The quest for business value drivers: applying machine learning to performance management

Pre print version, the final article is published at the following url:
https://www.tandfonline.com/doi/full/10.1080/09537287.2022.2157776?src=

The quest for business value drivers: applying Machine Learning to Performance Management

Abstract
The paper explores the potential role of Machine learning (ML) in supporting the development of a company's Performance Management System (PMS). In more details, it investigates the capability of ML to moderate the complexity related to the identification of the business value drivers (methodological complexity) and the related measures (analytical complexity). A second objective is the analysis of the main issues arising in applying ML to performance management. The research, developed through an action research design, shows that ML can moderate complexity by (1) reducing the subjectivity in the identification of the business value drivers; (2) accounting for cause-effect relationships between business value drivers and performance; (3) balancing managerial interpretability vs. predictivity of the approach. It also shows that the realisation of such benefits requires a combined understanding of the ML techniques and of the performance management model of the company to frame and validate the algorithm in light of the context in which the organisation operates. The paper contributes to the literature analysing the role of business analytics in the field of performance management and it provides new insights into the potential benefits of introducing an ML-based PMS and the issues to consider to increase its effectiveness.

1. Introduction

Performance Management Systems (PMSs) refer to the operational mechanisms aimed at driving the performance of the organization towards the expected goals. More specifically they deal with "the evolving formal and informal mechanisms, processes, systems, and networks used by organizations for conveying the key objectives and goals elicited by management, for assisting the strategic process and ongoing management through analysis, planning, measurement, control, rewarding, and broadly managing performance, and for supporting and facilitating organizational learning and change." (Ferreira and Otley 2009, p. 264). A growing number of studies have been discussing the role of PMSs in today's business environment where increased dynamism and organisational complexity call into question the effectiveness of established approaches (Nudurupati, Garengo and Bititci 2021, Otley and Soin 2014, Melnyk et al. 2014). With change often coming from unexpected directions, the fundamental tenet that the determinants of success can be known in advance and measures can be identified to manage their achievement is considered especially problematic (Bourne et al. 2018).

The identification of the business value drivers, i.e. the factors that affect the company's value, by driving the organisational performance and the analysis of their interactions have always been critical aspects of designing PMSs (Ferreira and Otley 2009, Bourne et al. 2000, Kaplan and Norton 1996, Simons 1995). On one side, the ability to align PMSs to organisational objectives (Gimbert, Bisbe and Mendoza 2010, Bhimani and Langfield-Smith 2007), capture their multiple drivers through diverse type of measures (Hall 2008, Ittner, Larcker and Meyer 2003) and define cause-effect relationships (Bisbe and Malagueno 2012, Hall 2011, Malina, Nørreklit and Selto 2007, Garengo, Biazzo and Bititci 2005) have been proved to deliver a number of organisational benefits (Lucianetti, Battista, Koufteros 2019, Micheli and Mura 2017, Bititci et al. 2015, Koufteros, Verghese and Lucianetti 2014, Wiersma 2009, Chenhall 2005). On the other one, their definition remains a challenging process (De Waal and Counet 2009, Ahn 2001), whose inappropriate management has been associated with a reduced effectiveness of PMSs (Franco-Santos and Otley 2018, Bourne, Kennerley and Franco-Santos 2005, Lipe and Salterio 2000, Nørreklit 2000).

The literature identifies a number of sources of complexity in the definition of business value drivers and related measures (Smith, Binns and Tushman 2010, Benson-Rea, Brodie and Sima 2013) and which can arise at two main

levels (Okwir et al. 2018): methodological and analytical[1]. More specifically, methodological complexity refers to problems associated with the identification of business value drivers. These arise from the multiple internal and external demands organisations typically face and which make their performance the result of a number of different factors (Benson-Rea, Brodie and Sima 2013, Santos, Belton and Howick 2002). On the other hand, analytical complexity concerns the specific selection of the related measures and the understanding of the interdependences between different variables and aspects of the business. These have been traditionally difficult to capture by PMSs given their non-linearity, inherent complexity and presence of lag effects (Kelly 2010, Johanson et al. 2006, Nørreklit 2000). Both methodological and analytical complexities are compounded by increased market dynamism and sophistication, which, on one side, increase the challenges of understanding what drives performance (Bisbe and Malagueno 2012) and on the other one requires the PMS not only to quickly adapt but also to help anticipate change (Bourne et al. 2018, Melnyk et al. 2014). At the same time, the huge and fast-growing amount of internally and externally generated data further increases such complexities (Nudurupati, Tebboune and Hardman 2016, Bhimani and Willcocks 2014).

Complexity in the definition and measurement of value drivers have been typically managed by relying on managerial wisdom (Ferreira and Otley 2009) and the use of top-down, subjective processes (Buytendijk, Hatch, and Micheli 2010, Ittner, Larcker and Meyer 2003, Malina and Selto 2001) with limited use of quantitative, data-driven approaches (Raffoni et al. 2018, López-Ospina et al. 2017, Ittner and Larcker 2005). However, these present some limitations (Campbell et al. 2015). For instance, the subjective selection of business value drivers and related measures has been associated with lack of strategic focus (Cheng and Humphreys 2012), questioning of selected drivers (Papalexandris, Ioannou, Prastacos 2004), misallocation of individual efforts (Burney, Henle and Widener 2009, Krishnan, Luft and Shields 2005), uncertainty over the valuation process (Ittner, Larcker and Meyer 2003). At the same time, recent studies show how interdependences among drivers and measures are generally overlooked in the implementation of PMSs (Lucianetti, Battista, Koufteros 2019, Silvi et al. 2015)

Based on the previous considerations and following recent calls on the use of business analytics (BA) to support PMSs implementation (Nudurupati, Garengo and Bititci 2021, Bourne et al. 2018, Mello, Leite, and Martins 2014), our study investigates whether methodological and analytical complexities in the identification of value drivers can be mitigated through the adoption of Machine Learning (ML) techniques.

ML includes different algorithms and approaches that are able to learn from past data without being specifically programmed to do so and present technical features which, from a theoretical standpoint, may address the above-mentioned sources of complexity. For instance, ML algorithms outperform conventional statistical approaches when there are a multitude of variables, their relationships are not linear (Granger and Terasvirta 1993), the value of the variables continuously evolves and data availability is extensive (Sokolova and Lapalme 2009). A key strength is their ability to learn from past data and adapt, which allow them to make predictions in conditions that mimic real-world, changeable and complex combinations of factors (Ryll and Seydens 2019, Syam and Sharma 2018). As explained above, the identification of value drivers falls in this class of issues, thus leading to hypothesize a potential role of ML in supporting the identification of the value drivers. The potential benefits of ML in dealing with complex, non-linear, evolving variables have already generated increasing interest in social sciences (Hindman 2015) and a number of ML applications can be found in business studies (Algorithmia 2020), especially in marketing and operations (Dzyabura and Yoganarasimhan 2018, Wuest et al. 2016). However, research also shows that in business contexts the application of BA (Sharma, Mithas, and Kankanhalli 2014) and ML more specifically presents several challenges, including data availability and quality (Nudurupati et al. 2011), interpretability (Meske et al. 2022) and validation of the results (Giboney et al. 2015). Hence, a second and related purpose of our work is to examine the issues that may arise when applying a ML approach for performance management purposes.

1 Okwir et al. 2018 framework identifies a third source of complexity namely technological complexity. However, it is not specifically linked to the identification and measurement of business value drivers but to the role of IT systems and platforms in supporting PMSs (Nudurupati et al. 2011, Nudurupati and Bititci 2005).

Our work adopts an action research approach (Kasanen, Lukka, and Siitonen 1993). Our choice was motivated by the contextual need of a detailed understanding of organisational practices and addressing an emerging challenge for which changes had not yet initiated.

The paper is organised as follows. Section 2 reviews the relevant literature on value drivers in the context of PMSs implementation and ML approaches. Section 3 presents the details of our methodological choices whilst Section 4 describes our empirical observations. The paper then ends with the discussion and our concluding remarks in Sections 5 and 6 respectively.

2. Theoretical framework

2.1. Challenges in the identification and measurement of business value drivers

A business value driver is a variable influencing a company's value (Rappaport 1998, Copeland, Koller, and Murrin 1999). The understanding and control of the most important business value drivers are crucial aspects to drive the strategy of the company towards the most relevant targets. The analysis of business value drivers has been developed in different fields of management research, from value-based management (Copeland, Koller, and Murrin 1999) to corporate finance (Damodaran 2012) and management control (Kaplan and Norton 1996; Ittner and Larcker 2001). Several studies have classified business value drivers. For instance, Rappaport (1998) highlights three main groups of drivers: operational, investment and financial, while Ittner and Larcker (2001) present 10 main value driver categories: financials, purchasers, employees, operational, quality, alliances, supply, environment, innovations and society. Aside from different classifications, extant studies present a number of commonalities in relation to the identification and measurement of business value drivers. In particular: (1) they are specific to the organisation and their choice of business model (Teece 2010, Lepak, Smith, and Taylor 2007), (2) they are complex and varied deriving from the multitude of logics that typically co-exist within organisations (Benson-Rea, Brodie, and Sima 2013, Smith, Binns, and Tushman 2010), (3) they have relational and dynamic properties which require a holistic rather than an element-based approach (Benson-Rea, Brodie, and Sima 2013, Nørreklit 2000).

In the field of performance management, the identification and measurement of business value drivers has always been interpreted as a foundational phase in the development of the PMSs (Simons 1995, Ferreira and Otley 2009). For instance, the well-established Ferreira and Otley's (2009) performance management framework considers the identification of the "key factors that are believed to be central to the organisation's overall future success" (2009, p. 268) one of the first questions to address. More broadly, there is strong consensus on the need to align the PMS to the objectives and goals of an organisation (Micheli and Manzoni 2010, Chenhall 2005, Bourne et al., 2000). Traditionally, business value drivers have been identified following a strategic planning process and based on what is considered to be important by the management concerned (Ferreira and Otley 2009). Experience and personal judgement guide the selection process whilst data analysis and quantitative approaches play a limited role (CIMA 2014, Silvi et al. 2012, Malina and Selto 2001). A similar process is followed in the setting of the related measures, which are typically specified through top-down, subjective approaches (Kelly 2010, Malina, Nørreklit, and Selto 2007, Ittner, Larker, and Meyer 2003) and cascaded down throughout the organisation. The effectiveness of such approaches has often been questioned and even more so in the current dynamic economic environment (Nudurupati, Garengo and Bititci 2021, Bourne et al. 2018, Melnyk et al. 2014).

According to Okwir et al. (2018) systematic literature review, issues related to the identification of business value drivers and their measurement can be linked to two forms of complexity: methodological and analytical. Methodological complexity is associated with the varied, often conflicting and dynamic nature of organisational objectives (Sundin, Grandlund, and Brown 2010), which challenges the managerial ability to identify the business value drivers (Benson-Rea, Brodie and Sima 2013, Santos, Belton and Howick 2002). Aside of bounded rationality arguments (Cyert and March 1963), their choice and understanding would reflect managers' mental models (Vandenbosch and Higgins 1996) and their own views and interpretation of organisational goals and what determines business success (Hall 2011, Ferreira and Otley 2009). The risk is for managers to invest their attention and measurement efforts on aspects that they think are important (Wong-On-Wing et al. 2007) based on their experience

and focus solely on information that support them. Issues related to confirmation bias have been found to be amplified when the related measures are grouped into pre-defined, subjective categories, drawing attention away from alternative explanations and potentially disruptive factors (Voelpel, Leibold, Eckhoff 2006). Analytical complexity, instead, derives from the difficulty in selecting appropriate measures for representing the value drivers and specifying the links between different aspects of performance. Relations between business value drivers and measures are inherently complex, non-linear, characterised by lag effects (Johanson et al. 2006, Kelly 2010) and, particularly in contemporary environments, dynamic (Melnyk et al. 2014). Measures selection and the definition of cause-effect relationships have always represented a critical aspect of the definition of PMSs (Franco-Santos, Bourne and Lucianetti 2012, Economist Intellingence Unit 2012, Hall 2008, Kaplan and Norton 1996). Among other aspects, the latter is considered a distinctive feature of comprehensive PMSs and a dimension against which such comprehensiveness is typically assessed (Lucianetti, Battista and Koufterors 2019, Gimbert, Bisbe and Mendoza 2010, Chenhall 2005). However, and given the considerations above, many studies observed that when multiple measures of performance are present, cause-effect relations are rarely defined (Silvi et al. 2015, Nørreklit 2000). When they are, formal and objective models are usually not specified and they are subjectively derived (López-Ospina et al. 2017, Malina, Nørreklit and Selto 2007, Ittner, Larker, and Meyer 2003, Ahn 2001).

At both methodological and analytical level, the limited use of quantitative approaches to support the identification and measurement of business value drivers and related measures have been associated with a number of negative effects impacting the effectiveness of PMSs. For instance, Papalexandris, Ioannou, Prastacos (2004) found that subjectivity in the definition of linkages between objectives and measures of performance caused continuous questioning of the PM model. Similarly, Burney, Henle and Widener (2009) provide evidence that the degree of PMS technical validity and the extent to which it reflects a causal model are positively associated with employee performance. Campbell et al. (2015) found that not statistically testing business value drivers limit the ability of the PMSs to uncover strategy shortcomings and the reasons for strategic problems. However, the lack of formal causal models and quantification of interdependences may also produce positive effects (Kolehmainen 2010) and its relevance may depend on the broader context and purpose of the PMS (Tayler 2010, Huelsbeck, Merchant, and Sandino 2011). In turbulent environments, formal models may introduce rigidity and a significant amount of resources may be required to constantly updating the system (Henri 2010, Papalexandris, Ioannou, Prastacos 2004). Hypothesised, rather than tested, relations could create a dialogue of control and ensure that the system reflect changes in the business value drivers and organisation's objectives (Malina, Nørreklit and Selto 2007). For instance, experimental studies found that specifying quantitative relations over qualitative ones do not improve overall performance (Kelly 2010). At the same time, fully tested relations may be too complex to set given the changing nature of organisations and the business environment (Johanson et al. 2006).

