.349	1.893	1.185	3.789	1.829	0.054	0.001	2.003	0.667
Lev_5	0.145	0.129	0.210	0.176	0.221	0.150	0.217	0.156	0.207	0.175
Lev_6	1.200	1.368	1.190	1.317	1.041	1.315	0.966	1.203	1.147	1.297
Lev_7	0.206	0.212	0.196	0.266	0.167	0.220	0.200	0.273	0.183	0.249
Lev_8	0.225	0.229	0.205	0.208	0.194	0.212	0.288	0.231	0.212	0.228
Lev_9	-3.630	-7.325	1.727	-4.063	2.922	2.873	0.382	0.004	1.750	-1.105
Lev_10	0.122	0.148	0.427	0.525	0.426	0.565	0.367	0.466	0.383	0.444
Lev_11	0.932	1.555	0.819	1.504	0.975	1.749	0.990	0.956	0.763	1.248
Lev_12	0.346	0.362	0.341	0.473	0.284	0.332	0.288	0.361	0.286	0.329
Lev_13	0.849	1.122	0.865	1.293	0.751	1.044	0.820	1.039	0.935	1.235
Lev_14	0.152	-0.048	0.325	0.105	0.403	0.067	0.386	0.112	0.335	0.076
Lev_15	0.039	0.010	0.050	0.023	0.051	0.001	0.061	0.025	0.036	0.016
Liq_1	0.152	0.119	0.551	0.476	0.536	0.445	0.557	0.460	0.462	0.427
Liq_2	0.195	0.139	0.219	0.154	0.149	0.107	0.172	0.118	0.185	0.129
Liq_3	0.071	0.078	0.091	0.078	0.067	0.052	0.070	0.065	0.078	0.072
Liq_4	0.111	0.082	0.158	0.112	0.100	0.068	0.089	0.077	0.109	0.086
Liq_5	2.788	1.251	1.337	0.609	1.383	0.583	1.565	0.395	1.684	0.521
Liq_6	0.148	0.116	0.153	0.131	0.149	0.123	0.158	0.128	0.127	0.120
Liq_7	0.410	0.393	0.389	0.395	0.422	0.401	0.430	0.420	0.418	0.432
Liq_8	0.771	0.622	0.799	0.622	0.775	0.721	0.783	0.678	0.765	0.711
Liq_9	0.590	0.607	0.611	0.605	0.578	0.599	0.570	0.580	0.582	0.573
Liq_10	-0.698	-1.454	0.235	-0.914	0.176	-1.136	0.090	-1.044	0.216	-0.659
Liq_11	0.044	-0.021	0.120	0.051	0.157	0.061	0.143	0.085	0.119	0.073
Liq_12	0.759	0.609	0.785	0.598	0.766	0.704	0.762	0.668	0.751	0.680
Liq_13	0.298	0.305	1.879	-0.314	1.751	-0.227	1.097	-0.180	-10.685	-7.993
Liq_14	0.052	0.002	0.064	0.012	0.041	0.005	0.059	0.013	0.030	0.007
Liq_15	0.027	-0.001	0.055	-0.010	0.040	0.008	0.048	0.017	0.035	0.010
Prof_1	-0.220	-0.349	-0.131	-0.296	-0.029	-0.346	-0.020	-0.166	-0.009	-0.117
Prof_2	-0.143	-0.195	-0.090	-0.199	0.000	-0.187	-0.020	-0.182	-0.006	-0.131
Prof_3	-0.012	-0.054	0.028	0.004	0.114	0.038	0.119	0.059	0.094	0.039
Prof_4	-0.207	-0.258	-0.100	-0.131	-0.016	-0.117	-0.038	-0.069	0.005	-0.034
Prof_5	-0.215	-0.403	-0.043	-0.153	0.041	-0.151	0.011	-0.169	0.070	-0.061
Prof_6	-0.425	-0.715	-0.117	-0.283	0.063	-0.300	0.005	-0.171	0.104	-0.061
Prof_7	-0.148	-0.323	-0.043	-0.255	0.035	-0.315	0.016	-0.148	0.020	-0.110
Prof_8	-0.079	-0.139	-0.022	-0.074	0.028	-0.088	0.010	-0.044	0.006	-0.022
Prof_9	-0.207	-0.242	-0.120	-0.139	-0.022	-0.101	-0.032	-0.061	-0.015	-0.049
Prof_10	2.218	2.228	1.916	2.025	2.154	2.066	2.473	2.333	2.569	2.552
Prof_11	15.254	7.796	2.394	0.392	3.240	2.841	2.235	1.321	3.710	-1.592
Prof_12	2.049	2.045	2.567	3.135	2.732	2.355	2.619	2.861	3.173	3.420
Prof_13	0.713	0.750	0.713	0.740	0.834	0.824	0.859	0.873	0.981	0.870

By analyzing the descriptive statistics of football clubs that are classified as non-distressed and distressed (Table 2 + Appendix 1), valuable insights can be obtained about their financial situation and the overall state of the football industry.

Referring to the leverage ratios, many football clubs struggle to generate sufficient results to meet their financial obligations due to high levels of indebtedness and weak results (e.g. Lev_3, Lev_9). This is a problem that affects both non-distressed and distressed clubs, but it can be particularly acute for those closer to financial distress. The level of debt is generally more significant for distressed clubs (e.g. Lev_2, Lev_12, etc.), but it should certainly be a point of

concern within the entire industry because of its magnitude. It is also common for football clubs to have total liabilities that exceed their total assets (Lev_13), resulting in negative equities.

Regarding liquidity ratios, non-distressed clubs tend to have a higher ability to generate funds from their current assets in order to meet financial obligations in the short-term (e.g. Liq_2, Liq_4, Liq_8, Liq_12). This stronger liquidity provides a more stable buffer for these clubs during financial instabilities.

In terms of profitability, it is notorious that neither non-distressed nor distressed clubs are profitable at all. The data suggests that football clubs often experience negative EBITDA, EBIT and net income (e.g. Prof_1, Prof_2, Prof_7, etc), even though they may generate outrageous revenues (Prof_13). In fact, it appears that non-distressed, despite being unprofitable on average, are just not as unprofitable as distressed clubs.

This analysis may support the hypothesis that football clubs prioritize sporting performance over financial stability, which can lead to bad financial practices and ultimately result in bankruptcy (Storm and Nielsen 2012). Table 3 presents some relevant aspects regarding the financial accounts of football clubs in the sample.

Table 3 – Percentage of clubs with negative performance between t-5 (2017) to t-1 (2021)

	t-1	t-2	t-3	t-4	t-5
Negative Equity	36%	30%	30%	31%	32%
Negative Working Capital	43%	41%	41%	41%	42%
Negative EBITDA	52%	48%	40%	36%	38%
Negative Net Income	76%	59%	49%	54%	47%

These descriptive statistics suggest that many football clubs would not be able to sustain their operations if they were subject to the same financial constraints as common businesses.

4. Methodology

This section includes and explains the methodologies used along the research. It highlights how the most important financial ratios for the study of financial distress were detected and how the logistic regression and the neural network model were built. In order to implement the following techniques, the programming language, Python, played a significant role (Appendix 5, 6, 7).

4.1. Financial Ratios Selection

To determine which financial ratios significantly influence the occurrence of financial distress in the football industry, a feature selection was conducted using Variance Inflator Factor and Mutual Information Classification. The financial ratios obtained will be used in both models as the most informative and less redundant.

4.1.1. Variance Inflator Factor

Before estimating the logistic regression and the neural network, it is important to address multicollinearity to avoid using superfluous and biased information. Multicollinearity arises when two or more independent variables in a model are highly correlated, not only with the dependent variable, but also with each other.

Multicollinearity typically does not significantly deteriorate the accuracy of a model. Although, failing to address it can reduce models' practical interpretability, as the coefficients assigned to each explanatory variable may lose credibility when applied into the real world (Midi et al. 2010). This issue will also reduce models' statistical power to identify which variables are statistically significant.

To overcome this problem a method called by Variance Inflator Factor (VIF) was applied. VIF measures how much the variance of a regression coefficient is inflated because of multicollinearity. Basically, it takes a financial ratio at a time as a target and the remaining ones as features, calculating multiple linear regressions. Then, each regression outcomes a R-squared (R^2) which is used to calculate the VIF value for each financial ratio using the following formula:

$$VIF_i = \frac{1}{(1 - R_i^2)} \qquad (2)$$

A high VIF value indicates that a ratio has an accentuated level of multicollinearity with the other ratios, therefore it should be ruled out as suggested by (Midi et al. 2010). According to Thompson et al. (2017), it is uncertain what the appropriate cutoff score is for eliminating

financial ratios, as it depends on the overall analysis. To decrease redundancy among the explanatory variables, the financial ratios with a VIF greater than 5 were removed from the study. However, this removal was not made at once. The VIF values were calculated multiple times, and each time, the financial ratio with the highest value was excluded. By constantly excluding the variable with the highest value, those which have remained, tend to be positively affected (VIF values may decrease) since we are reducing the overall multicollinearity.

4.1.2. Mutual Information Classification

After finding the financial ratios with the lowest degree of multicollinearity for each year, the next step focused on sorting the ratios according to their relevancy for the financial distress study. For this purpose, it was conducted a mutual information function. Mutual information is a measure of the amount of information that one variable contains about another variable (Chow and Huang 2005).

In machine learning, the mentioned mutual information function is used to measure the dependency between a set of variables (x) and a target variable (y). Contrary to VIF, mutual information is used to identify financial ratios which are strongly correlated to the target variable (financial distress), in order to make possible to recognize those which are the most useful for predicting financial distress across the football industry.

This function calculates the mutual information between each financial ratio and the target variable, returning the results in the form of a score for each ratio. These calculations are based on the Kullback-Leibler divergence which is a measure that compares the similarities and differences between the distributions of the two variables. Widely used in information theory (Vergara and Estévez 2014), the Kullback-Leibler divergence formula, where P and Q are the two probability distributions that are being compared, is defined below as:

$$D_{KL}(P||Q) = \sum_{x \epsilon X} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \qquad (3)$$

The higher the mutual information between the two variables, the more dependent they are on each other and the more useful they are likely to be for predicting the occurrence of financial distress. From the scores obtained, only the top 20% of financial ratios were retained in the study and considered as the most relevant and informative variables.

4.1.3. Statistical Tests

The next step hinges on assessing whether the group of distressed clubs and the group of non-distressed established à priori are statistically different from each other. To compare them, some statistical tests were elaborated such as a Levene's test and a t-test for equality of means.

The former pretends to check whether the variances between both groups are equal. Typically, this test is conducted before a comparison of means. If variances of the groups being compared are not equal, it may suggest they are not statistically comparable and bias the results (Gastwirth et al. 2009). The latter compares the means between both groups, which ideally should be statistically different in order to improve the accuracy of the machine learning algorithms.

