escola superior de tecnologia e gestão
instituto politécnico de leiria

Simulation Applications in Company Default Prediction

Master Degree in Corporate Finance

Ricardo Jorge Leandro Castelo

Leiria, September of 2023

Simulation Applications in Company Default Prediction

Master Degree in Corporate Finance

Ricardo Jorge Leandro Castelo

Dissertation under the supervision of Professor Magali Pedro Costa, and Professor Inês Margarida Cadima Lisboa

(Corrected version after public dissertation defense)

Leiria, September of 2023

Originality and Copyright

Acknowledgments

I hereby would like to thank my few friends and my close family for their support during the difficult and consuming task that is the development of a thesis under any subject. I would also like to thank Professor Magali Costa and Professor Inês Lisboa personally for their patience, understanding, and aid in the development and polish of this dissertation into its full potential, regardless of if there were ever plans lost along the way.

To all the future readers of this study, I thank you for taking the time to read this piece of work of mine, in which I proudly chose to challenge myself to learn, adapt, and manage, with all the help of those deserving of mention.

Abstract

This study applies a simulation methodology, Monte Carlo, to the field of corporate default prediction, where its presence is only superficial. It attempts to augment a famous model from a methodology already highly supported in traditional literature – the Z”-score of Altman (1983), created through multiple discriminant analysis – and transform it stochastically without the use of the highly complex intelligent models already available in literature. A sample of 20 000 Portuguese companies from the Agriculture, Forestry, Fishing, Mining and Construction sectors is analyzed, yielding results that support the Monte Carlo method as a strong competitor for simple approaches like the logit transformation. This helps to build the foundation for what may possibly be a path towards models easier to apply in practice for the average Micro, Small and Medium enterprise (MSME) and beyond. The evaluation to the model on this study also takes inspiration off the innovative points of view of Mitton (2021) and Zhang (2022) to scrutinize results through empirical tests of the model under a high number of parameter conditions, instead of relying heavily on statistical significance, which is often overrepresented and overvalued in literature, and easily manipulatable.

Keywords: simulation, default risk, prediction, Portugal, Monte Carlo, Z”-score

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

AI	Artificial Intelligence
AR	Accuracy Ratio
AUC	Area under the curve
BSM	Black-Scholes-Merton
CAP	Cumulative Accuracy Profile
CI	Confidence Interval
DPM(s)	[Corporate] Default Prediction Model(s)
EBIT	Earnings Before Interest and Taxes
ESTG	School of Technology and Management
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
GDP	Gross Domestic Product
INE	Instituto Nacional de Estatística (Portugal)
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MDA	Multiple Discriminant Analysis
MH	Metropolis-Hastings
MSME(s)	Micro, Small and Medium Enterprise(s)
ROC	Receiver Operating Characteristic
SA	Simulated Annealing
SIC	Standard Industrial Classification
SRRM	Square Root Recursive Method
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
Z”LT	Logit Transformed Z”-score
Z”MC_C	Monte Carlo Z”-score Simulated by Components (Ratios)
Z”MC_D	Monte Carlo Z”-score Simulated Directly

1. Introduction

Default prediction modeling is a field of financial literature that has been contested for almost 60 years now (Hillegeist et al., 2004; Balcaen & Ooghe, 2006; Bakshi et al., 2022; Kim et al., 2020; Sun et al., 2014). It encompasses various perspectives and approaches (Alaka et al., 2018; Aziz & Dar, 2006; Jackson & Wood, 2013; Kumar & Ravi, 2007), and has followed up with some of the best developments in computer science (Aziz & Dar, 2006; Kim et al., 2020), with models more and more complex as we approach the modern day (Kim et al., 2020; Sun et al., 2014).

The Monte Carlo simulation is a powerful tool that can aid in the calculations of the more traditional approaches that still strive due to their simplicity (Andrieu et al., 2003; Jackson & Wood, 2013; James, 1980; Kumar & Ravi, 2007; Sun et al., 2014). It was from the academic, personal, and professional contact with this tool that stemmed the intent to test its applications on the more traditional statistical techniques that are widespread in finance corporate default studies (Aziz & Dar, 2006; Balcaen & Ooghe, 2006; Dimitras et al., 1996; Jackson & Wood, 2013). Interestingly enough, this tool has not had much application in the default prediction field, which lifted the opportunity to test out its usefulness.

Until now, no direct applications of the Monte Carlo simulation have been executed to directly model default, only to assist either by bootstrapping samples for results (Abinzano et al., 2020; Bottazzi et al., 2011; Chava & Jarrow, 2004; Costa et al., 2022; Innocenti et al., 2022; Korteweg, 2011; Peláez et al., 2022; Reisz & Perlich, 2007; Xu et al., 2021) or assess the effects of parameter variation on said models (Hosseini & Hajiannejad, 2022). This study aims to take advantage of this opportunity to augment the pre-existing traditional literature on default prediction models using Monte Carlo simulation, in an attempt to bridge the practical and the powerful in order to present a new set of tools for the bettering of the micro, small and medium enterprise (MSME) company risk management as a whole, through something as simple as an excel sheet. More specifically, the objective is to test the usefulness of the Monte Carlo Simulation when assessing the default risk of MSME Portuguese companies, using more traditional and easy-to-apply models like the Z”-score of Altman (1983), by recreating those same conditions in a large data algorithm.

Much like in most countries of the world (Altman et al., 2016), most Portuguese companies are privately held: 91.94%, as of August 2022, to be exact (CEIC, 2022). Furthermore, 99.9% of Portuguese companies fall under the classification of micro, small or medium enterprises (INE & Pordata, 2021), with 96.1% solely falling under the micro category. The combination of these factors creates a need for statistically-based models of default prediction literature, in order to avoid the heavy dependency theoretical models have on market variables (Bakshi et al., 2022; Ramesh & Kumar, 2017), along with an avoidance for more complex models like the intelligent ones, as smaller companies would likely not have the sufficient resource structure to execute them. This study attempts to respond to these two needs, through the introduction of this new methodology.

Furthermore, this study aims to escape the common practices of the statistical significance overuse as a methodology validator, instead choosing to scrutinize the model through various modifications of parameters and conditions (Mitton, 2021; Zhang, 2022).

In terms of structure, this study will be organized as follows: Section 2 will develop on the current state of the art and attempt to justify any main choices required to proceed to Section 3; Section 3 will include information on data collection, the methodology and its development and base of assumptions, as well as the expected methodology for assessing the model’s predictive power and limitations; Section 4 will go over the application of the model and its results, peculiar findings, as well as the application of any tests and evaluation methods required; Section 5 will conclude and express any limitations found in the study, ideas for further developments, as well as some final regards.

2. State of the Art

Many have been the authors over the last 60 years to develop on the subject of corporate default, allowing it to become one of the major research domains within corporate finance (Hillegeist et al., 2004; Balcaen & Ooghe, 2006; Bakshi et al., 2022; Kim et al., 2020; Sun et al., 2014). In this section, the various developments from the reviewed literature will be explored, to attempt to develop the structural and methodological pillars of this study.

There are three fundamental literature ‘supports’ this study will develop on: (1) default prediction, (2) simulation application and (3) model evaluation. Each of these develops into their own sub-sections and decision criteria, which shall be explained below.

2.1. Default Prediction Models

Under the subject of default prediction modeling in literature, this study goes into detail specifically on model categorization, so one can better understand the framework and placement of the method developed ahead. Each of the following subsections will depict a different dimension through which default prediction modeling can be categorized.

2.1.1. Importance of Corporate Default Prediction

Before diving into the broad literature that encompasses corporate default, this study presents a brief explanation as to why this area is so important and why, given the grand amount of research already available, it still makes sense to try to innovate and add to its length.

Quoting Bottazzi et al. (2011, p. 374), “the 'death' of a firm represents the ultimate market response to the inability of an economic activity to survive the competitive pressure”. It is, hence, a sign of fragility of either specific firm strategies or of the underlying pressure of sectors, the market, and economy in general. Overall, default prediction is of interest not only to the company, but all stakeholders or decision-makers involved, namely, investors, creditors, banks, regulatory authorities, auditors, consultants, management, employees, and the government, among many others (Alaka et al., 2018; Dimitras et al., 1996; Du et al., 2020; Ketz, 1978; Kumar & Ravi, 2007; O'Leary, 1998; Shen et al., 2020; Sun et al., 2014, 2017). In fact, Bakshi et al. (2022) go as far as to name it the single "most significant credit event for a firm" (p. 397).

With that, it becomes clear that the area of default prediction has high economic significance, given the number of entities it affects (Jackson & Wood, 2013). A question remains, however: why is it so important for the stakeholders above? Well, the clearest of answers is that bankruptcy (the ultimate most extreme consequence of default) is costly, not just for the company, but to those stakeholders too (Bottazzi et al., 2011; Dimitras et al., 1996). Some examples of these costs are: (1) loans to companies have a heavy weight on banks' assets (Bonfim, 2009), so if the company enters bankruptcy, that cost will never have any return, and will in turn have an effect on the bank (Sun et al., 2017); (2) the failure of a business represents a high cost of tangible and intangible assets wasted (Bottazzi et al., 2011); (3) bankruptcy filings are associated with both legal and professional costs (Abinzano et al., 2020); (4) the failure of a business can produce substantial losses to creditors and stockholders that do not retrieve any of the invested value (Deakin, 1972; Jackson & Wood, 2013). Some researchers even mention how it is much less costly for a company to apply preventive regulation measures to detect default and to take corrective strategies if necessary than it is to go bankrupt (Altman, 1973; Aziz & Dar, 2006).

Company defaults may also be seen as having a 'domino' effect on their environment. Jackson and Wood (2013), for instance, compare the corporate default with a pebble being thrown into a lake, the shockwaves portraying its effects on partnered companies, industry, wider economy, and society itself (also reenforced by Kim et al., 2020). More generally, default causes an overall disruption of the supply chain and its productivity, when it happens (Abinzano et al., 2020). Hence why this field of research is deemed so important.

With the corporate default importance established, how can researchers and studies like this one aid in the protection of stakeholders from the consequences of the event? That much is achieved by the development of models that can accurately predict (forecast) default or classify defaulted companies in an *ex ante* setting. As Kim et al. (2020, p. 1) state, "default prediction models help in designing and improving the financial system". This is achieved by aiding the various agents in economy: (1) corporate default prediction models (DPMs) work as an extra tool for investors and institutions to evaluate their risk working with or investing in specific companies (Dimitras et al., 1996; Kim et al., 2020; Shen et al., 2020);

(2) DPMs can assist financial institutions (such as banks) in determining whether a client is liable for a loan or not (Du et al., 2020; O'Leary, 1998; Shen et al., 2020; Sun et al., 2017); (3) executives can use DPMs as tools to more stably manage their businesses by determining key indicators that affect it negatively (Kim et al., 2020; Sun et al., 2017); (4) bank regulators can use DPMs to help evaluate a bank's performance and take preventive or resolving measures (O'Leary, 1998); and (5) governments can use the timely knowledge offered by DPMs to better maintain market order and stability (Du et al., 2020). Overall, DPMs reveal the potential risk and financial defects of corporations, functioning as early warning systems that allow those responsible to take action and prevent or reduce considerable losses (Deakin, 1972; Dimitras et al., 1996; Du et al., 2020; Shen et al., 2020). For this reason, DPMs today are used in various fields across the economy, and it is stressed that more companies, especially the smaller sized ones, make use of these tools to allow the development of better management practices (Kim et al., 2020).

2.1.2. Default Definition

The first categorization dimension this study will grapple is the traditional dilemma of the default definition: what is considered default under the circumstances of DPMs. The truth about this long-lasting dilemma is that literature is inconsistent and fails still to this day to find consensus on how to define default properly (Balcaen & Ooghe, 2006; Costa et al., 2022; Dimitras et al., 1996; Du et al., 2020; Ohlson, 1980; Sun et al., 2014). There is little theoretical development on the subject and most studies choose arbitrarily based on their application of DPM or the data available (Balcaen & Ooghe, 2006; Chava & Jarrow, 2004; Sun et al., 2014). Facing this lack of basis, each researcher commonly applies their own criteria for default definition (Du et al., 2020).

First and foremost, it is important to address that, often in literature, the terms 'bankruptcy', 'insolvency', 'financial distress', 'failure' and 'default' (among others that may be missing in this section) are treated as similar, or even equal (Alaka et al., 2018; Altman et al., 2016; Costa et al., 2022; Dichev, 1998). In this study, the term 'default' will be addressed as the broader definition, which includes the many other terms mentioned above, hereinafter seen, instead, as different levels of default.

Through the reviewed literature, a good number of criteria were found, with little repetition, leaving out some exceptions. These are different conditions that for different studies must be met in order for a company to be declared as 'defaulted'. Although a large majority of the criteria found was utilized in modeling practice, some of them were simply mentioned as possible inclusions (i.e., in Chava and Jarrow, 2004, all criteria except for Bankruptcy Filing). A list of these criteria and associated sources follows:

(1) Bankruptcy Filing (Abinzano et al., 2020; Agarwal & Taffler, 2008; Altman, 1968, 1973, 1983; Altman et al., 2016; Beaver, 1966; Beaver et al., 2010; Blum, 1974; Chava & Jarrow, 2004; Correia et al., 2017; Deakin, 1972; Dichev, 1998; Duffie et al., 2007; Hillegeist et al., 2004; Jackson & Wood, 2013; Mensah, 1984; Nguyen, 2023; Ohlson, 1980; Ramesh & Kumar, 2017; Reisz & Perlich, 2007; Shumway, 2001; Sun et al., 2014; Wilcox, 1973; Zmijewski, 1984)
(2) Bond default (Beaver, 1966; Beaver, 1968; Beaver et al., 2010; Ramesh & Kumar, 2017; Sun et al., 2014)
(3) Credit overdue (Bonfim, 2009; Chava & Jarrow, 2004; Bottazzi et al., 2011; Nguyen, 2023; Sun et al., 2014)
(4) Overdrawn bank account (Beaver, 1966; Beaver, 1968; Ramesh & Kumar, 2017; Sun et al., 2014)
(5) Loan default (Edmister, 1972; Nguyen, 2023; Sun et al., 2014)
(6) Liabilities > Assets / Accounting Bankruptcy[1] (Beaver et al., 2010; Sun et al., 2014)
(7) Preferred stock dividend arrearage (Beaver, 1966; Beaver, 1968)
(8) Company de-listing (Dichev, 1998; Chava & Jarrow, 2004)
(9) Negative net profit for 2 successive years (Shen et al., 2020; Sun et al., 2017)
(10) Net assets per share lower than stock book value (Shen et al., 2020; Sun et al., 2017)
(11) Liquidation for the benefit of creditors (Deakin, 1972)
(12) Capital reconstructions (Balcaen & Ooghe, 2006)
(13) Forced disposals of parts of the firm (Balcaen & Ooghe, 2006)
(14) Informal government support (Balcaen & Ooghe, 2006)
(15) Loan covenant negotiations (Balcaen & Ooghe, 2006)
(16) Bank consideration that repayment of debt in full is unlikely (Bottazzi et al., 2011)
(17) Agreements with creditors to reduce debt (Sun et al., 2014)
(18) Audited operating income under a given threshold (Du et al., 2020)

---

[1] Nomenclature used by Sun et al. (2014).

(19) Audited net assets negative (Du et al., 2020)
(20) Mortgage foreclosure (Ramesh & Kumar, 2017)

As can be seen, the list of criteria is vast and, much like Balcaen and Oogue (2006) and Costa et al. (2022) pointed out, the top choice in literature seems to fall under bankruptcy filing, followed by a combination of that criteria with one other criteria or more[2]. From this broad list, the problem becomes one of classification and further analysis. On that matter, Duffie et al. (2007) explained a set of self-inclusive groups for general default criteria, in which each group fit into the next one, larger than the previous one. Their smallest group was composed of only legally bankrupt companies, which, in terms of classification, left out an option that, although not very common (Costa et al., 2022), has become a reality in literature – models that did not include the legal bankruptcy as a criteria for default. As such, this study's classification of criteria follows the perspective found in Costa et al. (2022), which divides the criteria into *ex post*, *ex ante*, and mixed[3].

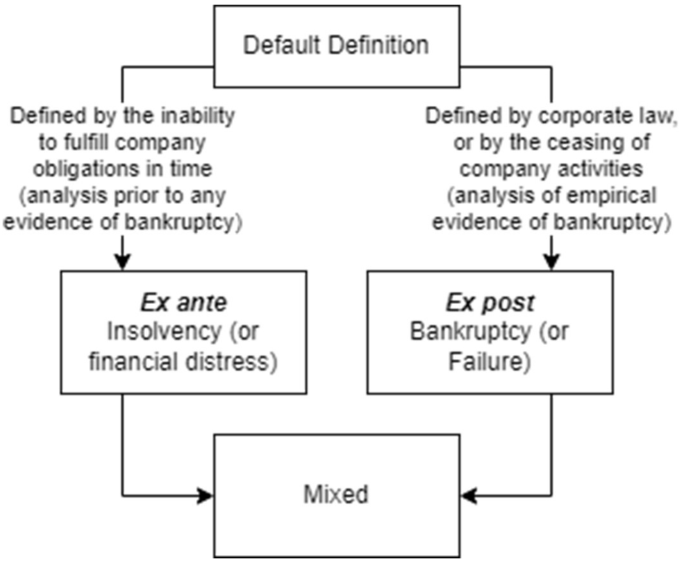


Figure 1 - Default Definition Types

Figure 1 summarizes this study's Costa et al. (2022) inspired logic for separating DPMs based on criteria utilized to define default. Beginning in the more subjective side of the

---

[2] The criteria may have specific advantages and disadvantages that have been pointed out by the authors mentioned. However, in an attempt not to lose focus on the analysis, this study will not introduce them, hence the reader is incentivized to delve into the original works if they are of interest.

[3] The terms '*ex ante*' and '*ex post*' in this study refer to default definition only and should not be confused with other applications of them, such as in Dichev (1998), which defines *ex ante* default risk as the risk a researcher aims to predict, and *ex post* default risk as the risk returned by a model.

classification, the *ex ante* models define default as the existence of financial distress signs (Costa et al., 2022). They define default under what was originally denominated as 'insolvency' – the inability of a company to fulfill its financial obligations as they mature (Beaver, 1966; Beaver et al., 2010; Bottazzi et al., 2011; Campbell et al., 2008; Edmister, 1972). In other words, insolvency is the situation in that a company has any certain kind of financial difficulties (Sun et al., 2014). This type of model, in other words, refers to models who use any of the criteria mentioned above except for that of Bankruptcy Filing. It poses some advantages over the bankruptcy filing criteria: (1) by utilizing *ex ante* criteria, the DPM works similarly to a set of warning signs and is more generalizable, since criteria can be chosen with that intent in mind (Costa et al., 2022); (2) seeing as insolvency comes first to bankruptcy (Du et al., 2020), *ex ante* models permit more timely decisions in DPMs (Costa et al., 2022); (3) given their broader sense of default, it is only logical that *ex ante* models can more easily find larger samples to train their models with; (4) bankruptcy is not an automatic consequence of insolvency (Wilcox, 1973), although insolvency can still cause deteriorating damage to companies, something that can very well be ignored by bankruptcy-only models and that *ex ante* models take into account. Regardless of these advantages, *ex ante* models are still clearly more subjective than *ex post,* which makes it difficult to find definitive criteria, which results in more inconsistency between models. Additionally, a few researchers believe there are still other warnings signs that can be discovered through *ex post* analysis (Beaver, 1968: what is apparent *ex post* may not be so apparent *ex ante*). Furthermore: (1) not assuming default as a definitive event like bankruptcy begs the question of whether a company can or cannot overcome insolvency and if previous insolvency events may be correlated also with default (Beaver, 1966); (2) the data argument can also work against this approach, seeing as certain criteria choices might make data collection more difficult or samples more limited (reason why Chava and Jarrow, 2004, chose to proceed considering only bankruptcy in their model)[4]; (3) using *ex ante* criteria eventually contaminates the model with defining default as the given criteria, even if empirically those criteria do not always mean default or insolvency (Balaen & Ooghe, 2006); that is, if the criteria used is somehow wrong (not actually portraying default, or portraying also other circumstances other than default), the model becomes a predictor of those wrong circumstances, and not just default.

---

[4] This is perhaps a disadvantage easily outdated by the ever-growing data availability of the modern world.

The direct alternative of Figure 1 to these issues is *ex post* models, i.e., models which take into consideration empirical observations of bankruptcy to declare a company defaulted (i.e., considering exclusively as default the criteria of Bankruptcy Filing). This definition is popular in literature because it offers an objective and straightforward definition (Balcaen & Ooghe, 2006). Additionally, DPMs modelled after *ex post* criteria will most of the times be able to predict insolvency as much as bankruptcy (Costa et al., 2022), while *ex ante* may not be as effective in predicting bankruptcy specifically without the right criteria. However, law does not escape from defining default vaguely either (Blum, 1974), which leads to some error space in which companies may have filed for bankruptcy not due to unavoidable financial issues, but for some sort of secondary interest (Balaen & Ooghe, 2006; Bottazzi et al., 2011; Costa et al., 2022). On the other hand, generalization is difficult with *ex post* models because different countries have different laws, different meanings for bankruptcy (Altman et al., 2016; Costa et al., 2022). There is also a common struggle when dealing with *ex post* models. Although such is considered in most *ex post* models, default does not necessarily lead to bankruptcy (Shen et al., 2020; Sun et al., 2014; Wilcox, 1973), some even hinting at the existence of default stages (e.g., Shen et al., 2020). In fact, Wilcox (1971) states that any company can recover from default, given enough capital or equity is granted to it.

It is clear, then, that both *ex post* and *ex ante* models have their benefits and downsides. A possible way to overcome the uncertainty and bias of both types of criteria is by attaining the best of both methods and considering default based on both *ex post* empirical and *ex ante* theoretical criteria. This is achieved through mixed criteria models, which is likely why most traditional literature on the insolvency side of DPMs chose mixed over *ex ante*. Table 1 summarizes the findings in default definition found in the corresponding bibliography, categorized according to Figure 1.