Building on the previous discussion, it is possible to conclude that business value drivers and the related measurement should present characteristics of flexibility and adaptability, especially in today's business context (Nudurupati, Garengo and Bititci 2021, Bourne et al. 2018). At the same time, the underlying complexity deriving by the multiplicity of objectives pursued by organisations and interdependencies of different dimensions of performance are only partially captured via conventional qualitative approaches (López-Ospina et al. 2017, Johanson et al. 2006).

2.2. Machine Learning for Performance Management

In a context of growing availability and capability to manage and analyse data (IDC 2019, Sheng, Amankwah-Amoah, and Wang 2019), an increasing number of studies have pointed out the potential of adopting mathematical, statistical, econometric and computer-based approaches to support the identification of business value drivers and related measurement (Raffoni et al. 2018, Appelbaum et al. 2017, Silvestro 2016). Such approaches can be considered a specific application of Business Analytics (BA), that represent the use of advanced quantitative models to analyse business data in order to support operational and strategic decisions (Davenport and Harris 2007). A key feature of BA is "being concerned with evidence-based problem recognition and solving that happen within the context of business situations" (Holsapple, Lee-Post, and Pakath 2014,134). With specific reference to business value drivers, extant contributions describe an array of potential benefits associated with BA applications: improved ability to identify the drivers of organisational performance (Appelbaum et al. 2017); clearer understanding of the impact of

specific variables and related interdependences (Raffoni et al. 2018, Silvestro 2016,); augmented capability to pick weak signals and anticipate change (Laguir et al. 2022, Lassila, Moilanen and Järvinen 2019). When fully integrated into the PMS, BA can be used to communicate strategy (Warren, Moffitt, and Byrnes 2015, Schläfke, Silvi, and Möller 2012) prescribe specific actions and behaviours (Schneider et al. 2015), optimize resource allocation (Raffoni et al. 2018). Laguir et al. (2022) found that PMSs have a mediating role in realising the benefits of developing analytical capabilities. The insights produced by advanced analytical techniques are used to sense emergent priorities, bolster their measurement systems and promote opportunity seeking-behaviours by challenging existing assumptions underlying the organisation's strategy.

Research in this field is relatively recent and whilst growing at a fast pace, is still theoretically and practically underdeveloped (Möller, Schäffer, and Verbeeten 2020, Appelbaum et al. 2017, Schneider et al. 2015). In particular, the application of computer-based algorithms like Machine Learning (ML) is relatively unexplored (Nielsen 2022), despite their increased pervasiveness in organisational life and implications for managerial decision making (Moll and Yigitbasioglu 2019). ML is an application of artificial intelligence (AI) that provides systems with the ability to automatically learn and improve from experience without being explicitly programmed. The process of learning begins with observations or data, to search for patterns in past data and support more effective future decisions (Mitchell 1997). ML and statistics are similar in many aspects but while statistical analysis is grounded in probability theory and distributions, ML is a set of mathematical functions, iteratively optimised, that are combined to best predict an outcome (Witten and Frank 2002). The algorithm can be "supervised" when applied to already-labelled data or "unsupervised" when the information used to train is neither classified nor labelled (Alloghani et al. 2020).

ML refers to a wide family of algorithms and approaches: from logistic regression (Kleinbaum et al. 2002) to decision trees (Kotsiantis 2013), from support vector machines (Zhang 2012, Steinwart and Christmann 2008,) to neural networks (Goodfellow, Bengio, and Courville, 2016, Haykin 1994).

These approaches are mainly known for their ability to predict a phenomenon once its past behaviour has been analysed. Accordingly, they are frequently ranked based on their ability to provide reliable predictions. On the other hand, the algorithms that are most effective in providing predictions are often "black boxes": they do not provide a representation of the phenomenon analysed, so the user gets a reliable prediction but without an explanation of the underlying phenomenon. This trade-off between predictivity and interpretability is widely discussed in ML literature (Shmueli 2010, Sharayu Rane 2018).

In recent years ML applied to business has developed quickly, mainly because of two factors: the increase in the data available from internal and external databases and the upsurge in calculation power that allows elaborating those data. A recent study developed by Algorithmia on 750 American companies (Algorithmia 2020) shows that 45% of the organisations apply some kind of ML approach and that many others are planning to invest in this field in the next 12 months. Growth is already very rapid, and the market for ML applications is forecasted to skyrocket in the coming years, with a CAGR of almost 40% in the period 2018-2024 (Reportlinker 2019). Several companies applying ML in their business processes have witnessed an upsurge in revenues and financial performance (Grover et al. 2018).

The fields of business management where ML is applied are very diverse, but two main sectors emerge (Helo and Hao 2021, Algorithmia 2020, Dzyabura and Yoganarasimhan 2018, Wuest et al. 2016;): marketing and operations. Marketing studies analyse ML as one of the potential drivers of sales and marketing (Syam and Sharma 2018) and several ML approaches have been developed to model customer behaviour (Martínez-López and Casillas 2009), to predict customer churn (Gordini and Veglio 2017) and to analyse customer preferences (Huang and Luo 2016). In Operations, ML has a well-known potential to support scheduling (Aytug et al. 1994) and predictive maintenance, but more interesting applications have been identified more recently, mainly thanks to the huge amount of data generated by the development of Industry 4.0 projects (Lolli et al. 2019, Wuest et al. 2016). As for the broad field of accounting and performance management, ML approaches are fairly unexplored (Nielsen 2022, Moll and Yigitbasioglu 2019, Sutton, Holt and Arnold 2016) and have been mainly focused on fraud detection (Bao et al. 2020) and predicting financial distress and bankruptcy (Jiang and Jones 2018).

Building on previous considerations, it is possible to argue that the definition of business value drivers and the identification of the related measures could benefit from the use of ML approaches. On one side PMSs aim to focus

organisational efforts on the drivers of business success. In so doing they need to be able to identify and monitor those drivers as well as predicting and anticipating changes, which are aspects only partially captured by conventional qualitative and quantitative approaches. On the other side, ML could enhance the diagnostic and anticipatory ability of PMSs. First, ML algorithms perform very well when the number of variables is high and the relationships between the independent and the dependent variable(s) is not linear. This is what often happens in the business world, where a huge number of interrelated and complex variables affect the performance of the company in a non-linear way (Granger and Terasvirta1993). Secondly, and considering that these relations are not static and change over time, ML's ability to continuously learn from new data would be able to uncover new patterns, thus providing effective predictions (Ryll and Seydens 2019; Caruana and Niculescu-Mizil 2006) and anticipate changes. Finally, ML algorithms perform particularly well compared to statistics when the amount of data is high (Sokolova and Lapalme 2009) but incomplete at the same time. The above technical features are especially relevant in the current business environment where huge sources of data are available but often not effectively exploited because of low quality levels (missing/unbalanced datasets) and a limited ability to analyse them.

Following the above reflections, the main research question of our work is to understand whether the use of ML could mitigate issues related to the complexity in identifying and measuring business value drivers, and in particular, aspects related to methodological complexity (i.e. the detection of business value drivers) and analytical complexity (i.e. the actual measurement of the drivers and identification of cause-effect relationships). In so doing our work also aims to address a second and related research question, which is what issues may arise in the adoption of a ML approach for performance management purposes. Research across different discipline areas shows that the development and application of BA tools alone do not allow organisations to achieve their full benefits (Ransbotham, Kiron, and Kirk 2016, Sharma, Mithas, and Kankanhalli 2014, Popovič et al. 2012) and there is evidence of limited implementations or even failure of BA initiatives (Fleming et al. 2018, Capgemini 2015). The PMS field is no exception and both conceptual works and the few empirical applications warn about a number of potential problems. In particular, lack of understanding of the underlying business model (Schläfke, Silvi, and Möller 2012), low data quality (Raffoni et al. 2018), limited analytical capabilities (Nudurupati, Tebboune, and Hardman 2016) and difficulty in processing the outcomes of BA models (Moll and Yigitbasioglu 2019, Bhimani and Willcocks 2014) are considered barriers to BA approaches in PMSs. In this context, the sophistication of ML algorithms constitutes an additional layer of complication compared to more traditional BA and potential benefits associated with reduced methodological and analytical complexity may be overcome by the lack of interpretability of the results (Meske et al. 2022). For instance, recent works demonstrated that when the results of advanced decision support systems contradict end-users' expectations, recommendations tend to be disregarded (Jensen et al 2010, Giboney et al. 2015).

3. Research method

The study was carried out through an action research (AR) approach. The peculiar feature of AR is that the researcher becomes part of the action related to the phenomenon being studied, with the aim "to bring together action and reflection, theory and practice, in participation with others, in the pursuit of practical solutions to issues" (Reason and Bradbury 2001, p. 1). With AR the researcher directly affects the reality (Jönsson and Lukka 2007), interacting with the members of the organisation (Reason 1999), thus making it possible to achieve a better understanding of the phenomenon (Parker 2012). AR generates a continuous interaction between theory and practice. The researcher approaches reality through the lens of theory and at the same time he/she reviews the literature based on the processes he/she is involved in and the practical outcomes obtained (Argyris, Putnam, and McLain Smith 1985; Jönsson and Lukka 2007; Van Aken 2004). Furthermore, being part of the process increases access to direct and indirect data sources and strengthens the relationships with the actors of the phenomenon being studied, increasing the possibility of interpreting the reality and developing innovative approaches (Van de Ven 2007; Van de Ven and Johnson 2006). Being actors of the process, the researchers obtain an easy access to the data while they are generated (e.g., Suomala Lyly-Yrjänäinen, and Lukka 2014; Lukka and Suomala, 2014), with the possibility to closely observe the process and the reactions generated by the development of new tools, approaches and frameworks (Korhonen et al. 2020).

In this, AR is particularly suited to dynamic situations where the process of change is the focus of the research (Coughlan and Coghlan 2002). AR approaches these situations with two goals: to support the solution of a practical problem and to provide an effective contribution to the development of the theory (Gummenson 2000). This was the reason why we selected this approach for the study. On the one hand, the case company was facing an emerging challenge for which change had not been initiated yet (i.e., the need to refocus the PMS to cope with a serious financial crisis). On the other one, there was an area of research that was practically and theoretically unexplored (i.e., whether ML could support the identification and measurement of the value drivers of the PMS). The intervention entailed technical problem solving and idea generation between the researchers and the case company staff to develop and implement a PMS more focused on key value drivers. Through their interventionist' role, the researchers were able to provide expertise on technical work and, at the same time, witness the process of change. Their active contribution allowed an in-depth investigation of the two research questions of the study: 1) assessing whether the adoption of a ML approach could contain the methodological and analytical complexity associated with the identification of the business value drivers and its related measurement b) examining the issues generated by employing ML for performance management purposes. Direct intervention proved extremely valuable especially for the second objective as the researchers had the opportunity to experience first-hand, without relying on interviews and other indirect data sources, the technical, cognitive and social problems associated with the development and adoption of a ML approach in a performance management context.

3.1. Research site and motivation

Financial Corporation (hereinafter FC) is a small company working in the field of debt collection. Founded in 1982, its core business is the collection of debts from the customers of banks, retail companies and public administrations. In the period 2016-2018, BC's financial performance deteriorated. Revenues generated through commissions ranging from 10 to 15% on collected debts dropped from € 16.2 to 14.6 million. For FC, the most relevant cost (80% of the total) is labour, which is mainly fixed. Therefore, the turnover decline significantly impacted financial performance: from a profit of €800,000 in 2016 to a loss of €300,000 in 2018. Whilst the number and amount of debts assigned to FC by its customers remained the same across the three years, the percentage of collected debts, i.e. the recovery rate (RR, the main performance measure monitored by the company), dropped significantly. The fall in RR was mainly due to the lower "quality" of the debts allocated by the customers, which followed the deteriorated financial situation of several companies and families in Italy.

In reaction to this decline in profitability, the company decided to act on its operating efficiency by revising the processes and investing in a more effective software. Three months after the implementation of the new procedures (March 2019), the RR and the financial performance were re-assessed. However, no significant improvement was observed, calling into question the relevance of the RR as a critical business value driver. Whilst a commonly considered and measured determinant of performance in the sector, the RR did not provide any visibility on the antecedents of the recovery capability (i.e., what factors increase/decrease the debt collection ability), how these factors could impact the value (effectiveness) and percentage (efficiency) of debt collection. At the same time, the ability of a single indicator to measure the value generated through debt collection was questionable. The GM was willing to revise the PMS but the redesign presented at least two orders of complexity: (1) methodological, factors driving value generation were not clear so, what variables should the PMS focus on? (2) analytical, how should variables be measured and how could the relationships between value drivers be assessed?

At that time the GM was attending a performance management course held by two of the researchers, to whom he explained the situation of the company to ask for support about changing the PMS. This was the start of the research collaboration which was soon followed by an initial focus group in FC headquarter to further discuss the problem. The GM also involved the controller of the company, a certified management accountant with extensive experience at FC and a background as an auditor in one of the Big Four. During the meeting they analysed the decline of the RR and reviewed some preliminary data provided by the controller. It emerged that given the very low average RR (only 7% of the debts were recovered), the understanding of which debts deserved attention and which processes and operators were the most effective to collect the money back should be given priority. Focusing on process efficiency

would have not let to any performance improvements if applied to credits with a very low likelihood of being collected. Thus, the following two key questions arose from the debate:

a) Which external and internal drivers affect the probability of collecting the debts? External drivers were interpreted as debtors' characteristics (e.g., gender, age, geographical area, occupation, etc.), while internal drivers referred to the staff members in charge of the collection process as well as the dynamics of the process itself.
b) Which process configuration maximises the likelihood of collecting the money for different kinds of debts/debtors?

When the research collaboration started, the monitoring of the performance was largely based on crosstabs development showing the RR for different kinds of debtors and for each operator. RR targets were then set for each operator and rewards were linked to the results achieved. However, the extant approach only allowed for a partial exploitation of the data available and for instance, did not consider which factors would impact the likelihood of debt collection by a specific operator (e.g., debtors characteristics, size of the debt). Given the above, a third researcher with a background as an expert data scientist in financial organisations was involved making the final composition of the "Research team" (RT) as follows: the three authors, the GM and the controller.

The factors that led the researchers to select FC as the research site were the limited complexity of the business model, which simplified the analysis, the full access to all the data and information available in FC and more generally the maximum support provided by the GM and the controller.