4.2. Logistic Regression Model

Once the most informative financial ratios were encountered and freed from redundancy and multicollinearity, a logistic regression model was employed. Logistic regression is a machine learning algorithm mainly used for classification problems when the output is categorical. It usually estimates the probability of an event occurring based on a specific set of independent variables.

To perform a logistic regression many assumptions must be taken into consideration: 1) the dependent variable must be dichotomous, taking 1 when a club is in financial distress and 0 when a club is financially healthy; 2) the independent variables should have little or no multicollinearity and should be meaningful to the analysis; 3) the independent variables should have a linear relationship with the log-odds of the dependent variable; 4) a large sample size to provide sufficient statistical power and reliable conclusions.

A logistic regression model will allow to evaluate whether the financial ratios classified as the most informative are able to classify correctly distressed and non-distressed clubs. Indeed, the estimation of a logistic regression embraces an enormous importance mainly in terms of interpretability. Its main goal relies on identifying and highlighting which financial ratios contribute the most to clubs' financial health and those which deteriorate it.

To find the logistic regression model that best fits with the available data, a machine learning approach denominated by sample splitting was conducted. This technique picks the original sample and splits it up into two subsequent sample groups defined as training data and testing data. This process is typically done by randomly assigning observations to each group.

The training data, as the denomination indicates, is an input dataset used to train a machine learning model. Basically, this data is conceived to discover some implicit data patterns and characteristics that are responsible for putting a football club in a situation of financial distress.

Using a supervised learning, the training will estimate the logistic regression coefficients and evaluate which financial ratios are more relevant and significant for that outcome. These coefficients (β) are estimated through a maximum likelihood estimation (MLE), a method that iteratively tests various values for betas to optimize for the best fit of log odds. Its iterations will try to maximize the log likelihood function that is the major goal of a logistic regression. Therefore, it will be achieved a logistic regression equation like the following one:

$$\log\left(\frac{y}{1-y}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n \qquad (4)$$

On the other hand, the testing data is a subset of the original sample used to check and evaluate whether the logistic regression model is accurate and powerful, in predicting financial distress. The logistic regression model generated during the training phase is then applied to the testing data sample to classify the football clubs in this subset as experiencing financial distress or not. The primary objective of this classification is to evaluate the accuracy of the model by comparing its results with the actual classifications of distressed and non-distressed clubs in the real-world.

To conduct this machine learning approach, from the original sample of 222 football clubs, 80% (178 clubs) were randomly assigned to the training data and 20% (44 clubs) were randomly assigned to the testing data. According to Hastie et al. (2009), the standard practice for allocating data for training and testing is to use a 70/30 split. However, due to the sample size in this research, using a larger proportion of the data for training may improve the model's ability to learn. The usage of a larger number of clubs in the training section is a common procedure, as the objective must be to feed the logistic regression model with as much data as possible to capture all the meaningful patterns of distressed clubs and of the football industry.

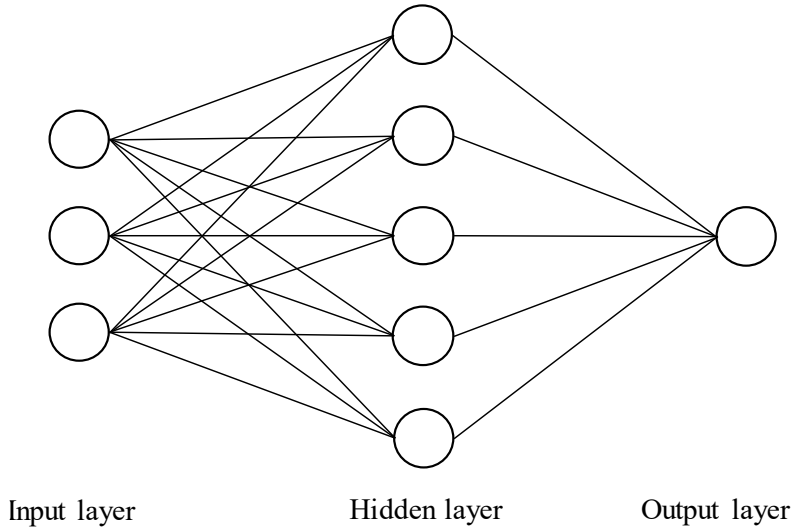
4.3. Neural Network

4.3.1. Neural Network Model

A neural network is a machine learning algorithm made up of a large number of interconnected processing nodes, also called "neurons", that work together to process complex data inputs and make predictions based on that data. These nodes communicate between themselves through weighted connections and are organized into layers, being the input layer responsible for receiving the data, the hidden layers for processing and transforming it and the output layer for producing the results.

There are several different types of neural networks. However, for this study was implemented a feedforward neural network called Multilayer Perceptron (MLP), ideal for binary classification problems (Zhang et al. 1999), which has at least three layers and where the data only flows in one direction, not looping back (Figure 1).

Figure 1 – Example of the architecture of a Multilayer Perceptron (MLP)



A neural network is a complex machine learning model which performance depends on its hyperparameters. These hyperparameters define the architecture of the network and must be set before training the model. Therefore, the main challenge relies on setting them carefully as they determine the networks' ability to learn patterns and achieve accurate results (Yu et al. 2014). They include, for instance, the number of layers, the number of neurons in each layer, the learning rate, the number of iterations ("epochs") in which the data will be trained, and the activation function used in each neuron. The hyperparameters manually assigned to the neural network are presented in Appendix 2 for visualization.

The data is processed along the neural network through a mathematical operation denominated as activation function. The activation function of a node defines the output of that node given a set of inputs, allowing the network to model non-linear relationships (Lacher et al. 1995). Different activation functions can be used in different layers or in different nodes (Sharma et al. 2020).

Similarly, to the logistic regression model, a training data and a testing data were randomly created in exactly the same terms to build the neural network model. As explained before, the training data will be responsible for capturing and learning some patterns which may justify why a football club is experiencing financial distress. The neural network will use the training data to adjust the values of its internal parameters in order to minimize the difference between

the predicted outputs and the real outputs. These parameters do not coincide with the hyperparameters already mentioned, once these parameters refer to the weights between nodes and biases that are learnt and adjusted during the training. This adjustment is typically done by an optimization algorithm such as stochastic gradient descent that adjusts the parameters reducing the estimation error.

4.3.2. GridSearch

Hyperparameters are fundamental when building a neural network because they may be the reason why the model is not working properly (Ding et al. 2008). Having the right values for those hyperparameters can ensure that the neural network is able to learn effectively and to make accurate predictions.

Therefore, a method for hyperparameter optimization defined as Grid Search was elaborated. This method uses a range of values for each hyperparameter in order to evaluate which combination of values is considered as the optimal one. To identify the optimal combination of hyperparameters, a 5-fold cross-validation approach was conducted, using the same process and goal described by Ding et al. (2008). Regardless, while the mentioned paper used a 10-fold cross-validation, this research used a 5-fold cross-validation due to the size of the training sample.

Grid search is usually applied during training to find the combination of parameters that maximizes the performance and accuracy of the neural network.

4.4. Predictive Ability

The performance of both machine learning models will depend on how accurate they are in predicting correctly non-distressed and distressed clubs. Figure 2 provides evidence of the potential outcomes of the models' classifications.

Figure 2 – Framework of the potential outcomes from classification

		Predicted Label: 0	Predicted Label: 1
Actual Label	0	Non-distressed clubs correctly predicted	Non-distressed clubs predicted as distressed
	1	Distressed clubs predicted as non-distressed	Distressed clubs correctly predicted

To evaluate model's accuracy, the metrics employed were Recall, Precision and F1-score (Meenu Sreedharan et al. 2020). The recall metric will be a fraction that relates the number of distressed (non-distressed) clubs correctly predicted with all the correct predictions that could have been made, while the precision will analyze the number of distressed (non-distressed) clubs correctly predicted with all the predictions made. Then, the measure that will define models' overall is the F1-score which is a metric that combines precision and recall into a single score:

$$F1\ score = \frac{2 * (Recall * Precision)}{Recall + Precision} \qquad (5)$$

5. Results

In this section, the performance of the logistic regression model and neural network will be evaluated. This analysis focuses mainly on understanding how the models perform in predicting financial distress 1 (t-1) to 5 years (t-5) prior to bankruptcy.

5.1. Financial Ratios Analysis

As mentioned before, to the original set of 43 financial ratios, a Variance Inflator Factor and then a Mutual Information function were employed. From this sample, all the financial ratios with a VIF value greater than 5 were discarded, and from those which had remained, only the 20% with the highest mutual information score were considered in both models. According to the performed analysis, the ratios listed in Table 4 are the most informative and relevant indicators of financial distress in the football industry for each year of the study period:

Table 4 - Financial ratios remained in the study after VIF and MI

Years	Most Informative Financial Ratios
t-1	Lev_3, Lev_10, Lev_15, Liq_13, Prof_2, Prof_5
t-2	Lev_2, Lev_7, Liq_10, Liq_13, Prof_5, Prof_6, Prof_8
t-3	Prof_1, Prof_2, Prof_5, Prof_6, Prof_7, Prof_8
t-4	Lev_2, Lev_9, Lev_11, Liq_14, Prof_1, Prof_7
t-5	Lev_2, Lev_7, Liq_10, Prof_5, Prof_6, Prof_7, Prof_8

The financial ratios mentioned above are believed to effectively capture the peculiar characteristics of the football industry. Furthermore, these ratios enable us to identify factors within the football industry that can be used to measure its levels of leverage, liquidity and profitability. At this point, it is worthy reiterating that despite these are the most relevant ratios, they do not guarantee an accurate prediction of financial distress.

Additionally, the results from Levene's tests for equality of variances and from t-tests for equality of means are reported in Table 5. These tests stand for evaluating whether the previous financial ratios diverge between distressed and non-distressed clubs for each period. According to Levene's test, the variances are equal between both groups for every timeline.

In relation to the t-test for equality of means, they demonstrated that t-3 is the only year for which the means are statistically different for both groups. The fact that groups have predominantly equal means and common characteristics may contribute to potential low accuracy rates due to its impact on models' ability to learn.

Table 5 – Results from Levene's tests and T-tests

	Levene's test for equality of variances		t-Test for equality of means	
Years	t-stat	p-value	t-stat	p-value
t-1	0.002	0.969	0.309	0.764
t-2	0.676	0.427	1.719	0.111
t-3	0.994	0.342	5.74	0.000***
t-4	0.451	0.517	0.301	0.770
t-5	0.082	0.780	1.676	0.119

*** Statistically significant at the 1% level

5.2. Logistic Regression Analysis

Logistic regression has a high inherent value due to its interpretability and predictive accuracy. The advantage of this type of model is that the output is determined by a linear combination of its inputs, making it possible to draw conclusions and obtain results that will not be possible with the neural network.