Table 1 - Summary of Default Criteria Review

Type of Criteria	Research Found to Apply it
Ex post	Abinzano et al. (2020), Agarwal and Taffler (2008), Altman (1968, 1973, 1983), Altman et al. (2016), Blum (1974), Chava and Jarrow (2004), Correia et al. (2017), Duffie et al. (2007), Hillegeist et al. (2004), Jackson and Wood (2013), Mensah (1984), Ohlson (1980), Reisz and Perlich (2007), Shumway (2001), Wilcox (1973), Zmijewski (1984)
Ex ante	Balcaen and Ooghe (2006), Beaver (1968), Bonfim (2009), Bottazzi et al. (2011), Du et al. (2020), Edmister (1972), Shen et al. (2020), Costa et al. (2022)
Mixed	Beaver (1966), Beaver et al. (2010), Deakin (1972), Dichev (1998), Ramesh and Kumar (2017), Sun et al. (2014)

2.1.3. Default Prediction Methodologies

Balcaen and Ooghe (2006) describe the classical paradigm of failure prediction as being “about supervised classification” (p. 71). They follow the idea that this classification is done by analyzing a set of companies and constructing a rule (or a set of rules) based on them that can accurately classify a company outside the dataset as either defaulted or non-defaulted. Here rises the criticism pointed towards more traditional DPMs: the dichotomous, discrete, non-overlapping and identifiable nature of traditional DPMs doesn’t reflect the true reality of company default (Jackson & Wood, 2013). Regardless of this, traditional models are still utilized and empirically tested, given their offer of simplicity to researchers and practitioners (Costa et al., 2022).

A ‘traditional’ approach is discussed because the subject of corporate default prediction has abundant literature and has suffered an extensive study especially ever since the pioneering work of Beaver (1966) – that is, for more than half a century (Jackson & Wood, 2013; Kim et al, 2020; Sun et al., 2014, 2017). In fact, literature offers a vast array of modeling techniques (Alaka et al., 2018; Aziz & Dar, 2006; Balcaen & Ooghe, 2006; Dimitras et al., 1996; Jackson & Wood, 2013; Shen et al., 2020), oftentimes new models emerging at a relatively fast rate (Alaka et al., 2018). This section will go into a concise overview of the various modeling techniques found to be used in the reviewed literature and their classification/categorization.

Main Categorizations

This study will initially divide the models into three categories: theoretical, statistical, and intelligent. This first separation is directly on par with that of Aziz and Dar (2006) and Jackson and Wood (2013), which both study the various methodology types found in DPMs. Alaka et al. (2018) and Kumar and Ravi (2007) also divide their mentioned methodologies into statistical and intelligent, though not including any of the below-mentioned theoretical models. In addition to the main separation, a few more subdivisions were executed. Inside theoretical models, similarly to Jackson and Wood (2013), we congregate the most common structural models of this review into the contingent claims models category, although still addressing each separately. Likewise, linear probability, Logit, and probit models were all inserted into the category of conditional probability models inside statistical methodologies, as in Balcaen and Ooghe (2006).

Table 2 presents a summary of the findings in terms of model general methodology types, for which will be given a short explanation further on this sub-section. Any extra models mentioned in literature outside of the scope of what was found will be shortly mentioned as extra models for each of the three main categories.

Theoretical Models

Theoretical models focus on the qualitative causes of failure, usually satisfying a theoretical argument proposed by the default prediction theory (Aziz & Dar, 2006). They are multivariate in nature, and usually employ appropriate statistical techniques to estimate the model with some quantitative support (Aziz & Dar, 2006). Since they are based on a theoretical foundation, these models tend to escape the problem of statistical over-fitting on the data they are trained in (Balcaen & Ooghe, 2006). In the other hand, that same theoretical basis usually means they run on crude simplifying assumptions (Wilcox, 1971). Table 3 provides a summary of the theoretical model findings in this study's review.

Table 2 – Default Prediction Methodologies

Theoretical Models	Gambler's Ruin Model		Hybrid Models	Ensemble Models
	Contingent Claims Models	Black-Scholes-Merton		
		Barrier Option Model		
		KMV Model		
Statistical Models	Univariate Analysis			
	Multiple Discriminant Analysis			
	Conditional Probability Models	Linear Probability		
		Logistic Regression (Logit)		
		Probit		
	Survival Analysis			
	Intensity-based Models			
	Cumulative Sums			
	Partial Adjustment Processes			
	Risk Index Models			
Intelligent Models	Fuzzy Logic			
	Decision Trees			
	Support Vector Machines			
	Neural Network Models			
	Data Envelopment Analysis			
	Rough Sets			
	Case-based Reasoning			
	Evolutionary Algorithms			

Contingent Claims Models

Contingent claims models (also known as Structural Models, or Firm Value Models, as in Ramesh and Kumar, 2017) base themselves on the specification of company value and the determination of default boundaries (or barriers) (Bakshi et al., 2022; Ramesh & Kumar, 2017). More specifically, contingent claims models assume the company enters a state of default when its asset value drops to a given level relative to its liabilities (Bakshi et al., 2022; Duffie et al., 2007; Ramesh and Kumar, 2017). As such, it models default under a large influence of the capital structure (Bakshi et al., 2022). With the setting of a barrier, these models allow their consideration of default to occur prior to debt maturity, and, for their modeling, they require a set of parameters to be estimated (Bakshi et al., 2022). Contingent claims models have their foundation and transparency at their favor, given they anchor data on debt and equity and rely on already established methods to determine the asset value and other variables for the model (Bakshi et al., 2022). Nevertheless, their commonly assumed flat term structure of default-free interest rates and tendency to predict either too high or too low spreads when compared to statistical methods may put that advantage to question (Bakshi et al., 2022). The application of contingent claims models is more common in the prediction of credit risk (Aziz & Dar, 2006), but a few models were consistently employed in the reviewed theoretical literature, especially in the debate between accounting-based and market-based models (Abinzano et al., 2020; Agarwal & Taffler, 2008; Hillegeist et al., 2004; Jackson & Wood, 2013; Reisz & Perlich, 2007). These models include: (1) the Black-Scholes-Merton (BSM) models, originally based on the work of Black and Scholes (1973) and Merton (1974), as mentioned by Jackson and Wood (2013), and the foundation for the structural models that would follow (Bakshi et al., 2022), by applying option pricing theory to the valuation of risky bonds and loans (Bonfim, 2009); (2) Barrier Option models, where the barrier is often shifted from the value of the debt liabilities in question to prior values (Jackson and Wood, 2013); (3) the KMV framework, developed originally by the proprietors of the KMV Corporation, eventually acquired by Moody's Corp (Bharath & Shumway, 2008; Bottazzi et al., 2011; Jackson and Wood, 2013).

Other Theoretical Models

The only other theoretical model type found explained in this review was the Balance Sheet Decomposition (Aziz & Dar, 2006; Jackson & Wood, 2013). These models work under the assumption that a company will try to maintain financial structure equilibrium and that, if a company's statements reflect significant changes in the asset and liability composition, it likely cannot maintain that equilibrium and is prone to failure (Aziz & Dar, 2006).

Table 3 - Theoretical Methodologies

Model	Description	Advantages	Disadvantages	Sources[5]
Gambler's Ruin Theory	• **Result Type:** Duration / Stochastic • Sees default as a discrete process. • Assumes default as a one-dimensional random walk with an absorbing state at one end – the ultimate failure. • Each company has an initial amount of capital that decreases or increases depending on cash-flows, zero being the absorbing state.	• Simple. • Theory-based. • Easy to estimate.	• Based on simplistic unproven theory.	• Aziz and Dar (2006) • Wilcox (1971, 1973)
Black-Scholes-Merton (BSM)	• **Result Type:** Duration / Stochastic • The equity of a company is viewed as a European call option with strike price equal to the face value of its debt, default being assumed if it is left to expire.	• Simple. • Result transparency. • Applies classic finance theory. • Little estimation required.	• Easy to outperform. • Unrealistic Assumptions. • Performance may stem from its non-linear function and not from its theory basis. • Requires market-based variables	• Bakshi et al. (2022) • Barath and Shumway (2008) • Bonfim (2009) • Hillegeist et al. (2004) • Jackson and Wood (2013)
Barrier Option Models	• **Result Type:** Duration / Stochastic • Variation of the BSM models where the strike price is shifted to values prior to the face value of debt.			
KMV Models	• **Result Type:** Duration / Stochastic • Particular application of the BSM model developed by the proprietors of the KMV Corporation.	• Simple. • Result transparency. • Applies classic finance theory. • Little estimation required. • Estimates empirical asset distributions from historical databases instead of assuming a normal distribution. • Allows for various maturity debt classes.	• Easy to outperform. • Unrealistic Assumptions. • Performance may stem from its non-linear function and not from its theory basis. • Requires market-based variables • Difficult to replicate due to its mostly proprietary nature.	• Bakshi et al. (2022) • Barath and Shumway (2008) • Bonfim (2009) • Hillegeist et al. (2004) • Jackson and Wood (2013) • Bharath and Shumway (2008) • Bottazzi et al. (2011) • Jackson and Wood (2013)

---

[5] The sources column on the table reverts to the references from the bibliography from which the information about each model was obtained, and not necessarily to the studies that applied the given methodology.

Statistical Models

Statistical Models are, as suggested by the name, based on statistical theory (Sun et al., 2014). They focus, unlike theoretical models, on the symptoms of default and on how to identify them, with data drawn mainly from company accounts (Aziz & Dar, 2006). Additionally, they are the main category to house a considerable number of univariate models along with the more common multivariate (Aziz & Dar, 2006). In fact, for the reviews both of Aziz and Dar (2006) and Jackson and Wood (2013), statistical methods are the prevailing choice of study in literature, Aziz and Dar (2006) go further by explaining how all other emerging DPMs have some form of a statistical heritage. The most common methods in this category are the classic cross-sectional statistical methods that resulted in numerous models focusing on company default classification (Balcaen & Ooghe, 2006), among which Multiple Discriminant Analysis (MDA) and Logistic Regression (Logit) stand out by a considerable margin (Aziz and Dar, 2006; Jackson & Wood, 2013).

The appreciation in literature for the statistical models stems from their simple, easy-to-use, and time-saving characteristics, along with a considerably more theoretical basis than the intelligent models that will be presented below (Kumar & Ravi, 2007; Sun et al., 2014). These, however, are still not free of restrictive assumptions much like their theoretical counterpart, from ratio distributions to the relationships between the independent and dependent variables, the inability to meeting which can lead to performance and efficiency losses (Alaka et al., 2018; Aziz & Dar, 2006; Dimitras et al., 1996; Mensah, 1984). Another issue pertaining to this type of models is their classic dichotomizing approach to default, which does not reflect its reality, along with their retrospective pattern recognition nature rather than solely predictive (Balcaen & Ooghe, 2006). Furthermore, these models: (1) are commonly static and inconsistent along the span of years they may be estimated for (Balcaen & Ooghe, 2006; Shumway, 2001); (2) are sensitive to multicollinearity (Alaka et al., 2018); (3) are prone to statistical over-fitting of the data and consequently to sample bias (Altman, 1973; Balcaen & Ooghe, 2006); (4) still fail to theoretically fundament most of their variable selection choices (Wilcox, 1971).

Table 4 presents a summary of the main findings on statistical models from the literature review of this study.

Conditional Probability Models

The Conditional Probability category includes the linear probability, logistic regression, and probit models, all of which make use of the non-linear maximum likelihood method to estimate the probability of failure conditional on a range of company characteristics (Alaka et al., 2018; Balcaen & Ooghe, 2006; Dimitras et al., 1996). All of these models offer a binary response to default, to which Kim et al. (2020) points out two main advantages over MDA: (1) they do not require any assumptions on the probability of default or the predictor distributions; (2) they can test the significance of individual independent variables. In fact, Jackson and Wood (2013) state that it was the common violations of the statistic assumptions of MDA that lead researchers to concentrate efforts on developing the conditional probability models further. This type of model is still viewed as part of the classification approach, rather than prediction: companies are assigned to the group they most closely resemble (Balcaen & Ooghe, 2006).

Other Statistical Models

Other statistical model types with less of a presence yet still found explained in the review include:

1. Intensity-based Models (Bakshi et al., 2022; Ramesh & Kumar, 2017) – mostly the term-structure models borrowed from interest-rate modeling literature (Ramesh & Kumar, 2017), but also used in areas such as portfolio risk, credit risk, and correlated risk modeling (Giesecke et al., 2011).
2. Cumulative Sums (Aziz & Dar, 2006; Jackson & Wood, 2013) - a cumulative time-series performance score that detects shifts in the distribution from a company's statement to another; when the accumulated score goes below zero, it signals that there was a change in the firm's condition (Aziz & Dar, 2006)
3. Partial Adjustment Processes (statistical model) – based on a theoretic rationale of the famous Koyck approach to estimate distributed-lag models, connected to the insolvency definition of failure through cash-flows (Aziz & Dar, 2006; Jackson & Wood, 2013)
4. Risk Index Models (Balcaen & Ooghe, 2006; Jackson & Wood, 2013) – described as a simple point system, including various ratios; the more important a ratio, the more points it can allocate to a firm, though the weights of each ratio are subjective. (Balcaen & Ooghe, 2006)

Table 4 - Statistical Methodologies

<table>
<tr><th>Model</th><th>Description</th><th>Advantages</th><th>Disadvantages</th><th>Sources[6]</th></tr>
<tr><td>Univariate (Linear) Analysis</td><td>• **Result Type:** Score
• Traditionally focused on financial ratio analysis.
• Pioneered by Beaver (1966), his one of the first DPMs to be developed.
• Predicts default based on the results of one ratio only.</td><td>• Simple to understand.
• Extremely easy to estimate.</td><td>• Inconsistencies between ratios.
• Ignores possible multi-variate predictors.
• Assumes a linear relationship between ratios and default.
• Decision-makers avoid decisions based on solely one ratio</td><td>• Aziz and Dar (2006)
• Balcaen and Ooghe (2006)
• Beaver (1966, 1968)
• Costa et al. (2022)
• Dambolena and Khoury (1980)
• Dimitras et al. (1996)
• Edmister (1972)
• Jackson and Wood (2013)
• Sun et al. (2014)</td></tr>
<tr><td rowspan="2">Linear Multiple Discriminant Analysis (MDA)</td><td rowspan="4">• **Result Type:** Score
• The method creates a linear (quadratic) combination of variables that best differentiates between two groups (defaulted and non-defaulted), that is, that offers the minimum overlap between the two groups. This line is found through a differential calculus procedure.
• Through one of more cut-off points, it determines whether an observation belongs to one group or another (default or non-default).
• Widespread by the traditional Z-score of Altman (1968).</td><td>• Easy to estimate.
• Can produce exceptionally good results for how simple it is.</td><td>• Assumes equal variance-covariance matrices for all groups.</td><td rowspan="4">• Alaka et al. (2018)
• Altman (1968)
• Altman et al. (1977, 2016)
• Aziz and Dar (2006)
• Balcaen and Ooghe (2006)
• Blum (1974)
• Campbell et al. (2008)
• Deakin (1972)
• Dichev (1998)
• Dimitras et al. (1996)
• Edmister (1972)
• Jackson and Wood (2013)
• Ketz (1978)
• Kim et al. (2020)
• Kim et al. (2020)
• Mensah (1984)
• Ohlson (1980)
• Sun et al. (2014)</td></tr>
<tr><td rowspan="2">• Simple to apply.
• Can produce exceptionally good results for how simple it is.</td><td rowspan="2">• Classifies, does not predict.
• Assumes normally distributed variables.
• Requires numerical predictor variables.
• Assumes discrete, non-overlapping and identifiable groups.
• Assumes i.i.d. predictor variables.
• Coefficients do not represent the importance of each predictor like in linear regression.
• Presents high deterioration of coefficients when applied out-of-sample.
• Requires large training samples.
• Requires the balancing of training samples.</td></tr>
<tr><td rowspan="2">Quadratic Multiple Discriminant Analysis (QMDA)</td></tr>
<tr><td>• Reduces misclassifications of linear MDA.</td><td>• High complexity.
• Large number of parameters.</td></tr>
</table>

---

[6] The sources column on the table reverts to the references from the bibliography from which the information about each model was obtained, and not necessarily to the studies that applied the given methodology.

Model	Description	Advantages	Disadvantages	Sources[6]
Linear Probability	• **Result Type:** Score • Linear Probability Models express the probability of default of a company as a dichotomous (binary) dependent variable that results from a linear function of a vector of predictors.	• Extremely easy to estimate (simple regression)	• Requires linear relationship between default and its predictors. • Assumes normally distributed and homoscedastic regression residuals with expected value at zero. • Sensitive to outliers, missing values, and multicollinearity.	• Aziz and Dar (2006) • Balcaen and Ooghe (2006) • Dimitras et al. (1996)
Logistic Regression (Logit)	• **Result Type:** Stochastic (probability) • Expresses the probability of default as a binary variable through the logarithm of the odds that the default event will occur, working under the assumption of a logistic distribution. • Pioneered by the O-score of Ohlson (1980)	• Results can be interpreted as a probability of default. • Appropriate for logistic and other non-linear problems.	• Requires a relatively large training sample. • Assumes discrete, non-overlapping and identifiable groups. • Extremely sensitive to outliers, missing values, and multicollinearity.	• Alaka et al. (2018) • Altman et al. (2016) • Aziz and Dar (2006) • Balcaen and Ooghe (2006) • Campbell et al. (2008) • Costa et al. (2022) • Dichev (1998) • Dimitras et al. (1996) • Hillegeist et al. (2004) • Jackson and Wood (2013) • Ohlson (1980) • Sun et al. (2014)
Probit Regression	• **Result Type:** Stochastic (probability) • Similar to the applications of logit models, though assuming a cumulative normal distribution instead of a logistic distribution.	• Results can be interpreted as a probability of default. • Appropriate for non-linear problems.	• Requires a relatively large training sample. • Requires considerably more computational effort than logit. • Assumes normally distributed error rates.	• Aziz and Dar (2006) • Balcaen and Ooghe (2006) • Bonfim (2009) • Dimitras et al. (1996)
Survival Analysis (Hazard Models)	• **Result Type:** Duration / Stochastic • Corporate default risk is a function of the latest financial data of a company and its age, exploiting its time-series for a better prediction of volatility and deterioration. • Non-defaulted companies are considered censored observations, defaulted companies are analyzed for their survival time, and the default is treated as an unexpected event.	• Incorporate volatility measures. • Maintain consistency along different periods of the same company. • Solves static modeling issues by having the dependent variable be the time spent on the healthy state.	• Have non-linear likelihood functions and time-varying covariates that make them difficult to estimate.	• Beaver et al. (2005) • Bonfim (2009) • Cox (1972) • Dimitras et al. (1996) • Kim et al. (2020) • Ramesh and Kumar (2017) • Shumway (2001)

Intelligent Models

Intelligent Models[7] are computer-based techniques (Alaka et al., 2018; Aziz & Dar, 2006) that focus on the symptoms of default (Aziz & Dar, 2006). They are drawn mainly from company accounts, usually multivariate in nature and were bred from technological advancement and information development (Aziz & Dar, 2006). This type of model may be very well seen as an automated offspring of the classic statistical approach (Aziz & Dar, 2006), albeit taking advantage of machine learning capabilities to attain sophistication at the cutting edge of advanced financial engineering (Aziz & Dar, 2006; Kim et al., 2020). Intelligent models have been used since the late 1980s for the prediction of corporate default for their non-parametric capabilities (Du et al., 2020), though only gaining more traction roughly two decades ago (Sun et al., 2014). They have evolved to serve essentially the same function as human intelligence and reasoning, learning with previous experience to overcome problem-solving tasks, in this case, the prediction of default (Aziz & Dar, 2006). According to the reviewed literature, the more commonly found intelligent methodologies were Neural Networks (Alaka et al., 2018; Du et al., 2020; Kim et al., 2020), Support Vector Machines (Du et al., 2020; Kim et al., 2020), Decisions Trees (Du et al., 2020; Kim et al., 2020) and Case-based Reasoning approaches (Du et al., 2020).

The most common intelligent models seem to carry in prior literature two main negative characteristics that are positively related to their performance – training timespans and lack of transparency. In theory, for intelligent iteration-based methods (such as Neural Networks), model accuracy tends to 100% when the iteration number tends to infinity, along with computational time (Sun et al., 2017). In other words, the more time the model takes to train, the better its accuracy performance. As for transparency, prior literature models seem to either be excellent classifiers or excellent scrutineers of the causes of default, which displays a clear trade-off between accuracy and interpretability (Alaka et al., 2018; Costa et al., 2022).

---

[7] Term used by Du et al. (2020) and Sun et al. (2014). They are also known in literature as Expert (Aziz & Dar, 2006), Data-driven (Costa et al., 2022; Kumar & Ravi, 2007), Artificial Intelligence (AI) (Aziz & Dar, 2006; Kim et al., 2020) and/or Machine Learning (Costa et al., 2022; Du et al., 2020; Kim et al., 2020; Kumar & Ravi, 2007) models.

In the last decades, intelligent methods have attracted more and more attention, with clear general advantages over most traditional statistical methods (Du et al., 2020; Sun et al., 2014). Intelligent models: (1) make no assumptions on data distributions (Du et al., 2020; Kumar & Ravi, 2007); (2) are generally non-parametric (Costa et al., 2022); (3) are not subject to the stringent assumptions of statistical methods (Sun et al., 2014); (4) can work with large amounts of data with considerably less effort (Du et al., 2020); (5) appear to perform generally better compared to other methods (Kim et al., 2020; Kumar & Ravi, 2007; Shen et al., 2020); (6) can constantly and iteratively improve on their own (Sun et al., 2014); (7) are very resistant to multicollinearity, apart from the example of case-based reasoning (Alaka et al., 2018); and (8) are more flexible and allow integration with other tools more easily than statistical tools (Alaka et al., 2018). All of these advantages, however, do not come without consequences. For instance, this type of models: (1) may suffer the issue of low interpretability, as stated above, which may cause the usefulness of the model to diminish, if it is unable to identify causes of default to the user; (2) are prone to eventual data protection limitations, as law progresses with AI (Kim et al., 2020); (3) still take the classification instead of the prediction/process approach in most cases (Kim et al., 2020); (4) have high complexity (Sun et al., 2014); (5) have high training spans (Alaka et al., 2018). This last limitation is likely to come cheaper than the cost of a loan going bad or a bank failing, as O'Leary (1998) has stated in traditional literature. Table 5 summarizes the findings on intelligent models, with the respective sources.

Other Intelligent Models

Beyond those expressed in Table 5, two models for which were not found specific advantages/disadvantages are mentioned in the reviewed literature:

(1) Fuzzy Logic – model imprecision and ambiguity in the data using fuzzy sets, also incorporating human experience knowledge into the model; they generate useful 'if-then' rules (Kumar & Ravi, 2007).

(2) Data Envelopment Analysis – uses linear programming to rank various companies according to some input and output variables; the model is known for its invariance to the scaling of variables (Kumar & Ravi, 2007).