3.2. Research process

The research process entailed a number of phases which, for descriptive purposes, can be represented as five main steps (Figure 1): a first understanding of the phenomenon under investigation (Phase 1), the collection of the available data (Phase 2), their analysis and correction (Phase 3) and the design of the ML model (Phase 4). Finally, once the model is developed and the results evaluated, they are translated into the organisation's performance management model. (Phase 5).

However, it should be remarked that AR is not a one-off process where one step neatly follows the previous one. Instead, AR is best described as an iterative process made of cycles of data analysis and action that benefit from previous cycles and generate insights on the researched topic (see Figure 1). Besides, each step tends to be characterised by a process of data-gathering, data-feedback and data-analysis (Coughlan and Coghlan 2002).

For instance, during the data collection phase, data were *gathered* through different approaches like interviews and original data. Then *feedback* about data reliability was obtained from the controller and the general manager and a first *analysis* allowed to understand the meaning of each variable and to classify internal and external drivers.

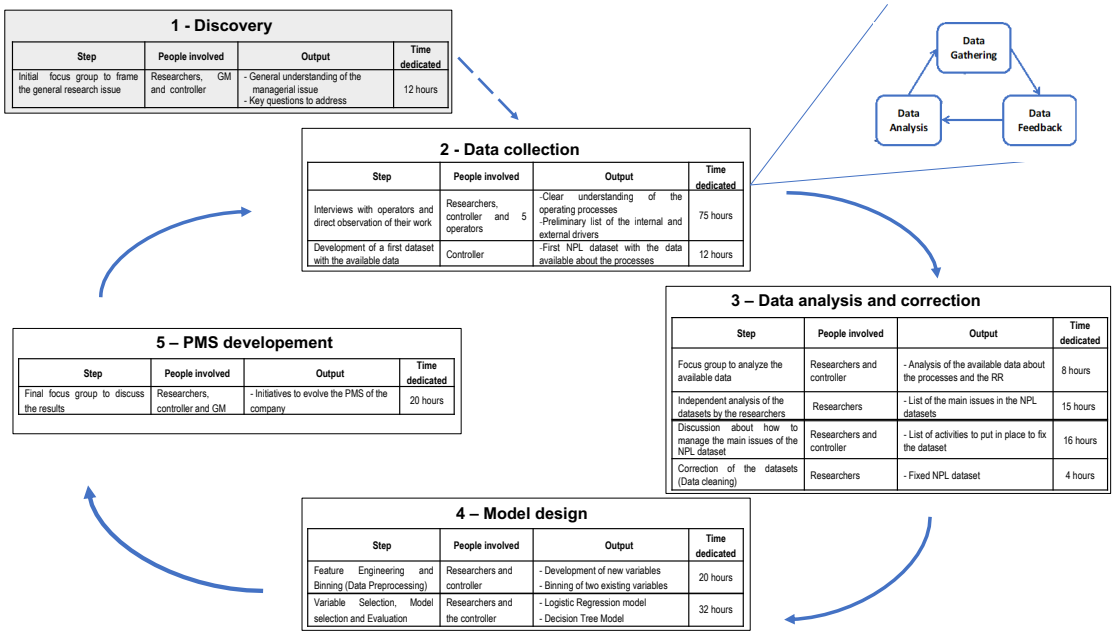

Figure 1: The main steps of the study

While the first phase has already been presented, the following paragraph describes the findings related to each of the remaining phases.

4. Findings

4.1 Data Collection

The research was conducted between April and October 2019. Data were collected through several methods in the context of AR. During the project, the research team frequently visited FC and qualitative data were gathered onsite through interviews, observation, focus groups, company documents and participation in management meetings (Denzin and Lincoln 2008; Yin 2008). As typical of AR, which tends to be highly abductive (Dubois and Gadde 2002), the data collection and analyses were rarely conducted as separate activities. Intervention implies continuous reflection, observation and collaboration to make sense of the practical problem and develop a contribution of theoretical significance. Semi-structured interviews were mainly used in the initial stage of the project. These were aimed at capturing interviewees' views of FC's strategy, their perceptions of critical performance variables and risks, their understanding of the company's PMS, as well issues or gaps in the existing approach. To begin with, interviews with five operators and the direct observations of their work allowed the researchers to understand how the operating processes were carried out and to develop a preliminary list of the internal and external drivers affecting the RR. Meanwhile, the controller collected all the available information about the processes developed and the outcome of each case managed in the last three years. Finally, a focus group involving the research team was held in order to discuss the available data and the meaning of every available variable. The original dataset was labelled as the "Non-Performing Loans" (NPL) dataset.

The NPL dataset consists of 20,218 records and 17 variables regarding both the debtor (gender, age, geographical region, solvent or not, etc.) and the specific debt (amount, aging, presence of a previous email to the debtor, etc.). These were subdivided into 10 numerical and 8 categorical variables, although some of the numerical ones were categorical in disguise. Details of all variables are presented in Table A1 in the Appendix.

4.2 Data analysis and correction

This phase of the process was technically developed through the so-called Exploratory Data Analysis (EDA). EDA in ML consists in analysing datasets to summarise their main characteristics, often with visual methods. A further purpose of EDA is to discover important associations between variables in the data set without necessarily using a statistical or ML model. EDA usually consists of two steps: data cleaning and feature engineering. In applying ML to the business context, these steps appeared essential to the RT: on one side, the RE agreed that data exploitation and quality of the expected insights was dependent on the quality of the inputs; on the others side, the understanding of the relevance of the value drivers required variables to be well organised and represented.

As for data cleaning, the initial assessment showed several issues despite the controller emphasising their completeness and reliability. Many typing errors were pointed out, together with several missing data or empty strings, and inconsistencies between different variables. For instance, for 124 records the variable "Amounts recovered" was zero, even if the variable "Amount recovered" was flagged. Table A2 in the Appendix summarises the main issues and the data-cleaning activities performed by the researchers after a joint evaluation with the controller. After cleaning the datasets from unreliable or missing data, 20,144 records remained.

Next, the Research Team was able to perform feature engineering (FE), which deals with the process of using domain knowledge to extract features from raw data. The main objective of FE is to improve the performance of ML algorithms in term of predictivity, computational performance and interpretability. In this phase, which typically entails considerations over how to extract more value from ML and revising existing variables or computing new ones, the researchers and the controller had extensive discussions about the meaningfulness of the RR as the main or even only performance measure used at FC. First, it was suggested to consider also the duration of the debt collection process (as a proxy of the cost) and not only the final output (the amount collected, which proxies the revenues). This would have allowed to weight positive outcomes against the efforts required in the debt collection process.

Hence, a new variable (labelled "Return on Effort" or ROE.L) was developed as follows:

$$\text{ROE. L} = \log\left(\frac{\text{Amount recovered}}{\text{procedure total duration}}\right)$$

ROE.L proxies the effectiveness of the debt recovery process by comparing the amount recovered to the effort required by the process[2].

Secondly, the RT agreed to revise some of the existing variables as they presented too many different potential values. This would have unnecessarily increased the complexity of the analysis and reduced the possibility to read and understand the final results. Then a binning procedure was applied to two variables to recode them into smaller number of intervals (bins). The variable "Operator" was binned into a new variable "Operator.B". The metric used for binning similar performing operators was the previously introduced ROE.L, whilst the algorithm applied was a decision tree. The results showed that different groups of operators presented very different performance in terms of ROE.L (Table 1).

[2] A second less relevant intervention was that from the variable "Amount recovery plan" we derived a new binary variable "Recovery plan agreed" equal to 1 when the amount was positive and equal to 0 when the amount was 0.

Operator Bin	Number of operators	Average ROE.L.
Bin 1	126	217.3
Bin 2	42	372.5
Bin 3	28	315.4
Bin 4	14	740.3
Bin 5	42	844,7
Bin 6	42	1,388.7
Bin 7	28	499.8
Total	322	625.6

Table 1: The bins the operators were clustered into

Similarly, "Year of birth" was binned into 10-year periods (decade) and 5-year periods. The ten-year variable proved to be more robust to rare levels (i.e., few very old or very young people) and was thus selected.

4.3 Model design

Once the final list of variables was obtained, the research team analysed the data in order to design the ML model. The whole process can be seen, for explanation purposes, as a sequence of three main steps: feature selection (the process of selecting a subset of most significant variables in the dataset), model selection and development (where the representation of the phenomenon take place) and the final evaluation and optimisation of the model.

4.3.1 Feature selection

In ML feature selection is aimed at reducing the number of variables to increase the interpretability of the results, reducing the time needed to train the algorithm and reducing overfitting (Cai et al. 2018). The interventionist researchers believed this to be a critical step towards the identification of the value drivers of the organisation. The identification of the variables with the highest impact on the performance of the organisation would have served as a preliminary assessment of its determinants.

First, the RT performed a correlation matrix to assess which variables were carrying the same kind of information. As expected, the variables "Solvent" and "Solvent with property" showed a very high correlation (CramerV test = 0.713). Consequently, only the former was kept. Furthermore, as presented in A1, "E-mail received" represented a simpler version of "E-mail status", so only the first was taken into consideration.

This phase of the analysis also evidenced a strong correlation between the dependent variable ("Procedure outcome") and the variable "Procedure total duration", thus suggesting the latter as a critical value driver. Such result was discussed in a meeting with the GM, from which it emerged that the actual causation relationship was inverted. The outcome of a call is not positive because the operator spends a lot of time with the debtor. On the contrary, operators spend lots of time with the debtor because the outcome of the call is positive, and they foresee a possibility to collect the money. This led the RT to exclude the variable "Total duration of the procedure" despite its statistical significance since it was not a predictor of the outcome. This instance showed how the algorithm may produce misleading results if not validated and interpreted in the context of the organisation.

Next, and in order to evaluate the significance of the remaining variables, the RT performed a Random Forest and analysed the mean decrease of the Gini Index as a measure of the classification error[3]. Figure 2 reports the mean Gini index decrease for the most relevant variables.

[3] The higher the decrease of the index when the variable is included, the higher the impact of the independent variable on the dependent one.

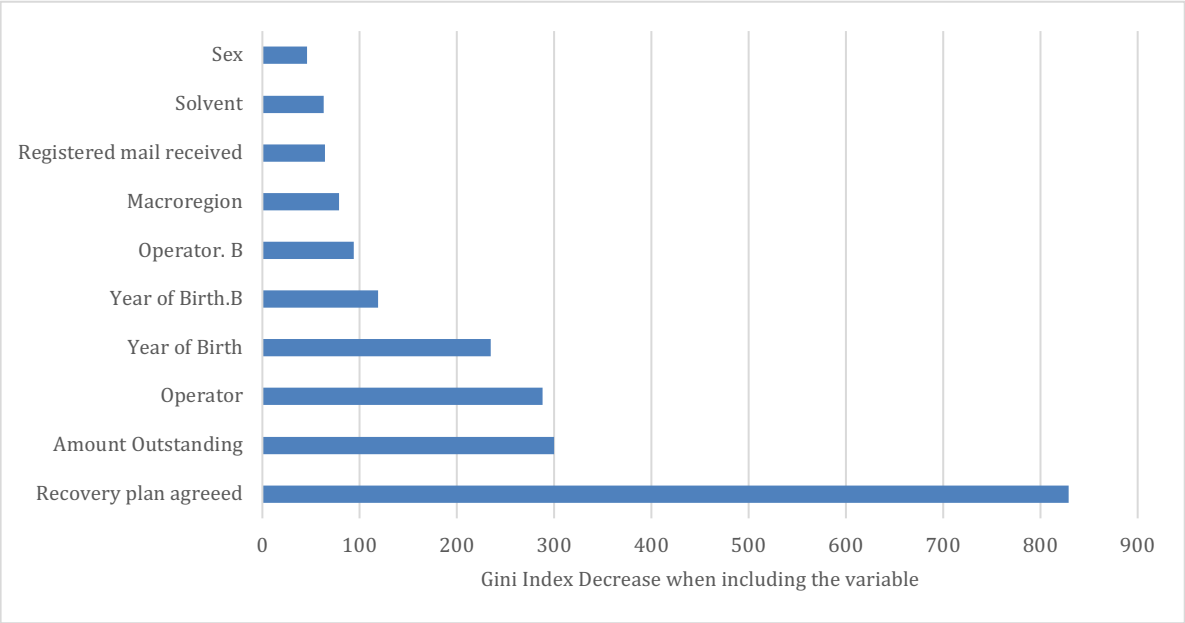

Figure 2: The mean Gini Index decrease for the most relevant variables

The results reported in Figure 2 were discussed during a meeting of the RT. What emerged from the analysis was that "Recovery plan agreed" was by far the variable with the highest impact on the dependent variable. Whilst it is intuitive that if the debtor agrees to a recovery plan the likelihood of receiving the payment increases sharply, the RT had to consider that this information was available only after at least one phone call, not before starting the recovery process. The variable was still included in the process but knowing the time-lag needed to obtain this information. Similarly, to the "Total duration of the procedure", the involvement of the GM and Controller proved critical in making business sense of the results of the feature selection. This phase of the analysis also evidenced that the binned variables ("Year of birth. B." and "Operator. B") whilst showing lower significance compared to the not-binned ones ("Year of birth" and "Operator"), were still quite relevant. The RT discussed whether to keep them or the not-binned variables in the analysis. They eventually agreed that the inclusion of the binned variables would have simplified the interpretation and communication of the results[4].

4.3.2 Model selection and development

The selection of the algorithm(s) to analyse the data was the next step of the process. In discussion over the different options, the GM made very clear that understandability of the model and its communication to managers and operators was his priority. In his view, that had to be preferred over the predictivity of the algorithm.

Therefore, the RT selected two ML algorithms able to provide high interpretability of the results: logistic regression and decision trees. For both the approaches a cross-validation (CV) was performed to ensure the model was able to generalise well out-of-sample and did not suffer from overfitting (Browne 2000)[5].

---

[4] At the end of the selection process the following independent variables were considered: "Procedure Outcome" as the dependent variable; "Recovery Plan Agreed", "Amount outstanding", "Operator. B", "Year of birth. B", "Solvent", "E-mail received", "Macroregion", "Sex" as independent variables.

[5] We used a K-fold validation with K=5. The general K-Fold algorithm shuffles the dataset randomly and splits it into k groups. Then, for each unique group it takes the group as a test dataset and takes the remaining groups as a training dataset. Then, it fits a model to the training set and evaluates it on the test set, then it retains the evaluation score (the performance measures specifically used will be discussed later in this section). Finally, it summarises the skill of the model using the mean of the performance scores achieved by each fold.