Table 6 – Logistic regression results for t-1

	ß	Std. Error	z	p-value
Constant	-0.0035	0.1495	-0.0231	0.9816
Prof_5	-0.5067	0.1871	-2.7074	0.0068***
Lev_15	-0.455	0.1717	-2.6504	0.0080***
Liq_13	-0.0463	0.1582	-0.2927	0.7698
Prof_2	-0.0181	0.1634	-0.1108	0.9117
Lev_3	-0.2009	0.1739	-1.1555	0.2479
Lev_10	0.0837	0.1503	0.5567	0.5778
No. Observations	222			
No. Iterations	5000			
Log-Likelihood	-128.77			
Pseudo R-Squared	0.107			

* Statistically significant at the 10% level
** Statistically significant at the 5% level
*** Statistically significant at the 1% level

Table 6 displays the estimated coefficients and significance levels for t-1. Exploring in-depth this table, it can be scrutinized the impact of the identified financial ratios into the financial distress event. For example, the explanatory variables Prof_5, Lev_15, Liq_13, Prof_2 and Lev_3 are negatively related to financial distress, while Lev_10 is the only one affecting it positively. It is important to notice that "negatively related" means that an increase in those ratios, will push clubs away from financial distress. Additionally, Prof_5 and Lev_15 are the

only financial ratios that are statistically significant at 1%, 5% and 10%, simultaneously, while the others are not significant at all.

By means of a classification/confusion matrix (Appendix 3), the model evidences an accuracy rate for t-1 of 71%. Also, it better predicts non-distressed clubs than distressed clubs, presenting accuracy rates of 75% and 67%, respectively. It implies that for t-1, the model had more difficulties comprehending and establishing relationships for distressed than for non-distressed clubs.

Table 7 exhibits the outcome from the logistic regression model adjusted to t-2. Prof_6 and Lev_7 have positive estimated coefficients, therefore, an increase on those ratios will put clubs closer to financial distress. In the other hand, Prof_5, Liq_13, Prof_8, Lev_2 and Liq_10 push away a scenario of distress. Across all these explanatory variables, Prof_5 and Prof_8 are statistically significant only at a 10% significance level, whilst Lev_7 is statistically relevant at 5% and 10%.

Table 7 – Logistic regression results for t-2

	ß	Std. Error	z	p-value
Constant	0.0098	0.1517	0.0647	0.9484
Prof_5	-0.4192	0.2273	-1.8445	0.0651*
Prof_6	0.0374	0.2253	0.1660	0.8682
Liq_13	-0.3732	0.2293	-1.6277	0.1036
Prof_8	-0.3570	0.2062	-1.7312	0.0834*
Lev_2	-0.3749	0.2365	-1.5857	0.1128
Liq_10	-0.0670	0.1507	-0.4447	0.6565
Lev_7	0.3960	0.1756	2.2547	0.0242**
No. Observations	222			
No. Iterations	6000			
Log-Likelihood	-127.10			
Pseudo R-Squared	0.118			

* Statistically significant at the 10% level

** Statistically significant at the 5% level

*** Statistically significant at the 1% level

In terms of accuracy, the logistic regression produced a predictive ability of 79% for non-distressed clubs and of 76% for distressed clubs. It stands out the improvement in relation to t-1, allowing the model to achieve an overall accuracy of 78%.

Furthermore, this machine learning algorithm have produced some of the most intriguing results for t-3. Analyzing the following Table 8, Prof_6, Prof_7 and Prof_8 are the only independent variables that are statistically meaningful. The first one at a significance level of 5% and 10% and the last two at 1%, 5% and 10%. Another common feature of the mentioned explanatory variables is the fact that all of them have a negative contribution to the financial distress occurrence, like the ratio Prof_5 as well. Additionally, Prof_1 and Prof_2 have the opposite effect.

Table 8 – Logistic regression results for t-3

	ß	Std. Error	z	p-value
Constant	0.2437	0.2540	0.9594	0.3374
Prof_5	-0.0906	0.4370	-0.2073	0.8358
Prof_6	-1.1489	0.4729	-2.4297	0.0151**
Prof_1	0.1612	0.3821	0.4218	0.6732
Prof_8	-1.5799	0.4960	-3.1854	0.0014***
Prof_7	-2.2212	0.6939	-3.2009	0.0014***
Prof_2	0.2366	0.3955	0.5983	0.5496
No. Observations	222			
No. Iterations	8000			
Log-Likelihood	-61.983			
Pseudo R-Squared	0,57			

* Statistically significant at the 10% level

** Statistically significant at the 5% level

*** Statistically significant at the 1% level

Surprisingly, the logistic regression model conducted for t-3 had an overall accuracy of 93% whose percentage was the highest regarding all years. The accuracy rate for distressed and non-distressed clubs were also very similar, being 92% and 94%, respectively.

According to Table 9, for t-4, all the estimated coefficients evidence negative values. These values indicate that an increase in the predictor variables is directly associated with a decrease in the probability of financial distress, mitigating it. The most relevant ratios are Prof_7 and Liq_14, being statistically significant at 1%, 5% and 10%, and Lev_11 which is also meaningful but only at a 10% significance level.

Table 9 – Logistic regression results for t-4

	ß	Std. Error	z	p-value
Constant	0.0508	0.1649	0.3000	0.7581
Prof_7	-1.0713	0.2929	-3.6575	0.0003***
Lev_9	-0.1038	0.1644	-0.6318	0.5275
Lev_2	-0.1952	0.1699	-1.1489	0.2506
Prof_1	-0.0861	0.2399	-0.3589	0.7196
Lev_11	-0.3011	0.1740	-1.7303	0.0836*
Liq_14	-0.6794	0.1915	-3.5472	0.0004***
No. Observations	222			
No. Iterations	8000			
Log-Likelihood	-61.983			
Pseudo R-Squared	0,57			

* Statistically significant at the 10% level

** Statistically significant at the 5% level

*** Statistically significant at the 1% level

The model exhibits an overall accuracy rate of 73% for a 4-year period prior to bankruptcy. In this case, a deterioration on the prediction of distressed clubs is noticeable since it only achieves a 63% rate. The accuracy rate for non-distressed clubs remains on 79% which implies the model continues to have less issues in classifying this type of clubs.

Table 10 – Logistic regression results for t-5

	ß	Std. Error	z	p-value
Constant	0.0278	0.1581	0.1760	0.8603
Prof 5	-0.3262	0.2651	-1.2308	0.2184
Prof_6	0.0176	0.2457	0.0717	0.9428
Prof_8	-0.0996	0.2207	-0.4513	0.6518
Liq_10	-0.0330	0.1495	-0.2205	0.8255
Lev_2	-0.2613	0.1643	-1.5902	0.1118
Prof_7	-0.5829	0.2355	-2.4752	0.0133**
Lev_7	0.5168	0.1697	3.0449	0.0023***
No. Observations	222			
No. Iterations	6000			
Log-Likelihood	-120.74			
Pseudo R-Squared	0.163			

* Statistically significant at the 10% level

** Statistically significant at the 5% level

*** Statistically significant at the 1% level

Lastly, the logistic analysis for the 5-year period prior to bankruptcy (t-5) may be found in Table 10. Prof_5, Prof_8, Liq_10, Lev_2 and Prof_7 have negative betas which means that all these ratios contribute negatively to a financial distress event, as they will reduce its probability. In addition, the coefficients assigned to Prof_6 and Lev_7 are positive. In terms of relevancy, Prof_7 and Lev_7 are statistically significant, the former at 5% and 10%, while the latter at 1%, 5% and 10%.

For, t-5, the model showed an accuracy rate of 60%. The low predictive ability of the model in classifying distressed clubs, 53%, had a significant impact on the model's overall accuracy rate. Regarding non-distressed clubs, the model classifies them correctly 65% of the times.

The results from the logistic regression model may be seen summed up in Table 11.

Table 11 – Summary of logistic regression results

Logistic Regression				
Years	Distressed Clubs	Non-distressed clubs	Overall accuracy	Significant Financial Ratios
t-1	67%	75%	71%	Lev_15, Prof_5
t-2	76%	79%	78%	Lev_7, Prof_5, Prof_8
t-3	92%	94%	93%	Prof_6, Prof_7, Prof_8
t-4	63%	79%	73%	Lev_11, Liq_14, Prof_7
t-5	53%	65%	60%	Lev_7, Prof_7

In addition, the estimated logistic regressions models that predict financial distress for 1 (t-1) to 5 (t-5) prior to bankruptcy are presented below:

$$\rightarrow FD_{t-1} = -0.0035 - 0.5067Prof_5 - 0.455Lev_{15} - 0.0463Liq_{13} - 0.0181Prof_2 - 0.2009Lev_3 + 0.0837Lev_{10} \quad (6)$$

$$\rightarrow FD_{t-2} = 0.0098 - 0.4192Prof_5 + 0.0374Prof_6 - 0.3732Liq_{13} - 0.3570Prof_8 - 0.3749Lev_2 - 0.0670Liq_{10} + 0.3960Lev_7 \quad (7)$$

$$\rightarrow FD_{t-3} = 0.2437 - 0.0906Prof_5 - 1.1489Prof_6 + 0.1612Prof_1 - 1.5799Prof_8 - 2.2212Prof_7 + 0.2366Prof_2 \quad (8)$$

$$\rightarrow FD_{t-4} = 0.0508 - 1.0713Prof_7 - 0.1038Lev_9 - 0.1952Lev_2 - 0.0861Prof_1 - 0.3011Lev_{11} - 0.6794Liq_{14} \quad (9)$$

$$\rightarrow FD_{t-5} = 0.0278 - 0.3262Prof_5 + 0.0176Prof_6 - 0.0996Prof_8 - 0.0330Liq_{10} - 0.2613Lev_2 - 0.5829Prof_7 + 0.5168Lev_7 \quad (10)$$

5.3. Robustness Test

In perspective, the logistic regression showed interesting results, however, the main concern is how well the developed model will generalize and perform when facing new real-world data, different from the one used in this research, ensuring the same level of accuracy.

One way to determine the validity and robustness of the logistic regression model would be to apply it to a new dataset that it has not seen before, performing an out-of-sample testing (Bellovary et al. 2007). However, due to lack of information, most clubs do not have available financial accounts prior to 2017.

Therefore, to examine the stability of this model, the logistic equations from t-1 (6), t-2 (7) and t-3(8) were matched with the financial information from 2017 to 2019. The reason behind this strategy focuses not only on analyzing the adaptability of the model but also on determining whether the covid scenario has had a negative impact on its development. For practical application of this test, clubs classified as distressed will be considered in such condition only by the end of 2019. This classification also includes clubs that have exhibited negative income over the 3-year period (2017-2019). Thereby, among the 222 football clubs in the sample, 57 were classified à priori as distressed while 165 were classified as non-distressed.