Table 5 - Intelligent Methodologies

Model	Description	Advantages	Disadvantages	Sources[8]
Neural Networks	• Created to imitate how the neural system of the human brain works. • Formed by layers of simple interconnected nodes that make a complex decision tree. • Go through an iterative process on a training algorithm.	• No specific functional form. • Flexible parameters and algorithms. • Strong network structure mapping ability. • Not affected by statistical assumptions.	• Results are difficult to interpret. • Weak under high levels of data noise. • Require large training samples. • Prone to over-fitting the data. • Produce different models on each training span. • Strong computational power required.	• Alaka et al. (2018) • Aziz and Dar (2006) • Costa et al. (2022) • Jackson and Wood (2013) • Kim et al. (2020) • Kumar and Ravi (2007) • Sun et al. (2014)
Support Vector Machines	• Find the optimal separating hyperplane using a highly non-linear mapping of vectors into a high-dimensional feature space. • Define a binary outcome utilizing the closest support vectors to that hyperplane.	• Handle smaller sample sizes. • High flexibility. • Avoid over-fitting.	• Results are difficult to interpret. • Less predictive capability when faced with noise or a lot of different characteristics. • Allows some misclassification to avoid overfitting.	• Alaka et al. (2018) • Costa et al. (2022) • Kim et al. (2020) • Sun et al. (2014)
Decision Trees	• Solves classifications problems by charting decision rules in a binary tree structure. • Has a set of pre-determined heuristics. • Complex sequence of dichotomous decisions or models.	• The generated 'if-then' rules work as default warning signs. • Intuitive and easy to interpret. • Do not require the setting of parameters before training.	• Require large training samples. • Difficult to compare evaluated companies. • Prone to over-fitting the data.	• Alaka et al. (2018) • Dimitras et al. (1996) • Kim et al. (2020) • Kumar and Ravi (2007) • Sun et al. (2014)
Case-based Reasoning	• Classifies a company based on similarities with other companies from a sample. • Justifies its decision based on cases from the sample.	• Handles smaller sample sizes. • Does not require training. • Easy to interpret.	• Cannot handle non-linear problems.	• Alaka et al. (2018) • Aziz and Dar (2006) • Sun et al. (2014)
Rough Set Theory	• Separates objects with similar attributes into elementary sets. • Creates a table of conditions and attributes, and classifies companies based on upper or lower approximation to them.	• Handles smaller sample sizes. • Easy to understand and interpret. • Made to encompass non-quantitative features.	• Difficult to implement. • Different for different samples and decision-maker knowledges. • Difficult to generalize.	• Alaka et al. (2018) • Aziz and Dar (2006) • Dimitras et al. (1996) • Kumar and Ravi (2007) • Sun et al. (2014)
Evolution Algorithms	• Generic population-based meta-heuristic algorithms inspired by biological evolution theory able to solve highly non-linear non-convex optimization problems.	• Can extract decision rules from data and define cut-offs for each variable. • Easy to interpret.	• Produce different models on each training span. • Yield the best results when hybridized.	• Alaka et al. (2018) • Sun et al. (2014)

---

[8] The sources column on the table reverts to the references from the bibliography from which the information about each model was obtained, and not necessarily to the studies that applied the given methodology.

Hybrid Models

Hybrid models are models that result from the combination of two or more methodologies (Jackson & Wood, 2013; Alaka et al., 2018; Shen et al., 2020), intelligent algorithms, or even other hybrid models (Sun et al., 2014). These became more popular in the late 2000s to 2010s, seemingly maintaining a good appeal in literature (Sun et al., 2014). Various authors incentivize this practice in the modeling of default prediction (Alaka et al., 2018; Jackson & Wood, 2013; Kumar & Ravi, 2007; Shen et al., 2020). Kumar and Ravi (2007), for instance, state how the clear ability for intelligent methodologies to perform better than the traditional models is in itself an incentive for the testing of augmentation on the traditional DPMs with the emerging AI methods. Jackson and Wood (2013) test this, finding that one can in fact augment the performance of both the Z-score (Altman, 1968) and the O-score (Ohlson, 1980) with the use of neural network algorithms. The benefits of creating hybrid models include both better performance than the standalone methods (Alaka et al., 2018; Kumar & Ravi, 2007) and the transcendence of method specific limitations and disadvantages (Kumar & Ravi, 2007; Sun et al., 2014). The big downside of developing hybrid models is the longer training timespans, given the added complexity (Alaka et al., 2018).

Ensemble Models

In this section, we introduce ensemble models, also known as only ‘ensembles’ (Du et al., 2020; Sun et al., 2014), as part of the hybrid models. As Du et al. (2020) mention, the idea of ensemble is meant to improve machine learning models by combining multiple methodologies, which in its core means hybridizing, but in a specific way. A technique is called an ensemble if the outputs of several predictors are combined to obtain a single one (Sun et al., 2014). This means the ensemble divides the initial problem into smaller sub-problems solved by separate algorithms, from which each output is combined to form a final decision or value (Du et al., 2020). This type of model has been widely used and usually shows better performance than the constituents of it (Du et al., 2020). Ensembles can be executed at various levels, some even utilizing specific sampling algorithms in order to feed newly generated samples into the model in training (Sun et al., 2014). Another example of this is the application of random forests, a specific ensemble algorithm that generates multiple decision trees and returns average results from the combination (Du et al., 2020;

Kim et al., 2020). Some of the benefits specific to ensembles and not common in all hybrids are, as in Du et al. (2020), the fact that multiple runs of an algorithm can help perfect it and escape local minima, and that it may give rise to new hypothesis spaces researchers were not aware existed.

2.1.4. Sampling of Default

Almost all of the methods of corporate default prediction are capable of providing consistent accuracy rates using any data set, provided that data was drawn reliably (Aziz & Dar, 2006). Hence, it is important for DPMs to be trained with samples representative of the population (Balcaen & Ooghe, 2006).

Data Balancing

In the overall corporate population, default is a rare event (Bakshi et al., 2022; Beaver, 1968; Du et al., 2020; Duffie et al., 2007; Hillegeist et al., 2004; Shumway, 2001), which makes estimating it still a challenge (Bakshi et al., 2022). In other words, default samples are highly imbalanced (Kim et al., 2020; Sun et al., 2014). Imbalanced samples occur when the data inside a data set differs greatly in terms of distribution and/or sub-sample size (Shen et al., 2020). Shen et al. (2020) denominate this default phenomenon as holistic imbalance, distinguishing it from another type of imbalance – internal sample category imbalance – which portrays the possibility of different company categories (such as industries) presenting different default rates[9] inside the sample. Analyzing samples by industry separately avoids this last type of imbalance (Costa et al., 2022). As for the first type of imbalanced data, there are a few issues relating to its use: (1) some modeling methodologies, such as discriminant analysis, are sensitive to unequal data dispersion (Alaka et al., 2018); (2) non-default data may overwhelm default data, making it difficult for the model to learn how to distinguish between the two (O'Leary, 1998; Shen et al., 2020); (3) if default data has too few observations, it may not allow the model to fully learn the right characteristics of default, especially if that data is prone to noise (Shen et al., 2020).

---

[9] The default rate is the proportion of defaulted companies in the total number of companies in the sample.

There are two main ways to treat imbalanced samples: one can either do it at the data level, or the algorithm level (Du et al., 2020). On the data level, a researcher focuses on manipulating the original sample to obtain a better distribution of characteristics (Du et al., 2020; Shen et al., 2020). This can be achieved through over-sampling (increasing the number of minor class sub-samples in the data), under-sampling (decreasing the number of major class sub-samples in the data), or hybrid/mixed sampling (a combination of both) (Shen et al., 2020). The algorithm level of treating data imbalance focuses on improving the existing classification algorithm by setting different weights for different kinds of sub-samples (Du et al., 2020; Shen et al., 2020). Data level treatment of imbalanced data was commonly performed in literature in the form of matched paired sampling, which consists in creating a sample with the same amount of defaulted and non-defaulted observations by matching each non-defaulted(defaulted) observation with a defaulted(non-defaulted) one with other similar characteristics (Altman, 1968; Altman et al., 1977; Beaver, 1966, 1968; Dambolena & Khoury, 1980; Du et al., 2020; Jackson & Wood, 2013; Sun et al., 2017; Wilcox, 1973; Zmijewski, 1984). This matching was commonly made by characteristics such as age, size, and/or industry (Altman et al., 1977; Balcaen & Ooghe, 2006; Mensah, 1984; Sun et al., 2017; Wilcox, 1973).

Data balancing does not come free of issues, however. Training models in samples with higher default rates than the population may lead to choice-based sample bias and default upward estimation (Balcaen & Ooghe, 2006; Hillegeist et al., 2004; Stein, 2007; Zmijewski, 1984). If said models are tested in the unrealistically balanced sample, the evaluation may overstate the *ex post* accuracy of the model and invalidate its application (Balcaen & Ooghe, 2006; Du et al., 2020). Even if researchers evaluate the balanced trained model with an imbalanced sample, a loss in performance is still expected (Shen et al., 2020). Regardless, it is important to note that when Zmijewski (1984) tested and found bias in the balancing of samples, he did not find significant effects of this bias in model default estimations. Zmijewski (1984) presents some methods to avoid this bias, but, given the results, it remains unclear whether they are worth the extra effort.

Data Completion

The complete data criterion is a common criterion for sampling decisions in literature which limits the entry of observations into the data sample according to the availability of information the researcher deems necessary for the estimation of the DPM – it is directly related to the existence and treatment of missing values (Blum, 1974; Ketz, 1978; Kwon & Lee, 2018; Ohlson, 1980; Shen et al., 2020; Zmijewski, 1984). The usage of this criterion is prone to the sample selection bias, by reducing the original sample to one which can be applied in the specific DPM, therefore making the DPM only applicable to that type of sample; this may occur in terms of years of data available, company types, variables available, etc. (Balcaen & Ooghe, 2006; Beaver, 1966; Bottazzi et al., 2011; Zmijewski, 1984). To avoid this bias, there are a few methods for dealing with missing values used in literature besides the more common exclusion of observations: (1) replacing missing values with random values, which additionally helps train the model against noise (Balcaen & Ooghe, 2006); (2) replacing missing values with mean values for the corresponding variable/indicator (Bonfim, 2009; Du et al., 2020); (3) replacing missing values with the nearest value (e.g. previous month, previous year, etc.) (Kwon & Lee, 2018; Nguyen, 2023); (4) keeping the missing values as simply missing values (Kwon & Lee, 2018; Nguyen, 2023).

Outlier Treatment

When dealing with outliers in DPM independent variables, it is very uncommon for researchers to mention their outlier treatment at all (Adams et al., 2019; Mitton, 2021). When outlier treatment is mentioned, a few methods are found, methods which this study orders (by descending presence in literature) below:

1. **Winsorization** – the most common type of outlier treatment in corporate default, which consists of replacing the values considered outliers for the value of given percentiles closest to them (Adams et al., 2019; Mitton, 2021). Winsorization is commonly applied as the substitution of all values under the 1st percentile for that percentile and all values over the 99th percentile for that percentile as well (Altman et al., 2016; Beaver et al., 2005; Bharath & Shumway, 2008; Bonfim, 2009; Chava & Jarrow, 2004; Correia et al., 2017; Dichev, 1998; Mitton, 2021; Shumway, 2001). Other applications include the replacement of those percentiles for the 5th and the 95th instead (Campbell et al., 2008).
2. **Retaining** – the most common method after winsorization (Mitton, 2021), which consists in maintaining the outliers for estimation, in some cases being applied as a

sort of robustness test on the treatment of outliers (Abinzano et al., 2020; Chava & Jarrow, 2004; Shumway, 2001).

(3) **Trimming** (or **dropping**) – consists in removing observations which present outliers in any of their variables, ignoring the other useful information they may harbor (Adams et al., 2019; Sun et al., 2017).

There are some additional methods that can help reduce outliers, such as the transformation of variables through normalization, or industry averaging (Alaka et al., 2018), or even into binary variables that inform the DPMs of when certain variables breach pre-determined thresholds (Adams et al., 2019; Edmister, 1972). Variable normalization can be achieved through logarithmic, square root or lognormal transformations (Balcaen & Ooghe, 2006). Whether outliers should be treated or not is not commonly discussed (Mitton, 2021). Outliers can be the most important data in a sample, since they reflect unusual facts or events that may raise suspicion in terms of default (Adams et al., 2019). However, when theory does suggest a certain effect, the presence of outliers should not be able to shake hypothesis down (Adams et al., 2019). One must keep in mind, however, that certain treatments, such as winsorization, can have a disruptive effect on variables as if an entirely new independent variable was generated (Mitton, 2021).

2.1.5. Model Evaluation

Corporate default model evaluation literature can be divided into two main aspects of the predictive ability of a model: power and calibration (Agarwal & Taffler, 2008; Reisz & Perlich, 2007; Stein, 2007). Power describes how well a model discriminates between a defaulting company and a non-defaulting company (Reisz & Perlich, 2007; Stein, 2007). Calibration describes how well the predicted probabilities of a model agree with the empirical data (Stein, 2007). Some authors also mention the concept of goodness-to-fit in terms of model performance, where the methods used are usually associated with calibration (Balcaen & Ooghe, 2006; Chava & Jarrow, 2004). According to Stein (2007), predictive power should be the focus before calibration, given that it is generally far easier to calibrate a powerful model to true default rates than it is to make a well calibrated model more powerful. Sun et al. (2014) point out the importance of caution when assessing model accuracy through calibration on imbalanced data, given that, in a set with only 1% default rate, a model that assumes all companies non-defaulting can obtain a 99% accuracy rate. Additionally, poor calibration does not necessarily mean that a model is not able to carry information about the true probability of default (Hillegeist et al., 2004). On the subject of data balance and evaluation, Stein (2007) mentions that what can cause the most variance in performance results is the default rate of a sample, hence making it an interesting idea to perform tests on various samples with different default rates. Besides the general performance analysis, literature also commonly performed so-called robustness tests, consisting in making slight modifications to model or the evaluation methodology and evaluating the impact those changes have on the overall model performance (Blum, 1974; Costa et al., 2022; Hillegeist et al., 2004; Mitton, 2021; Shen et al., 2020; Tafakori et al., 2021). With this introduction, this sub-section will go over the various types of evaluation methods for DPMs found in the reviewed literature. Table 6 summarizes the recurrent evaluation methodologies found by this study.

Table 6 - Evaluation Methodologies Summary

Methodologies			Sources
Performance	General Simplistic Performance Measures	Pearson Correlation	Hillegeist et al. (2004)
		Spearman Correlation	Hillegeist et al. (2004)
		Decile Analysis	Beaver et al. (2005), Bharath and Shumway (2008), Shumway (2001)
	Predictive (Discriminatory) Power Evaluation	Contingency Tables (Confusion Matrices)	Altman (1968), Altman (1973), Beaver (1968), Bonfim (2009), Du et al. (2020), Jackson and Wood (2013), Ketz (1978), Stein (2007)
		Cumulative Accuracy Profile (CAP) Plots	Abinzano et al. (2020), Duffie et al. (2007), Nguyen (2023), Stein (2007)
		Receiver Operating Characteristic (ROC) Curves	Agarwal and Taffler (2008), Altman et al. (2016), Chava and Jarrow (2004), Du et al. (2020), Jackson and Wood (2013), Reisz and Perlich (2007), Shen et al. (2020)
		F-ratio / F-measure[10]	Altman (1973), Blum (1974), Deakin (1972), Shen et al. (2020), Xu et al., (2021)
		ACC Measure	Du et al. (2020)
		G-measure	Shen et al. (2020)
	Calibration	Likelihood Ratios	Beaver (1968), Bonfim (2009), Hillegeist et al. (2004), Reisz and Perlich (2007), Tafakori et al. (2021)
		R-squared	Costa et al. (2022), Ketz (1978)
		Pseudo R-squared	Bonfim (2009), Campbell et al. (2008), Costa et al. (2022), Hillegeist et al. (2004), McFadden (1973), Mensah (1984)
		Calibration Graph	Reisz and Perlich (2007)

---

[10] Largely inconsistent between authors and hence not further developed in this section.

Methodologies		Sources
Robustness Checks	Model Variation	Blum (1974), Costa et al. (2022), Hillegeist et al. (2004), Mitton (2021), Shen et al. (2020), Tafakori et al. (2021), Zhang (2022)
	Relative Information Content Tests	Agarwal and Taffler (2008), Hillegeist et al. (2004)
	Noise Insertion	Balcaen and Ooghe (2006), Stein (2007)
	Sample Sensitivity Analysis	Chava and Jarrow (2004), Stein (2007)

Performance Analysis

There are a few simplistic approaches to the evaluation of the performance of a DPM a researcher can perform ahead of more complex methods. Hillegeist et al. (2004), for instance, calculated the simple Pearson and Spearman correlations between the score results of their model and the empirical bankruptcy binary variable, as a simplistic and initial measure of performance (the higher the correlation, the better the performance). Studies like Shumway (2001), Beaver et al. (2005), and Bharath and Shumway (2008) performed a little more complex evaluations, placing each company on a different decile of their probabilistic model results and accounting for the percentage of defaults that were considered by the model in the worst decile(s) as a measure of performance.

Discriminatory / Predictive Power Analysis

Contingency Tables

One of the early methods used to test the predictive power of DPMs was contingency tables (also known as confusion matrices) (Altman, 1968, 1973; Beaver, 1968; Bonfim, 2009; Du et al., 2020; Jackson & Wood, 2013; Ketz, 1978; Stein, 2007). The concept requires the selection of an optimal cutoff point of the model score or probability result, which commonly minimizes the number of misclassifications (Altman, 1968, 1973; Beaver, 1968; Bottazzi et al., 2011). From it, a table is built which describes the misclassifications and the correct classifications of the model by group (default and non-default), as in Table 7.

Table 7 - Contingency Table Example

Contingency Table		Empirical Observation	
		Defaulted	Non-Defaulted
Model Predictions	Defaulted	TP	FP
	Non-Defaulted	FN	TN

A researcher would ideally intend for the forecasts of the model to fall on the diagonal of the contingency table, where true positives and true negatives lie (Beaver, 1968). True positives (TP) and true negatives (TN) occur when the model correctly classifies a defaulting and non-defaulting company, respectively (Altman, 1968; Beaver, 1968; Stein, 2007). Likewise, false positives (FP) and false negatives (FN) occur when the model misclassifies a company as respectively defaulting and non-defaulting (Altman, 1968; Beaver, 1968; Stein, 2007). A few ratios that can be calculated from this type of table are expressed in Table 8.

The distinction of groups based on a cut-off and the construction of contingency tables also produces two types of errors for a model when evaluating a company – type I errors and type II errors (Balcaen & Ooghe, 2006; Beaver, 1966). Type I errors refer to the false negatives (FN), when a model misclassifies a defaulting company as healthy, while type II errors refer to the false positives (FP), when a model misclassifies a healthy company as defaulting (Alaka et al., 2018; Balcaen & Ooghe, 2006; Jackson & Wood, 2013). There is a strong consensus on that type I errors are much more costly than type II errors, since they are associated to distress and/or bankruptcy costs (Agarwal & Taffler, 2008; Alaka et al., 2018; Altman, 1973; Beaver, 1966, 1968; Beaver et al., 2010; Bottazzi et al., 2011; O'Leary, 1998). The existence of uneven costs has many authors insisting on their consideration when modeling default (Alaka et al., 2018; Balcaen & Ooghe, 2006; Beaver, 1968; Ohlson, 1980; Reisz & Perlich, 2007; Shen et al., 2020), hence the avoidance of the use of contingency tables and other simplistic methods alike (Balcaen & Ooghe, 2006; Ohlson, 1980; Reisz & Perlich, 2007). To avoid it, multi-cutoff methods, the designation of grey areas or the evaluation with marginal costs in mind are adopted while evaluating methodologies (Edmister, 1972; Shen et al., 2020).

Table 8 - Hit/Miss Ratios

Ratio	Formulas		Sources
True Positive Rate (Sensitivity / Hit Rate / TPR)	$TPR = \frac{TP}{TP + FN}$	$TPR = 1 - FNR$	Du et al. (2020), Jackson and Wood (2013), Reisz and Perlich (2007), Stein (2007)
True Negative Rate (Specificity / TNR)	$TNR = \frac{TN}{TN + FP}$	$TNR = 1 - FPR$	
False Positive Rate (False Alarm Rate / FPR)	$FPR = \frac{FP}{TN + FP}$	$FPR = 1 - TNR$	
False Negative Rate (FNR)	$FNR = \frac{FN}{TP + FN}$	$FNR = 1 - TPR$	

Stein (2007) argues that, ideally, the researcher will want to minimize the FP and FN or maximize the TPR to 1. There are a few limitations to this approach: (1) it limits the analysis to a single cutoff point (Altman, 1973); (2) a single optimal cutoff point may not necessarily exist (several cutoffs may be optimal) (Blum, 1974); (3) the analysis is based on a dichotomous choice, while the default-related decision making is typically a more continuous decision choice (Hillegeist et al., 2004; Reisz & Perlich, 2007); (4) assumes equal misclassification costs (Hillegeist et al., 2004); (5) error minimizing cutoff selection may result in misleading estimates of model accuracy (Jackson & Wood, 2013); and (6) there is no consensus as to how one should select the optimal cutoff point for analysis.

Power Curves

Power curves fix the issue of single cutoff points for contingency tables – they allow the researcher to visualize and further analyze results for multiple cutoffs at once (Jackson & Wood, 2013; Stein, 2007). Two types of power curves can be extracted from literature – cumulative accuracy profile (CAP) plots and receiver operating characteristic (ROC) curves. Figure 2 offers an example of both a CAP and ROC curves. ROC curves appear the most common in this review, yet their functionality is relatively similar.

CAP curves are created by plotting, for each rating category, the proportion of defaults classified in the same or lower rating (in the y-axis) against the proportion of companies classified in the same or lower rating (in the x-axis) (Abinzano et al., 2020; Stein, 2007). The rating categories / cutoffs must be ordered from ‘worst’ to ‘best’, with the same applying for ROC curves (Stein, 2007). CAP curves display the proportion of defaults taken from a sample for every percentage of the worst classified companies one would remove, considering the estimations of the model (Stein, 2007).