Furthermore, during this phase, the unbalanced nature of our dataset was taken into consideration. This situation was derived by the low frequency of positive outcomes compared to the negative ones. In other words, u the dependent variable ("Procedure Outcome") was equal to 1 in 7% of the cases (minority class), while it was equal to 0 in the remaining 93% of the cases (majority class). This represented an issue with the chosen algorithms as standard classifier ones like Decision Tree and Logistic Regression have a bias towards minority classes. They tend to predict the majority class data only, while the characteristics of the minority class are treated as noise and frequently ignored. Consequently, there was a high probability of misclassification. This led to the rebalancing the overall dataset by oversampling the minority class with probability 0.5 by bootstrapping a reduced dataset of records. This is a specific feature of machine learning compared with "traditional" statistics, where this kind of invasive approach would not be admitted.

4.3.3 Model evaluation

Model evaluation seeks to estimate how well the selected model will perform on unseen data. The model was applied on the test dataset to compare the predictions with the actual data and fill the so called "Confusion matrix", which, in FC case highlighted:

- How many of the debts for which the algorithm predicted a positive outcome were actually paid (True Positives or TP) or not (False Positives or FP).
- How many of the debts for which the algorithm predicted a negative outcome were actually paid (False Negatives or FN) or not (True Negatives or TN).

Figure 3 reports the confusion matrix for the decision tree and the logistic regression after rebalancing the dataset.

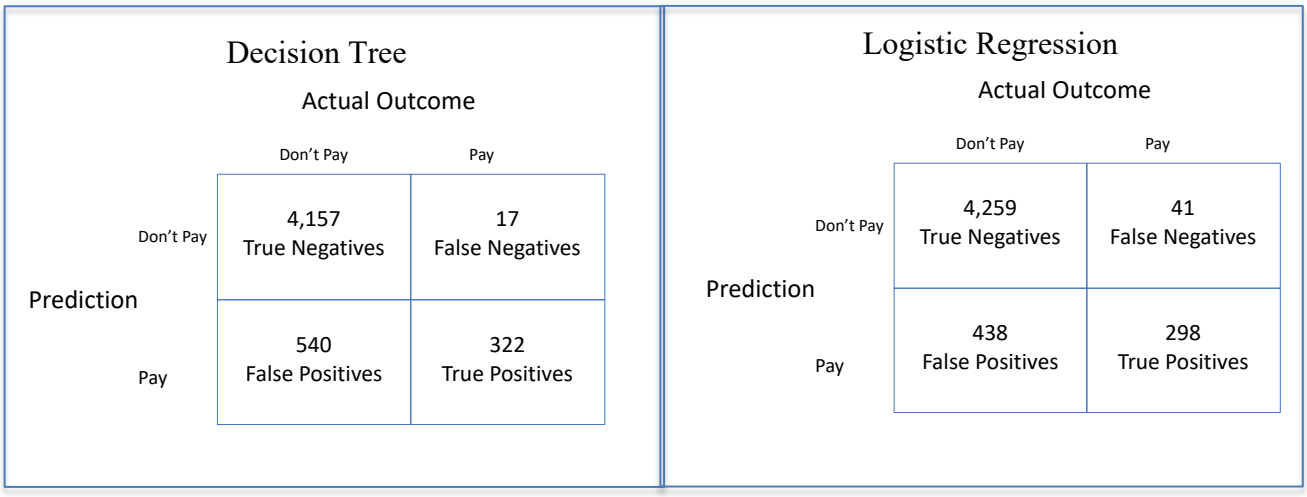

Figure 3: The confusion matrix of the two algorithms applied

The most basic evaluation metrics that can be obtained from a confusion matrix is the already-mentioned accuracy, i.e. the ratio between the values correctly predicted and the total cases:

$$\text{Accuracy of a model} = (\text{TP} + \text{TN})/(\text{TP} + \text{FN} + \text{FP} + \text{TN})$$

However, when this was discussed within the RT, it emerged that accuracy would have not been an appropriate approach to rank different algorithms at FC. This because of the unbalanced nature of the dataset and the focus of the company on a specific result of the dependent variable. In other words, the effectiveness of the model in supporting the performance management process at FC depended on its capability to drive the behaviours towards positive outcomes ("Procedure Outcome" = 1), not in maximising the total accuracy. As a paradox, an algorithm predicting

that no debtor will pay would get an accuracy equal to 93% (=100%-7% of positive outcomes that would be neglected by the algorithm). However, it would not help the company in identifying the debtors to prioritise or design more effective processes.

The interventionist researchers realized that for FC a more appropriate measure was the Area Under the Curve (AUC)[6]. In Table 2 Accuracy and AUC of the two models are noted, both before and after the rebalancing of the dataset.

Unbalanced dataset			Balanced dataset		
Logistic Regression	Accuracy:	0.930	Logistic regression	Accuracy:	0.905
	AUC:	0.684		AUC:	0.893
Decision Tree	Accuracy:	0.934	Decision Tree	Accuracy:	0.889
	AUC:	0.711		AUC:	0.917

Table 2: The performance of the models developed

It was evident that after rebalancing the sample the accuracy and the AUC of the two algorithms was very close to 1 so both logistic regression and decision could be considered good models. While rebalancing the dataset led to a small reduction of the accuracy, it generated an upsurge of the AUC, i.e., the capability of the model to truly distinguish between positive and negative outcomes. Since AUC was the most important performance measure to focus on, then the Decision Tree performed slightly better than the Logistic Regression (AUC=0.917 v 0.893).

When the results of the evaluation were presented to the GM and the controller, they complained about the complexity of the concept of AUC to measure the performance of the models. They explained they needed easier measures, because "we are managers, not mathematicians" (GM quotation). Their objective was to understand the reliability of the models when predicting a positive outcome of the procedure. Basically, the percentage of True Positives out of the total number of predicted positives (True Positives + False Positives). This specific measure in ML language is labelled as Precision. In order to use a language better fitting the needs of the managers, AUC values were therefore replaced by Precision values. The results were 0.879 for the Logistic Regression and 0.95 for the Decision Tree on the Rebalanced Dataset, so again the Decision Tree outperformed the Logistic Regression.

4.3.4 The final models

The final models provided by the two algorithms were discussed in a focus group with the GM and the controller of the company to understand the potential support they could provide to the performance management initiatives.

Figure 4 reports the results of the logistic regression. If the debtor was not classified as "Insolvent" and had already received the e-mail, the probability of a positive outcome increased. On the other hand, debtor sex and age did not play any statistically significant role, while the geographical area negatively affected the result only if the debtor lived in specific areas. Furthermore, and importantly, the size of the debt negatively affected the likelihood of collecting the money. Focusing on internal drivers, the relevant role of the operators of Bin 7, followed by Bins 4 and 3 in increasing the probability of a positive outcome was uncovered.

[6] AUC measures the extent to which a model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting each class. An excellent model has an AUC close to 1, which means it has a good measure of separability. When AUC is equal to 0.5, it means that the model is not able to achieve any class separability. If AUC=0 the model provides completely wrong predictions, predicting negative outcomes as positive and vice versa.

Variables	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.612175	0.679726	-5.314	1.07e-07	***
E-mail received	0.598941	0.125717	4.764	1.90e-06	***
Solvent.FNP	1.344008	0.137113	9.802	< 2e-16	***
Solvent.FSI	1.150077	0.313971	3.663	0.000249	***
Operator B.2	0.361281	0.281671	1.283	0.199619	
Operator B.3	0.808555	0.313516	2.579	0.009909	**
Operator B.4`	0.975813	0.267596	3.647	0.000266	***
Operator B.5`	0.265868	0.239698	1.109	0.267353	
Operator B.6`	0.431423	0.229781	1.878	0.060444	.
Operator B.7`	3.527787	0.325287	10.845	< 2e-16	***
Macroregion “Northern”	0.047179	0.149075	0.316	0.751639	
Macroregion “Southern”	-0.352583	0.152880	-2.306	0.021095	*
Amount Outstanding	-0.224603	0.077397	-2.902	0.003708	**
Sex	0.062196	0.123438	0.504	0.614359	
Year of Birth.B	-0.006631	0.004246	-1.562	0.118374	
Recovery Plan Agreed	5.178253	0.177808	29.123	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4773.5 on 3999 degrees of freedom
Residual deviance: 1814.5 on 3984 degrees of freedom
AIC: 1846.5

Figure 4: The outcome of the logistic regression

Once the results of the Logistic Regression were analysed, the outcome of the Decision Tree was discussed (Figure 5).

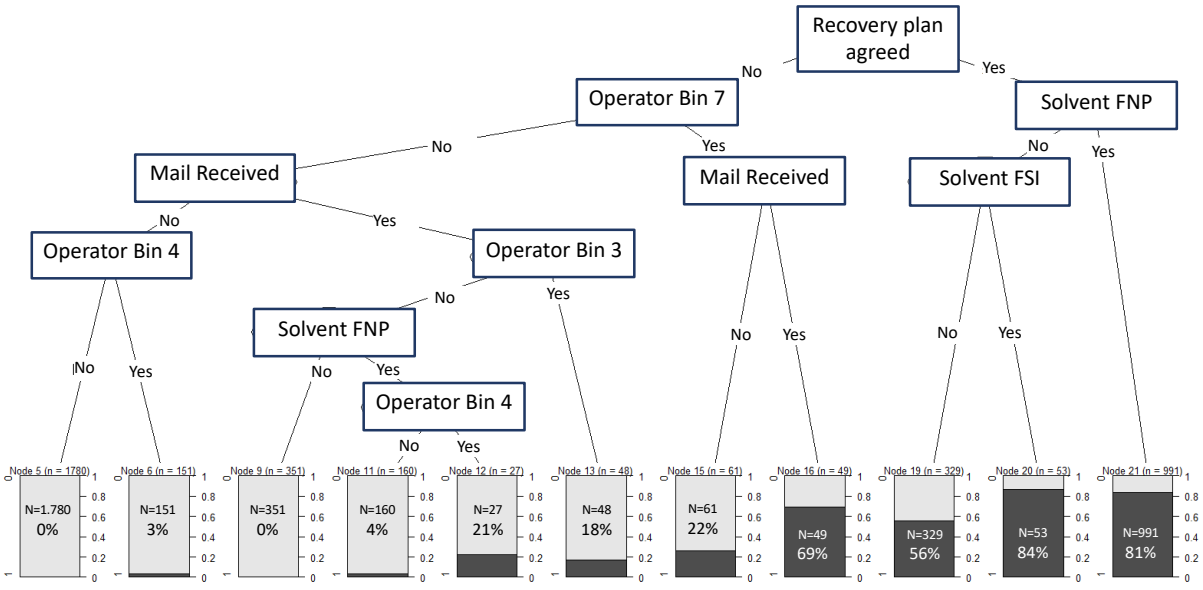

Figure 5: The decision tree

Looking at the tree, the role of the variable "Recovery plan agreed" is confirmed. Once the debtor agreed to the plan, the likelihood to receive the payment became much higher, especially if the debtor was solvent or their liquidity unknown (Nodes 20 and 21, with over 80% of payments in the rebalanced sample).

However, as previously stated, when the company started the procedure, the operator did not know if the debtor would agree to the plan, so for FC the left-hand side of the tree was the most significant one. Here, a relevant role is played by the operators of Bin 7 and by the e-mail sent to the debtor before the phone call. When the two factors were jointly present, the percentage of collected debts rose (Node 16, 69%), but the frequency, even in the rebalanced sample, was very low (only 49 occurrences). On the contrary, when neither of the two factors were present (Nodes 5 and 6, with over 1,900 occurrences in the rebalanced sample) the percentage of collected debt was close to 0. Furthermore, the tree highlighted the positive performance of the operators of Bin 4 in managing both "easy" cases (E-mail received and Solvent FNP, Node 12, RR=21%) and much more complex situations (E-mail not received in Node 6). In this latter case the RR was only 3%, but still much higher than the other Bins in the same conditions (0% in 1,780 cases of Node 5).

4.4 Performance Management System development

During the discussion, managers' attention was soon captured by decision tree which they were able to make sense of without significant explanations by the RT. They appreciated its simplicity and clarity and they immediately started discussing each node of the tree to understand the insights provided by the analysis and how they could be employed to manage the performance of the organisation. Conversely, the outcome of the logistic regression was basically neglected.

The results of the analysis were taken further. First, it was agreed to focus on the relevance of the recovery plan and of the e-mail sent before the phone call. Accordingly, the company decided to better frame the procedure for sending the e-mail to the debtors and to propose a recovery plan. Secondly, they decided to measure and reward the performance of the operators with different weights for "easy" cases (e-mail received, or even recovery plan agreed, solvent debtor) and "complex" cases (e-mail not received, insolvent debtor living in Southern Italy). This result completely changed their perspective on operators' performance. Looking at the ROE.L. displayed in Table 1, it seemed that Bins 5 and 6 were the most effective because they showed the highest values. On the contrary, looking at the outcomes of the ML models, it became evident that Bins 7, 4 and 3 were the most effective and played a relevant role in improving the company's performance. Bins 5 and 6 had a higher ROE.L simply because they managed a higher percentage of "easy" cases. Furthermore, the joint analysis of Table 1 and the two ML models, provided "shocking" results in the eyes of the GM and of the controller. Bin 1, which included almost 40% of the operators, showed a very low effectiveness (the lowest ROE.L among all the Bins) and a negative impact on the performance of the company. Accordingly, the GM decided to redeploy part of the Operators of the ineffective Bin 1 to boost the restructured activity of sending e-mails to the debtors. The remaining part of the Operators of Bin 1 would instead be dedicated to managing the easiest cases, where the recovery plan had already been accepted.

As for the general structure of the PMS system, it was decided to select and assign cases to operators based on the likelihood of collecting the money as calculated by the ML algorithm. The RR for the cluster of cases assigned to each operator predicted by the decision tree became the quarterly target of the operator and the starting point of the PMS. At the end of the quarter, the actual RR of each operator had to be compared with the target and a variance analysis performed to understand the reasons for the positive or negative discrepancies between targets and results. Furthermore, the model is not static, because the data generated by the operations month after month will be used to train the algorithm, and consequently to revise the model and the targets, if needed.