For 2019, using the estimated logistic regression (6), the model reported an accuracy rate of 77%. Scrutinizing this rate, the classification accuracy of non-distressed and distressed clubs set on 86% and 45%, respectively. Despite an improvement in the overall accuracy, the rate for distressed clubs have deteriorated in relation to the one presented before (67%). This deterioration may derive from the fact that equation (6) was developed in a scenario where the Covid-19 pandemic had significantly impacted the overall financials of clubs. With that being said, the estimated logistic regression had more issues in predicting financial distress for 2019.

Moving forward to 2018, the model (7) documented a predictive ability of 71%, predicting correctly 80% of non-distressed clubs and 47% of distressed clubs. In this case, the overall accuracy is lower than the model accuracy for t-2 (78%). Once again, the previous designed logistic regression for t-2 may be biased due to the impact of covid, explaining the low precision in predicting distress.

Lastly, the equation (8) was carried out for 2017. Its overall accuracy accounted for 86%, the highest among the 3 years. In fact, not only t-3 was the year with the highest accuracy in this robustness test, but also in the estimation of the logistic regression model (93%). For non-distressed and distressed clubs, the accuracy rates stood on 88% and 75%, correspondingly. For t-3, the robustness test corroborates the fact that the model can predict distress with a greater probability of success when compared to the others.

Results from the test are synthetize in Table 12. It is important to note that, for each period, the overall accuracy achieved in the robustness test is highly inflated by a higher number of non-distressed clubs in the sample. This imbalance leads to an overestimation of the real accuracy.

Table 12 – Logistic regression's robustness test results

Robustness Test			
Years	Distressed Clubs	Non-distressed clubs	Overall accuracy
t-1	45%	86%	77%
t-2	47%	80%	71%
t-3	75%	88%	86%

5.4. Neural Network Analysis

Generally, neural networks such as a Multilayer Perceptron (MLP) are known for not having the most suitable degree of interpretability when compared with other machine learning algorithms (Jones et al. 2017). This lack of interpretability arises from the fact that neural networks' inner workings in processing input data are too complex, establishing several non-linear and abstract relationships. Additionally, some decisions/predictions took by neural networks may not be easy to understand because some parameters such as the weights may not have a clear intuitive meaning as most of these values are determined by training the data and optimization algorithms. Indeed, its main advantage hinges on its out-of-sample predictive success (Jones et al. 2017).

Therefore, once a MLP do not directly provide information about which input variables are the most important, the performance of the conducted neural networks from t-1 to t-5 will be assessed mainly in terms of their accuracy in predicting financial distress across football clubs (Appendix 4).

There are two types of results evaluated by the neural network model, without the Grid Search and with it. The first type takes only into account a neural network with the hyperparameters manually assigned (Appendix 2), while the second applies the optimization technique to the same network. In Table 13, we may verify the results vary slightly along with the usage of Grid Search. Despite this technique plays an optimization role, for some periods it worsened models' performance (t-2 and t-5). This happens due to an overfitting problem which emerges when a model is too complex for the underlying data, not because the training data is simple, but because there is not enough of it. Therefore, the model ends up memorizing the data rather than learning from it (Hastie et al. 2009).

Bearing that in mind, the predictive ability for each year will be evaluated through the most favorable scenario between them.

Table 13 – Neural network model best scenario

	(1) Without Grid Search			(2) With Grid Search			
Years	Distressed Clubs	Non-distressed clubs	Accuracy	Distressed Clubs	Non-distressed clubs	Accuracy	Best Scenario
t-1	54%	68%	62%	62%	71%	67%	(2)
t-2	80%	84%	82%	71%	75%	73%	(1)
t-3	89%	92%	91%	89%	92%	91%	(2)
t-4	70%	79%	76%	77%	82%	80%	(2)
t-5	68%	70%	69%	51%	55%	53%	(1)

As reported above, the neural network algorithm reached its highest accuracy for t-3, exhibiting an overall rate of 91%. This includes a 92% success rate for correctly classifying non-distressed clubs and an 89% for distressed clubs. On the other hand, and surprisingly, t-1 was the year with the lowest accuracy, 67%, as result of a 71% rate for healthy and a 63% rate for unhealthy clubs. For t-2 and t-4, the model presented accuracy rates of 82% and 80%, respectively. Both periods demonstrated identical predictive abilities, although t-2 showed slightly higher rates for non-distressed and distressed clubs, at 84% and 80% accordingly, compared to 82% and 77% for t-4. Lastly, for t-5, the neural network has correctly predicted the status of non-distressed clubs in 70% of the cases, while only predicting it correctly for distressed clubs in 68% of the times, leading to an overall accuracy of 69%.

6. Discussion of Results

According to previous literature on the financial distress topic, models' predictive ability tends to shrink throughout time, evidencing better performances in the years preceding bankruptcy (Bellovary et al. 2007). Nevertheless, this research paper does not corroborate it, since t-1 was one of the worst years in predicting financial distress. This situation accrues from the covid outbreak. Overall, the pandemic has had a negative impact on clubs' financial accounts. While in 2020, some clubs have experienced a slight reduction in expenditure due to the uncertainty surrounding the pandemic, in 2021, many have reported record-high expenditures despite the unfavorable scenario of recession, which has further exacerbated their financial situation, as mentioned in section 2. These events may have caused issues with the models' ability to accurately classify clubs based on their performance specially at time t-1. Both models classified at least 40% of the clubs as distressed in t-1, indicating the profound effect of the pandemic on the industry as a whole. It is worth noting that the financial ratios highlighted for t-1 and t-2 may not accurately reflect the typical industry characteristics across football, as the Covid-19 pandemic had a significant impact on it.

Therefore, upon reviewing the remaining ratios from t-3 to t-5 for the event of financial distress, it appears that the most relevant ratios are related to profitability (e.g. Prof_6 and Prof_7) and leverage (e.g. Lev_7 and Lev_9), setting aside the liquidity ones. There are a significant number of ratios that use net income as numerator, rather than EBITDA or EBIT. Those financial ratios may highlight the costs associated with high levels of debt, which can be seen as a determining factor in distinguishing distressed from non-distressed clubs. The level of debt is also frequently analyzed in various ratios due to its magnitude and impact on clubs. For example, Lev_7 is often used to assess the balance between clubs' long-term debt and total assets.

Regarding the logistic regression, its robustness test matches the overall accuracy achieved by the model for t-1, t-2 and t-3. However, particularly for t-1 and t-2, the robustness test evidenced low accuracy rates for predicting distressed clubs, which were set on 45% (2019) and 47% (2018), respectively. These results suggest that this part of the model has some limitations when applied to real-world and unseen data, due to the pandemic effect.

Despite the usage of different methodologies, when comparing the performance of the logistic regression to that of the neural network, it is notable that both models follow similar lines of prediction. In terms of the overall accuracy, the logistic regression outperforms in t-1 and t-3 and underperforms in t-2, t-4 and t-5. It appears that as we go further back in time, the neural network tends to show higher accuracy compared to the logistic regression. This may confirm

the perception that neural networks are capable of diving deeper into processing data than logistic regressions.

Surprisingly, both models demonstrated a significant accuracy rate for t-3, with 93% for the regression and 91% for the network. They also exhibited a high predictive ability in classifying correctly distressed clubs, 92% and 89%, respectively. This aligns with the findings from the t-test in section 5, which showed that the means of non-distressed and distressed groups were only statistically different for t-3.

Additionally, both models tend to perform better at predicting the outcomes of non-distressed than distressed clubs, which is very common in this type of studies (Bellovary et al. 2007). A contributing factor to this is that clubs were initially classified as distressed based on a proxy, rather than being considered clubs that have actually filed for bankruptcy. Despite the fact that more detailed information would have made the overall findings even more insightful and accurate, the models demonstrated a positive ability to capture patterns and identify distress in the football industry.

7. Limitations and Future Research

7.1. Limitations

Firstly, the major drawback of this research paper hinges on the fact that football clubs are not compelled to disclosure their financial accounts on a regular basis which compromises their level of transparency. With that being said, the data retrieved from Orbis and complemented by clubs' financial reports is subject to certain limitations. For instance, the data contains missing values for some clubs' accounts and a couple of them did not have enough available information for the entire time horizon (2017-2021). These issues were treated, but they surely had implications for the accuracy of the models in differentiating distressed from non-distressed groups.

Secondly, the implemented machine learning models usually require a larger sample size in order to provide more accurate and stable results. A larger sample would increase the quality of the training data in terms of capturing more effectively patterns associated to the football industry.

Thirdly, to classify distressed and non-distressed clubs a priori, a proxy was utilized since there was no available real-world information about a sample of clubs that had already filed for bankruptcy. Thus, for some cases, this classification may bring some misunderstandings as it may not properly reflect the real financial conditions of certain clubs.

Lastly, while this research only took into consideration financial ratios as financial distress predictors, there are other variables related to sportive performance and clubs' governance that could have been used to improve models' accuracy. Despite this limitation, the research findings are still valuable as one of the main objectives was to scrutinize the financial drivers of financial distress in the football industry and evaluate the effect of clubs operating at the limits of their financial capacity on their survival.

7.2. Future Research

A well-reasoned point about limitations is that they should not be seen as something purely negative and unfavorable. Not only they are important to provide context and perspective, but also to feed up future research so as to not let some subjects die.

The potential for research on predicting financial distress in the football industry has a huge spectrum, with the success of such research largely dependent on the amount and quality of available information. Once there is transparent information, interesting and more reliable results can be achieved. There are several potential areas of investigation and controversial

factors that may emerge related to financial distress and to the overall financial side of the football industry.

Future research can dig deeper into the complex and multifaceted causes of financial distress in the football industry by considering a wider range of variables that were not addressed in the present research. This study may act as a useful foundation for that purpose.

Moreover, there are several engaging topics related to financial distress in the football industry that might be worth to explore, including a more detailed analysis of clubs' capital structure, the effectiveness of financial restructuring strategies on clubs that have filed for bankruptcy, the impacts of financial distress on a club's local community and why some clubs are bailed out from bankruptcy while others are not.

Lastly, in recent years, an increasing number of private equities and hedge funds have acquired stakes in football clubs. Since these investors are profit-orientated, it may be interesting to investigate the financial prospects of the football industry over the next couple years and their impact on the acquired clubs.

8. Conclusion

The topic of financial distress is of great importance in a variety of fields because of its potential consequences for business participants, including the football industry. This is not only relevant for decision-making, but also for protecting creditors, investors, shareholders, employers and local communities against failure.

This dissertation presents an analysis of financial distress in the football industry performing two machine learning models, a logistic regression and a neural network. The aims of the study are to gain a better understanding of the financial performance of football clubs, how it can be characterized and evaluate how the models perform in such a disruptive and special industry.