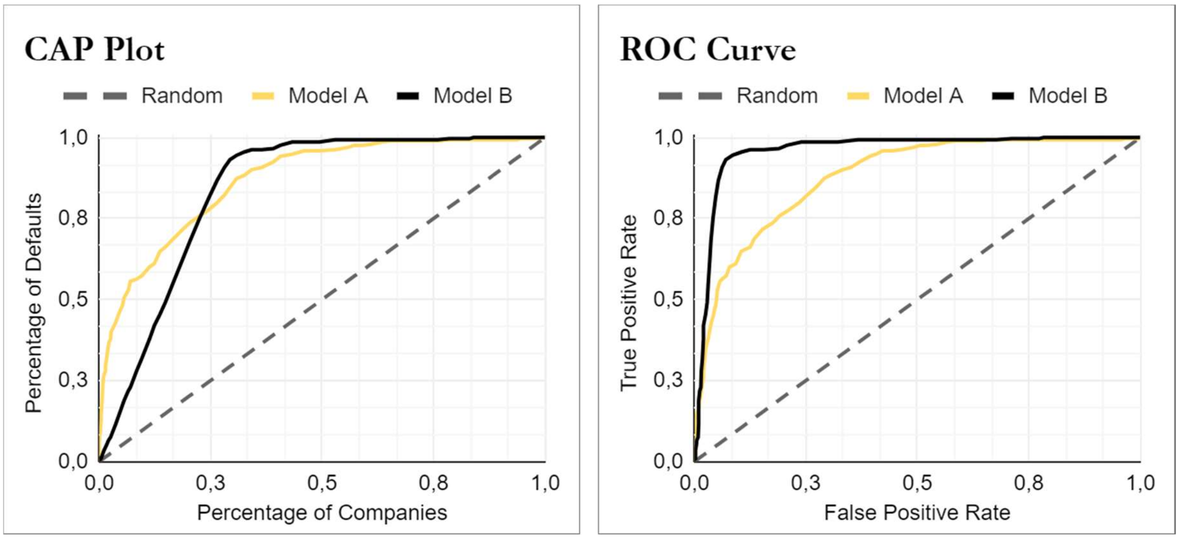


Figure 2 - Power Curve Examples[11]

ROC curves are created by plotting, for each cutoff point, the TPR (on the y-axis) against the FPR (on the x-axis) (Du et al., 2020; Jackson & Wood, 2013; Reisz & Perlich, 2007; Stein, 2007). Overall, ROC curves display the proportion of defaults correctly classified for every amount of non-defaults incorrectly considered defaulting, based on the model estimated (Stein, 2007). Additionally, Stein (2007) argues that the slope of the ROC curve at each point can also be considered a likelihood ratio of the probability of default to non-default for that specific model score.

Comparability between models using the CAP and ROC curves is relatively straightforward. If the curve of a model is completely wrapped on the top left by the curve of another model (as in the ROC curve of Figure 2), then it can be concluded that the model with the wrapping curve has the best predictive power (Model B being the best, according to the ROC curve in Figure 2) (Du et al., 2020; Stein, 2007). If the two curves intersect (as in the CAP curve of Figure 2), it becomes more difficult to assert which of the models is best in general (Du et al., 2020), the interpretation being that the model with a higher curve on lower cutoffs may be favorable for identifying default among worst companies, while the other would be better to identify default on medium-high quality companies (Stein, 2007).

Derived from the power curves, researchers commonly calculate the area under the curve (AUC), as a measure of power performance (Abinzano et al., 2020; Agarwal & Taffler, 2008;

---

[11] Elaborated by the author, based on the graph results displayed by Stein (2007), Reisz and Perlich (2007), Jackson and Wood (2013), Abinzano et al. (2020), and Du et al. (2020). The examples in question have no relation whatsoever to the results of this study and were fabricated purely for the purpose of visualization.

Altman et al., 2016; Du et al. 2020; Duffie et al., 2007; Jackson & Wood, 2013; Reisz & Perlich, 2007). For a ROC curve, it represents the probability of the predicted score of a randomly chosen defaulting company being greater than the predicted score of a randomly chosen non-defaulting company (Reisz & Perlich, 2007). Agarwal and Taffler (2008) utilize the Wilcoxon statistic to estimate the AUC, though, given the curve is discretely built, Jackson and Wood (2013) utilize the simpler trapezoid rule to estimate it, and Du et al. (2020) follow an equivalent approach by utilizing the following formula:

$$AUC = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1} - x_i) * (y_i + y_{i+1}) \qquad (1)$$

where the curve is formed by a sequential connection of points whose coordinates are $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ (Du et al., 2020). From the AUC, an accuracy ratio (AR) is commonly calculated, by subtracting the random model AUC from the estimated model AUC and multiplying the result by 2 (formula (2)) (Abinzano et al., 2020; Agarwal & Taffler, 2008; Altman et al., 2016; Shen et al., 2020). The AR is a ratio that varies between 0 and 1, 0 pertaining to a model no better than a random one and 1 pertaining to the power perfect model (Abinzano et al., 2020; Altman et al., 2016; Jackson & Wood, 2013).

$$AR = 2 * (AUC - 0.5) \qquad (2)$$

Edmister (1972) states that, by reporting the classification results for all cutoff values of a model, one overcomes the specification of prior probabilities and misclassification costs, the same being traditionally achieved through the creation of the ‘grey area’ for company classification (as was done originally by Altman, 1968). However, in order to compare two models, their outputs need to have the same ranges, making it difficult for score outputting models to be compared to stochastic/probabilistic outputting models using power curves (Agarwal & Taffler, 2008; Hillegeist et al., 2004; Jackson & Wood, 2013; Reisz & Perlich, 2007). For this reason, the scores of models, like the Z-score of Altman (1968), are commonly transformed into probabilities utilizing the logistic transformation method (Agarwal & Taffler, 2008; Correia et al., 2017; Hillegeist et al., 2004; Jackson & Wood, 2013; Reisz & Perlich, 2007). For models that output a score of financial quality instead of a measure of default, that is, that output a score that is negatively related to default, the researcher must take care to invert the output scores of the model before transforming it (Hillegeist et al., 2004). In linear functions, such as the Z-score of Altman (1968), inverting the signs of the original model parameters should suffice (Hillegeist et al., 2004). The logistic transformation method is as follows:

$$Probability\ of\ Default = \frac{e^{score}}{1 + e^{score}} \tag{3}$$

One must keep in mind that this transformation is not flawless and is not strictly correct for score outputting models (Hillegeist et al., 2004; Jackson & Wood, 2013). The inverse transformation was also applied to a probabilistic model in literature by Hillegeist et al. (2004), though this led to the issue of more outlier generation.

Other Measures

Few other simple measures were found that can be included in power analysis, though with small presence in the reviewed literature. Table 9 presents these extra models, a short explanation on them, their calculation, and respective sources.

Table 9 - Extra Model Power Measures

Measure	Interpretation	Calculation	Sources
ACC Measure	Accuracy (ACC) measure of the probability of correctly classifying the sample into financial distress.	$ACC = \frac{TP + TN}{TP + TN + FP + FN}$	Du et al. (2020) Xu et al. (2021)
G-measure[12]	Geometric mean of the TPR and TNR. Higher values mean higher predictive power.	$G = \sqrt{TPR * TNR}$	Shen et al. (2020)

Calibration Analysis

Most of the measures of calibration for DPMs in literature are likelihood-based (Bonfim, 2009; Campbell et al., 2008; Reisz & Perlich, 2007; Stein, 2007). Normally, either likelihood itself is used (Reisz & Perlich, 2007), or a type or ratio or test is created (Bonfim, 2009; Campbell et al., 2008; Ohlson, 1980). The issue with likelihood is that it may only be

---

[12] The TPR and TNR in the formula refer to the True Positive Rate and the True Negative Rate presented in **Table 8**, respectively.

calculated for stochastic models (those that output probabilities as their default risk estimate) (Reisz & Perlich, 2007). Typically, given it eases computation, the log-likelihood is utilized instead of the direct likelihood calculation (Stein, 2007). Stein (2007) presents a generalized formula for default model log-likelihood calculation:

$$\ell(model) = \sum_{i=1}^{n} y_i * \ln[p(x_i)] + (1 - y_i) * \ln\,[1 - p(x_i)] \qquad (4)$$

where $p(x_i)$ is the probability of default predicted by the model for observation/company i and y_i is 1 if company i is defaulting and 0 if not (Stein, 2007). A better model is one with the largest total (log-)likelihood (Reisz & Perlich, 2007). From the log-likelihood of a model, researchers commonly derive a pseudo-R^2, originally proposed by McFadden (1973) for the logistic regression (Bonfim, 2009; Campbell et al., 2008; Ohlson, 1980). Traditional regression already outputs the R^2 (Ketz, 1978; Costa et al., 2022), and this is a measure of similar interpretation – the higher the value of the pseudo-R^2, the higher the goodness-to-fit of the model to the data (Bonfim, 2009; McFadden, 1973). The formula of the pseudo-R^2 is as follows:

$$Pseudo - R^2 = 1 - \frac{\ell(model)}{\ell(0)} \qquad (5)$$

where $\ell(model)$ is the log-likelihood of the estimated model and $\ell(0)$ is the log-likelihood of the model but assuming the vector of variables to be 0, that is, assuming the model is equal to its constant component only (Bonfim, 2009; Campbell et al., 2008; McFadden, 1973; Ohlson, 1980).

Robustness Checks

Robustness checks usually assess the resistance of a model to change, either in data, model components or evaluation methodologies, to determine whether model performance was due to any related conditions (Costa et al., 2022; Mitton, 2021; Stein, 2007; Tafakori et al., 2021). Mitton (2021) proposes what he denominates of ‘specification checks’, which consist in remaking a model with every possible combination of methods and creating a result distribution. The random forest method for decision trees is a good example of how this can be applied (Costa et al., 2022). The same can be done at the sample level, by varying the training and/or evaluation sample for the model (Beaver et al., 2010; Jackson & Wood, 2013). A good model would maintain its predictive quality stable and with the least spread in distribution along the changes (Beaver et al., 2010; Jackson & Wood, 2013; Ohlson, 1980). Zhang (2022) goes a step further by stating that the best way to test the robustness of a model is by multiple repeated experiments, multi-sample testing and multiple party testing, in the sense that various studies with different authors should be made on it. All of this, however, can be extremely time-consuming.

Similar to the model variation mentioned, some authors mention the gradual corruption of the samples used in the model with more and more noise and the registering of how its performance is affected (Bakshi et al., 2022; Stein, 2007). Bakshi et al. (2022) argue that models must account for the presence of possible noise in the estimation of parameters.

Another type of test commonly applied to DPMs is the relative information content test, which examines whether a variable (or a set of them) provides more information compared to the others (Hillegeist et al., 2004). These tests, however, require the estimation of an extra Hazard Model (Agarwal & Taffler, 2008; Hillegeist et al., 2004).

Validation Sample

This study already discussed the existence of a training sample (a sample where the model is trained) and a validation sample (a sample used to validate and evaluate the model), especially on the subject of intelligent models. The two samples do not necessarily have to be different, yet in-sample tests (testing the model on the same sample it was trained on) are prone to the model overfitting the noise in the data and giving estimates with biased significant results (Chava & Jarrow, 2004). Given most models have a statistical basis, which leads them to develop a certain specificity towards training samples, out-of-sample testing (testing the model with a different sample from the training one) is very important to avoid over-fitting bias and provide a more challenging test to the predictive ability of the model (Altman, 1973; Balcaen & Ooghe, 2006; Beaver et al., 2005, 2010; Edmister, 1972). In this sense, there are a few validation sample related methods found in literature:

(1) **Hold-out Sample** – splitting a total sample randomly in two to originate both the training sample and the validation sample (denominated of hold-out sample) from the

same origin (Altman et al., 1977; Blum, 1974; Edmister, 1972; Jackson & Wood, 2013; Sun et al., 2014). Du et al. (2020) also denominate their approach of hold-out sample, yet they only take 30% of the original total sample.

(2) **Leave-one-out** (or Lachenbruch's U-method) – a method that involves holding out one observation from the sample, estimating a model with the rest of the observations and then evaluating the results on that single observation, and repeating this for every single observation in the original total sample (Dambolena & Khoury, 1980; Mensah, 1984; Sun et al., 2014). The big issue with this method is the higher computational power required, as well as it may be very difficult or time-consuming to apply to more complex models (Sun et al., 2014).

(3) **K-fold Cross-validation** – consists in splitting the total sample randomly into k smaller equally-sized samples (denominated folds) and taking one out to serve as the validation sample, with the rest serving as training (Jackson & Wood, 2013; Sun et al., 2014; Xu et al., 2021). The most interesting point of this method is that it can technically have the same result as any of the aforementioned, since it equals the hold-out method for $k = 2$ and the leave-one-out method for $k = N$, with N being the number of observations in the total sample (Sun et al., 2014).

Beyond the usage of these methods, various authors also incentivize the use of out-of-time samples, meaning the validation of models with tendentially more recent data (Altman et al., 1977; Balcaen & Ooghe, 2006; Mensah, 1984; Stein, 2007). Out-of-time sample testing helps to understand how the model is affected by what has more recently been called 'default concept drift' (Shen et al., 2020; Sun et al., 2017). This idea refers to changes in the distribution or the relationship between default and its features (Shen et al., 2020) and has been the focus of some intelligent models in literature, through time windowing algorithms (Sun et al., 2017).

Edmister (1972) utilizes a relatively straight-forward method for calculating the sample bias of a model estimation, a procedure which creates a so-called 'scrambled' sample. The idea is to reorganize the default variable among the observations randomly, re-estimate the model on that scrambled sample and evaluate the power of the model (Edmister, 1972). If the model power is significantly different from that of a random model, then the estimation is biased to the sample (Edmister, 1972).

Statistical Significance

Given, once again, the statistical basis of most models, researchers typically test for model parameter and result significance (Beaver et al., 2005; Bonfim, 2009; Campbell et al., 2008; Hillegeist et al., 2004; Jackson & Wood, 2013; O'Leary, 1998; Zhang, 2022). Mitton (2021) and Zhang (2022) both argue that statistical significance is often overrepresented and overvalued in literature. Zhang (2022) argues that scientific research can only be falsified, not confirmed, and therefore models should be put to test under more conditions and situations, instead of attempting to achieve statistical significance. He emphasizes, following Benjamin et al. (2018), that a stricter p-value should be established by researchers when evaluating statistical significance, so as to avoid p-hacking – the idea of making modeling and evaluation choices so as to obtain significant results according to the p-value (Mitton, 2021; Zhang, 2022).

Important Notes on Model Evaluation

(1) **What defines a good DPM?** Aziz and Dar (2006) defend that the usefulness of a DPM is ultimately an empirical matter. Ketz (1978), Bakshi et al. (2022) and Zhang (2022) defend the scientific approach of modeling corporate default prediction, as in taking DPMs through various types of empirical testing and data. Bonfim (2009) states that a good DPM is one that, with the least set of parameters and features, allows for a comprehensive explanation of default to the point of accurate results.

(2) **What amount of performance is better performance?** Chava and Jarrow (2004), Stein (2005), and Agarwal and Taffler (2008) defend how small differences in performance when comparing models can in fact have significant economic impacts for the user.

(3) **Performance depreciation:** an efficient well-performing model will eventually fall victim of its own success, since the market will begin using it to detect default, which will in turn reduce the defaults in the market to only those the model cannot predict (Jackson & Wood, 2013). Likewise, the use of a good model to predict default may also lead companies to default, given the withdrawal of stakeholders from companies marked as so by the model, and hence perpetuating those characteristics as default characteristics (Jackson & Wood, 2013).

2.2. Simulation Application

The simulation section will develop on an introductory view of the applications that simulation (more specifically Monte Carlo and its variants, as well as the Bootstrap) has currently in the corporate finance area of literature.

2.2.1. Monte Carlo Simulation

The Monte Carlo (MC) method is a powerful technique for performing calculations that may become too complex for a more traditional approach (James, 1980). In general, it can be any technique that makes use of random numbers to approximate the solution for a given problem (James, 1980; Kwak & Ingall, 2007). It has been in literature from at least since the 1930's (Andrieu et al., 2003), the first ever document on its methodology being published in 1949 (Andrieu et al., 2003; Metropolis & Ulam, 1949). Its name stems from the famous Monte Carlo gambling casino (James, 1980; Kwak & Ingall, 2007). More than the method itself, are known various of its algorithms, that work as its core (Andrieu et al., 2003).

Let it be considered that a model or function for a given system or situation is mathematically constructed, one which depends on a set of variables that can be represented each by a probability function (James, 1980; Kwak & Ingall, 2007). The principle of MC simulation is to calculate a large number of independent and independently distributed (i.i.d.) variable input scenarios for that model or function and approximate the solution through the sample obtained (Andrieu et al., 2003; Borgonovo & Gatti, 2013). In other words, each variable fed into the model is generated randomly N times (trials/iterations) and the results are accumulated in a sample R that characterizes the distribution of the possible results (Kwak & Ingall, 2007; Oberle, 2015; Nguyen, 2023). The simulation above is considered a direct simulation when the sample R can be said to belong to the population being studied. The central limit theorem, from the statistical theory, helps the researcher approximate the distribution of the mean estimate that can be made from the sample R. It states that, regardless of the distribution of the random variables in question, the result of the sum of a large number of independent and independently distributed variables will always be normally distributed (i.e., follow a Gaussian or Normal distribution) (James, 1980; Oberle, 2015). In mathematical expression, if $\bar{Y}$ is the mean of a sample $S = \{Y_1, Y_2, \dots, Y_n\}$, where

$S \in Y$, then for n sufficiently large, $\bar{Y} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, μ and σ representing the mean and standard deviation of Y, respectively; this facilitates the application of confidence intervals and other statistical analysis (James, 1980; Oberle, 2015).

The MC simulation has been used in literature for various purposes, some of which include theoretical investigation (James, 1980), simulating networks of correlated company risk (Blanchet et al., 2015), risk profiling (Borgonovo & Gatti, 2013), unspecified bias tests (Duffie et al., 2007), aiding intensity-based model estimation (Giesecke et al., 2010), portfolio risk management (Giesecke et al., 2010), assessing model performance/robustness (Jackson & Wood, 2013; Kumar & Ravi, 2007; Kwon & Lee, 2018; Robey & Barcikowski, 1988), project risk management (Borgonovo & Gatti, 2013; Kwak & Ingall, 2007; Riemer & Wagner, 2016) and sample creation (Kumar & Ravi, 2007), not to mention other areas outside finance (Innocenti et al., 2022; Kwak & Ingall, 2007). Furthermore, and accordingly to some of the examples given, MC is not limited only to the scenario propagation usage, but may also aid in sampling procedures (e.g., approximating a target distribution) (Blanchet et al., 2015; James, 1980), as well as optimization procedures (Andrieu et al., 2003; James, 1980).

In the discussion of whether this methodology should be used in a situation, it poses some advantages: (1) in sampling procedures, it positions the samples in regions of high probability (Andrieu et al., 2003); (2) it can aggregate any number of risks expressed by any probability distribution (Gleißner, 2019); (3) can handle extremely high dimensional[13] problems, converging regardless of dimensionality (James, 1980; Kwak & Ingall, 2007); (4) can handle extremely non-linear problems (James, 1980; Kwak & Ingall, 2007; Kwon & Lee, 2018); and (5) it is an extremely powerful tool to understand and quantify the potential effects of uncertainty in models or systems (Kwak & Ingall, 2007); can work even under limited data (O'Leary, 1998). One must keep its limitations in mind, however: (1) regardless of working with stochastic inputs, MC discretizes the processes by generating finite intervals, which may lead to discretization bias (Giesecke et al., 2010); although this bias can be easily reduced by increasing the number of iterations, it comes at the cost of increased time and computational effort (Giesecke et al., 2010); (2) the application of MC simulation

---

[13] High dimensional problems are categorized by considering a high number of variables (referred to as dimensions), hence dimensionality relating to the amount of variables.

is limited by how well a researcher can formulate the problem in question, which also depends on its mathematical properties (James, 1980); (3) MC algorithms converge[14] relatively slowly, having low efficiency (James, 1980; Riemer & Wagner, 2016); (4) utilizing MC does not necessarily help to improve the theory to solve problems (James, 1980); (5) managers may avoid utilizing MC due to lack of thorough understanding of the method (Kwak & Ingall, 2007; Riemer & Wagner, 2016); (6) requires relatively high computational power and takes significant time compared to standard methods (Kwak & Ingall, 2007; Riemer & Wagner, 2016). In the issue of computational power and time consumption, a few studies already carry high efficiency MC methods through quantum computer speed-ups (Kaneko et al., 2021), though these computers are not the common management tool. Additionally, seeing as the number of iterations clashes in disadvantages (1) and (6), the statements of Bukaçi et al (2016) are supported in the importance of assessing the necessary minimum number of iterations that guarantee either less variance, error, or bias, also since accuracy is supposed to increase with each iteration (Oberle, 2015).

2.2.2. Number of Iterations

Oberle (2015) states that, for every MC simulation, there is a calculable number of iterations that provide an approximation to any desired variation. Raftery and Lewis (1970) state that, for running the MC algorithms, it is necessary to determine for how long the simulation needs to run. In literature, we find common practices of 10 000 iterations (Borgonovo & Gatti, 2013; Jackson & Wood, 2013; Liang et al., 2011; Nguyen, 2023), with a minimum of 100 and maximum of 10 000 000 iterations executed[15]. A few methods are found that attempt to offer a calculatable approach, which will follow below.

For all following methods, the n represents the number of iterations and $z_{\alpha/2}$ represents the score of a standard normal distribution (a Gaussian/normal distribution with mean 0 and standard deviation 1) such that $\alpha/2$ is the area under the curve to the right of $z_{\alpha/2}$. The latter calls for the assumptions above met with the central limit theorem (Bukaçi et al., 2016; James, 1980; Oberle, 2015).

---

[14] A sequence *{A}* is said to converge to the intended value B if for any arbitrarily small distance k, an element of that sequence *{A}* can be found such that all succeeding elements of *{A}* are guaranteed to be within k or less of B. (James, 1980).