5. Discussion

The aim of this paper was to understand whether ML could moderate methodological and analytical complexity in the search of the main business value drivers. A second objective was the examination of the potential issues arising in applying ML to performance management. To achieve the above, we built on topical literature on the design and

implementation of PMS and the use of business analytics in this context with a specific focus on ML approaches. By adopting an action research design, we examined Financial Corporation's attempt to introduce a data-driven approach to the identification and measurement of the business value drivers by employing ML techniques. Examining this change while it was being developed allowed us to gain deep insights into such processes generating findings that a distant observation could have hardly produced (Lyly-Yrjänäinen et al. 2017). These are next discussed in relation to our two main research questions.

5.1 The role of ML in mitigating methodological and analytical complexity

Our first research question stemmed from the observation that the identification and measurement of business value drivers is characterized by high levels of complexity, of which methodological and analytical ones constitute the core of PMS development (Okwir et al. 2018, Santos, Kelly 2010, Belton and Howick 2002). In this respect, one of the main issues of conventional approaches is that the crucial drivers of performance are not always identified (Kaplan and Norton 1996) and their impact on each other and the overall performance is often overlooked (CIMA 2014, Franco-Santos, Bourne and Lucianetti 2012) and not adequately measured (Silvi et al. 2015, López-Ospina et al. 2017).

In FC, initial attempts to revise the PMS were not effective since what was believed to be the main business value driver, namely efficiency in the debt recovery process, did not impact performance in terms of recovery rate. At the same time, the measurement focus on the RR expressed as the amount of debt collected over the total amount only provided a macro understanding of the FC's effectiveness in the recovery process. For instance, it did not consider the efforts required in the collection process (e.g., the time and therefore cost necessary to recover the debts) as well as providing an understanding of the antecedents of such results. In this context, the application of ML proved first beneficial in terms of moderating methodological complexity, where the descriptive and diagnostic features of the algorithm lessened the difficulty of identifying the key drivers of business performance among multiple potential options (Benson-Rea, Brodie and Sima 2013, Santos, Belton and Howick 2002). Data processing capacity of ML and its ability to learn from past data supported the identification of patterns and relations without the need of a pre-defined theoretical model (Ryll and Seydens 2019; Caruana and Niculescu-Mizil 2006). As such ML algorithms allow to capture the complexity of real business situations where multiple, possibly conflicting factors influence the organisational performance (Granger and Terasvirta 1993) and where the understanding of cause-effect relations is rarely unambiguous (Kelly 2010, Johanson et al. 2006). Compared to more conventional subjective approaches where managerial wisdom plays a major role (Ferreira and Otley 2009, Malina, Nørreklit, and Selto 2007, Ittner, Larker, and Meyer 2003) ML appeared to moderate issues of bounded rationality (Cyert and March 1963) as well as reducing confirmation biases (Voelpel, Leibold, Eckhoff 2006). In other words, ML algorithms supported an identification of the drivers of organisational performance less influenced by managers mental models and interpretations of what they believed to drive business success (Hall 2011, Wong-On-Wing et al. 2007, Vandenbosch and Higgins 1996). At FC this became quite evident when managers were put in front of the results of the logistic regression and decision tree to find that focusing attention on pre-sending emails to debtors was identified as one of the key factors driving the likelihood of money recovery. Therefore, ML algorithms may reduce the risks of focusing the PMS and organisational resources on non-relevant factors (Kaplan and Norton 1996). These results do not imply that human sensemaking does not play a role in the process. The experience of the managers, their knowledge of the organization and the business constitute critical factors in exploiting the full potential of ML for PMS purposes. Every phase of the research process, from the "discovery phase" to the "performance management" one, was characterised by a series of iterative cycles where data gathering and analysis implied a continuous dialogue with the managers about the performance model, the evaluation of the value drivers and the analysis of the ML model developed, the validation of the results. These aspects were as relevant as the analytical capabilities required to clean, organise and analyse the available data.

Secondly, ML moderated issues related to analytical complexity by clarifying and quantifying how different variables impacted the RR. In particular, ML showed both predictive and prescriptive abilities (Schneider et al. 2015). From a predictive perspective, ML identified how the RR depended on the ability to agree on a recovery plan. Whilst this variable was not known in advance and therefore could not be used in the PMS, ML identified other patterns that

increased the likelihood of debt collection and therefore could predict how specific variables contribute to it. For instance, the critical role of Bin 7 operators and sending an email ahead of the phone call were uncovered by the algorithm. At the same time, the algorithm showed a low frequency of such occurrence indicating that a key driver was overlooked and not measured. From a prescriptive perspective, ML proved beneficial to identify actions that could influence the desired outcome. In FT, the results provided by the decision tree triggered the revision of how resources were allocated and of the measurement system more broadly, thus reducing the negative impact of misallocation of individual and organisational efforts (Burney, Henle and Widener 2009, Krishnan, Luft, and Shields 2005). By quantifying and making the effect of specific types of operators and actions (e.g. sending emails ahead of the phone call) visible, it was possible to: (1) assign cases to operators on the basis of the likelihood of RR; (2) re-assign the least effective group of operators to either the management of easy cases or boost the process of email sending; (3) adjust the measurement and reward of operators performance based on the proportion of easy and difficult cases managed.

These results suggest that the adaptability and capability of ML to continuously learn from past data do not necessarily generate rigidity in the PMS, contrarily to previous literature analysing the use of business analytics for performance management (Kelly 2010, Malina, Nørreklit and Selto 2007). This may reduce the costs of updating the system both in terms of time and efforts required but also in terms of risks that lack of flexibility would introduce (Henri 2010, Papalexandris, Ioannou, Prastacos 2004). Second, and in connection with the previous point, the results indicate that the introduction of a data-driven approach does not restrict opportunities for a creative dialogue (Quattrone 2016, Kolehmainen 2010), which is one of the limitations frequently associated with such approaches. This was evident throughout the phases of development of the algorithm and in the discussions that took place when the outcomes where shared with the management team.

5.2 Issues in ML implementation for performance management

The previous considerations highlighted the ability of ML to address key issues in the design of PMS and, in particular, those that are specifically related to the framing of the measurement model. However, whilst ML appear "technically" able to deliver the aforementioned benefits, our empirical investigation offers ample evidence that such outcome is not immediate or to be taken for granted.

To start with, the adoption of a ML approach in FC raised questions over the framing of the business problem (Schläfke, Silvi, and Möller 2012) and required a managerial validation of the inputs and outputs of the algorithm at various stages of the process (Moll and Yigitbasioglu 2019). Hence, the definition of what business problem the model should have answered, e.g., what are the drivers of the likelihood of debt collection, represented a necessary starting point. For instance, applying a more sophisticated approach to analyse the determinants of process efficiency would have been misleading if such increased understanding had been applied to debts with low chances of recovery. Similarly, some of the variables that emerged as key drivers of the RR during the feature selection phase could not be considered relevant from either a business and/or performance management perspective. For example, the strong correlation shown by the correlation matrix between the "procedure total duration" and the "procedure outcome" was immediately dismissed by the GM as a possible predictor of the likelihood of debt recovery since causation, as previously discussed, was in fact the other way round. The use of the Random Forest technique presented similar issues when the variable "recovery plan agreed" emerged as the main driver of the "procedure outcome". These instances confirm that a mechanical application of business analytics does not automatically generate actionable insights (Sharma, Mithas, and Kankanhalli 2014). Further, they indicate that an interpretation disconnected from the business model of the organisation and where data is automatically given to decision makers without a substantial participation into their fabrication can be misleading (Quattrone 2016). Whilst this is an issue shared by other business analytics (Raffoni et al. 2018), complex algorithms like ML worsen validation and interpretation problems due to their pattern identification functionalities and self-learning mechanisms (Shmueli 2010). Our empirical investigation adds to this debate by showing that the identification of business value drivers via ML should not be seen as a form of "automation", i.e. the algorithm taking over a human task. Rather, it should be interpreted from a perspective of augmentation, that is of humans closely collaborating with the algorithm to perform the task (Raisch and Krakowski

2021). FC's experience suggests that for ML to provide meaningful and actionable insights, data cannot be simply consumed and considered mere inputs to the decision-making process. The exercise of judgement, questioning and debating remains essential to lever ML functionalities for performance management (Quattrone 2016). In addition, it shows how such intervention benefited from the cooperation and involvement of different kinds of expertise and skills and in particular: the analytical competences of the ML expert, the capabilities of the controller and performance management experts to assess and frame the processes at the basis of organisational performance and the knowledge of the organisation and business environment brought by the managers. Hence it provides evidence for the need of multiple expertise to develop and make business sense of complex algorithms (Raffoni et al. 2018, Ransbotham, Kiron, and Kirk 2016, Popovič et al. 2012).

Secondly, and related to the previous point, the complexity and self-learning functionalities of ML approaches raise concerns over the actual interpretability of the outputs produced by the algorithm from a managerial perspective (Meske et al. 2022). In discussing the ability of ML to moderate analytical complexity, we emphasised the predictive ability of ML in terms of identifying the antecedents of the likelihood of debt recovery as well as its capacity to adapt over time. In the context of managing business performance, these are considered essential qualities for PMSs to identify weak signals and anticipate change (Nudurupati, Garengo and Bititci 2021, Bourne et al. 2018). However, predictivity in ML comes at the cost of interpretability with the best predictive approaches typically showing low degrees of interpretability. In this context, FC's experience helped to shed light on this dichotomy and the management of the related tensions. First, these emerged in the choice of the ML approach when neural networks (which have the highest predictive ability) where excluded in favour of the logistic regression and the decision tree to ensure a higher ability to unpack the value drivers and their impact. Secondly, and less obviously, such dichotomy became more relevant when discussing the actual predictivity of the model and its measurement. For instance, the unbalanced nature of the dataset made accuracy expressed as the ratio between values correctly predicted over total cases rather meaningless in FC context, with AUC, i.e., the ability of the algorithm to distinguish between classes of outcomes, being more appropriate to understand the drivers of debt recovery. However, AUC itself proved challenging from a managerial perspective where the ability of the model to predict "true" positive outcomes was perceived as the most relevant feature. Hence, the substitution of AUC with a measure of "precision" helped to encourage managerial engagement with the results of the algorithm. These findings suggest that whilst predictivity is an important feature of ML in performance management, it is not one that would be prioritised over the interpretability of the model. More broadly, and in line with recent contributions, they indicate that the ability to explain the algorithm outcomes represent a prerequisite for an accountable and trustworthy use of such approaches (Meske et al. 2020, Miller 2019). The relevance of interpretability was further supported by the clear managerial preference for the decision tree when it came to the visualisation of the results of the algorithm. Whilst comparable in terms of predictivity and actual outputs, the decision tree, unlike the logistic regression, was able to clearly describe the outcome generated by different combinations of variables. This allowed managers to understand which levers to act upon in order to maximise performance, and how to predict and assess the outcome of a given combination of external and internal variables. As a first exploratory result, the decision tree appears therefore more able to moderate methodological and analytical complexity in the identification and measurement of the value drivers by balancing high exploratory capability, predictivity and interpretability compared to logistic regression. This result is consistent with previous research underlying the role of reporting and visualisation in reducing the knowledge gap between managers and data scientists for BA applications (LaValle et val. 2011), and particularly for ML applications (Reis et al. 2020).

Thirdly, previous considerations over the need for different sets of expertise and interpretability of the algorithm have implications in terms of skills and knowledge requirements. Our case showed how team members had to work together at various stages to develop and validate the ML algorithm. Here the process was supported by the two initial researchers who shared research experience in performance management and business analytics. Their cross knowledge of the subjects was critical in finding common communication channels and bridging the different expertise of the team members. This, in our view, has at least two important consequences. First, the application of ML to performance management would benefit from team members basic understanding of each other subject areas. ML experts (data scientist) need to be able to explain the basics of the algorithm to favour engagement and interpretability

of the inputs and outputs of the model, PM experts need to clarify how to frame the performance of the company and managers and end-users more in general need to describe the organisation and its business environment (La Valle et al. 2011, Stubbs 2011). In line with previous studies, our work reiterates the context-dependency of business analytics applications and relevance of developing and interpreting them in relation to the organisational business model (Quattrone 2016, Schläfke, Silvi and Möller 2012). At the same time, sharing a common language would benefit end-user engagement with BA results, which, as shown by FC case, very much depends on their ability to understand the potential of the approach and interpret the results (Raffoni et al. 2018, Giboney et al. 2015, Popovič et al. 2012). Secondly, and in consideration of their established role in supporting performance management decisions, controllers emerge as potential key figures in bridging data science expertise and business knowledge. However, and in line with recent literature on the changing role of controllers in the current digital environment (Oesterreich et al. 2019), our case study suggests the need for controllers to develop new analytical capabilities to both exploit the potential of advanced analysis tools and frame and validate their outcomes (Kokina et al. 2021)

Lastly, our paper shows that to be able to adopt a ML approach in PMS, the information infrastructure must be able to provide significant support. Whilst this is common to other BA applications (Raffoni et al. 2018, Bhimani and Willcocks 2014, Nudurupati et al. 2011), it is worth reminding its essential role in terms of ensuring data availability and most importantly data quality. In the FC case the controller was convinced the organisation had an extensive, solid dataset. However, when the researchers made their initial assessment, a significant amount of missing or unreliable data were found. This required significant data cleaning and even greater efforts in terms of data pre-processing where variables are analysed, possibly recoded or binned in new ones more suitable for the analysis. However, when considering data quality and availability, ML applications present additional risks. Bias in the original dataset, as well as under or over representation of certain categories, would be aggravated by the self-learning mechanism of the algorithm. This may lead to statistical bias and the learning of correlations that do not exist in the real world (Lapuschkin et al., 2019).

6. Conclusions

This paper investigated the capability of ML to moderate the methodological and analytical complexity that characterise the identification of value drivers and the main issues generated by the process. We adopted an action-research approach and examined the development of a ML approach in the making. The contribution of our work to the performance management literature is threefold.