This research suggests that football clubs often struggle to manage their finances properly. The descriptive statistics as well as the coefficients from the logistic regressions illustrate that these clubs are heavily indebted, have negative equity and consistently generate negative results. It appears that clubs that were initially thought to be financially stable and those that were considered distressed experience financial difficulties. This implies that football clubs prioritize other factors over financial stability and profitability as mentioned by Storm and Nielsen (2012). These findings also suggest that the intended goals of FFP regulations may not have been achieved, raising questions about their effectiveness.

The logistic regression and neural network models demonstrate a satisfactory predictive power across the football industry, despite covid-19 pandemic, industry-specific challenges and all the mentioned limitations of this research. These factors make it difficult to compare the results from this research with those from other studies. However, when we compare the results of these models with the performance of the models used by Alaminos and Fernández (2019), it is not accurate or coherent to assert they perform worse.

To conclude, a question that may arise from this dissertation is “Even though the number of bankruptcies among these clubs is relatively low, why is the prediction of financial distress important in the context of professional football?”. This research aims to highlight the current state of the football industry with the goal of promoting more transparency and encouraging regulators to act. For instance, they can use some of the financial ratios selected as the most relevant and informative to monitor football clubs in a similar way how banks use covenants to manage their relationships with borrowers. It also intends to motivate managers to change their financial policies and to inform investors about the inherent risks of the industry. By doing so, it aims to provide evidence of long-term unsustainable difficulties in the football industry if no changes are made.

9. References

Ahtiainen, Santeri, and Henry Jarva. 2022. "Has UEFA's Financial Fair Play Regulation Increased Football Clubs' Profitability?" *European Sport Management Quarterly* 22 (4). https://doi.org/10.1080/16184742.2020.1820062.

Alaminos, David, and Manuel Ángel Fernández. 2019. "Why Do Football Clubs Fail Financially? A Financial Distress Prediction Model for European Professional Football Industry." *PLoS ONE* 14 (12). https://doi.org/10.1371/journal.pone.0225989.

Altman, Edward I. 1968. "FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY." *The Journal of Finance* 23 (4). https://doi.org/10.1111/j.1540-6261.1968.tb00843.x.

Altman, Edward I. 2000. "Predicting Financial Distress of Companies: Revisiting The Z-Score And Zeta Models. Updated From E. Altman, Financial Ratios, Discriminant Analysis and The Prediction of Corporate Bankruptcy, Journal of Banking & Finance 1."

Barajas, Ángel, and Plácido Rodríguez. 2010. "Spanish Football Clubs' Finances: Crisis and Player Salaries." *International Journal of Sport Finance*.

Barboza, Flavio, Herbert Kimura, and Edward Altman. 2017. "Machine Learning Models and Bankruptcy Prediction." *Expert Systems with Applications* 83. https://doi.org/10.1016/j.eswa.2017.04.006.

Beaver, William H. 1966. "Financial Ratios As Predictors of Failure." *Journal of Accounting Research* 4. https://doi.org/10.2307/2490171.

Bellovary, Jodi L., Don E. Giacomino, and Michael D. Akers. 2007. "A Review of Bankruptcy Prediction Studies: 1930-Present." *Journal of Financial Education* 33 (Winter). https://doi.org/10.1017/CBO9781107415324.004.

Chow, Tommy W.S., and Di Huang. 2005. "Estimating Optimal Feature Subsets Using Efficient Estimation of High-Dimensional Mutual Information." *IEEE Transactions on Neural Networks* 16 (1). https://doi.org/10.1109/TNN.2004.841414.

Ding, Yongsheng, Xinping Song, and Yueming Zen. 2008. "Forecasting Financial Condition of Chinese Listed Companies Based on Support Vector Machine." *Expert Systems with Applications* 34 (4). https://doi.org/10.1016/j.eswa.2007.06.037.

Dobson, Stephen, and John Goddard. 2011. *The Economics of Football*. *The Economics of Football*. https://doi.org/10.1017/cbo9780511973864.

Drut, Bastien, and Gaël Raballand. 2012. "Why Does Financial Regulation Matter for European Professional Football Clubs?" *International Journal of Sport Management and Marketing* 11 (1–2). https://doi.org/10.1504/IJSMM.2012.045488.

Evans, Richard, Geoff Walters, and Sean Hamil. 2022. "Gambling in Professional Sport: The Enabling Role of 'Regulatory Legitimacy.'" *Corporate Governance (Bingley)* 22 (5). https://doi.org/10.1108/CG-07-2021-0251.

Franck, Egon. 2010. "Private Firm, Public Corporation or Member's Association Governance Structures in European Football." In *International Journal of Sport Finance*. Vol. 5.

Francois, Aurélien, Nadine Dermit-Richard, Daniel Plumley, Rob Wilson, and Natacha Heutte. 2022. "The Effectiveness of UEFA Financial Fair Play: Evidence from England and France, 2008–2018." *Sport, Business and Management: An International Journal* 12 (3). https://doi.org/10.1108/SBM-03-2021-0024.

Gastwirth, Joseph L., Yulia R. Gel, and Weiwen Miao. 2009. "The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice." *Statistical Science* 24 (3). https://doi.org/10.1214/09-STS301.

Guzmán, Isidoro, and Stephen Morrow. 2007. "Measuring Efficiency and Productivity in Professional Football Teams: Evidence from the English Premier League." *Central European Journal of Operations Research* 15 (4). https://doi.org/10.1007/s10100-007-0034-y.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *Springer Series in Statistics The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. *Springer*. Vol. 2nd.

Hodge, Victoria J., and Jim Austin. 2004. "A Survey of Outlier Detection Methodologies." *Artificial Intelligence Review*. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9.

Jones, Stewart, David Johnstone, and Roy Wilson. 2015. "An Empirical Evaluation of the Performance of Binary Classifiers in the Prediction of Credit Ratings Changes." *Journal of Banking and Finance* 56. https://doi.org/10.1016/j.jbankfin.2015.02.006.

Jones, S., Johnstone, D., & Wilson,R. 2017. "Predicting Corporate Bankruptcy: An Evaluation of Alternative Statistical Frameworks." *Journal of Business Finance and Accounting* 44 (1–2). https://doi.org/10.1111/jbfa.12218.

Lacher, R. C., Pamela K. Coats, Shanker C. Sharma, and L. Franklin Fant. 1995. "A Neural Network for Classifying the Financial Health of a Firm." *European Journal of Operational Research* 85 (1). https://doi.org/10.1016/0377-2217(93)E0274-2.

Meenu Sreedharan, Ahmed M. Khedr, and Magdi el Bannany. 2020. "A Multi-Layer Perceptron Approach to Financial Distress Prediction with Genetic Algorithm." *Automatic Control and Computer Sciences* 54 (6). https://doi.org/10.3103/S0146411620060085.

Midi, Habshah, S. K. Sarkar, and Sohel Rana. 2010. "Collinearity Diagnostics of Binary Logistic Regression Model." *Journal of Interdisciplinary Mathematics* 13 (3). https://doi.org/10.1080/09720502.2010.10700699.

Neale, Walter C. 1964. "The Peculiar Economics of Professional Sports." *Quarterly Journal of Economics* 78 (1). https://doi.org/10.2307/1880543.

Ohlson, James A. 1980. "Financial Ratios and the Probabilistic Prediction of Bankruptcy." *Journal of Accounting Research* 18 (1). https://doi.org/10.2307/2490395.

Peeters, Thomas, and Stefan Szymanski. 2014. "Financial Fair Play in European Football." *Economic Policy* 29 (78). https://doi.org/10.1111/1468-0327.12031.

Plumley, Daniel, Girish Mohan Ramchandani, and Robert Wilson. 2019. "The Unintended Consequence of Financial Fair Play: An Examination of Competitive Balance across Five European Football Leagues." *Sport, Business and Management: An International Journal* 9 (2). https://doi.org/10.1108/SBM-03-2018-0025.

Rousseeuw, Peter J., and Mia Hubert. 2011. "Robust Statistics for Outlier Detection." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1 (1). https://doi.org/10.1002/widm.2.

Sayari, Naz, and Can Simga Mugan. 2017. "Industry Specific Financial Distress Modeling." *BRQ Business Research Quarterly* 20 (1). https://doi.org/10.1016/j.brq.2016.03.003.

Sharma, Siddharth, Simone Sharma, and Anidhya Athaiya. 2020. "ACTIVATION FUNCTIONS IN NEURAL NETWORKS." *International Journal of Engineering Applied Sciences and Technology* 04 (12). https://doi.org/10.33564/ijeast.2020.v04i12.054.

Shumway, Tyler. 2001. "Forecasting Bankruptcy More Accurately: A Simple Hazard Model." *Journal of Business* 74 (1). https://doi.org/10.1086/209665.

Storm, Rasmus K., and Klaus Nielsen. 2012. "Soft Budget Constraints in Professional Football." *European Sport Management Quarterly* 12 (2). https://doi.org/10.1080/16184742.2012.670660.

Szymanski, Stefan. 2017. "Entry into Exit: Insolvency in English Professional Football." *Scottish Journal of Political Economy* 64 (4). https://doi.org/10.1111/sjpe.12134.

Szymanski, Stefan, and Daniel Weimar. 2019. "Insolvencies in Professional Football: A German Sonderweg?" *International Journal of Sport Finance* 14 (1). https://doi.org/10.32731/IJSF.141.022019.05.

Thompson, Christopher Glen, Rae Seon Kim, Ariel M. Aloe, and Betsy Jane Becker. 2017. "Extracting the Variance In Flation Factor and Other Multicollinearity Diagnostics from Typical Regression Results." *Basic and Applied Social Psychology* 39 (2). https://doi.org/10.1080/01973533.2016.1277529.

Union of European Football Associations, UEFA (2018) Club Licensing and financial fair play regulations, Edition 2018.

Union of European Football Associations, UEFA (2022) The European Club Footballing Landscape, Edition 2022.

Vergara, Jorge R., and Pablo A. Estévez. 2014. "A Review of Feature Selection Methods Based on Mutual Information." *Neural Computing and Applications*. https://doi.org/10.1007/s00521-013-1368-0.

Wilson, Robert, Daniel Plumley, and Girish Ramchandani. 2013. "The Relationship between Ownership Structure and Club Performance in the English Premier League." *Sport, Business and Management: An International Journal* 3 (1). https://doi.org/10.1108/20426781311316889.

Yu, Qi, Yoan Miche, Eric Séverin, and Amaury Lendasse. 2014. "Bankruptcy Prediction Using Extreme Learning Machine and Financial Expertise." *Neurocomputing* 128. https://doi.org/10.1016/j.neucom.2013.01.063.

Zhang, Guoqiang, Michael Y. Hu, B. Eddy Patuwo, and Daniel C. Indro. 1999. "Artificial Neural Networks in Bankruptcy Prediction: General Framework and Cross-Validation Analysis." *European Journal of Operational Research* 116 (1). https://doi.org/10.1016/S0377-2217(98)00051-4.