[15] These values do not include bootstrap iterations.

a) Population Proportion Confidence Interval Approach (Oberle, 2015)

Oberle (2015) utilized the Wilson score method to obtain an *a priori* confidence interval for the population proportion p and with it define a stop condition for the simulation. The formulation goes as follows:

$$\frac{z_{\alpha/2} * \sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + \frac{\left(z_{\alpha/2}^2\right)}{n}} < \Delta \quad , \qquad \hat{q} = 1 - \hat{p}, \tag{6}$$

where $\hat{p}$ is the natural estimator of population proportion of success and is given by the fraction of successes, and $1 - \alpha$ is the confidence level, and Δ is half the length for the intended confidence interval, respectively. Once the condition above was met, it being checked for every iteration, the simulation would stop.

b) Mean Confidence Interval Approach (Oberle, 2015)

Additionally, Oberle (2015) also offers a calculation for the mean confidence interval.

$$n = \frac{z_{\alpha/2}^2 * s^2}{\Delta^2} \,, \qquad with \qquad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1} \,, \tag{7}$$

where x_i is observation number i of the sample X with mean $\bar{X}$. The calculations are relatively simple compared to the previous proposition, though at the cost of approximating the population variance σ^2 by assuming the variance of the sample s^2 instead. A similar stop condition to the above is not given, but can be easily calculated, being presented as follows:

$$\frac{z_{\alpha/2} * s}{\sqrt{n}} < \Delta \tag{8}$$

As Oberle (2015) states, as more iterations are performed, more refined estimates are obtained from s, until the algorithm converges. Once again, meeting the condition above would stop the simulation.

c) Mean Percent Confidence Interval Approach (Bukaçi et al., 2016)

Bukaçi et al. (2016) stray slightly from the approach of Oberle (2015). Instead, they calculate a formula to assess the minimum number of iterations through a maximum error value. This error represents half the length of the intended interval in percentage of the estimate of the mean $\bar{X}$, which is measured in integer percentage (twelve percent being 12). The calculations are as follow:

$$n = \left[\frac{100 * z_{\alpha/2} * s}{\bar{X} * \; E}\right]^2, \qquad E = \frac{100 * z_{\alpha/2} * s}{\bar{X} * \sqrt{n}} \tag{9}$$

with s being the standard deviation of the estimated sample. The difference between this approach and the simple one by Oberle (2015) is more related to interpretability, though determining the confidence interval in percentage of the estimate does seem easier to understand *a priori*.

2.2.3. Distribution Generation

Random Variable Dependency

MC Simulation basis describes that, in its process, independent and independently distributed random variables are generated (James, 1980). If the variables in question are independent, they will have no correlation (James, 1980). In that sense, for two independent variables, $Var(X+Y) = Var(X) + Var(Y)$, where $Var(X)$ is the variance of the random variable X and $E(XY) = E(X) * E(Y)$, where $E(X)$ is the expected value of the random variable X (James, 1980).

In the occasion where the variables being replicated through simulation are not independent and must take into account correlations, generating i.i.d. random variables may lead to an under- or over-estimation of the results (Touran & Wiser, 1992). James (1980) and Touran and Wiser (1992) reference a relatively simple algorithm to generate pseudo-random correlated variables, under the assumption they follow a Gaussian distribution (James, 1980; Touran & Wiser, 1992). James (1980) denominates this method of 'square root method'. Let *C* be a matrix of generated random values of n different random variables with a standard normal distribution and considering the intended correlation and *Z* their respective correlation matrix. In the method explained by both studies (James, 1980; Touran & Wiser, 1992), there is an $n \times n$ triangular matrix *X* such that $Z = X * X^T$, X^T being the transpose of *X*. In it, *C* can be obtained from the following equation:

$$C^T = X * N^T \Leftrightarrow C = (X * N^T)^T \tag{10}$$

The matrix *N* includes a set of *n* i.i.d. distributed variables (all following a standard normal distribution), one for each column, with the rows being the number of generated values equal for each variable. The same applies to *C*, though with the correlated transformed values. The variables in the columns should respect the same order in all the matrices in equation (10). As for the *X*, James (1980) and Touran and Wiser (1992) calculate it using the logic in Figure 3.

$$Z = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n,1} & \sigma_{n,2} & \cdots & \sigma_{n,n} \end{bmatrix} \qquad X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,n} \end{bmatrix}$$

$$\sigma_{i,j} = \sigma_{j,i} \text{ and } \sigma_{i,i} = 1$$

$$x_{i,j} = \begin{cases} \dfrac{\sigma_{i,1}}{\sqrt{\sigma_{1,1}}} & , \quad 1 \le i \le n \quad , \quad j = 1 \\ \sqrt{\sigma_{i,i} - \displaystyle\sum_{j=1}^{i-1} x_{i,j}^2} & , \quad 1 < j, \qquad j = i \\ \dfrac{\sigma_{i,j} - \sum_{m=1}^{j-1} x_{i,m} * x_{j,m}}{x_{j,j}} & , \quad 1 < j \le i - 1 \\ zero \quad , & \qquad otherwise \end{cases}$$

Figure 3 – Square Root Recursive Method for Simulating Correlated Random Variables

The resulting *X* matrix can then be transformed into matrix *C* through formula (10). After this, *C* must be de-standardized, that is, transformed back from a standard normal distribution for each of its composing variables. This can be executed by reversing the standardization formula below (James, 1980; Oberle, 2015):

$$\frac{Y-\mu}{\sigma} = C \sim N(0,1) \qquad \text{for} \qquad Y \sim N(\mu,\sigma) \tag{11}$$

A researcher can validate if the X matrix was correctly built by checking if the calculated $Z' = X * X^T$ is equal to the observed Z matrix (James, 1980; Touran & Wiser, 1992).

Producing Discretely Distributed Random Numbers

One can produce discretely distributed Z random numbers using a uniformly generated number $X \sim U(0,1)$ and taking the random value from the first position (in ascending order) in the cumulative distribution of Z for which the density is greater than X (James, 1980).

Sampling from a Continuous Distribution

To randomly sample according to a one-dimensional distribution with a smooth function, one can determine its inverse cumulative function and feed it with a uniformly generated number $X \sim U(0,1)$, a process which will return the value for which the function would have that cumulative distribution (James, 1980).

3. Methodology

3.1. The Model

The objective of the methodology of this study is to test the augmentation of a statistical model easily applicable in practice through the Monte Carlo simulation methodology, in an attempt to attain more explanatory power over default, while still keeping the model practical and widely applicable, in terms of company size, industry, etc. Following the objective of applicability, the choice of model for this study should be applicable to as many companies as possible. Additionally, intelligent models do not benefit from the applications of simulation, seeing as some algorithms are already more complex methodologies than Monte Carlo simulation, and therefore a statistical or theoretical model is to be chosen. The top modeling techniques in the remaining literature which may apply to the most companies are the statistical approaches MDA and Logit (Alaka et al., 2018; Balcaen & Ooghe, 2006; Jackson & Wood, 2013; Mensah, 1984), with relatively the same performance (Alaka et al., 2018; Altman et al., 2016). Additionally, Altman et al. (2016) poses the third option, which conveniently satisfies the objective of applicability (having been applied internationally) and combines the two above options – the Z”-score LR model, a Logit estimation with the same variables as the Z”-score originally proposed by Altman (1983). The original function of the Z”-score by Altman (1983), originally estimated using MDA, is as follows:

$$Z" = 3.25 + 6.56 * X_1 + 3.26 * X_2 + 6.72 * X_3 + 1.05 * X_4\,, \qquad (12)$$

where[16]

$$X_1 = \frac{Working\ Capital}{Total\ Assets},$$

$$X_2 = \frac{Retained\ Earnings}{Total\ Assets},$$

$$X_3 = \frac{EBIT}{Total\ Assets},$$

$$X_4 = \frac{Total\ Equity}{Total\ Liabilities}.$$

According to Altman (1983) and Altman et al. (2016), this is the most widely applicable Z-score model, as is not dependent on market variables nor more industry specific features.

---

[16] All values being book values, so as to keep the model applicable to non-listed companies.
EBIT stands for earnings before interest and taxes.

The Logit version of Z”-score utilized by Altman et al. (2016) is relatively similar to a Logit transformation of the Z”-score (Agarwal & Taffler, 2008; Correia et al., 2017; Hillegeist et al., 2004; Jackson & Wood, 2013; Reisz & Perlich, 2007), albeit being re-estimated using the logistic regression (Altman et al., 2016). Its function is as follows:

$$Z"LR = \frac{1}{1+e^{Z"}} \tag{13}$$

With the two possible models described, it is only left to address the Monte Carlo application to be executed. The objective of this application is seeing how MC simulation can augment either one of the traditional models above through the addition of uncertainty and scenario propagation. Both of the models suffer from stationarity, in that they either have it as an assumption (i.e., MDA) or treat all observations as a single year and do not incorporate default as a process (Balcaen & Ooghe, 2006; Mensah, 1984; Shen et al., 2020). The scheme of the method of this study is described in Figure 4. For each company, a set of data before default is introduced and used in the MC algorithm to calculate a set of *N* model result random variables for each year until default (default included). The result is a sample of size *N* comprised of a set of score values obtained for the year of analysis which describe how the company quality probabilities are distributed, precisely as in the basic MC approach (Andrieu et al., 2003; Borgonovo & Gatti, 2013). From this distribution, we can obtain various descriptive statistics of the scores, such as the average, the variance, the distribution, or, most interestingly, measures of the probability of that score being equal, lower, or higher than a pre-selected score. In other words, it may offer a measure of default risk, given a pre-determined threshold.

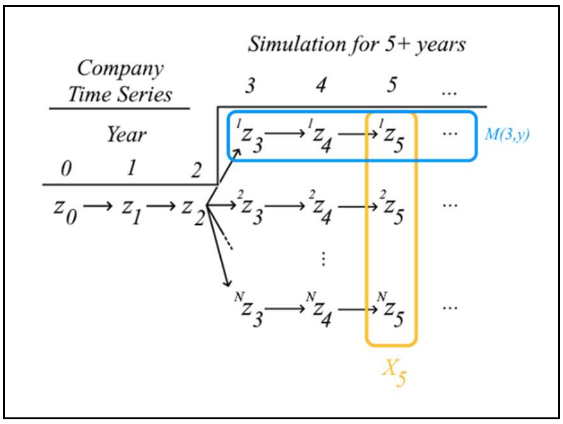


Figure 4 - MC Simulation Scheme

Given the above, this methodology may serve as a substitute for the Logit transformation which usually is applied to score-outputting models (Agarwal & Taffler, 2008; Correia et al., 2017; Hillegeist et al., 2004; Jackson & Wood, 2013; Reisz & Perlich, 2007). For that reason, it makes more sense to apply the MC simulation method to MDA alone, seeing as how one may then compare its results to the results of the Logit estimated MDA, or the Logit transformed MDA. The Logit transformed MDA ($Z"LT$) would be the result of manipulating the scores obtained through the Z"-score for each company utilizing the aforementioned formula (3), as follows:

$$Z"LT = \frac{e^{Z"}}{1 + e^{Z"}} \tag{14}$$

Before any model transformation, however, re-estimation of the parameters of the models is advised (Agarwal and Taffler, 2008; Altman et al., 2016; Balcaen and Ooghe, 2006; Hillegeist et al., 2004). For MDA, this can be equally achieved through a simple linear regression of the binary observed values of default (Blum, 1974; Dimitras et al., 1996; Ketz, 1978; Sun et al., 2014). Important to note is that, following re-estimation, a re-definition of the grey-area of the Z"-score is not required, the determination of one cut-off point sufficing to give a threshold for the MC simulation to consider in the determination of the default probabilities. This because the algorithm will offer a percentage of companies below the cut-off point selected, which will always be higher the more the values of a company are lower; and because the choice of prediction power evaluator for the analysis will be the ROC curve AUCs, which overcomes single-cutoff point limitations for the probability results.

In the methodology, $g(M_k)$ will represent the distribution from where the values of the Z"-score for year i (represented by z_i), will be generated. In this case, M_k stands for the 'memory' of the model and represents a vector of previous values of z (or its components) on the time series, and k is the number of years used as data in order to generate values for z_i. There is an important distinction within the way this study will generate z_i. It can be generated directly from the previous values of the Z"-score in the time series (the direct approach), or by generating the variables inputted on the Z"-score and calculating the resulting z_i from them (the component approach). Both approaches will be tested in this study, so as to understand if one can obtain better results by dividing the simulation into the four different variables/components of the Z"-score. The direct approach, requiring only the simulation of one random variable, will be less computationally intensive, while the component approach will require four times more effort. If no significant improvements are found in choosing the component approach over the direct approach, the latter will hence be advisable. In addition to a lesser computational effort, the direct approach, simulating only one variable, avoids the possible issue of variable correlation. This is because the Monte Carlo crude application generates i.i.d. random variables (Andrieu et al., 2003; Borgonovo & Gatti, 2013) for variables that, in reality, can be significantly correlated, even though the variables from the Z"-score (Altman, 1983), much like the original Z-score (Altman, 1968), were chosen each from a different category of ratios, in theory helping to diminish the possible multicollinearities (Altman, 1968; Altman et al., 2016). In the other hand, utilizing the component approach may explore a more substantial number of scenarios, since it will propagate for the tendency of components separately, each able to evolve in quite different directions.

Through the utilization of a time series and memory approach on the calculation of future scores, it is expected that this methodology can overcome the stationarity of the more traditional models (Balcaen & Ooghe, 2006; Mensah, 1984; Shen et al., 2020) and approach default more like a process, as it more realistically is (Bonfim, 2009; Shen et al., 2020; Sun et al., 2014; Wilcox, 1971). This does not overcome the possible concept drift (Shen et al., 2020; Sun et al., 2017), however, if the parameters of the Z"-score still remain the same regardless of the year.

The remaining aspect of the simulation to be discussed is the distribution that $g(M_k)$ follows. This consists in assuming a certain continuous distribution for the values of the components or the score. This can be done by testing the goodness-to-fit of each distribution to the actual sample distribution of the different values found and can be achieved easily through the use of a software like @Risk, which encompasses a plethora of different distribution types. From there, the parameters for the distribution, be it the mean, standard deviation, or any other, are obtained from the vector M_k and the simulation is run by generating random values from the chosen distributions adjusted to the calculated respective parameters.

As an additional note, the Z-score and its variations are usually estimated for a given period using data of prior periods (Altman, 1968; Altman, 1983; Altman et al. 2016), so as to be considered a prediction. The methodology used in this research does not need this lag on variables because it will generate a high number of scenarios for the variables or the z-score itself for those periods that follow. In that sense, the MC Z”-score will estimate the distribution of the quality of the company in years to come, making a better use of the MDA methodology for what it was originally designed – classification (Balcaen & Ooghe, 2006). Through this approach, since the prediction tool is not MDA but MC, the Z”-score parameters will be estimated using financial data and default data from the same year.

Number of Iterations

The methods presented in Section 2.2.1 to calculate the minimum number of iterations (*n*) for a given level of interval confidence and width all utilize the sample variance to develop stop signals for the algorithm. In the methodology of this study, it does not make sense to utilize any of those methods to calculate the intended *n*, since the sole purpose of it is to spread the distribution to its maximum deviation for a given period, and not to achieve the more accurate representation of a population. That is because utilizing variance-based methods on an algorithm that variates values to analyze risk will only lead to exponentially higher *n* values the longer the horizon is extended.

3.2. Data and Sampling

This study focuses on the introduction of its methodology on the Portuguese company case. Around 99.9% of Portuguese companies fall under the classification of micro, small or medium enterprises (MSME) (INE & Pordata, 2021) and, hence, it can be expected for the samples obtained from the population to be of relatively small size when compared to the common literature (Blum, 1974; Bonfim, 2009). In that sense, an original sample of 90 000 Portuguese companies was selected, composed of 9 subsamples of 10 000 companies each, each subsample corresponding to a division of the Standard Industrial Classification (SIC) system (*SIC Manual*, 2023). This subsample size was meant to be in accordance with the recent studies reviewed that analyze single country samples, considering not all the observations collected will be usable in estimation. Traditional studies reviewed have an average and median sample size of approximately 473 and 138, respectively, while studies found of the 21st century share an average and median of 6 860 and 2 770, respectively[17]. The sample size cannot be stretched much further, or the simulation becomes too computationally heavy to apply in useful time to all the subsamples.

In order to reduce category imbalance (Shen et al., 2020), the subsamples are analyzed separately, and a model is estimated for each one (Costa et al., 2022). This is because all industry subsamples have the same number of observations, when in reality, the population will not be uniformly distributed along the defined categories.

The data is obtained yearly since that is the periodicity available and mandatory for the most companies. As for the data period, the reviewed literature presents an average of approximately 20 periods studied (with a mode of 7 periods and a median of 15) and a higher focus on more traditional periods (roughly 76% of all the periods studied by the reviewed literature are prior to 2000). Given these measures, this study selects the years between 2000 and 2022 for analysis (inclusive), which allows for: (1) modernizing the sample when compared to most non-intelligent model studies; (2) having a study period longer than the reviewed average; (3) after estimation out-of-sample testing due to an extended number of years to analyze.

The data on the state of the company (which will offer information on default) is obtained from the ORBIS database, along with the remaining general and financial information required. The samples are obtained randomly from each of the industry segmented database populations, so as to maintain the default rate of each subsample as close as possible to their industry specific default rate. The divisions are as follow, according to source:

(1) Division A - Agriculture, Forestry, and Fishing
(2) Division B – Mining
(3) Division C – Construction
(4) Division D – Manufacturing
(5) Division E – Transportation, Communications, Electric, Gas, and Sanitary Services
(6) Division F – Wholesale Trade

---

[17] (Abinzano et al., 2020; Agarwal & Taffler, 2008; Altman et al., 2016; Beaver et al., 2005; Bharath & Shumway, 2008; Bonfim, 2009; Bottazzi et al., 2011; Chava & Jarrow, 2004; Correia et al., 2017; Costa et al., 2022; Du et al., 2020; Duffie et al., 2007; Hillegeist et al., 2004; Korteweg, 2011; Kwon & Lee, 2018; Mitton, 2021; Peláez et al., 2022; Reisz & Perlich, 2007; Shen et al., 2019; Shumway, 2001; Sun et al., 2017; Tafakori et al., 2021)

(7) Division G – Retail Trade
(8) Division H – Finance, Insurance, and Real Estate
(9) Division I – Services
(10) Division J – Public Administration

The Division B of the SIC classification had less than 2 000 companies in the database and hence was combined with the Division A (seeing as both were primary type sectors). This was done prior to obtaining the random sample from the combined industry group, so as to not influence its randomization. No subsample was collected for the Division J due to the unavailability of substantial data on them for Portuguese companies in the database (less than 6 000 companies and data only available for less than 50 companies in every year of the period of analysis). Although the data was all collected for analysis, due to computational limitations, only the Divisions A+B and C were analyzed, reducing the total sample to 20 000 companies, which is still on par with the common literature. This combination of divisions therefore encompasses the Portuguese companies of Agriculture, Forestry, Fishing, Mining and Construction sectors.

For analysis later on, companies were additionally classified by size. Company classification by size is determined based off of the Portuguese law determination (Portugal 2020, 2021), as follows:

(1) Micro Company: has less than 10 employees, and either its yearly total of sales and services are under 2 000 000 € (2 million euros), or its total assets are under 2 000 000 €.
(2) Small Company: is not considered a micro company, has 50 or less employees, and either its yearly total of sales and services are under 10 000 000 € (10 million euros) or its total assets, are under 10 000 000 €.
(3) Medium-sized Company: is not considered nor a micro nor a small company, has less than 250 employees, and either its yearly total of sales and services are under 50 000 000 € (50 million euros), or its total assets under 43 000 000 € (43 million euros).
(4) Large Company: all the companies that are not considered nor micro, nor small, nor medium, under the above conditions.

Default Definition

To maintain the consideration of default objective, more widely applicable and in accordance with Altman (1983) and Altman et al. (2016), this study considers the *ex post* definition of default – bankruptcy. Companies are considered defaulted for a given year if they declared one of the following in the ORBIS database in said year: (1) Dissolved (with no further information as to why); (2) In liquidation; (3) Active (insolvency proceedings); (4) Bankruptcy. The data after any of these declarations is not necessarily considered as default data, which can serve as form of noise test on the model. Unavailability of data prior to default is used in the simulation to test longer horizons of default prediction. The year of default considered in the criteria above is manipulated as follows: if a company entered the status of “Active (dormant)” in the database directly before any of the status considered default and no data exists for it on the database from the point it was declared dormant, then the default date is set to the date in which the dormant status was declared instead; this because if a defaulted company was kept in a state of dormancy and no changes in its financials were executed until its default, default was likely already imminent when the company became frozen.

3.3. The Algorithm

Let N be the year for which one wishes to predict the probability of default of a given company – hereby referred to as the year of analysis. The algorithm used in this study will have two main parameters to determine from where and with what it will be generating predictions and attempting to offer the user an answer: h and k. k has been presented before as the number of years of memory that are fed into $g(M_k)$ to output predicted values for the future. h stands for the horizon, namely the distance from where the algorithm will be generating its predictions. In other words, h is the number of years the algorithm will generate in front of the memory to predict default. Say a manager wished to know, in 2009, what would be its company default risk in year 2012, using data from 2005 through 2009, inclusive. The algorithm, in this case, would have the parameters $k = 5$ and $h = 3$. Its important to note that the memory of the algorithm functions pretty much like a moving average, meaning the predictions made for year 2011 would come from a distribution considering the years 2006 through 2010, inclusive, always maintaining the 5 years of memory as pre-determined by k. Only the predictions for the year N are counted for assessing the default risk of the company in the end. In the above example, $N = 2012$. What the fully sized algorithm below will perform is an execution of these predictions for each $N \in \{2002, 2003, 2004, \dots, 2022\}$, for which every combination of h and k will be tested, for as long as it does not exceed the period of years from 2000 through 2022. The reason why N will not be evaluated as 2000 or 2001 is because a minimum of 2 years of memory is necessary for every type of model that will be tested to work. Additionally, both k and h will be capped at 10, for consistency with the reviewed literature on default and to avoid extending to too much computational effort.

The Z”-score utilized to calculate the predictions from the algorithm results has its parameters fit to a ordinary least squares multiple linear regression (using the LINEST function available in Excel) for each year and correspondent different combination of h and k, where each ratio is an independent variable, and the dependent variable, given that the Z”-score is a measure of financial quality, is a dichotomous variable equal to -1 when the company is defaulted, and 1 otherwise, in an attempt to have zero represent the state of uncertainty. With this, it is also expected that the Logit transformation of the Z”-score ($Z''LT$) yields more useful results, since its function is centers scores at 0 for a probability of 50%. The results of the algorithm are samples of N iterations for each company in each year they could validly be generated. Due to computational limitations, N had to be set to 100, otherwise the data was extremely time consuming to process correctly. From each of these samples, the percentage of observations with Z”-score values under 0 in the total of the 100 gave the default probability results for correspondent company.

3.4. Evaluation

As mentioned prior to this section, the evaluation of the model power for the results of this study will be performed utilizing the ROC curve AUCs, as they are more present in literature and allow for an all-in-one variable cut-off evaluation. The ROC curves will not be presented, as the amount of different curves would make the visual summary too extensive. The AUC summary tables will be presented instead. This study will follow the evaluation philosophies of Mitton (2021) and especially Zhang (2022) of testing a model not based off of statistical significance, but of parameter, condition, and situation scrutiny. In other words, the model will be evaluated for a set of combinations of horizons, memories, and years, for which there will be a different re-calibration of score parameters and correlations each. The results will be presented through averages, variances, and other useful measures.