First, it adds to the literature investigating the internal complexity that emerges in the implementation of PMS and in particular the challenges faced in the identification and measurement of the value drivers the PMS should be based on (Okwir et al. 2017, Buytendijk, Hatch, and Micheli 2010, Sundin, Grandlund, and Brown 2010). By providing a detailed empirical account of the shift from the adoption of a conventional qualitative approach to the implementation of a sophisticated ML method, it shows how the latter can reduce the difficulty and risks faced by organisations when selecting the value drivers but also how the variables and interdependences can be measured and quantified. As such, our results suggests that the technical features of ML may enhance the diagnostic and predictive ability of PMS as well as contribute to its adaptability over time (Bourne et al. 2018, Melnyk et al. 2014).

Second, it contributes to the recent and ongoing debate on the adoption of business analytics in performance management (Raffoni et al. 2018, Schläfke, Silvi and Möller 2012) and to the best of our knowledge, it is the first work exploring the application of ML in this context. As such, it makes a methodological contribution by providing an in-depth empirical application of the development of a ML approach for performance measurement purposes. In addition, and in our view more importantly, our work shows how the benefits generated by the application of ML follow an augmenting rather than an automation logic. In other words, knowledge and insights produced by the application of ML are not the result of a mere mechanical computation but the ones of a hybrid approach where different forms of human interventions and skills are required at various stages of the development process (Robert, Giuliani and Gurau 2022, Raisch and Krakowski 2021). The observed shift from predictivity to interpretability is emblematic and suggests that the challenge for the implementation of ML and possibly other forms of artificial

intelligence in this context lies in the organisational and managerial readability of their inputs and outcomes (Meske et al. 2020, Miller 2019).

Third, and linked to the previous point, our work extends the literature on the changing role of the controller in the digital age (Oesterreich et al. 2019). Our results add to the discussion about the role that controllers can play in a context of increased use of data and analytical tools and the type of expertise and skills that may be required in the future. On one side, controllers' knowledge of the organisation and performance management model appear to constitute the perfect bridge between data experts and decision-makers in terms of framing and validation of business analytics applications (Kokina et al. 2021). At the same time, our case study clearly demonstrates that this bridging role tends to be beyond the skills of a typical controller and that such function benefits from the presence of cross expertise.

Our paper has also several managerial implications. First, it sheds light on the potential role of ML in supporting the definition of the organisational measurement model and the detailed empirical application clearly identifies the different phases to follow to implement such approaches. In so doing, it also makes managers aware of the associated risks and in particular of the importance of developing and interpreting ML inputs and results in relation to the business model of the organisation (Raffoni et al. 2018, Harford 2014). One additional point from a practice perspective is that ML is able deal with imperfections and unbalances in datasets, an aspect that is very common in the practice of organisations and that in many cases constitutes a barrier to the adoption of business analytics approaches based on statistical models. A clear example in this sense is represented by the rebalancing of the sample where a significant amount of new, "artificial" data was generated and which would have been unacceptable from a statistical perspective. In our view, this more "practical" approach to business analytics may prove especially valuable for effective applications in real settings.

However, we also note that our work presents some limitations, most of which in our view constitute future research opportunities. To start with, and given its exploratory nature, this paper is based on a single case study which mainly draws on an internal dataset. Considering the rise in digitisation and the availability of external, user-generated information, future research could look at whether and how the support of ML to the identification and measurement of value drivers changes in more complex and digitalised settings (Nudurupati, Garengo and Bititci 2021, Moll and Yigitbasioglu 2019). Secondly, our work mainly focuses on the adoption rather than use of a ML approach to support the framing of the measurement model which leaves many aspects of the performance measurement process unexplored. For instance, future works could investigate the interplay between the technical and social complexity that may emerge from ML implementation and related tensions (Okwir et al. 2018). Considering that subjectivity in the selection of value drivers has often been associated with the questioning of the PMS model and reward system, research could look at the implications that adopting more evidence-based approaches may generate in terms of employee trust and performance (Franco-Santos and Otley 2018). Further, interpretability of advanced algorithms like ML constitutes a challenging aspect in the application of ML in performance management and business settings in general. Understanding how different ML approaches can provide a balance between interpretability and predictivity as well as further exploring the augmenting nature of ML applications undoubtedly represent relevant research opportunities (Meske et al. 2020, Moll and Yigitbasioglu 2019, Schneider et al. 2015). Finally, and linked to the previous point, the investigation of business analytics in the context of PMS lacks theoretical development. Considering how the extensive introduction of business analytics in organisational practice generate hybrid decision making processes, future work may benefit from the use of theories that are able to overcome the dualism between technology and human actors and accommodate both aspects jointly (Burger, White and Yearworth 2019). For instance, our understanding of the implications of employing ML in performance management may benefit from the employment of Actor-Network Theory (Latour, 2007), Sociomateriality (Orlikowski, 2010) and Cyborgs (Haraway, 2018).

References

Ahn, H. 2001."Applying the balanced scorecard concept: an experience report." *Long Range Planning* 34 (4): 441-461. doi:10.1016/S0024-6301(01)00057-7.

Algorithmia. 2020: “State of enterprise machine learning”, available at: https://algorithmia.com/state-of-ml. Last access: 17/12/2021.

Alloghani, M., D. Al-Jumeily, J. Mustafina, A. Hussain and A. Aljaaf. 2020. “A systematic review on supervised and unsupervised machine learning algorithms for data science.”, in Berry W., Mohamed A., Wah Yap B. (Eds.), *Supervised and unsupervised learning for data science*, 3-21. New York: Springer. doi:10.1007/978-3-030-22475-2_1.

Appelbaum, D., A. Kogan, M. Vasarhelyi and Z. Yan. 2017. “Impact of business analytics and enterprise systems on managerial accounting.” *International Journal of Accounting Information Systems* 25: 29-44. doi:10.1016/j.accinf.2017.03.003.

Argyris, C., R. Putnam, and D. McLain Smith. 1985. *Action Science*. San Francisco: Jossey-Bass.

Aytug, H., S. Bhattacharyya, G.J. Koehler, and J.L. Snowdon. 1994. “A review of machine learning in scheduling” *IEEE Transactions on Engineering Management* 41 (2): 165-171. doi:10.1109/17.293383.

Bao, Y., B. Ke, B. Li, Y.J. Yu, and J. Zhang. 2020. “Detecting Accounting Fraud in Publicly Traded US Firms Using a Machine Learning Approach.” *Journal of Accounting Research* 58 (1): 199-235. doi:10.1111/1475-679X.12292.

Bhimani, A., and K. Langfield-Smith. 2007. “Structure, formality and the importance of financial and non-financial information in strategy development and implementation.” *Management Accounting Research* 18 (1): 3-31. doi:10.1016/j.mar.2006.06.005.

Bhimani, A., and L. Willcocks. 2014. “Digitisation, ‘Big Data’ and the Transformation of Accounting Information.” *Accounting and Business Research*, 44 (4): 469-490. doi:10.1080/00014788.2014.910051.

Bisbe, J. and R. Malagueño. 2012. "Using strategic performance measurement systems for strategy formulation: Does it work in dynamic environments?." *Management Accounting Research* 23 (4): 296-311. doi:10.1016/j.mar.2012.05.002.

Bititci, U., P. Garengo, A. Ates, and S.S. Nudurupati. 2015. “Value of maturity models in performance measurement.” *International Journal of Production Research* 53 (10): 3062-3085. doi:10.1080/00207543.2014.970709.

Bititci, U., P. Garengo, V. Dörfler, and S. Nudurupati. 2012. “Performance Measurement: Challenges for Tomorrow.” *International Journal of Management Reviews* 14 (3): 305-327. https://doi.org/10.1111/j.1468-2370.2011.00318.x.

Bourne, M., J. Mills, M. Wilcox, A. Neely, and K. Platts. 2000. “Designing, implementing and updating performance measurement systems.” *International Journal of Operations & Production Management* 20 (7): 754-771. doi:10.1108/01443570010330739.

Bourne, M., M. Franco-Santos, P. Micheli, and A. Pavlov. 2018. “Performance measurement and management: a system of systems perspective.” *International Journal of Production Research* 56 (8): 2788-2799. doi:10.1080/00207543.2017.1404159.

Bourne, M., M. Kennerley, and M. Franco-Santos. 2005. “Managing through measures: a study of impact on performance.” *Journal of Manufacturing Technology Management*, 16 (4): 373-395. doi:10.1108/17410380510594480.

Browne, M. W. 2000. “Cross-validation methods.” *Journal of Mathematical Psychology* 44 (1): 108-132. doi:10.1006/jmps.1999.1279.

Burger, K., L. White, and M. Yearworth. 2019. “Developing smart operational research with hybrid practice theories.” *European Journal of Operational Research* 277 (3): 1137-1150. doi:10.1016/j.ejor.2019.03.027.

Burney, L.L., C.A. Henle, and S.K. Widener. 2009. “A path model examining the relations among strategic performance measurement system characteristics, organizational justice, and extra-and in-role performance.” *Accounting, Organizations and Society* 34 (3-4): 305-321. doi:10.1016/j.aos.2008.11.002.

Buytendijk, F., T. Hatch, and P. Micheli. 2010. “Scenario-Based Strategy Maps.” *Business Horizons* 53 (4): 335-347. doi:10.1016/j.bushor.2010.02.002.

Cai, J., J. Luo, S. Wang, and S. Yang. 2018. “Feature selection in machine learning: A new perspective.” *Neurocomputing* 300: 70-79. doi:10.1016/j.neucom.2017.11.077.

Campbell, D., S.M. Datar, S.L. Kulp, and V.G. Narayanan. 2015. “Testing strategy with multiple performance measures: Evidence from a balanced scorecard at Store24.” *Journal of Management Accounting Research* 27(2): 39-65. doi:10.2308/jmar-51209.

Capgemini. 2015. “Cracking the Data Conundrum: How Successful Companies Make Big Data Operational.” Accessed 10 January, 2022. https://www.capgemini.com/gb-en/wp-content/uploads/sites/3/2019/01/Cracking-the-Data-Conundrum-How-Successful-Companies-Make-Big-Data-Operational.pdf.

Caruana, R., and A. Niculescu-Mizil. 2006. “An empirical comparison of supervised learning algorithms.” Proceedings of the 23rd international conference on Machine learning: 161-168. doi:10.1145/1143844.1143865.

Cheng, M.M., and K.A. Humphreys. 2012. “The differential improvement effects of the strategy map and scorecard perspectives on managers' strategic judgments.” *The Accounting Review* 87 (3): 899-924. doi:10.2308/accr-10212.

Chenhall, R. H. 2005. “Integrative Strategic Performance Measurement Systems, Strategic Alignment of Manufacturing, Learning and Strategic Outcomes: An Exploratory Study.” *Accounting, Organizations and Society* 30 (5): 395-422. doi:10.1016/j.aos.2004.08.001.

CIMA - Chartered Institute of Management Accountants. 2014. “Big Data. Readying Business for the Big Data Revolution.” Accessed 15 January, 2022. http://www.cgma.org/Resources/Reports/DownloadableDocuments/CGMA-briefing-big-data.pdf.

Copeland, T., T. Koller, and J. Murrin. 1999. *Measuring and managing the value of companies*. New York: John Wiley & Sons.

Coughlan, P., and D. Coghlan. 2002. “Action research for operations management.” *International Journal of operations & Production Management* 22 (2): 220-240. doi:10.1108/01443570210417515.

Cyert, R. M., and J.G. March. 1963. *A behavioral theory of the firm*, Englewood Cliffs, NJ: Prentice Hall,

Damodaran, A. 2012. *Investment valuation: Tools and techniques for determining the value of any asset*. New York: John Wiley & Sons.

Davenport, T. H., and J. G. Harris. 2007. *Competing on Analytics. The New Science of Winning*. Boston: Harvard Business Press.

De Waal, A.A., and H. Counet. 2009. “Lessons learned from performance management systems implementations.” *International Journal of Productivity and Performance Management* 58 (4): 367-390. doi:10.1108/17410400910951026.

Denzin, N. K., and Y. S. Lincoln. 2008. *Collecting and Interpreting Qualitative Materials*. 3rd ed. London: Sage.

Dubois, A., and L.E. Gadde. 2002. “Systematic Combining: An Abductive Approach to Case Research.” Journal of Business Research, 55 (7): 553-560. doi: 10.1016/S0148-2963(00)00195-8.

Dzyabura, D., and H. Yoganarasimhan. 2018. “Machine learning and marketing”. In Mizik, N., & Hanssens, D. M. (Eds.). (2018). *Handbook of marketing analytics: Methods and applications in marketing management, public policy, and litigation support*. Cheltenham, UK: Edward Elgar Publishing. doi: 10.4337/9781784716752.00023.

Economist Intelligence Unit. 2012. “The Deciding Factor: Big Data & Decision Making”. Capgemini, Accessed 14 July, 2021.
https://www.capgemini.com/wpcontent/uploads/2017/07/The_Deciding_Factor__Big_Data___Decision_Making.pdf .

Ferreira, A., and D. Otley. 2009. “The Design and Use of Performance Management Systems: An Extended Framework for Analysis.” *Management Accounting Research* 20 (4): 263-282. doi:10.1016/j.mar.2009.07.003.

Fleming O., T. Fountaine, N. Henke and T. Saleh. 2018. “Ten red flags signaling your analytics program will fail”. Accessed 5 January, 2022. https://www.mckinsey.com/business-functions/quantumblack/our-insights/ten-red-flags-signaling-your-analytics-program-will-fail.

Franco-Santos, M., L. Lucianetti, and M. Bourne. 2012. "Contemporary performance measurement systems: A review of their consequences and a framework for research". *Management Accounting Research* 23 (2): 79-119. doi:10.1016/j.mar.2012.04.001.

Franco-Santos, M. and D. Otley. 2018. "Reviewing and theorizing the unintended consequences of performance management systems." *International Journal of Management Reviews* 20 (3): 696-730. doi:10.1111/ijmr.12183.

Garengo, P., S. Biazzo, and U.S. Bititci. 2005. "Performance measurement systems in SMEs: A review for a research agenda." *International Journal of Management Reviews* 7 (1): 25-47. doi: 10.1111/j.1468-2370.2005.00105.x.

Giboney, J.S., S.A. Brown, P.B. Lowry, and J.F. Nunamaker Jr. 2015. "User acceptance of knowledge-based system recommendations: Explanations, arguments, and fit". *Decision Support Systems* 72: 1-10. doi:10.1016/j.dss.2015.02.005.