Zmijewski, Mark E. 1984. "Methodological Issues Related to the Estimation of Financial Distress Prediction Models." *Journal of Accounting Research* 22. https://doi.org/10.2307/2490859.

10. Appendixes

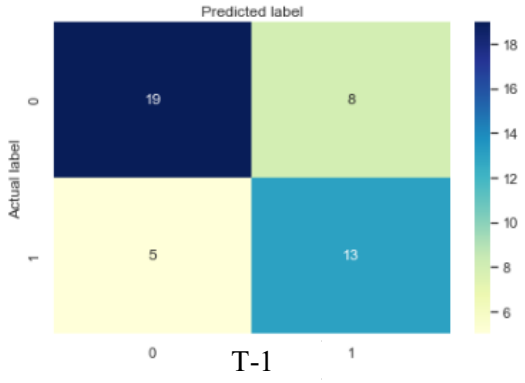
Appendix 1 – Standard Deviations from non-distressed and distressed clubs from t-1 to t-5

	T-1		T-2		T-3		T-4		T-5	
Variables	Non-distressed	Distressed	Non-distressed	Distressed	Non-distressed	Distressed	Non-distressed	Distressed	Non-distressed	Distressed
Lev_1	6.768	5.622	3.561	5.383	6.802	6.141	5.570	8.609	4.208	2.560
Lev_2	16.693	6.988	4.964	4.487	3.175	3.784	4.086	3.827	7.118	6.882
Lev_3	9.377	10.954	11.759	12.587	19.256	17.943	27.378	22.992	19.605	23.408
Lev_4	4.221	3.728	3.981	5.659	7.168	9.161	4.129	3.164	4.743	4.043
Lev_5	0.163	0.127	0.134	0.117	0.137	0.103	0.143	0.124	0.133	0.119
Lev_6	0.608	0.551	0.482	0.414	0.711	0.530	0.602	0.475	0.576	0.467
Lev_7	0.153	0.141	0.198	0.209	0.128	0.141	0.163	0.196	0.140	0.133
Lev_8	0.074	0.067	0.068	0.058	0.074	0.064	0.082	0.071	0.075	0.051
Lev_9	53.223	74.966	17.358	23.345	8.415	13.296	6.406	8.117	12.829	15.923
Lev_10	0.076	0.086	0.302	0.338	0.278	0.334	0.228	0.224	0.211	0.235
Lev_11	1.367	2.474	1.229	2.250	2.051	3.227	1.779	0.965	0.961	1.476
Lev_12	0.251	0.265	0.337	0.456	0.208	0.227	0.239	0.250	0.202	0.234
Lev_13	0.426	0.547	0.638	0.935	0.310	0.457	0.455	0.481	0.732	0.827
Lev_14	0.434	0.392	0.617	0.562	0.720	0.618	0.642	0.526	0.649	0.510
Lev_15	0.072	0.060	0.093	0.082	0.082	0.070	0.080	0.067	0.071	0.072
Liq_1	0.101	0.088	0.349	0.374	0.335	0.327	0.361	0.297	0.287	0.317
Liq_2	0.202	0.159	0.190	0.161	0.129	0.127	0.180	0.132	0.175	0.142
Liq_3	0.059	0.065	0.066	0.063	0.050	0.053	0.063	0.058	0.063	0.067
Liq_4	0.116	0.087	0.147	0.122	0.096	0.085	0.093	0.082	0.101	0.094
Liq_5	20.815	6.040	3.186	3.930	3.569	3.450	3.001	7.626	5.656	4.966
Liq_6	0.098	0.083	0.097	0.103	0.093	0.090	0.102	0.083	0.079	0.090
Liq_7	0.113	0.102	0.106	0.119	0.112	0.100	0.114	0.106	0.107	0.111
Liq_8	0.408	0.368	0.397	0.361	0.286	0.281	0.354	0.309	0.281	0.269
Liq_9	0.113	0.102	0.106	0.119	0.112	0.100	0.114	0.106	0.107	0.105
Liq_10	7.167	5.197	3.782	5.110	5.605	5.236	6.018	5.900	3.865	3.138
Liq_11	0.125	0.102	0.112	0.110	0.090	0.108	0.107	0.105	0.101	0.101
Liq_12	0.405	0.364	0.397	0.334	0.291	0.278	0.352	0.322	0.275	0.262
Liq_13	5.575	2.203	22.885	8.860	5.149	2.664	4.528	2.836	48.684	32.228
Liq_14	0.098	0.098	0.114	0.101	0.080	0.076	0.072	0.070	0.066	0.064
Liq_15	0.069	0.073	0.125	0.110	0.072	0.058	0.060	0.056	0.059	0.058
Prof_1	0.279	0.314	0.282	0.373	0.295	0.372	0.182	0.235	0.202	0.248
Prof_2	0.201	0.210	0.214	0.289	0.176	0.189	0.175	0.261	0.211	0.249
Prof_3	0.131	0.100	0.142	0.130	0.119	0.109	0.123	0.100	0.119	0.113
Prof_4	0.129	0.114	0.118	0.086	0.104	0.077	0.095	0.079	0.086	0.101
Prof_5	0.312	0.294	0.193	0.202	0.158	0.150	0.239	0.249	0.256	0.261
Prof_6	0.448	0.452	0.370	0.298	0.275	0.245	0.269	0.234	0.303	0.286
Prof_7	0.221	0.306	0.211	0.339	0.163	0.316	0.152	0.212	0.158	0.236
Prof_8	0.089	0.095	0.077	0.080	0.062	0.089	0.073	0.086	0.072	0.079
Prof_9	0.116	0.119	0.137	0.110	0.101	0.078	0.102	0.079	0.104	0.107
Prof_10	1.335	1.278	0.932	0.972	0.961	0.997	1.363	1.138	1.391	1.533
Prof_11	86.580	24.770	6.146	8.273	10.039	11.614	18.268	23.597	12.619	15.512
Prof_12	2.614	2.333	4.238	6.311	3.770	2.789	3.684	3.691	5.391	5.790
Prof_13	0.253	0.248	0.331	0.321	0.238	0.278	0.321	0.291	0.497	0.440

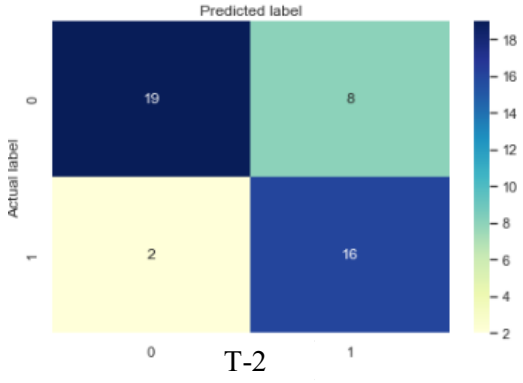
Appendix 2 – Hyperparameters manually defined for the neural network

Hyperparameter	Description	Function/value assigned
Activation function	These functions are responsible for establishing non-linear relationships transforming some given inputs into the required outputs. They include ReLU, sigmoid and tanh	Input and hidden layer: ReLU Output layer:Sigmoid
Batch size	Defines the number of samples formed from the training data that are processed to update the neural network's weights. At the end, they are compared to estimate the error and improve model's performance	10
Dropout rate	Determines the portion of neurons that will be dropped out at each training step	0.0
Learning rate	Sets how fast the weights are updated along the training	0.01
Loss Function	Function that is used to calculate the error between the predicted output of the network and the true output	Binary cross-entropy
Number of epochs	Establishes how many times the training data passes trough the training process	20
Number of neurons	Neurons are responsible for receiving, processing and producing data.	Input layer:6; Hidden layer:12; Output layer:1
Numbers of hidden layers	Layers located exactly between the input layer and the output layer	1
Optimization algorithm	Responsible for modifying the weights from the neural network based on the gradient of the loss function	Adam

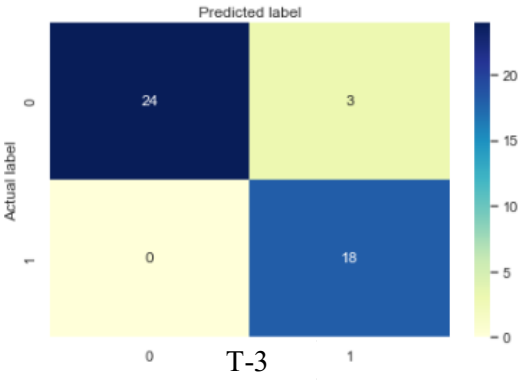
Appendix 3 - Confusion matrices for the logistic regression from t-1 until t-5



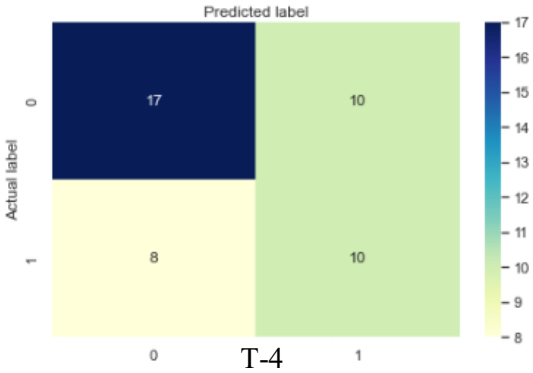
T-1



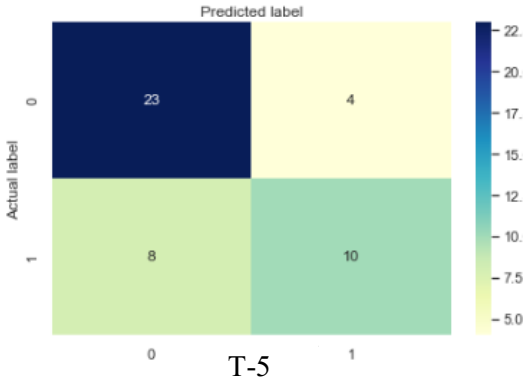
T-2



T-3

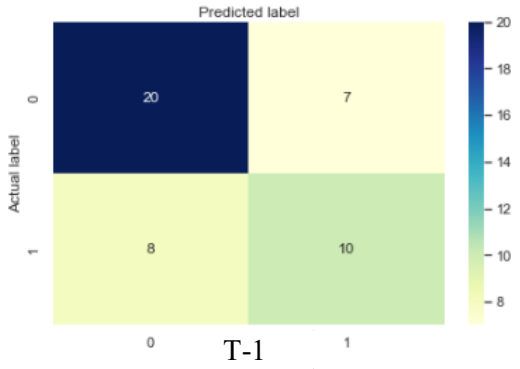


T-4

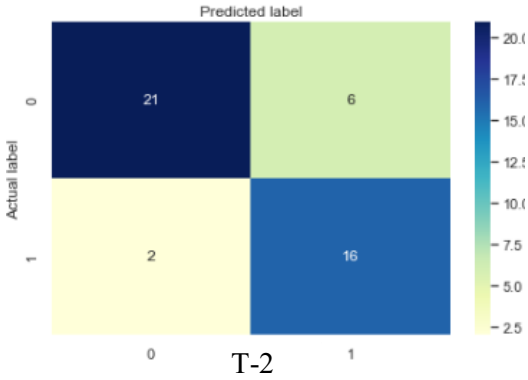


T-5

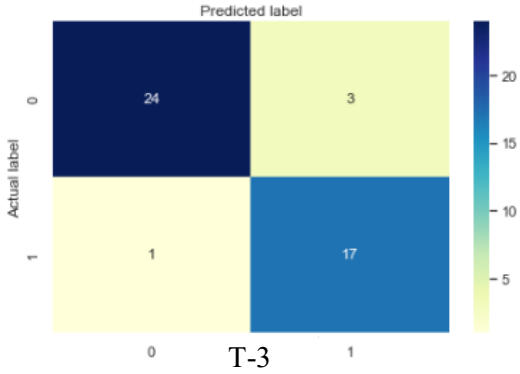
Appendix 4 – Confusion matrices for the neural network model from t-1 to t-5