Although prediction power is the main evaluator of the model, calibration will also be evaluated through the use of the Pseudo-R^2 (Bonfim, 2009; Campbell et al., 2008; McFadden, 1973), based on the log-likelihood presented in equation (4)[18].

All sizes of companies are maintained in the evaluation, and results are obtained by size too in order to understand both the effects of the larger companies in the model and the differences between using the model in each of the types of companies. In summary, the models MC Component Approach Z”-score (Z”MC_C), MC Direct Approach Z”-score (Z”MC_D) and Z”LT are all evaluated and compared in terms of predictive power and calibration, by horizon, memory, and company size, for each of the SIC divisions separately.

---

[18] In the cases where the probability returned by the model for a given company default is 100% or 0%, **Equation (4)** returns an error for ln(0); this was overcome by replacing the values inside the natural logarithm for a value different from zero, but close enough to zero (10^{-100}) for the computer, so as to avoid the error and still obtain a liable likelihood value.

4. Execution and Results

4.1. Pre-execution Analysis

Before executing the models, some statistics are important to calculate, which are present in this subsection. For information purposes, first and foremost, the number of companies that have available information to calculate all ratios for the given year of the period of this study is analyzed, along with the respective yearly default rate. Furthermore, for purposes of pseudo-random variable generation, the Pearson correlation values between the ratios for each SIC division is also analyzed. Company size weight on each of the divisions is also expressed for analysis.

Regarding the recalibration of the Z”-score through multiple linear regression, the R^2 values relating to the linear regressions for each year are analyzed per division. These R^2 values are expressed as averages of all R^2 obtained for each of the available logical combinations of horizons and memory values for the model in each of the years.

4.1.1. Default Rates

Table 10 expresses, for every year of analysis, the number of companies that, for that year, had enough data for estimation and simulation along with the number of which are defaulted, with the respective default rate. As expected from an unbalanced default sample, default rates on Table 10 are extremely low and reinforce the idea that the event of default, in itself, is rare. This can, of course, affect the ability of the models to capture the characteristics of default correctly (O'Leary, 1998; Shen et al., 2020). The intent here, though, is to test the algorithm in an as realistic as possible environment, so as to maximize its usefulness in practice, following the logic of Balcaen and Ooghe (2006) and Du et al. (2020) on the matter of data imbalance. Therefore, no over-sampling is performed.

Table 10 - Sample Default Rates

Division	Division A+B			Division C		
Total Valid Companies	6232			5414		
Year	No	D.C.	D.R.	No	D.C.	D.R.
2000	0	0	0,00%	0	0	0,00%
2001	0	0	0,00%	0	0	0,00%
2002	0	0	0,00%	0	0	0,00%
2003	0	0	0,00%	1	0	0,00%
2004	61	0	0,00%	140	1	0,71%
2005	69	1	1,45%	177	0	0,00%
2006	112	1	0,89%	297	3	1,01%
2007	178	7	3,93%	459	17	3,70%
2008	517	16	3,09%	1371	58	4,23%
2009	523	12	2,29%	1284	19	1,48%
2010	551	13	2,36%	1241	30	2,42%
2011	560	15	2,68%	1139	54	4,74%
2012	678	10	1,47%	1118	50	4,47%
2013	922	8	0,87%	1189	39	3,28%
2014	2969	54	1,82%	2706	88	3,25%
2015	3316	34	1,03%	2829	59	2,09%
2016	3571	26	0,73%	2913	55	1,89%
2017	3900	47	1,21%	2938	39	1,33%
2018	4092	37	0,90%	2984	31	1,04%
2019	4278	41	0,96%	3083	28	0,91%
2020	4383	38	0,87%	3168	23	0,73%
2021	4488	29	0,65%	3339	40	1,20%
2022	201	9	4,48%	159	19	11,95%

For this table: “No” stands for the number of valid companies; “D.C.” for the number of companies that are defaulted; “D.R.” for the default rate.

The total number of valid companies for each division described in Table 10 refers to the number of companies from the original 10 000 that have enough available information for each of the ratios for the algorithm to use them. This was due to the fact that, to be analyzed for a given year, a company had to have all ratios available for a consecutive period of 2 years prior to it, at least. The percentage of validated companies from the original samples is averaged at 58.23%, with Division A+B having a validation rate of 62.32% and Division C having a validation rate of 54.14%.

4.1.2. Ratio Correlation

Figure 5 presents the Pearson correlation matrices for the Z”-score ratios that will be generated in the component approach of the methodology, per division. From the analysis of Figure 5, it is clear that some of the ratios are extremely correlated, while others are only mildly or not at all correlated. Regardless of whether all of the ratios are correlated enough to affect the simulation, one must keep in mind that, in order to generate the pseudo-random correlated variables on par with the given correlation matrices, the Square Root Recursive Method (James, 1980; Touran & Wiser, 1992) must be utilized. This method has the limitation of working only with normally distributed variables and, so, the following algorithm must assume a normal distribution for the correlated generated variables. With that in mind, however, it would be inconsistent to consider only some of the variables as normally distributed while assessing the distribution for the rest by other means. Besides, it is unclear to what extent the simulation should ignore low correlation between variables. Following these lines of thought, the decision is to consider all variables generated as normally distributed, with sample mean and standard deviation obtained from the memory of the model. Furthermore, all variables are generated through the Square Root Recursive Method (SRRM) (James, 1980; Touran & Wiser, 1992), regardless of correlation, for a matter of consistency.

A+B

	X1	X2	X3	X4
X1	1,000	0,011	0,787	0,000
X2	0,011	1,000	0,996	0,000
X3	0,787	0,996	1,000	0,000
X4	0,000	0,000	0,000	1,000

C

	X1	X2	X3	X4
X1	1,000	0,455	-0,294	0,000
X2	0,455	1,000	0,238	0,000
X3	-0,294	0,238	1,000	0,000
X4	0,000	0,000	0,000	1,000

Note: $X_1 = \frac{Working\ Capital}{Total\ Assets}$, $X_2 = \frac{Retained\ Earnings}{Total\ Assets}$,

$X_3 = \frac{EBIT}{Total\ Assets}$, and $X_4 = \frac{Total\ Equity}{Total\ Liabilities}$

(following Altman, 1983)

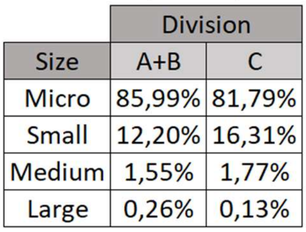
Figure 5 - Pearson Correlations per SIC Division

The calculations of the X matrices for the SRRM[19] could be done directly from the correlation matrices shown in Figure 5, though, in an attempt to replicate the use of the algorithm in a single company environment with only a given memory as data for the calculations of those correlations, the matrices X are calculated solely with the memory data inputs for each of the year, horizon and memory combinations and, hence, it would be unproductive to display a list of them here. All of them were pre-validated for simulation purposes. Although correlation issues do not apply to the direct approach of the model, making the SRRM only necessary for the component approach, again, for the sake of consistency and comparability, the direct approach is also performed assuming the Z"-score distribution as normal.

4.1.3. Size Disparity

Companies inside divisions, as expected from the data already available on Pordata (INE & Pordata, 2021), are more predominant the smaller they are, with micro companies encompassing a great part of the total, as portrayed in the Table 11 below. Estimation and evaluation occurred regardless of company size, with the intent to test the model against any possible company value, along with noise, seeing as large companies, per example, are extremely rare in the sample, but likely have ratios very different from the majority.

Table 11 - Percent of Yearly Data by Size and Division



	Division	
Size	A+B	C
Micro	85,99%	81,79%
Small	12,20%	16,31%
Medium	1,55%	1,77%
Large	0,26%	0,13%

4.1.4. Parameter Re-estimation

When it comes to parameter re-estimation calibration values for the calculated parameters of the Z"-score for each of the year-memory-horizon combinations, there are some clear tendencies distinguishable throughout Table 12:

---

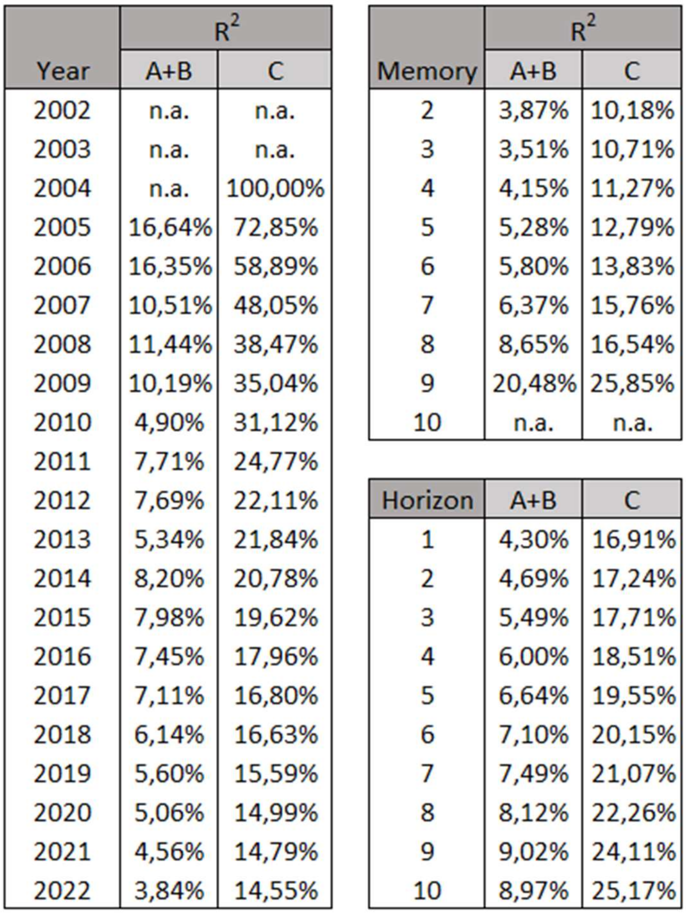
[19] Across calculations of the matrices X, certain cells often return an error due to the components inside square roots equaling zero. This was easily resolved by replacing the error values with zero without affecting the validation of the respective matrices.

(1) The **R^2 values decrease with the years**, which can be explained by both the following ideas: (a) the fact that the number of company years to analyze increases with the year makes it more difficult to fit a linear function as perfectly through all the data as with less data, as well as it means a higher chance of outliers and noise in the data, which directly affect the goodness-to-fit of linear regressions; (b) the ratios utilized were originally chosen as the most adequate to predict default in 1983 (Altman, 1983), which means that these ratios may have lost the descriptive ability of default they had along the years.

(2) The **R^2 values increase with the memory utilized**, which is a telltale sign that one can obtain a better recalibration of the Z"-score values by selecting companies with more data available, even if there are fewer, instead of more companies with less data available.

(3) The **R^2 values generally increase with the horizon utilized**, which implies that, although only slightly and interestingly enough, the Z"-score obtains better recalibration the more the data used is lagged.

Table 12 - Recalibration Regression R-squared Values by Year, Memory and Horizon for each Division



	R^2	
Year	A+B	C
2002	n.a.	n.a.
2003	n.a.	n.a.
2004	n.a.	100,00%
2005	16,64%	72,85%
2006	16,35%	58,89%
2007	10,51%	48,05%
2008	11,44%	38,47%
2009	10,19%	35,04%
2010	4,90%	31,12%
2011	7,71%	24,77%
2012	7,69%	22,11%
2013	5,34%	21,84%
2014	8,20%	20,78%
2015	7,98%	19,62%
2016	7,45%	17,96%
2017	7,11%	16,80%
2018	6,14%	16,63%
2019	5,60%	15,59%
2020	5,06%	14,99%
2021	4,56%	14,79%
2022	3,84%	14,55%

	R^2	
Memory	A+B	C
2	3,87%	10,18%
3	3,51%	10,71%
4	4,15%	11,27%
5	5,28%	12,79%
6	5,80%	13,83%
7	6,37%	15,76%
8	8,65%	16,54%
9	20,48%	25,85%
10	n.a.	n.a.

Horizon	A+B	C
1	4,30%	16,91%
2	4,69%	17,24%
3	5,49%	17,71%
4	6,00%	18,51%
5	6,64%	19,55%
6	7,10%	20,15%
7	7,49%	21,07%
8	8,12%	22,26%
9	9,02%	24,11%
10	8,97%	25,17%

The additional commentary that can be made regarding the results obtained from the recalibration is that, in general, recalibrating the Z”-score utilizing multiple linear regression through the methodology presented offers poor goodness-to-fit results. Furthermore, Division C presents higher goodness-to-fit values and can be expected to obtain better results. The high results of the earlier years of Division C cannot be considered high goodness-to-fit results, since they depend on relatively small samples with extremely low default rates. As recalibration is not necessarily the key of evaluating a model, neither is it the focus of this study, the simulation proceeds with the obtained models, regardless of their R^2 values. Once more, given the sheer amount of different parameters estimated, these will not be displayed in this report.

4.2. Results

4.2.1. Computational Effort

As the methodology of this study requires a considerable amount of computational power, making the simulations time consuming, which is one of the largest limitations to its execution, Table 13 presents a summary of some essential representations of the effort required to perform the tasks included in the algorithm.

Table 13 - Computational Demands of the Simulation per Division

Division	Maximum Validated Observations	Total Iterations Generated	Total Time Consumed (without breaks)
A+B	6 232	160 800	27h36m
C	5 414	189 945	35h32m

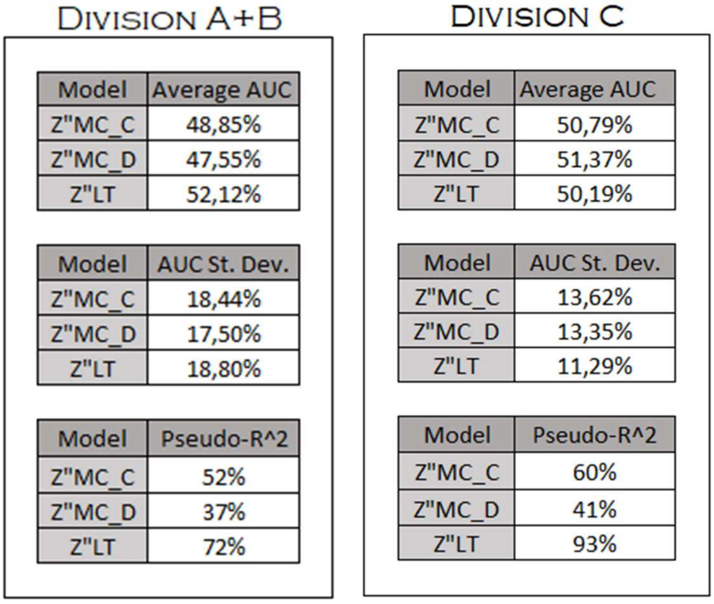
To the time expressed above adds up the unregistered time of transforming the high amounts of data into a reduced sample of predicted default probabilities.

4.2.2. General Results

The general average and variance of the predictive power of each of the models, along with calibration results are presented in Table 14. According to the average AUCs obtained for each of the models, all three seem to present very mediocre predictive power, around the AUC of a random model (which in theory would be 50%). At an initial analysis, this indicates that the application of MC simulation onto the Z”-score is at least as good of a method to create stochastic probabilities from the Z”-score as the logit transformation. What can be concluded from the standard deviation values is that the models with higher variance of the AUC generally present higher the AUC average, which could mean that the difference between the averages is consequence of higher tailed AUC distributions. This will be further developed per division to understand if this affirmation stands. As for calibration, the winning model is the logit transformation of the simple Z”-score (Z”LT) for both the divisions, by a considerable margin of at least 20%. It is interesting to compare the differences between the Pseudo-R^2 and the initial calibration given by the regression R^2, in that the former is more than 4 times larger than the maximum average obtained yearly for the linear regressions in Table 12, though the comparison is not as direct, given that the summary calibration refers to all the data, instead of just a year, let alone that the Pseudo-R^2 is only an approximation to the actual R^2 that would be calculated in a regression. Once again, however, what truly depicts if a model can or cannot distinguish default correctly is not calibration, but predictive power (Hillegeist et al., 2004; Stein, 2007; Sun et al., 2014). In that sense and, seeing as the MC simulation suffered considerable limitations to be computationally executable, this only implies that, with higher number of iterations and less limits to the model, one can expect better results off of the methodology of this study.

In terms of the direct and component approaches of the methodology, the two divisions show that different types of companies (in this case, industry specific types) may have different uses for each of the approaches, with Division C showing better results for the direct approach, while Division A+B advocates for the use of the component approach. Crossing this analysis with the respective correlation matrices of Figure 5 (which expressed higher correlation values among ratios in Division A+B as compared to Division C), one can conclude that for industries with higher correlation between the ratios, the component approach yields the best results, and that the direct approach is more useful in the absence of significant correlation.

Table 14 - General Model Performance Results



DIVISION A+B

Model	Average AUC
Z"MC_C	48,85%
Z"MC_D	47,55%
Z"LT	52,12%

Model	AUC St. Dev.
Z"MC_C	18,44%
Z"MC_D	17,50%
Z"LT	18,80%

Model	Pseudo-R^2
Z"MC_C	52%
Z"MC_D	37%
Z"LT	72%

DIVISION C

Model	Average AUC
Z"MC_C	50,79%
Z"MC_D	51,37%
Z"LT	50,19%

Model	AUC St. Dev.
Z"MC_C	13,62%
Z"MC_D	13,35%
Z"LT	11,29%

Model	Pseudo-R^2
Z"MC_C	60%
Z"MC_D	41%
Z"LT	93%

Z"MC_C → Component Approach Monte Carlo Z"-score Model
Z"MC_D → Direct Approach Monte Carlo Z"-score Model
Z"LT → Logit Transformed Z"-score

With the general results off the way, the analysis will continue below by breaking down the various division nuances that may explain (or fail to) what was concluded above.

4.2.3. Division A+B

Figure 6 gives a visual summary of the distribution differences of the AUCs of each model, for Division A+B. It presents the clear difficulty in distinguishing the predictive power of the three different models in it, which, once again, implies that the MC simulation can be just as effective in obtaining default risk results from a Z"-score as a simpler model like Z"LT.

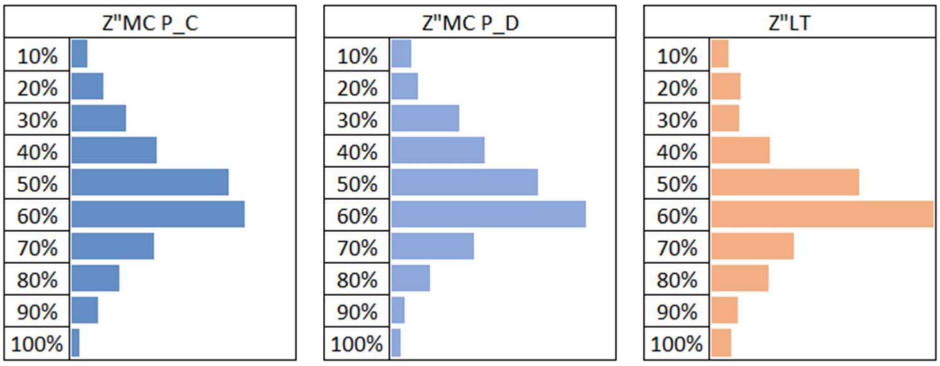
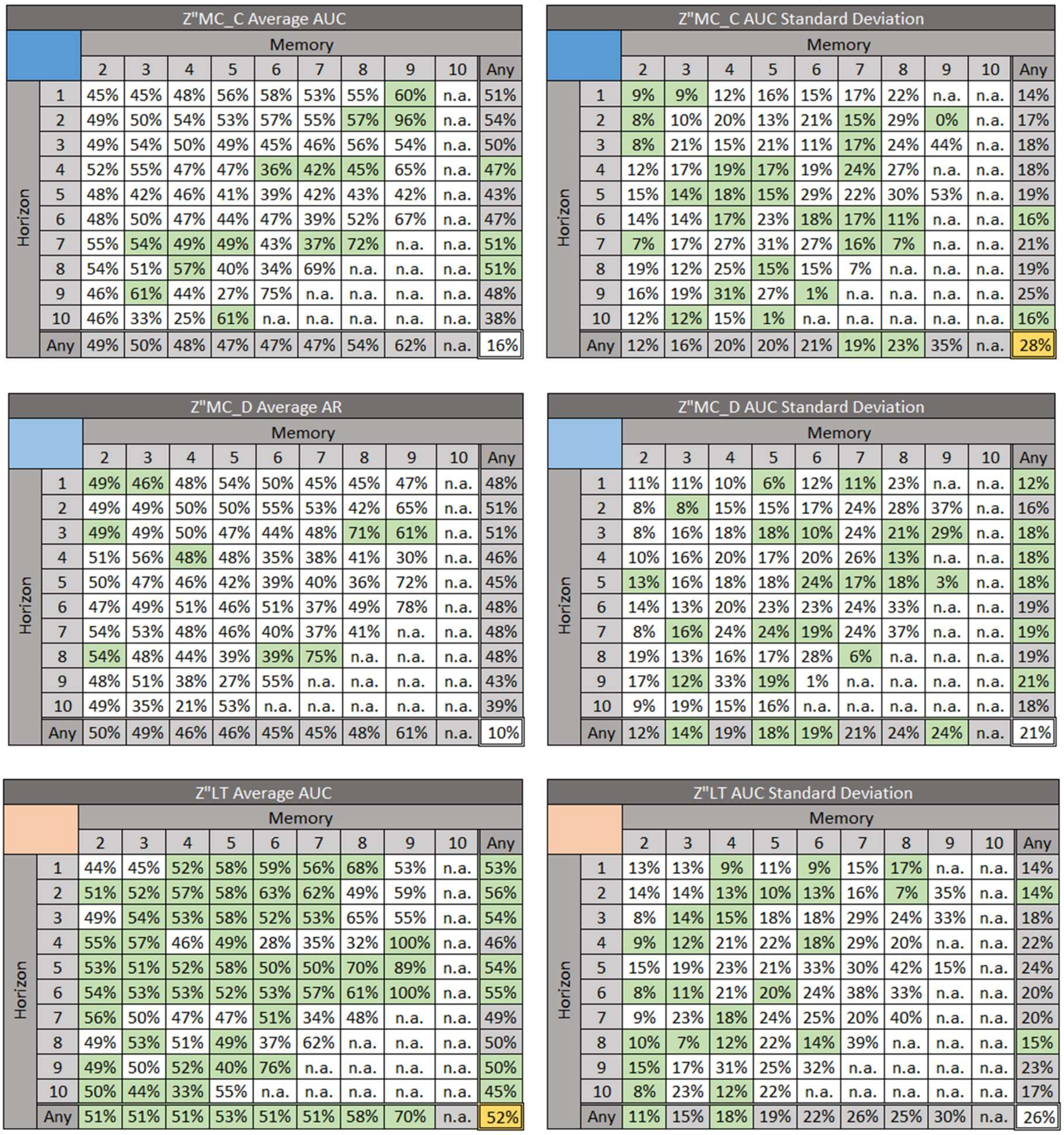


Figure 6 - AUC Distribution per Model for Division A+B

There is also a noticeable although mild difference in kurtosis between the three distributions, especially between Z"MC_C and Z"LT, meaning Z"LT classifies more companies in the 50%-60% interval than the rest of the models. Table 15 scrutinizes the differences in the average AUC and the respective standard deviation for each model. The cells painted green infer that the respective model obtained the best results out of the three for that specific combination of horizon and memory. The cell on the bottom right of each individual sub-table of Table 15 represents the percentage of combinations $h \times m$ for which the model was the best of the three.