Gimbert, X., J. Bisbe, and X. Mendoza. 2010. "The Role of Performance Measurement Systems in Strategy Formulation Processes." *Long Range Planning*, 43: 477-497. doi:10.1016/j.lrp.2010.01.001.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep learning*. Cambridge, MA: MIT press

Gordini, N., and V. Veglio. 2017. "Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry." *Industrial Marketing Management* 62: 100-107. doi:10.1016/j.indmarman.2016.08.003.

Granger, C. W., and T. Terasvirta. 1993. *Modelling non-linear economic relationships*. Oxford, UK: OUP Catalogue.

Grover, V., R.H. Chiang, T.P. Liang, and D. Zhang. 2018. "Creating strategic business value from big data analytics: A research framework." *Journal of Management Information Systems* 35 (2): 388-423. doi:10.1080/07421222.2018.1451951.

Gummesson, E. 2000. *Qualitative Methods in Management Research*, 2nd ed. Thousand: Sage.

Hall, M. 2008. "The effect of comprehensive performance measurement systems on role clarity, psychological empowerment and managerial performance." *Accounting, Organizations and Society* 33 (2-3): 141-163. doi:10.1016/j.aos.2007.02.004.

Hall, M. 2011. "Do comprehensive performance measurement systems help or hinder managers' mental model development?" *Management Accounting Research* 22 (2): 68-83. doi:10.1016/j.mar.2010.10.002.

Haraway, D. 2018. "Staying with the trouble for multispecies environmental justice." *Dialogues in Human Geography* 8(1): 102-105. doi:10.1177/2043820617739208.

Harford, T. 2014. "Big data: A big mistake?" *Significance* 11 (5): 14-19. doi: 10.1111/j.1740-9713.2014.00778.x.

Haykin, S. 1994. *Neural networks: a comprehensive foundation*. Hoboken: Prentice Hall.

Helo, P., and Hao, Y., 2021. "Artificial intelligence in operations management and supply chain management: an exploratory case study." *Production Planning & Control*: 1-18. doi:10.1080/09537287.2021.1882690.

Henri, J.F. 2010. "The periodic review of performance indicators: an empirical investigation of the dynamism of performance measurement systems." *European Accounting Review* 19 (1): 73-96. doi:10.1080/09638180902863795.

Hindman, M. 2015. "Building better models: Prediction, replication, and machine learning in the social sciences." *The Annals of the American Academy of Political and Social Science* 659 (1): 48-62. doi:10.1177/0002716215570279.

Holsapple, C. W., A. Lee-Post, and R. Pakath. 2014. "A Unified Foundation for Business Analytics." *Decision Support Systems* 64: 130-141. doi:10.1016/j.dss.2014.05.013.

Huang, D., and L. Luo. 2016. "Consumer preference elicitation of complex products using fuzzy support vector machine active learning." *Marketing Science* 35 (3): 445-464. doi:10.1287/mksc.2015.0946.

Huelsbeck, D.P., K.A. Merchant, and T. Sandino. 2011. "On testing business models." *The Accounting Review* 86 (5): 1631-1654. doi:10.2308/accr-10096.

IDC. 2019. "IDC Future Scape: Worldwide IT Industry 2020 Predictions." IDC (International Data Corporation), Excerpt. Accessed 1 February, 2022. https://www.idc.com/getdoc.jsp?containerId=US45599219&pageType.

Ittner, C. D., and D. F. Larcker. 2001. "Assessing Empirical Research in Managerial Accounting: A Value-Based Management Perspective." *Journal of Accounting and Economics* 32(1-3): 349-410. doi:10.1016/S0165-4101(01)00026-X.

Ittner, C. D., D. F. Larcker, and M. W. Meyer. 2003. "Subjectivity and the Weighting of Performance Measures: Evidence from a Balanced Scorecard." *The Accounting Review* 78 (3): 725-758. doi:10.2308/accr.2003.78.3.725.

Ittner, C.D. and Larcker, D.F. 2005. "Moving from strategic measurement to strategic data analysis." In *Controlling strategy: management, accounting, and performance measurement*, edited by C.S. Chapman, 86-105. New York: Oxford University Press.

Jensen, M. L., P.B. Lowry, J.K. Burgoon, and J.F. Nunamaker. 2010. "Technology dominance in complex decision making: The case of aided credibility assessment." *Journal of Management Information Systems* 27 (1): 175-202. doi:10.2753/MIS0742-1222270108.

Jiang, Y., and S. Jones. 2018. "Corporate distress prediction in China: a machine learning approach." *Accounting & Finance* 58 (4): 1063-1109 doi:10.1111/acfi.12432.

Johanson, U., M. Skoog, A. Backlund, and R. Almqvist. 2006. "Balancing dilemmas of the balanced scorecard." Accounting, Auditing & Accountability Journal 19 (6): 842-857. doi:10.1108/09513570610709890.

Jönsson, S., and K. Lukka. 2007. "There and Back Again: Doing Interventionist Research in Management Accounting." In *Handbook of Management Accounting Research*, edited by C.S. Chapman, A.G. Hopwood, and M.D Shields, 373-397. Amsterdam: Elsevier Ltd.

Kaplan, R. S., and D. P. Norton. 1996. *The Balanced Scorecard: Translating Strategy Into Action*. Boston: Harvard Business School Press.

Kasanen, E., K. Lukka, and A. Siitonen. 1993. "The Constructive Approach in Management Accounting Research." *Journal of Management Accounting Research* 5 (Fall): 241-264.

Kelly, K. 2010. "Accuracy of relative weights on multiple leading performance measures: Effects on managerial performance and knowledge." *Contemporary Accounting Research* 27 (2): 577-608. doi: 10.1111/j.1911-3846.2010.01017.x.

Kleinbaum, D. G., K. Dietz, M. Gail, M. Klein, and M. Klein. 2002. *Logistic regression*. New York: Springer-Verlag.

Kokina, J., R. Gilleran, S. Blanchette, and D. Stoddard. 2021. "Accountant as digital innovator: Roles and competencies in the age of automation." *Accounting Horizons* 35 (1): 153-184. doi:10.2308/HORIZONS-19-145.

Kolehmainen, K., 2010. "Dynamic strategic performance measurement systems: balancing empowerment and alignment." *Long Range Planning* 43 (4): 527-554. doi:10.1016/j.lrp.2009.11.001.

Korhonen, T., E. Selos, T. Laine, and P. Suomala. 2020. "Exploring the programmability of management accounting work for increasing automation: an interventionist case study." *Accounting, Auditing & Accountability Journal* 34 (2): 253-280 doi:10.1108/AAAJ-12-2016-2809.

Kotsiantis, S. B. 2013. "Decision trees: a recent overview." *Artificial Intelligence Review* 39 (4): 261-283. doi:10.2308/accr-10096.

Koufteros, X., A. Verghese, and L. Lucianetti. 2014. "The effect of performance measurement systems on firm performance: A cross-sectional and a longitudinal study." *Journal of Operations Management* 32 (6): 313-336. doi:10.1007/s10462-011-9272-4.

Krishnan, R., J.L. Luft, and M.D. Shields. 2005. "Effects of accounting-method choices on subjective performance-measure weighting decisions: Experimental evidence on precision and error covariance." *The Accounting Review* 80 (4): 1163-1192. doi:10.2308/accr.2005.80.4.1163.

Laguir, I., S. Gupta, I. Bose, R. Stekelorum, and L. Laguir. 2022. "Analytics capabilities and organizational competitiveness: Unveiling the impact of management control systems and environmental uncertainty." *Decision Support Systems* 156, 113744. doi:10.1016/j.dss.2022.113744.

Lapuschkin, S., S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.R. Müller. 2019. "Unmasking clever hans predictors and assessing what machines really learn." *Nature Communications*, 10 (1096): 1–8. doi:10.1038/s41467-019-08987-4.

Lassila, E.M., S. Moilanen, and J.T. Järvinen. 2019. "Visualising a "good game": analytics as a calculative engine in a digital environment." *Accounting, Auditing & Accountability Journal* 32 (7): 2142-2166. doi:10.1108/AAAJ-11-2017-3252.

Latour, B. 2007. *Reassembling the social: An introduction to actor-network-theory*. Oxford: OUP.

LaValle, S., E. R. Lesser, M. Shockley, M. S. Hopkins, and N. Kruschwitz. 2011. "Big Data, Analytics and the Path From Insights to Value." *MIT Sloan Management Review* 52 (2): 21-32.

Lepak D.P., K.G. Smith, and M.S. Taylor. 2007. "Value creation and value capture: a multilevel perspective." *Academy of Management Review* 32: 180-194. doi:10.5465/amr.2007.23464011.

Lipe, M.G. and S.E Salterio. 2000. "The balanced scorecard: Judgmental effects of common and unique performance measures." *The Accounting Review* 75 (3): 283-298. doi:10.2308/accr.2000.75.3.283.

Lolli, F., E. Balugani, A. Ishizaka, R. Gamberini, B. Rimini, and A. Regattieri, 2019. "Machine learning for multi-criteria inventory classification applied to intermittent demand." *Production Planning & Control* 30 (1): 76-89. doi:10.1080/09537287.2018.1525506.

López-Ospina, H., L.E. Quezada, R.A. Barros-Castro, M.A Gonzalez, and P.I. Palominos. 2017. "A method for designing strategy maps using DEMATEL and linear programming." *Management Decision* 55 (8): 1802-1823. doi:10.1108/MD-08-2016-0597.

Lucianetti, L., V. Battista, and X. Koufteros. 2019. "Comprehensive performance measurement systems design and organizational effectiveness." *International Journal of Operations & Production Management* 39 (2): 326-356. doi:10.1108/IJOPM-07-2017-0412.

Lukka, K., and P. Suomala. 2014. "Relevant interventionist research: balancing three intellectual virtues." *Accounting and Business Research*, 44 (2): 204-220. doi:10.1080/00014788.2013.872554.

Lyly-Yrjänäinen, J., P. Suomala, T. Laine, and F. Mitchell. 2017. *Interventionist management accounting research: Theory contributions with societal impact*. London: Routledge.

Malina, M. A., and F. H. Selto. 2001. "Communicating and Controlling Strategy: An Empirical Study of the Effectiveness of the Balanced Scorecard." *Journal of Management Accounting Research* 13 (1): 47-90. doi:10.2308/jmar.2001.13.1.47.

Malina, M.A., H.S. Nørreklit, and F.H. Selto. 2007. "Relations among measures, climate of control, and performance measurement models." *Contemporary Accounting Research* 24 (3): 935-982. doi:10.1506/car.24.3.10.

Martínez-López, F. J., and J. Casillas. 2009. "Marketing Intelligent Systems for consumer behaviour modelling by a descriptive induction approach based on Genetic Fuzzy Systems." *Industrial Marketing Management* 38 (7): 714-731. doi:10.1016/j.indmarman.2008.02.003.

Maureen Benson-Rea, R. J. Brodie, H. Sima, 2013. "The plurality of co-existing business models: Investigating the complexity of value drivers." *Industrial Marketing Management* 42 (5): 717-729. doi:10.1016/j.indmarman.2013.05.011.

Mello, R., L. R. Leite, and R. A. Martins. 2014. "Is Big Data the Next Big Thing in Performance Measurement Systems." *Proceedings of the Industrial and Systems Engineering Research Conference*, Montreal, Canada, May 3.

Melnyk, S. A., U. Bititci, K. Platts, J. Tobias, and B. Andersen. 2014. "Is Performance Measurement and Management Fit for the Future?" *Management Accounting Research* 25 (2): 173-186. doi:10.1016/j.mar.2013.07.007.

Meske, C., E. Bunde, J. Schneider, and M. Gersch. 2022. "Explainable artificial intelligence: objectives, stakeholders, and future research opportunities." *Information Systems Management* 39 (1): 53-63. doi:10.1080/10580530.2020.1849465.

Micheli, P. and M. Mura. 2017. "Executing strategy through comprehensive performance measurement systems." *International Journal of Operations & Production Management* 37 (4): 423-443. doi:10.1108/IJOPM-08-2015-0472.

Micheli, P., and J. F. Manzoni. 2010. "Strategic Performance Measurement: Benefits, Limitations and Paradoxes." *Long Range Planning* 43 (4): 465-476. doi:10.1016/j.lrp.2009.12.004.

Miller, T. 2019. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial intelligence*, 267: 1-38. doi:10.1016/j.artint.2018.07.007.

Mitchell, T.M. 1997. *Machine Learning*. Mcgraw-hill, N.Y.

Moll, J., & O. Yigitbasioglu. 2019. “The role of internet-related technologies in shaping the work of accountants: New directions for accounting research.” *The British Accounting Review* 51 (6): 1-20. doi:10.1016/j.bar.2019.04.002.

Möller, K., U. Schäffer, and F. Verbeeten. 2020. “Digitalization in management accounting and control: an editorial.” *Journal of Management Control* 31(1): 1-8. doi:10.1007/s00187-020-00300-5#Bib1.

Nielsen, S. 2022. “Management accounting and the concepts of exploratory data analysis and unsupervised machine learning: a literature study and future directions.” *Journal of Accounting & Organizational Change* ahead of print, doi.org/10.1108/JAOC-08-2020-0107.

Norreklit, H. 2000. “The balance on the balanced scorecard a critical analysis of some of its assumptions.” *Management Accounting Research*, 11(1): 65-88. doi: 10.1006/mare.1999.0121.

Nudurupati, S. S., and U. S. Bititci. 2005. “Implementation and Impact of IT-Supported Performance Measurement Systems.” *Production Planning & Control* 16 (2): 152-162. doi:10.1080/09537280512331333057.

Nudurupati, S. S., P. Garengo, P., and U.S. Bititci. 2021. “Impact of the changing business environment on performance measurement and management practices.” *International Journal of Production Economics* 232: 1-15. doi:10.1016/j.ijpe.2020.107942.

Nudurupati, S. S., S. Tebboune, and J. Hardman. 2016. “Contemporary performance measurement and management (PMM) in digital economies.” *Production Planning & Control* 27(3): 226-235. doi:10.1080/09537287.2015.1092611.

Nudurupati, S. S., U. S. Bititci, V. Kumar, and F. T. S. Chan. 2011. “State of the Art Literature Review on Performance Measurement.” *Computers and Industrial Engineering* 60 (2): 279-290. doi:10.1016/j.cie.2010.11.010.