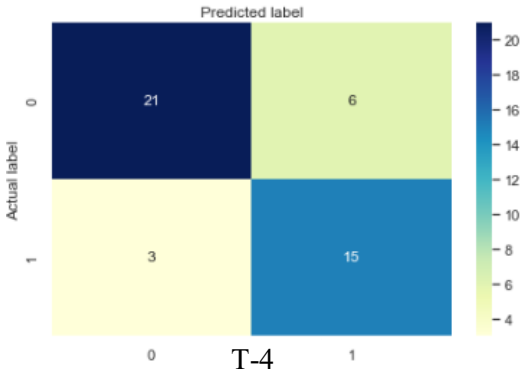
T-1



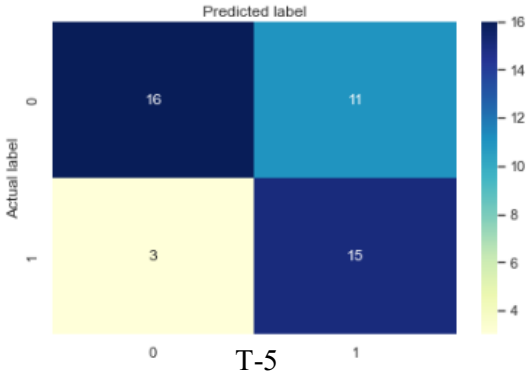
T-2



T-3



T-4



T-5

Appendix 5 – Python code for cleaning and treating the data

```
import pandas as pd
import os
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression
from imblearn.over_sampling import RandomOverSampler
from statsmodels.stats.outliers_influence import variance_inflation_factor
from statsmodels.tools.tools import add_constant
import warnings
import seaborn as sns
warnings.filterwarnings('ignore')
# identify outliers with interquartile range
from numpy.random import seed
from numpy.random import randn
from numpy import percentile


#get the path
path = os.getcwd()
files = os.listdir(path)
files


#create a dataframe
df = pd.DataFrame()


data = pd.read_excel('FinalRatios.xlsx','2021')
df = df.append(data)



df.describe()

df.groupby(by=["FD"]).mean()

df.groupby(by=["FD"]).std()


list1 =df.isnull().sum()

list1.sort_values()


duplicateRows = df[df.duplicated()]
#view duplicate rows
duplicateRows


#Check Data type
df.info()


#Remove and transform rows with NA
data=df
data = data.dropna(how='all')
```

```
data = data.fillna(data.mean())
data.isnull().sum()
data = data.dropna()


# Creating a histogram to see if the classes are balanced
fig, ax = plt.subplots(figsize =(10, 7))
ax.hist(data['FD'])
# Show plot
plt.show()


out= data.iloc[:,1:48]


# Check Outliers
fig = plt.figure()
for column in out:
    plt.figure()
    sns.boxplot(out[column])


fig = plt.figure()
for column in out:
    plt.figure()
    sns.distplot(out[column])


#Treat some outliers
#Two possible ways
#1st IQR
for col in out.columns:
    median = out[col].mean()
    Q1=out[column].quantile(0.25)
    Q3=out[column].quantile(0.75)
    IQR=Q3-Q1
    lwr_bound = Q1-(1.5*IQR)
    upr_bound = Q3+(1.5*IQR)

    newcol = []
    for val in out[col]:
        if val < lower or val > upper:
            newcol.append(median)
        else:
            newcol.append(val)

    out[col] = newcol


# 2nd using mean e std
for col in out.columns:
    mean = out[col].mean()
    std = out[col].std()

    N = 1
    upper = mean + N*std
    lower = mean - N*std

    median = out[col].median()
```

```
    newcol = []
    for val in out[col]:
        if val < lower or val > upper:
            newcol.append(median)
        else:
            newcol.append(val)

    out[col] = newcol

#Check if there are still outliers
fig = plt.figure()
for column in out:
    plt.figure()
    sns.boxplot(out[column])


new_data=out
new_data['FD']=data['FD']


#Check the correlation for some variables

corr = new_data.corr()
sns.heatmap(corr, vmin=-1, vmax=1,
            xticklabels=corr.columns.values,
            yticklabels=corr.columns.values,annot=True)
plt.show()


corr["FD"].sort_values(ascending=False)


new_data.columns


#SAVE treated df
new_data.to_csv('data_clean.csv',index=False)
```

Appendix 6 – Python code for the implementation of the logistic regression

```
#Imports
import pandas as pd
import os
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression
from imblearn.over_sampling import RandomOverSampler
from statsmodels.stats.outliers_influence import variance_inflation_factor
from statsmodels.tools.tools import add_constant
import warnings
import seaborn as sns
warnings.filterwarnings('ignore')
from sklearn.model_selection import train_test_split
sns.set(style="white")
sns.set(style="whitegrid", color_codes=True)
from sklearn import metrics
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report


data = pd.read_csv('data_clean_2021.csv')
data


Variance Inflator Factor
vif_data = pd.DataFrame()
vif_data["ratio"] = data.iloc[:,1:43].columns
vif_data["VIF"] = [variance_inflation_factor(data.iloc[:,1:43].values, i)
                   for i in range(len(data.iloc[:,1:43].columns))]


vif_data.sort_values("VIF",ascending=False)


vif_data["VIF"][vif_data["VIF"]>5].count()


x_vif = data.iloc[:,:43]
y = data['FD']

# threshold of 5
thres = 5


# looping the VIF
while True:
    Cols = range(x_vif.shape[1])

    vif = np.array([variance_inflation_factor(x_vif.values, i) for i in
Cols])
    if all(vif < thres):
        break
```

```
    else:
        Cols = np.delete(Cols,np.argmax(vif))
        x_vif = x_vif.iloc[:,Cols]

x_vif.columns

# new data according to vif
new_data = data[['Net_Income/Equity', 'Net_income/Capital_employed',
       'Net_income/Total_assets', 'EBITDA_margin',
       'Cash_flow/Operating_revenue', 'EBIT/Interest_Paid',
       'Payable_Accounts/Revenue', 'Equity/Non_Current_Liabilities',
       'Equity/Total_Assets', 'Working_Capital/Total_Assets',
       'Net_Assets/Fixed_Assets', 'Net_Income/Revenue',
       'Cash/Total_Liabilities', 'Net_Debt/EBITDA', 'Debt-to-equity',
       'EBIT/Total_Assets', 'Cash/Current_Liabilities',
       'Total_Debt/Fixed_Assets', 'Long_Term_debt/Total_Assets',
       'Working_Capital/Total_Debt', 'Working_Capital/Total_Liabilities',
       'Working_Capital/Net_Income', 'Current_Liabilities/Fixed_Assets',
       'Revenue/Equity', 'Revenue/CurrentAssets', 'Revenue/Fixed_Assets',
       'Working_Capital/Revenue', 'Current_Assets/Equity',
       'Cash/Total_Assets','FD']]


# Train/Test split
# Import `train_test_split` from `sklearn.model_selection`
from sklearn.model_selection import train_test_split

# Specify the data
X = new_data.iloc[:,:29]

# Specify the target labels and flatten the array
y = np.ravel(new_data.FD)

# Split the data up in train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20,
random_state=42)


#Oversampling Logistic Regression
import imblearn
import collections
from collections import Counter

# define oversampling strategy
oversample = RandomOverSampler(sampling_strategy='minority',random_state=0)
# fit and apply the transform
X_over, y_over = oversample.fit_resample(X_train, y_train)
# summarize class distribution
print(Counter(y_over))


#Scaler
# Import `StandardScaler` from `sklearn.preprocessing`
from sklearn.preprocessing import StandardScaler

# Define the scaler
scaler = StandardScaler().fit(X_over)
```

```
# Scale the train set
X_train = scaler.transform(X_over)

# Scale the test set
X_test = scaler.transform(X_test)
```

Mutual Information

```
#Selecting features using univarite Feature Selection
#Select top features based on mutual_info_classif
#np.random.seed(16)
selector = SelectPercentile(mutual_info_classif,percentile=20)
selector.fit(X_train, y_over)
X.columns[selector.get_support()]


print(X.columns.get_loc('Net_income/Capital_employed'))
print(X.columns.get_loc('Working_Capital/Total_Liabilities'))
print(X.columns.get_loc('Working_Capital/Net_Income'))
print(X.columns.get_loc('EBIT/Total_Assets'))
print(X.columns.get_loc('EBIT/Interest_Paid'))
print(X.columns.get_loc('Payable_Accounts/Revenue'))


X_train_1 = X_train[:,[1,20,21,15,5,6]]
X_test_1 = X_test[:,[1,20,21,15,5,6]]
```

Logistic Regression application

```
logit_model=sm.Logit(y_over,sm.add_constant(X_train_1))
result=logit_model.fit()
print(result.summary2())


result_df = pd.DataFrame({"Coefficient": result.params, "p-value":
result.pvalues})
print(result_df)


# instantiate the model (using the default parameters)
logreg = LogisticRegression(random_state=16)

# fit the model with data
logreg.fit(X_train_1, y_over)

y_pred = logreg.predict(X_test_1)


cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
cnf_matrix


target_names = ['No_Default', 'Default']
print(classification_report(y_test, y_pred, target_names=target_names))


class_names=[0,1] # name  of classes
fig, ax = plt.subplots()
tick_marks = np.arange(len(class_names))
plt.xticks(tick_marks, class_names)
```

```
plt.yticks(tick_marks, class_names)
# create heatmap
sns.heatmap(pd.DataFrame(cnf_matrix), annot=True, cmap="YlGnBu" ,fmt='g')
ax.xaxis.set_label_position("top")
plt.tight_layout()
plt.title('Confusion matrix', y=1.1)
plt.ylabel('Actual label')
plt.xlabel('Predicted label')
```