Table 15 - Division A+B AUCs by Model, Horizon and Memory



Z"MC_C Average AUC											
		Memory									
		2	3	4	5	6	7	8	9	10	Any
Horizon	1	45%	45%	48%	56%	58%	53%	55%	60%	n.a.	51%
	2	49%	50%	54%	53%	57%	55%	57%	96%	n.a.	54%
	3	49%	54%	50%	49%	45%	46%	56%	54%	n.a.	50%
	4	52%	55%	47%	47%	36%	42%	45%	65%	n.a.	47%
	5	48%	42%	46%	41%	39%	42%	43%	42%	n.a.	43%
	6	48%	50%	47%	44%	47%	39%	52%	67%	n.a.	47%
	7	55%	54%	49%	49%	43%	37%	72%	n.a.	n.a.	51%
	8	54%	51%	57%	40%	34%	69%	n.a.	n.a.	n.a.	51%
	9	46%	61%	44%	27%	75%	n.a.	n.a.	n.a.	n.a.	48%
	10	46%	33%	25%	61%	n.a.	n.a.	n.a.	n.a.	n.a.	38%
	Any	49%	50%	48%	47%	47%	47%	54%	62%	n.a.	16%

Z"MC_C AUC Standard Deviation											
		Memory									
		2	3	4	5	6	7	8	9	10	Any
Horizon	1	9%	9%	12%	16%	15%	17%	22%	n.a.	n.a.	14%
	2	8%	10%	20%	13%	21%	15%	29%	0%	n.a.	17%
	3	8%	21%	15%	21%	11%	17%	24%	44%	n.a.	18%
	4	12%	17%	19%	17%	19%	24%	27%	n.a.	n.a.	18%
	5	15%	14%	18%	15%	29%	22%	30%	53%	n.a.	19%
	6	14%	14%	17%	23%	18%	17%	11%	n.a.	n.a.	16%
	7	7%	17%	27%	31%	27%	16%	7%	n.a.	n.a.	21%
	8	19%	12%	25%	15%	15%	7%	n.a.	n.a.	n.a.	19%
	9	16%	19%	31%	27%	1%	n.a.	n.a.	n.a.	n.a.	25%
	10	12%	12%	15%	1%	n.a.	n.a.	n.a.	n.a.	n.a.	16%
	Any	12%	16%	20%	20%	21%	19%	23%	35%	n.a.	28%

Z"MC_D Average AR											
		Memory									
		2	3	4	5	6	7	8	9	10	Any
Horizon	1	49%	46%	48%	54%	50%	45%	45%	47%	n.a.	48%
	2	49%	49%	50%	50%	55%	53%	42%	65%	n.a.	51%
	3	49%	49%	50%	47%	44%	48%	71%	61%	n.a.	51%
	4	51%	56%	48%	48%	35%	38%	41%	30%	n.a.	46%
	5	50%	47%	46%	42%	39%	40%	36%	72%	n.a.	45%
	6	47%	49%	51%	46%	51%	37%	49%	78%	n.a.	48%
	7	54%	53%	48%	46%	40%	37%	41%	n.a.	n.a.	48%
	8	54%	48%	44%	39%	39%	75%	n.a.	n.a.	n.a.	48%
	9	48%	51%	38%	27%	55%	n.a.	n.a.	n.a.	n.a.	43%
	10	49%	35%	21%	53%	n.a.	n.a.	n.a.	n.a.	n.a.	39%
	Any	50%	49%	46%	46%	45%	45%	48%	61%	n.a.	10%

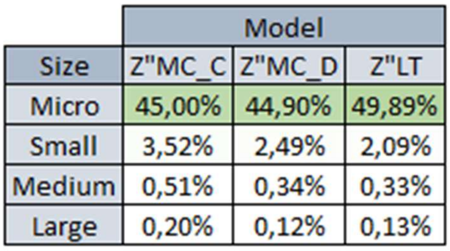
Z"MC_D AUC Standard Deviation											
		Memory									
		2	3	4	5	6	7	8	9	10	Any
Horizon	1	11%	11%	10%	6%	12%	11%	23%	n.a.	n.a.	12%
	2	8%	8%	15%	15%	17%	24%	28%	37%	n.a.	16%
	3	8%	16%	18%	18%	10%	24%	21%	29%	n.a.	18%
	4	10%	16%	20%	17%	20%	26%	13%	n.a.	n.a.	18%
	5	13%	16%	18%	18%	24%	17%	18%	3%	n.a.	18%
	6	14%	13%	20%	23%	23%	24%	33%	n.a.	n.a.	19%
	7	8%	16%	24%	24%	19%	24%	37%	n.a.	n.a.	19%
	8	19%	13%	16%	17%	28%	6%	n.a.	n.a.	n.a.	19%
	9	17%	12%	33%	19%	1%	n.a.	n.a.	n.a.	n.a.	21%
	10	9%	19%	15%	16%	n.a.	n.a.	n.a.	n.a.	n.a.	18%
	Any	12%	14%	19%	18%	19%	21%	24%	24%	n.a.	21%

Z"LT Average AUC											
		Memory									
		2	3	4	5	6	7	8	9	10	Any
Horizon	1	44%	45%	52%	58%	59%	56%	68%	53%	n.a.	53%
	2	51%	52%	57%	58%	63%	62%	49%	59%	n.a.	56%
	3	49%	54%	53%	58%	52%	53%	65%	55%	n.a.	54%
	4	55%	57%	46%	49%	28%	35%	32%	100%	n.a.	46%
	5	53%	51%	52%	58%	50%	50%	70%	89%	n.a.	54%
	6	54%	53%	53%	52%	53%	57%	61%	100%	n.a.	55%
	7	56%	50%	47%	47%	51%	34%	48%	n.a.	n.a.	49%
	8	49%	53%	51%	49%	37%	62%	n.a.	n.a.	n.a.	50%
	9	49%	50%	52%	40%	76%	n.a.	n.a.	n.a.	n.a.	50%
	10	50%	44%	33%	55%	n.a.	n.a.	n.a.	n.a.	n.a.	45%
	Any	51%	51%	51%	53%	51%	51%	58%	70%	n.a.	52%

Z"LT AUC Standard Deviation											
		Memory									
		2	3	4	5	6	7	8	9	10	Any
Horizon	1	13%	13%	9%	11%	9%	15%	17%	n.a.	n.a.	14%
	2	14%	14%	13%	10%	13%	16%	7%	35%	n.a.	14%
	3	8%	14%	15%	18%	18%	29%	24%	33%	n.a.	18%
	4	9%	12%	21%	22%	18%	29%	20%	n.a.	n.a.	22%
	5	15%	19%	23%	21%	33%	30%	42%	15%	n.a.	24%
	6	8%	11%	21%	20%	24%	38%	33%	n.a.	n.a.	20%
	7	9%	23%	18%	24%	25%	20%	40%	n.a.	n.a.	20%
	8	10%	7%	12%	22%	14%	39%	n.a.	n.a.	n.a.	15%
	9	15%	17%	31%	25%	32%	n.a.	n.a.	n.a.	n.a.	23%
	10	8%	23%	12%	22%	n.a.	n.a.	n.a.	n.a.	n.a.	17%
	Any	11%	15%	18%	19%	22%	26%	25%	30%	n.a.	26%

It is once again clear from Table 15 that Z”LT holds the highest AUC for most combinations for Division A+B, and the highest AUC cells are too spread out on the rest of the models to allow for a separation of model quality by horizon and memory. This may also be due to the 100 iterations limitation, which leaves the model AUC distribution more prone to randomness than the actual real effects of using MC simulation properly. The deviations suffer from the same interpretation issue – there is no clear distinction in what cases a model is better than the other, only that Z”MC_D obtained the least deviation of all, by a smaller margin than the Z”LT did for the average of AUC.

Table 16 - Model Average AUC by Size (Division A+B)



	Model		
Size	Z"MC_C	Z"MC_D	Z"LT
Micro	45,00%	44,90%	49,89%
Small	3,52%	2,49%	2,09%
Medium	0,51%	0,34%	0,33%
Large	0,20%	0,12%	0,13%

Lastly, AUC averages are analyzed by company size in Table 16, which depicts for the majority (micro companies) exactly the same as the total average, with Z”LT leading the results, and Z”MC_D having the lowest ratings. Interestingly enough, for all companies other than micro, though with much smaller sample and, hence, certainty, Z”MC_C leads the results, which implies that MC simulated Z”-scores may overpower simpler estimations for companies larger than micro.

4.2.4. Division C

Figure 7 gives a visual summary of the distribution differences of the AUCs of each model, now for Division C. Similarly to the general analysis and the Division A+B, the AUC distributions are almost indistinguishable from each other, even more than in the previous division.

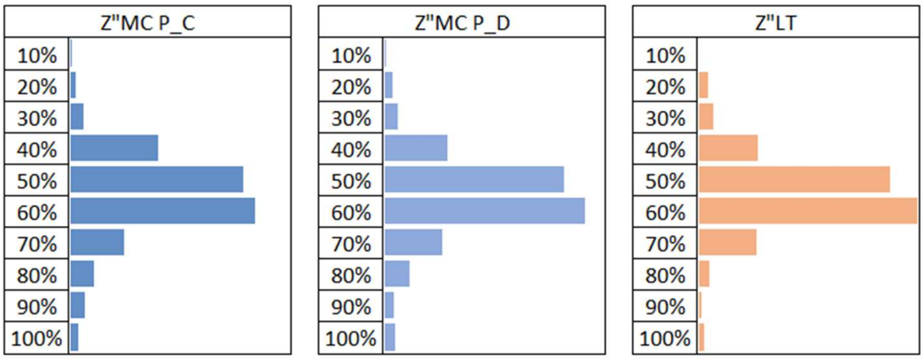
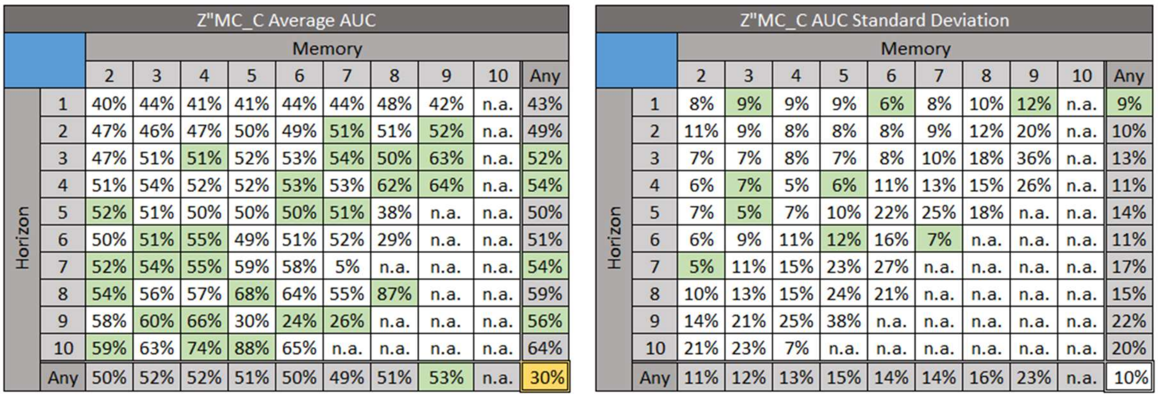


Figure 7 - AUC Distribution per Model for Division C

Table 17 presents the AUC averages per combination of $h \times m$, once more in an attempt to better understand predictive power variations. Yet again, the cells painted green infer that the respective model obtained the best results out of the three for that specific combination of horizon and memory; and the cell on the bottom right of each individual sub-table of Table 17 represents the percentage of combinations $h \times m$ for which the model was the best of the three. In the Division C case, although the Z"MC_D obtained better overall average results, the Z"MC_C obtained better averages for most combinations of $h \times m$, though the distribution is, once again, quite random and difficult to determine any patterns. As for the standard deviation of the AUCs, the method with the least variation of results is the Z"LT, on the absolute contrary to Division A+B.

Table 17 - Division C AUCs by Model, Horizon and Memory

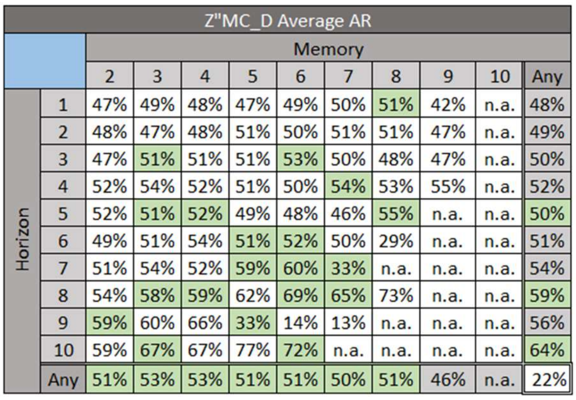


Z"MC_C Average AUC

		Memory									
		2	3	4	5	6	7	8	9	10	Any
Horizon	1	40%	44%	41%	41%	44%	44%	48%	42%	n.a.	43%
	2	47%	46%	47%	50%	49%	51%	51%	52%	n.a.	49%
	3	47%	51%	51%	52%	53%	54%	50%	63%	n.a.	52%
	4	51%	54%	52%	52%	53%	53%	62%	64%	n.a.	54%
	5	52%	51%	50%	50%	50%	51%	38%	n.a.	n.a.	50%
	6	50%	51%	55%	49%	51%	52%	29%	n.a.	n.a.	51%
	7	52%	54%	55%	59%	58%	5%	n.a.	n.a.	n.a.	54%
	8	54%	56%	57%	68%	64%	55%	87%	n.a.	n.a.	59%
	9	58%	60%	66%	30%	24%	26%	n.a.	n.a.	n.a.	56%
	10	59%	63%	74%	88%	65%	n.a.	n.a.	n.a.	n.a.	64%
	Any	50%	52%	52%	51%	50%	49%	51%	53%	n.a.	30%

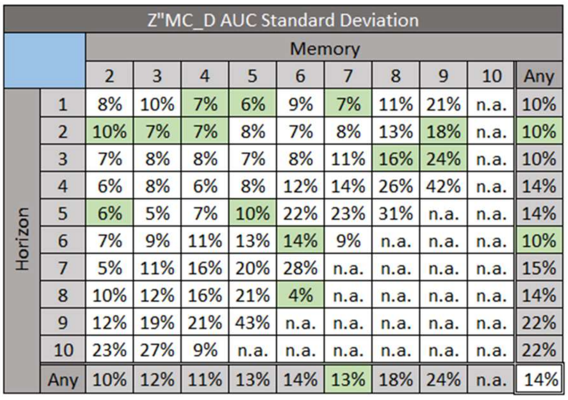
Z"MC_C AUC Standard Deviation

		Memory									
		2	3	4	5	6	7	8	9	10	Any
Horizon	1	8%	9%	9%	9%	6%	8%	10%	12%	n.a.	9%
	2	11%	9%	8%	8%	8%	9%	12%	20%	n.a.	10%
	3	7%	7%	8%	7%	8%	10%	18%	36%	n.a.	13%
	4	6%	7%	5%	6%	11%	13%	15%	26%	n.a.	11%
	5	7%	5%	7%	10%	22%	25%	18%	n.a.	n.a.	14%
	6	6%	9%	11%	12%	16%	7%	n.a.	n.a.	n.a.	11%
	7	5%	11%	15%	23%	27%	n.a.	n.a.	n.a.	n.a.	17%
	8	10%	13%	15%	24%	21%	n.a.	n.a.	n.a.	n.a.	15%
	9	14%	21%	25%	38%	n.a.	n.a.	n.a.	n.a.	n.a.	22%
	10	21%	23%	7%	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	20%
	Any	11%	12%	13%	15%	14%	14%	16%	23%	n.a.	10%



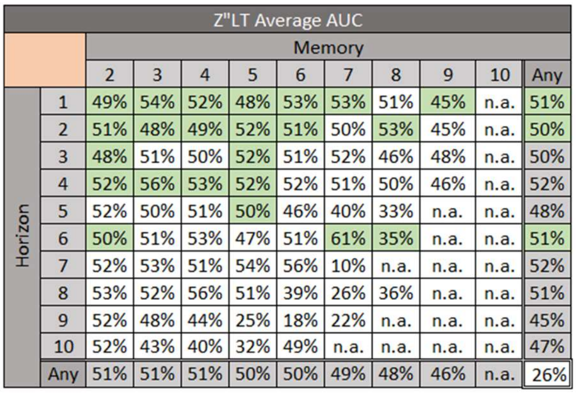
Z"MC_D Average AR

		Memory									
		2	3	4	5	6	7	8	9	10	Any
Horizon	1	47%	49%	48%	47%	49%	50%	51%	42%	n.a.	48%
	2	48%	47%	48%	51%	50%	51%	51%	47%	n.a.	49%
	3	47%	51%	51%	51%	53%	50%	48%	47%	n.a.	50%
	4	52%	54%	52%	51%	50%	54%	53%	55%	n.a.	52%
	5	52%	51%	52%	49%	48%	46%	55%	n.a.	n.a.	50%
	6	49%	51%	54%	51%	52%	50%	29%	n.a.	n.a.	51%
	7	51%	54%	52%	59%	60%	33%	n.a.	n.a.	n.a.	54%
	8	54%	58%	59%	62%	69%	65%	73%	n.a.	n.a.	59%
	9	59%	60%	66%	33%	14%	13%	n.a.	n.a.	n.a.	56%
	10	59%	67%	67%	77%	72%	n.a.	n.a.	n.a.	n.a.	64%
	Any	51%	53%	53%	51%	51%	50%	51%	46%	n.a.	22%



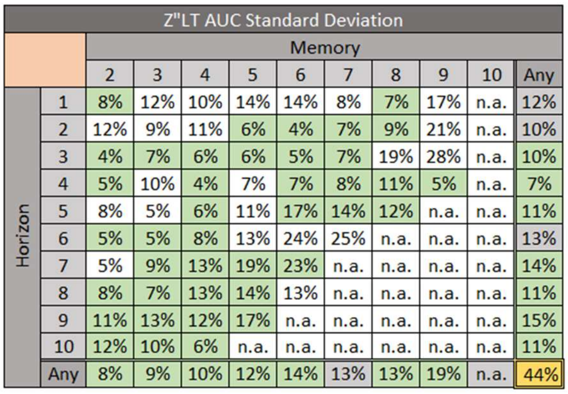
Z"MC_D AUC Standard Deviation

		Memory									
		2	3	4	5	6	7	8	9	10	Any
Horizon	1	8%	10%	7%	6%	9%	7%	11%	21%	n.a.	10%
	2	10%	7%	7%	8%	7%	8%	13%	18%	n.a.	10%
	3	7%	8%	8%	7%	8%	11%	16%	24%	n.a.	10%
	4	6%	8%	6%	8%	12%	14%	26%	42%	n.a.	14%
	5	6%	5%	7%	10%	22%	23%	31%	n.a.	n.a.	14%
	6	7%	9%	11%	13%	14%	9%	n.a.	n.a.	n.a.	10%
	7	5%	11%	16%	20%	28%	n.a.	n.a.	n.a.	n.a.	15%
	8	10%	12%	16%	21%	4%	n.a.	n.a.	n.a.	n.a.	14%
	9	12%	19%	21%	43%	n.a.	n.a.	n.a.	n.a.	n.a.	22%
	10	23%	27%	9%	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	22%
	Any	10%	12%	11%	13%	14%	13%	18%	24%	n.a.	14%



Z"LT Average AUC

		Memory									
		2	3	4	5	6	7	8	9	10	Any
Horizon	1	49%	54%	52%	48%	53%	53%	51%	45%	n.a.	51%
	2	51%	48%	49%	52%	51%	50%	53%	45%	n.a.	50%
	3	48%	51%	50%	52%	51%	52%	46%	48%	n.a.	50%
	4	52%	56%	53%	52%	52%	51%	50%	46%	n.a.	52%
	5	52%	50%	51%	50%	46%	40%	33%	n.a.	n.a.	48%
	6	50%	51%	53%	47%	51%	61%	35%	n.a.	n.a.	51%
	7	52%	53%	51%	54%	56%	10%	n.a.	n.a.	n.a.	52%
	8	53%	52%	56%	51%	39%	26%	36%	n.a.	n.a.	51%
	9	52%	48%	44%	25%	18%	22%	n.a.	n.a.	n.a.	45%
	10	52%	43%	40%	32%	49%	n.a.	n.a.	n.a.	n.a.	47%
	Any	51%	51%	51%	50%	50%	49%	48%	46%	n.a.	26%

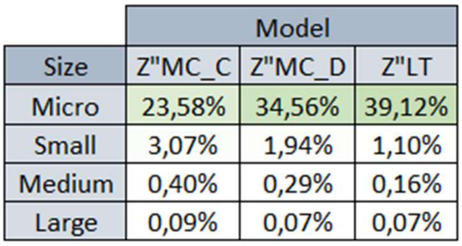


Z"LT AUC Standard Deviation

		Memory									
		2	3	4	5	6	7	8	9	10	Any
Horizon	1	8%	12%	10%	14%	14%	8%	7%	17%	n.a.	12%
	2	12%	9%	11%	6%	4%	7%	9%	21%	n.a.	10%
	3	4%	7%	6%	6%	5%	7%	19%	28%	n.a.	10%
	4	5%	10%	4%	7%	7%	8%	11%	5%	n.a.	7%
	5	8%	5%	6%	11%	17%	14%	12%	n.a.	n.a.	11%
	6	5%	5%	8%	13%	24%	25%	n.a.	n.a.	n.a.	13%
	7	5%	9%	13%	19%	23%	n.a.	n.a.	n.a.	n.a.	14%
	8	8%	7%	13%	14%	13%	n.a.	n.a.	n.a.	n.a.	11%
	9	11%	13%	12%	17%	n.a.	n.a.	n.a.	n.a.	n.a.	15%
	10	12%	10%	6%	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	11%
	Any	8%	9%	10%	12%	14%	13%	13%	19%	n.a.	44%

As for the final evaluation of the AUC averages by company size as presented in Table 18, the results mimic the behaviors found in Division A+B, having the best model for micros be the Z”LT, while the Z”MC_C yields the best results for companies of small size or above.