Oesterreich, T. D., F. Teuteberg, F., Bensberg and G. Buscher. 2019. “The controlling profession in the digital age: Understanding the impact of digitisation on the controller's job roles, skills and competences.” *International Journal of Accounting Information Systems* 35(C): 1-23. doi:10.1016/j.accinf.2019.100432.

Okwir, S., S.S. Nudurupati, M. Ginieis, and J. Angelis. 2018. “Performance measurement and management systems: a perspective from complexity theory.” *International Journal of Management Reviews* 20 (3): 731-754. doi:10.1111/ijmr.12184.

Orlikowski, W.J. 2010. “The sociomateriality of organisational life: considering technology in management research.” *Cambridge Journal of Economics* 34 (1): 125-141. doi:10.1093/cje/bep058.

Otley, D. and K. Soin. 2014. *Management control and uncertainty*, Palgrave Macmillan, London.

Papalexandris, A., G. Ioannou, and G.P. Prastacos. 2004. “Implementing the balanced scorecard in Greece: a software firm’s experience.” *Long Range Planning* 37(4): 351-366. doi:10.1016/j.lrp.2004.05.002.

Parker, L. D. 2012. “Qualitative Management Accounting Research: Assessing Deliverables and Relevance.” Critical Perspectives on Accounting 23 (1): 54-70. doi:10.1016/j.cpa.2011.06.002.

Popovič, A., R. Hackney, P.S. Coelho, and J. Jaklič. 2012. “Towards Business Intelligence Systems Success: Effects of Maturity and Culture on Analytical Decision Making.” *Decision Support Systems* 54 (1): 729-739. doi:10.1016/j.dss.2012.08.017.

Quattrone, P. 2016. “Management accounting goes digital: Will the move make it wiser?” *Management Accounting Research* 31: 118-122. doi:10.1016/j.mar.2016.01.003.

Raffoni, A., F. Visani, M. Bartolini and R. Silvi. 2018. “Business performance analytics: exploring the potential for performance management systems.” *Production Planning & Control* 29 (1): 51-67. doi:10.1080/09537287.2017.1381887.

Raisch, S. and S. Krakowski. 2021. “Artificial intelligence and management: The automation–augmentation paradox”. *Academy of Management Review* 46 (1): 192-210. doi:10.5465/amr.2018.0072.

Ransbotham, S., D. Kiron, and P. Kirk Prentice. 2016. “Beyond the Hype: The Hard Work Behind Analytics Success, The 2016 Data & Analytics Report by MIT Sloan Management Review & SAS.” Accessed 14 July, 2021. http://sloanreview.mit.edu/projects/the-hard-work-behind-data-analytics-strategy/.

Rappaport, A. 1998. *Creating shareholder value: a guide for managers and investors*. New York: Simon and Schuster.

Reason, P. 1999. “Integrating action and reflection through co-operative inquiry.” *Management learning* 30 (2): 207-225. doi:10.1177/1350507699302007.

Reason, P., and H. Bradbury (Eds.). 2001. *Handbook of action research: Participative inquiry and practice*. New York: Sage.

Reis, C., P. Ruivo, T. Oliveira, and P. Faroleiro. 2020. “Assessing the drivers of machine learning business value.” *Journal of Business Research* 117: 232-243. doi:10.1016/j.jbusres.2020.05.053.

Reportlinker. 2019. “Machine Learning as a Service Market: Global Industry Analysis, Trends, Market Size, and Forecasts up to 2024.”. Accessed 14 September, 2021. Available at www.reportlinker.com

Robert, M., P. Giuliani, and C. Gurau. 2022. “Implementing industry 4.0 real-time performance management systems: the case of Schneider Electric.” *Production Planning & Control* 33 (2-3): 244-260. doi:10.1080/09537287.2020.1810761.

Ryll, L., and S. Seidens. 2019. “Evaluating the Performance of Machine Learning Algorithms in Financial Market Forecasting: A Comprehensive Survey”. ArXiv preprint:1906.07786. doi:10.48550/arXiv.1906.07786.

Santos, S.P., V. Belton, and S. Howick. 2002. “Adding value to performance measurement by using system dynamics and multicriteria analysis.” *International Journal of Operations & Production Management* 22 (11): 1246-1272. doi:10.1108/01443570210450284.

Schläfke, M., R. Silvi, and K. Möller. 2012. “A Framework for Business Analytics in Performance Management.” *International Journal of Productivity and Performance Management* 62 (1): 110-122. doi:10.1108/17410401311285327.

Schneider, G.P., J. Dai, D.J. Janvrin, K. Ajayi, and R.L. Raschke. 2015. “Infer, predict, and assure: Accounting opportunities in data analytics.” *Accounting Horizons* 29 (3): 719-742. doi:/10.2308/acch-51140.

Sharma, R., S. Mithas, and A. Kankanhalli. 2014. “Transforming Decision-Making Processes: A Research Agenda for Understanding the Impact of Business Analytics on Organizations.” *European Journal of Information Systems* 23 (4): 433-441. doi:10.1057/ejis.2014.17.

Sheng, J., J. Amankwah-Amoah, and X. Wang. 2019. “Technology in the 21st century: New challenges and opportunities.” *Technological Forecasting and Social Change* 143: 321-335. doi:10.1016/j.techfore.2018.06.009.

Shmueli, G. 2010. “To explain or to predict?,” *Statistical Science* 25 (3): 289–310. doi:10.1214/10-STS330.

Silvestro, R. 2016. “Do you Know what Really Drives your Business’s Performance?” *MIT Sloan Management Review* Summer: 1-10.

Silvi, R., M. Bartolini, A. Raffoni, and F. Visani. 2012. “Business Performance Analytics: Level of Adoption and Support Provided to Performance Measurement Systems.” *Management Control* 3 (Special Issue): 117-142. doi:10.3280/MACO2013-SU3006.

Silvi, R., M. Bartolini, A. Raffoni, and F. Visani. 2015. “The Practice of Strategic Performance Measurement Systems: Models, Drivers and Information Effectiveness.” *International Journal of Productivity and Performance Management* 64 (2): 194-227. doi:10.1108/IJPPM-01-2014-0010.

Simons, R. 1995. *Levers of Control. How Managers Use Innovative Control Systems to Drive Strategic Renewal*. Boston: Harvard Business School Press.

Smith, W.K., A. Binns, and M.L. Tushman. 2010. “Complex business models: Managing strategic paradoxes simultaneously.” *Long Range Planning* 43 (2-3): 448-461. doi:10.1016/j.lrp.2009.12.003.

Sokolova, M., and G. Lapalme. 2009. “A systematic analysis of performance measures for classification tasks.” *Information Processing & Management* 45 (4): 427-437. doi:10.1016/j.ipm.2009.03.002

Steinwart, I., and A. Christmann. 2008. *Support vector machines*. New York: Springer Science & Business Media.

Stubbs, E. 2011. *The Value of Business Analytics*. New Jersey: John Wiley & Sons Inc.

Sundin, H., M. Granlund and D.A. Brown. 2010. “Balancing multiple competing objectives with a balanced scorecard.” *European Accounting Review* 19 (2): 203-246. doi:10.1080/09638180903118736.

Suomala, P., J. Lyly-Yrjänäinen, and K. Lukka. 2014. “Battlefield Around Interventions: A Reflective Analysis of Conducting Interventionist Research in Management Accounting.” *Management Accounting Research* 25 (4): 304-314. doi:10.1016/j.mar.2014.05.001.

Sutton, S. G., M. Holt, and V. Arnold. 2016. "The reports of my death are greatly exaggerated—Artificial intelligence research in accounting." *International Journal of Accounting Information Systems* 22: 60-73. doi:10.1016/j.accinf.2016.07.005.

Syam, N., and A. Sharma. 2018. "Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice." *Industrial Marketing Management* 69: 135-146. doi:10.1016/j.indmarman.2017.12.019.

Tayler, W.B., 2010. "The balanced scorecard as a strategy-evaluation tool: The effects of implementation involvement and a causal-chain focus." *The Accounting Review* 85 (3): 1095-1117. doi:10.2308/accr.2010.85.3.1095

Teece, D.J. 2010. "Business models, business strategy and innovation." *Long Range Planning* 43: 172-194. doi:10.1016/j.lrp.2009.07.003.

Van Aken, J. 2004. "Management Research Based on the Paradigm of the Design Sciences: the Quest for Field-tested and Grounded Technological Rules." *Journal of Management Studies* 41(2): 219-246. doi:10.1111/joms.2004.41.issue-2.

Van de Ven, A. H. 2007. *Engaged Scholarship: A Guide for Organizational and Social Research*. Oxford: Oxford University Press.

Van de Ven, A. H., and P. E. Johnson. 2006. "Knowledge for Theory and Practice." *Academy Management Review*, 31(4): 802-821. doi: 10.5465/amr.2006.22527385.

Vandenbosch, B. and C. Higgins. 1996. "Information acquisition and mental models: An investigation into the relationship between behaviour and learning." *Information Systems Research* 7 (2): 198-214. doi:10.1287/isre.7.2.198.

Voelpel, S. C., M. Leibold, and R.A. Eckhoff. 2006. "The tyranny of the Balanced Scorecard in the innovation economy". *Journal of Intellectual Capital* 7 (1): 43-60. doi:10.1108/14691930610639769.

Warren, J. D., K. C. Moffitt, and P. Byrnes. 2015. "How Big Data will Change Accounting." *Accounting Horizons* 29 (2): 397-407. doi:10.2308/acch-51069.

Wiersma, E. 2009. "For which purposes do managers use Balanced Scorecards?: An empirical study." *Management Accounting Research* 20 (4): 239-251. doi:10.1016/j.mar.2009.06.001.

Witten, I. H., and E. Frank. 2002. "Data mining: practical machine learning tools and techniques with Java implementations". ACM SIGMOD Record 31(1): 76-77. doi:10.1145/507338.507355.

Wong-On-Wing, B., L. Guo, W. Li, and D. Yang. 2007. "Reducing conflict in balanced scorecard evaluations." *Accounting, Organizations and Society* 32 (4-5): 363-377. doi:10.1016/j.aos.2006.05.001.

Wuest, T., D. Weimer, C. Irgens, and K.D. Thoben. 2016. "Machine learning in manufacturing: advantages, challenges, and applications." *Production & Manufacturing Research* 4(1): 23-45. doi:10.1080/21693277.2016.1192517.

Yin, R. K. 2008. *Case Study Research: Design and Methods*. Beverly Hills, CA: Sage Publications.

Zhang, Y. 2012. "Support vector machine classification algorithm and its application." In *International Conference on Information Computing and Applications*, Springer, Berlin, Heidelberg: 179-186. doi:10.1007/978-3-642-34041-3_27.

Appendix 1

Name	Type	Description
Amount outstanding	Numerical	Amount of money (€) to be recovered
E-mail status	Categorical	Current status of the e-mail sent to contact the debtor (Categories.: Sent/In Transit/Rejected/Wrong Address/Undelivered/Delivered)
E-mail received	Categorical	Synthetic status of the e-mail sent to contact the debtor (0=not delivered, 1=delivered)
Payment before phone call	Categorical	Payment obtained before the phone call. (0=no; 1=yes)
Amount Recovered before phone call	Numerical	Amount of money (€) recovered before the phone call.
Total duration of the procedure	Numerical	Total amount of time (minutes) spent by an operator in the procedure aimed to recover the money.
Operator	Categorical	Numerical ID identifying the operator in charge of the procedure
Operator Contact Status	Categorical	The detailed current status of the procedure (Categories: Contacted/Awaiting Documentation/Wrong Phone Number/Call Planned/Refused to Pay/ Agreed to Pay).
Procedure Outcome	Categorical	A binary variable summarizing the final status of the procedure (0=Not paid; 1=Paid)
Amount Recovery Plan	Numerical	The amount (€) of the recovery plan to which the debtor has agreed.
Amount Recovered	Numerical	The amount or money (€) recovered so far.
Year of birth	Numerical	The year of birth of the debtor inferred from his/her personal tax code.
Sex	Categorical	The gender of the debtor (M=Male, F=Female)
Province	Categorical	The Italian province where the debtor lives.
Macroregion	Categorical	The Italian Macroregion where the debtor lives (Categories: Northern Italy, Central Italy, Southern Italy and Islands).
Solvent	Categorical	The solvency of the debtor as inferred by the company. (Categories: YES (FSI)= the debtor has a full-time job; NOT KNOWN (FNP)= the debtor salary cannot be determined accurately (i.e. freelancer, part-time/occasional workers). NO (FNO)= the available data about the liquidity of the debtors are negative.
Solvent with Property	Categorical	The solvency of the debtor according to his/her ownership of real estate. (Categories: YES= the debtor owns any real estate; NO= the debtor does not own any real estate.

Table A1: The 17 variables of the NPL dataset

Issue	Data cleaning activity
"Amount Outstanding" equal to 0 or very low	19 records with "Amount Outstanding" lower than €10 were removed
"Amount recovered" higher than the "Amount outstanding"	The company explained that the mismatch was due to interests charged although no variable is available in the database for the exact amount and nature of these infrequent interests.
The variable "E-mail status" had 2,788 values recorded as empty strings.	All the 2,788 cases are encoded as "E-mail not delivered" in the corresponding summary variable "E-mail received". The company confirmed that "E-mail received" was to be preferred.
The dependent variable "Procedure Outcome" treats as success also partial debt recovery for 933 records	We decided to stick with the company logic of considering a success even partial debt recovery.
The variable "Operator" has some levels with a non-significant number of records (i.e. level 0 with only 1 record) and 5 NAs.	This issue was addressed by using a binning procedure discussed later.
The variable "Province" has 3,188 missing values.	The variable was dropped from the analysis given that the related variable "Macroregion" is not descriptive enough to provide meaningful imputation.
The variables "Payment before phone call" and "Amount Recovered before phone call" have 124 records flagged as successful but the corresponding "Amount recovered" is always zero.	While theoretically these records could be dropped, they still convey information about the type of debtors likely to pay, so it was decided to set "Amount recovered" equal to "Amount recovered before phone call", while the two variables "Payment before phone call" and "Amount Recovered before phone call" were dropped.
60 records have a value equal to 0 for the variables "Total Duration of the procedure", with a corresponding "Procedure Outcome" equal to 0.	The records were eliminated from the dataset.

Table A2: The actions taken to clean the dataset