Appendix 7 – Python code for the implementation of the neural network

```
#Imports
import pandas as pd
import os
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression
from imblearn.over_sampling import RandomOverSampler
from statsmodels.stats.outliers_influence import variance_inflation_factor
from statsmodels.tools.tools import add_constant
import warnings
import seaborn as sns
warnings.filterwarnings('ignore')
from sklearn.model_selection import train_test_split
sns.set(style="white")
sns.set(style="whitegrid", color_codes=True)
from sklearn import metrics
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report


data = pd.read_csv('data_clean_2021.csv')
data.shape


np.random.seed(16)


# Train/Test split
# Import `train_test_split` from `sklearn.model_selection`
from sklearn.model_selection import train_test_split

# Specify the data
X = data.iloc[:,:43]

# Specify the target labels and flatten the array
y = np.ravel(data.FD)

# Split the data up in train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20,
random_state=42)


#Oversampling Logistic Regression
import imblearn
import collections
from collections import Counter

np.random.seed(16)
# define oversampling strategy
oversample = RandomOverSampler(sampling_strategy='minority',random_state=0)
# fit and apply the transform
X_over, y_over = oversample.fit_resample(X_train, y_train)
```

```
# summarize class distribution
print(Counter(y_over))


#Scaler
# Import `StandardScaler` from `sklearn.preprocessing`
from sklearn.preprocessing import StandardScaler

# Define the scaler
scaler = StandardScaler().fit(X_over)

# Scale the train set
X_train = scaler.transform(X_over)

# Scale the test set
X_test = scaler.transform(X_test)


np.random.seed(16)
print(X.columns.get_loc('Net_income/Capital_employed'))
print(X.columns.get_loc('Working_Capital/Total_Liabilities'))
print(X.columns.get_loc('Working_Capital/Net_Income'))
print(X.columns.get_loc('EBIT/Total_Assets'))
print(X.columns.get_loc('EBIT/Interest_Paid'))
print(X.columns.get_loc('Payable_Accounts/Revenue'))


X_train = X_train[:,[1,32,34,23,7,9]]
X_test = X_test[:,[1,32,34,23,7,9]]
```

Neural Network without Grid Search

```
# Import `Sequential` from `keras.models`
from keras.models import Sequential

# Import `Dense` from `keras.layers`
from keras.layers import Dense

np.random.seed(16)
# Initialize the constructor
model = Sequential()

# Add an input layer
model.add(Dense(6, activation='relu', input_shape=(6,)))

# Add one hidden layer
model.add(Dense(12, activation='relu'))

# Add an output layer
model.add(Dense(1, activation='sigmoid'))


# Model output shape
model.output_shape

# Model summary
model.summary()
```

```
# Model config
model.get_config()

# List all weight tensors
model.get_weights()


model.compile(loss='binary_crossentropy',
              optimizer='adam',
              metrics=['accuracy'])

model.fit(X_train, y_over,epochs=20, batch_size=10, verbose=1)


score = model.evaluate(X_test, y_test,verbose=1)

print(score)


y_pred=(model.predict(X_test) >= 0.5).astype("int32")


# Import the modules from `sklearn.metrics`
from sklearn.metrics import confusion_matrix, precision_score,
recall_score, f1_score, accuracy_score

# Confusion matrix
print(confusion_matrix(y_test, y_pred))

# Precision
print(precision_score(y_test, y_pred))

# Recall
print(recall_score(y_test, y_pred))

#accuracy
print(accuracy_score(y_test,y_pred))


target_names = ['No_Default', 'Default']
print(classification_report(y_test, y_pred, target_names=target_names))


neural_cf_matrix =confusion_matrix(y_test, y_pred)


class_names=[0,1] # name  of classes
fig, ax = plt.subplots()
tick_marks = np.arange(len(class_names))
plt.xticks(tick_marks, class_names)
plt.yticks(tick_marks, class_names)
# create heatmap
sns.heatmap(pd.DataFrame(neural_cf_matrix), annot=True, cmap="YlGnBu"
,fmt='g')
ax.xaxis.set_label_position("top")
plt.tight_layout()
plt.title('Confusion matrix', y=1.1)

plt.ylabel('Actual label')
plt.xlabel('Predicted label')
```

Defining the GridSearch for further application

```
# Importing the necessary packages
from sklearn.model_selection import GridSearchCV, KFold
from keras.models import Sequential
from keras.layers import Dense
from keras.wrappers.scikit_learn import KerasClassifier
from keras.optimizers import Adam


def create_model():
# create model
    model = Sequential()
    model.add(Dense(6, activation='relu', input_shape=(6,)))
    model.add(Dense(12, activation='relu'))
    model.add(Dense(1, activation='sigmoid'))
    adam=Adam(lr=0.01)
    model.compile(loss='binary_crossentropy', optimizer=adam,
metrics=['accuracy'])
    return model


# create model
model = KerasClassifier(build_fn = create_model,verbose = 0)
# Define the grid search parameters
batch_size = [10,20,40]
epochs = [10,50,100]
# Make a dictionary of the grid search parameters
param_grid = dict(batch_size = batch_size,epochs = epochs)
# Build and fit the GridSearchCV
grid = GridSearchCV(estimator = model,param_grid = param_grid,cv =
KFold(),verbose = 10)
grid_result = grid.fit(X_train,y_over)


print('Best : {}, using
{}'.format(grid_result.best_score_,grid_result.best_params_))
means = grid_result.cv_results_['mean_test_score']
stds = grid_result.cv_results_['std_test_score']
params = grid_result.cv_results_['params']
for mean, stdev, param in zip(means, stds, params):
  print('{},{} with: {}'.format(mean, stdev, param))


#Gridsearch learning rate & dropout rate
from keras.layers import Dropout

# Defining the model

def create_model(learning_rate,dropout_rate):
    model = Sequential()
    model.add(Dense(6, activation='relu', input_shape=(6,)))
    model.add(Dropout(dropout_rate))
    model.add(Dense(12, activation='relu'))
    model.add(Dropout(dropout_rate))
    model.add(Dense(1,activation = 'sigmoid'))

    adam = Adam(lr = learning_rate)
```

```
    model.compile(loss = 'binary_crossentropy',optimizer = adam,metrics =
['accuracy'])
    return model

# Create the model

model = KerasClassifier(build_fn = create_model,verbose = 0,batch_size =
20,epochs = 100)

# Define the grid search parameters

learning_rate = [0.001,0.01,0.1]
dropout_rate = [0.0,0.1,0.2]

# Make a dictionary of the grid search parameters

param_grids = dict(learning_rate = learning_rate,dropout_rate =
dropout_rate)

# Build and fit the GridSearchCV

grid = GridSearchCV(estimator = model,param_grid = param_grids,cv =
KFold(),verbose = 10)
grid_result = grid.fit(X_train,y_over)


# Summarize the results

print('Best : {}, using
{}'.format(grid_result.best_score_,grid_result.best_params_))
means = grid_result.cv_results_['mean_test_score']
stds = grid_result.cv_results_['std_test_score']
params = grid_result.cv_results_['params']
for mean, stdev, param in zip(means, stds, params):
    print('{},{} with: {}'.format(mean, stdev, param))


# Defining the model

def create_model(neuron1,neuron2):
    model = Sequential()
    model.add(Dense(neuron1, activation='relu', input_shape=(6,)))
    model.add(Dropout(0.0))
    model.add(Dense(neuron2, activation='relu'))
    model.add(Dropout(0.0))
    model.add(Dense(1,activation = 'sigmoid'))

    adam = Adam(lr = 0.01)
    model.compile(loss = 'binary_crossentropy',optimizer = adam,metrics =
['accuracy'])
    return model

# Create the model

model = KerasClassifier(build_fn = create_model,verbose = 0,batch_size =
20,epochs = 100)
```

```
# Define the grid search parameters

neuron1 = [4,8,12,16,20,24,28,32,36,40]
neuron2 = [2,4,8,10,12,14,16,18,20]

# Make a dictionary of the grid search parameters

param_grids = dict(neuron1 = neuron1,neuron2 = neuron2)

# Build and fit the GridSearchCV

grid = GridSearchCV(estimator = model,param_grid = param_grids,cv =
KFold(),verbose = 10)
grid_result = grid.fit(X_train,y_over)


# Summarize the results

print('Best : {}, using
{}'.format(grid_result.best_score_,grid_result.best_params_))
means = grid_result.cv_results_['mean_test_score']
stds = grid_result.cv_results_['std_test_score']
params = grid_result.cv_results_['params']
for mean, stdev, param in zip(means, stds, params):
    print('{},{} with: {}'.format(mean, stdev, param))
```

Neural Network with Grid Search

```
# Import `Sequential` from `keras.models`
from keras.models import Sequential

# Import `Dense` from `keras.layers`
from keras.layers import Dense

np.random.seed(16)
# Initialize the constructor
model = Sequential()

# Add an input layer
model.add(Dense(20, activation='relu', input_shape=(6,)))
model.add(Dropout(0.0))


# Add one hidden layer
model.add(Dense(4, activation='relu'))
model.add(Dropout(0.0))

# Add an output layer
model.add(Dense(1, activation='sigmoid'))


# Model output shape
model.output_shape

# Model summary
model.summary()
```

```
# Model config
model.get_config()

# List all weight tensors
model.get_weights()


adam = Adam(lr = 0.01)
model.compile(loss='binary_crossentropy',
              optimizer=adam,
              metrics=['accuracy'])

model.fit(X_train, y_over,epochs=100, batch_size=20, verbose=1)


score = model.evaluate(X_test, y_test,verbose=1)

print(score)


y_pred=(model.predict(X_test) > 0.5).astype("int32")


# Import the modules from `sklearn.metrics`
from sklearn.metrics import confusion_matrix, precision_score,
recall_score, f1_score, accuracy_score

# Confusion matrix
print(confusion_matrix(y_test, y_pred))

# Precision
print(precision_score(y_test, y_pred))

# Recall
print(recall_score(y_test, y_pred))

#accuracy
print(accuracy_score(y_test,y_pred))


target_names = ['No_Default', 'Default']
print(classification_report(y_test, y_pred, target_names=target_names))


neural_cf_matrix =confusion_matrix(y_test, y_pred)


class_names=[0,1] # name  of classes
fig, ax = plt.subplots()
tick_marks = np.arange(len(class_names))
plt.xticks(tick_marks, class_names)
plt.yticks(tick_marks, class_names)
# create heatmap
sns.heatmap(pd.DataFrame(neural_cf_matrix), annot=True, cmap="YlGnBu"
,fmt='g')
ax.xaxis.set_label_position("top")
plt.tight_layout()
plt.title('Confusion matrix', y=1.1)
```

```
plt.ylabel('Actual label')
plt.xlabel('Predicted label')
```