Table 18 - Model Average AUC by Size (Division C)



	Model		
Size	Z"MC_C	Z"MC_D	Z"LT
Micro	23,58%	34,56%	39,12%
Small	3,07%	1,94%	1,10%
Medium	0,40%	0,29%	0,16%
Large	0,09%	0,07%	0,07%

5. Conclusions and Suggestions

This study aimed to introduce a new application of the MC simulation into the widely applicable statistical DPM methodology – Z”-score (Altman, 1983) – and test its predictive power under various parameterizations against a simple and widely used transformation of this type of scores into stochastic default probabilities (Agarwal & Taffler, 2008; Correia et al., 2017; Hillegeist et al., 2004; Jackson & Wood, 2013; Reisz & Perlich, 2007).

With the results of this study, it can be concluded that utilizing Monte Carlo simulation to augment a model such as the Z”-score of Altman (1983) allows it to be transformed into a stochastic model with at least as much predictive power as its simple logit transformation (Z”LT). The obtained and compared averages of ROC curve AUCs for the logit transformed Z”-score model (Z”LT) and the Monte Carlo simulated Z”-score models – the direct approach (Z”MC_D) and the component approach (Z”MC_C) – are relatively similar for the methodology utilized, and only for Division A+B does Z”LT obtain the best results. Since MC simulation has the tendency to improve the higher the number of iterations (Andrieu et al., 2003; Borgonovo & Gatti, 2013; James, 1980), then the results only imply the predictive power obtained by the Z”LT model as the minimum for this methodology.

The results of this study also indicate that the direct approach of the MC simulation model is better applied under low correlation between ratios, while the component approach yields better results for highly correlated ratio data.

In terms of company size disparities in the results, this study finds little evidence, though one that may be interesting to analyze, that the Monte Carlo simulation approach can perform consistently better than the simpler stochastic transformation methods (Z”LT) when applied to companies larger than those considered micro in Portugal.

There were various limitations to the present application of the Z”MC models in this study, mostly due to computational costs and limits. Applying the analysis in a faster and more capable computer or software could allow the it to yield better and more reliable results, as Monte Carlo Simulation improves the higher the iteration number. This would allow the 100-iteration limit to be extended, along with more variation to the model being possible to test. Furthermore, the act of applying this methodology outside Excel would not invalidate its use on it, since it would simply be to accelerate the processing of the high number of companies. Also due to the extra computational power and time it would require, the cut-off of zero to transform the MC scenario samples into a probability of default was not evaluated as it should, and it may very well influence the results considerably. Last but not least, under better computational conditions, the industries that were not tested could easily be added into the analysis, or at least it would be interesting to have them be studied for the results of this methodology.

An additional augment to the analysis performed on this study would be the application of the Z"MC abroad, to analyze the variations of its application in different countries with different market conditions.

Beyond sample and computational power, a limitation that this study found was the generation of correlated variables without the assumption of normality, which could be an interesting point of further development, let it be in terms of generating other distribution types with the intended correlation, or simply ignoring the correlation all together to analyze how that would affect the results.

To finalize, our use of the multiple linear regression was merely a practical simplification of the recalibration process of an MDA methodology and recalibrating it utilizing the original model type could prove more advantageous, or even choosing different methodologies altogether, such as the logit model.

Bibliography

Abinzano, I., Corredor, P., & Martinez, B. (2020). Measuring credit risk in family firms. *BRQ Business Research Quarterly*, *25*(3), 265–282. https://doi.org/10.1177/2340944420941857

Adams, J., Hayunga, D., Mansi, S., Reeb, D., & Verardi, V. (2019). Identifying and treating outliers in finance. *Financial Management*, *48*(2), 345–384. https://doi.org/10.1111/fima.12269

Agarwal, V., & Taffler, R. (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking and Finance*, *32*(8), 1541–1551. https://doi.org/10.1016/j.jbankfin.2007.07.014

Alaka, H., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O., & Bilal, M. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems With Applications*, *94*, 164–184. https://doi.org/10.1016/j.eswa.2017.10.040

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, *23*(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Altman, E. I. (1973). Predicting Railroad Bankruptcies in America. *The Bell Journal of Economics and Management Science*, *4*(1), 184–211. https://doi.org/10.2307/3003144

Altman, E., I. (1983). *Corporate Financial Distress: A Complete Guide to Predicting, Avoiding, and Dealing with Bankruptcy (Frontiers in Finance Series)* (1st ed.). Wiley.

Altman, E. I., Haldeman, R. G., & Narayanan, P. (1977). ZETA^{TM} analysis: A new model to identify bankruptcy risk of corporations. *Journal of Banking &Amp; Finance*, *1*(1), 29–54. https://doi.org/10.1016/0378-4266(77)90017-6

Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., & Suvas, A. (2016). Financial Distress Prediction in an International Context: A Review and Empirical Analysis of Altman's *Z*- Score Model. *Journal of International Financial Management and Accounting*, *28*(2), 131–171. https://doi.org/10.1111/jifm.12053

Álvarez-Díaz, M., Villanueva-Villar, M., & Rivo-López, E. (2022). Unraveling the profile of domestic cultural travelers in Spain: An analysis based on bootstrap p values. *Journal of Hospitality and Tourism Management*, *51*, 397–400. https://doi.org/10.1016/j.jhtm.2022.04.012

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An Introduction to MCMC for Machine Learning. *Machine Learning*, *50*(1/2), 5–43. https://doi.org/10.1023/a:1020281327116

Aziz, M. A., & Dar, H. A. (2006). Predicting corporate bankruptcy: where we stand? *Corporate Governance*, *6*(1), 18–33. https://doi.org/10.1108/14720700610649436

Bakshi, G., Gao, X., & Zhong, Z. (2022). Decoding Default Risk: A review of modeling approaches, findings, and estimation methods. *Annual Review of Financial Economics*, *14*(1), 391–413. https://doi.org/10.1146/annurev-financial-111720-090709

Balcaen, S., & Ooghe, H. (2006). 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. *British Accounting Review*, *38*(1), 63–93. https://doi.org/10.1016/j.bar.2005.09.001

Beaver, W. H. (1966). Financial Ratios As Predictors of Failure. *Journal of Accounting Research*, *4*, 71–112. https://doi.org/10.2307/2490171

Beaver, W. H. (1968). Market Prices, Financial Ratios, and the Prediction of Failure. *Journal of Accounting Research*, *6*(2), 179–192. https://doi.org/10.2307/2490233

Beaver, W. H., Correia, M., & McNichols, M. F. (2010). Financial Statement Analysis and the Prediction of Financial Distress. *Foundations and Trends in Accounting*, *5*(2), 99–173. https://doi.org/10.1561/1400000018

Beaver, W. H., McNichols, M. F., & Rhie, J. W. (2005). Have Financial Statements Become Less Informative? Evidence from the Ability of Financial Ratios to Predict Bankruptcy. *Review of Accounting Studies*, *10*(1), 93–122. https://doi.org/10.1007/s11142-004-6341-9

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., . . . Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Bharath, S. T., & Shumway, T. (2008). Forecasting Default with the Merton Distance to Default Model. *Review of Financial Studies*, *21*(3), 1339–1369. https://doi.org/10.1093/rfs/hhn044

Black, F., & Scholes, M. S. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, *81*(3), 637–654. https://doi.org/10.1086/260062

Blanchet, J., Li, J., & Shi, Y. (2015). Stochastic Risk Networks: Modeling, analysis and Efficient Monte Carlo. *Social Science Research Network*. https://doi.org/10.2139/ssrn.2012987

Blum, M. (1974). Failing Company Discriminant Analysis. *Journal of Accounting Research*, *12*(1), 1–25. https://doi.org/10.2307/2490525

Bonfim, D. (2009). Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics. *Journal of Banking &Amp; Finance*, *33*(2), 281–299. https://doi.org/10.1016/j.jbankfin.2008.08.006

Borgonovo, E., & Gatti, S. (2013). Risk analysis with contractual default. Does covenant breach matter? *European Journal of Operational Research*, *230*(2), 431–443. https://doi.org/10.1016/j.ejor.2013.04.047

Bottazzi, G., Grazzi, M., Secchi, A., & Tamagni, F. (2011). Financial and economic determinants of firm default. *Journal of Evolutionary Economics*, *21*(3), 373–406. https://doi.org/10.1007/s00191-011-0224-6

Bukaçi, E., Korini, T., Periku, E., Allkja, S., & Sheperi, P. (2016). Number of iterations needed in Monte Carlo Simulation using reliability analysis for tunnel supports. *Int. Journal of Engineering Research And Applications*, *6*(6), Part-3. https://www.researchgate.net/publication/304331698_Number_of_Iterations_Needed_in_Monte_Carlo_Simulation_Using_Reliability_Analysis_for_Tunnel_Supports

Campbell, J. Y., Hilscher, J., & Szilagy, J. (2008). In Search of Distress Risk. *The Journal of Finance*, *63*(6), 2899–2939. https://doi.org/10.1111/j.1540-6261.2008.01416.x

CEIC. (2022, August 23). *Portugal*. CEIC Data. Retrieved December 13, 2022, from https://www.ceicdata.com/

Chava, S., & Jarrow, R. A. (2004). Bankruptcy Prediction with Industry Effects. *Review of Finance*, *8*(4), 537–569. https://doi.org/10.1093/rof/8.4.537

Chen, H. (2022). Prediction and Analysis of Financial Default Loan Behavior Based on Machine Learning Model. *Computational Intelligence and Neuroscience*, *2022*, 1–10. https://doi.org/10.1155/2022/7907210

Correia, M., Kang, J., & Richardson, S. (2017). Asset volatility. *Review of Accounting Studies*, *23*(1), 37–94. https://doi.org/10.1007/s11142-017-9431-1

Costa, M., Lisboa, I., & Gameiro, A. (2022). Is the Financial Report Quality Important in the Default Prediction? SME Portuguese Construction Sector Evidence. *Risks*, *10*(5), 98. https://doi.org/10.3390/risks10050098

Cox. (1972). Regression models and Life-Tables. *Journal of the Royal Statistical Society Series B-methodological*, *34*(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Dambolena, I. G., & Khoury, S. J. (1980). Ratio Stability and Corporate Failure. *The Journal of Finance*, *35*(4), 1017–1026. https://doi.org/10.1111/j.1540-6261.1980.tb03517.x

Deakin, E. B. (1972). A Discriminant Analysis of Predictors of Business Failure. *Journal of Accounting Research*, *10*(1), 167–179. https://doi.org/10.2307/2490225

Dichev, I. D. (1998). Is the Risk of Bankruptcy a Systematic Risk? *The Journal of Finance*, *53*(3), 1131–1147. https://doi.org/10.1111/0022-1082.00046

Dimitras, A., Zanakis, S., & Zopounidis, C. (1996). A survey of business failures with an emphasis on prediction methods and industrial applications. *European Journal of Operational Research*, *90*(3), 487–513. https://doi.org/10.1016/0377-2217(95)00070-4

Du, X., Li, W., Ruan, S., & Li, L. (2020). CUS-heterogeneous ensemble-based financial distress prediction for imbalanced dataset with ensemble feature selection. *Applied Soft Computing*, *97*, 106758. https://doi.org/10.1016/j.asoc.2020.106758

Duffie, D., & Glynn, P. (1995). Efficient Monte Carlo Simulation of Security Prices. *The Annals of Applied Probability*, *5*(4), 897–905. https://doi.org/10.1214/aoap/1177004598

Duffie, D., Saita, L., & Wang, K. (2007). Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, *83*(3), 635–665. https://doi.org/10.1016/j.jfineco.2005.10.011

Edmister, R. O. (1972). An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction. *The Journal of Financial and Quantitative Analysis*, *7*(2), 1477–1493. https://doi.org/10.2307/2329929

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Efron, B. (1982). *The Jackknife, the Bootstrap, and other resampling plans*. SIAM.

Efron, B., & Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, *1*(1), 54–75. https://doi.org/10.1214/ss/1177013815

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap* (1st ed.). Chapman and Hall/CRC.

Emery, G. W., & Cogger, K. O. (1982). The Measurement of Liquidity. *Journal of Accounting Research*, *20*(2), 290–303. https://doi.org/10.2307/2490741

Eren, Y., Kucukdemiral, I. B., & Ustoglu, I. (2017). Introduction to optimization. In *Elsevier eBooks* (pp. 27–74). https://doi.org/10.1016/b978-0-08-101041-9.00002-8

Franklin, L. A., & Gary, W. (1991). Bootstrap confidence interval estimates of c_{pk}: an introduction. *Communications in Statistics - Simulation and Computation*, *20*(1), 231–242. https://doi.org/10.1080/03610919108812950

Giesecke, K., Kakavand, H., Mousavi, M., & Takada, H. (2010). Exact and Efficient Simulation of Correlated Defaults. *SIAM Journal on Financial Mathematics*, *1*(1), 868–896. https://doi.org/10.1137/090778055

Giesecke, K., Kim, B., & Zhu, S. (2011). Monte Carlo algorithms for default timing problems. *Management Science*, *57*(12), 2115–2129. https://doi.org/10.1287/mnsc.1110.1411

Gilleland, E. (2020). Bootstrap Methods for Statistical Inference. Part II: Extreme-Value Analysis. *Journal of Atmospheric and Oceanic Technology*, *37*(11), 2135–2144. https://doi.org/10.1175/jtech-d-20-0070.1

Gleißner, W. (2019). Cost of capital and probability of default in value-based risk management. *Management Research Review*, *42*(11), 1243–1258. https://doi.org/10.1108/mrr-11-2018-0456

Guilmeau, T., Chouzenoux, E., & Elvira, V. (2021). Simulated Annealing: a Review and a New Scheme. *2021 IEEE Statistical Signal Processing Workshop (SSP)*. https://doi.org/10.1109/ssp49050.2021.9513782

Hillegeist, S. A., Keating, E. K., Cram, D. P., & Lundstedt, K. G. (2004). Assessing the Probability of Bankruptcy. *Review of Accounting Studies*, *9*(1), 5–34. https://doi.org/10.1023/b:rast.0000013627.90884.b7

Hosseini, S. R., & Hajiannejad, A. (2022). The Effect of Statistical Distribution of Financial Ratios on Altman Model Values using Monte Carlo Simulation. *Journal of Accounting Knowledge*, *13*(3), 161–193. https://doi.org/10.22103/jak.2022.18028.3545

INE & Pordata. (2021). *Small and medium-sized enterprises: total and by size*. Pordata. Retrieved February 28, 2023, from https://www.pordata.pt/en/portugal/small+and+medium+sized+enterprises+total+and+by+size-2927

Innocenti, S., Matte, P., Fortin, V., & Bernier, N. (2022). Analytical and Residual Bootstrap Methods for Parameter Uncertainty Assessment in Tidal Analysis with

Temporally Correlated Noise. *Journal of Atmospheric and Oceanic Technology*, *39*(10), 1457–1481. https://doi.org/10.1175/jtech-d-21-0060.1

Jackson, R. H., & Wood, A. (2013). The performance of insolvency prediction and credit risk models in the UK: A comparative study. *The British Accounting Review*, *45*(3), 183–202. https://doi.org/10.1016/j.bar.2013.06.009

James, F. (1980). Monte Carlo theory and practice. *Reports on Progress in Physics*, *43*(9), 1145–1189. https://doi.org/10.1088/0034-4885/43/9/002

Kaneko, K., Miyamoto, K., Takeda, N., & Yoshino, K. (2021). Quantum speedup of Monte Carlo integration with respect to the number of dimensions and its application to finance. *Quantum Information Processing*, *20*(5), 1–24. https://doi.org/10.1007/s11128-021-03127-8

Ketz, J. E. (1978). The Effect of General Price-Level Adjustments on the Predictive Ability of Financial Ratios. *Journal of Accounting Research*, *16*, 273–284. https://doi.org/10.2307/2490438

Kim, H., Cho, H., & Ryu, D. (2020). Corporate Default Predictions Using Machine Learning: Literature Review. *Sustainability*, *12*(16), 6325. https://doi.org/10.3390/su12166325

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science*, *220*(4598), 671–680. https://doi.org/10.1126/science.220.4598.671

Korteweg, A. G. (2011). Markov Chain Monte Carlo methods in Corporate finance. *Social Science Research Network*. https://doi.org/10.2139/ssrn.1964923

Kulesa, A., Krzywinski, M., Blainey, P., & Altman, N. (2015). Sampling distributions and the bootstrap. *Nature Methods*, *12*(6), 477–478. https://doi.org/10.1038/nmeth.3414

Kwak, Y. H., & Ingall, L. (2007). Exploring Monte Carlo Simulation Applications for Project Management. *Risk Management*, *9*(1), 44–57. https://doi.org/10.1057/palgrave.rm.8250017

Kwon, T. Y., & Lee, Y. (2018). Industry specific defaults. *Journal of Empirical Finance*, *45*, 45–58. https://doi.org/10.1016/j.jempfin.2017.10.002

Liang, J., Zhou, P., Zhou, Y., & Ma, J. (2011). Valuation of Credit Default Swap with Counterparty Default Risk by Structural Model. *Applied Mathematics*, *2*(1), 106–117. https://doi.org/10.4236/am.2011.21012

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, 105–142. https://ci.nii.ac.jp/naid/10007461386

Mensah, Y. M. (1984). An Examination of the Stationarity of Multivariate Bankruptcy Prediction Models: A Methodological Study. *Journal of Accounting Research*, *22*(1), 380–395. https://doi.org/10.2307/2490719

Merton, R. C. (1974). ON THE PRICING OF CORPORATE DEBT: THE RISK STRUCTURE OF INTEREST RATES*. *Journal of Finance*, *29*(2), 449–470. https://doi.org/10.1111/j.1540-6261.1974.tb03058.x

Metropolis, N., & Ulam, S. M. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, *44*(247), 335–341. https://doi.org/10.1080/01621459.1949.10483310

Mitton, T. (2021). Methodological Variation in Empirical Corporate Finance. *The Review of Financial Studies*, *35*(2), 527–575. https://doi.org/10.1093/rfs/hhab030

Nguyen, H. (2023). An empirical application of Particle Markov Chain Monte Carlo to frailty correlated default models. *Journal of Empirical Finance*, *72*, 103–121. https://doi.org/10.1016/j.jempfin.2023.03.003

Oberle, W. F. (2015). Monte Carlo Simulations: Number of Iterations and Accuracy. *US Army Research Laboratory*, 1–34. https://apps.dtic.mil/dtic/tr/fulltext/u2/a621501.pdf

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, *18*(1), 109–131. https://doi.org/10.2307/2490395

O'Leary, D. E. (1998). Using Neural Networks to Predict Corporate Failure. *Intelligent Systems in Accounting, Finance and Management*, *7*(3), 187–197. https://doi.org/10.1002/(SICI)1099-1174(199809)7:3<187::AID-ISAF144>3.0.CO;2-7

Peláez, R., Cao, R., & Vilar, J. M. (2022). Bootstrap Bandwidth Selection and Confidence Regions for Double Smoothed Default Probability Estimation. *Mathematics*, *10*(9), 1523–1547. https://doi.org/10.3390/math10091523

Portugal 2020. (2021, November 2). *PME Pequenas e Médias Empresas - PT2020*. https://portugal2020.pt/glossario/pme-pequenas-e-medias-empresas/

Raftery, A. E., & Lewis, S. M. (1970). The Number of Iterations, Convergence Diagnostics and Generic Metropolis Algorithms. *ResearchGate*. https://www.researchgate.net/publication/2598914

Ramesh, A., & Kumar, C. B. S. (2017). Structure and Intensity Based Approach in Credit Risk Models: A Literature Review. *International Journal of Economics and Financial Issues*, *7*(3), 609–612. https://www.econjournals.com/index.php/ijefi/article/download/4721/pdf/12883

Ravi Kumar, P., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. *European Journal of Operational Research*, *180*(1), 1–28. https://doi.org/10.1016/j.ejor.2006.08.043

Reisz, A. S., & Perlich, C. (2007). A market-based framework for bankruptcy prediction. *Journal of Financial Stability*, *3*(2), 85–131. https://doi.org/10.1016/j.jfs.2007.02.001

Riemer, S., & Wagner, C. (2015). An efficient approach for calculating default probabilities for cash-flow based project finance with reserve account. *Computing and Visualization in Science*, *17*(4), 203–216. https://doi.org/10.1007/s00791-015-0258-7

Robey, R. R., & Barcikowski, R. S. (1988). The Number of Iterations in Monte Carlo Studies of Robustness. *Iterations in Robustness Studies*. https://files.eric.ed.gov/fulltext/ED298155.pdf

Serrano-Cinca, C. (1996). Self organizing neural networks for financial diagnosis. *Decision Support Systems*, *17*(3), 227–238. https://doi.org/10.1016/0167-9236(95)00033-x

Shchygol, G., Yakovlev, A., Trnka, T., Van Duin, A. C. T., & Verstraelen, T. (2019). ReaxFF Parameter Optimization with Monte-Carlo and Evolutionary Algorithms: Guidelines and Insights. *Journal of Chemical Theory and Computation*, *15*(12), 6799–6812. https://doi.org/10.1021/acs.jctc.9b00769

Shen, F., Liu, Y., Wang, R., & Zhou, W. (2020). A dynamic financial distress forecast model with multiple forecast results under unbalanced data environment. *Knowledge Based Systems*, *192*, 105365. https://doi.org/10.1016/j.knosys.2019.105365

Shumway, T. (2001). Forecasting Bankruptcy More Accurately: A Simple Hazard Model. *The Journal of Business*, *74*(1), 101–124. https://doi.org/10.1086/209665

SIC Manual. (2023). U.S. DEPARTMENT OF LABOR. Retrieved August 18, 2023, from https://www.osha.gov/data/sic-manual

Stein, R. (2007). Benchmarking default prediction models: pitfalls and remedies in model validation. *The Journal of Risk Model Validation*, *1*(1), 77–113. https://doi.org/10.21314/jrmv.2007.002

Stein, R. M. (2005). The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing. *Journal of Banking and Finance*, *29*(5), 1213–1236. https://doi.org/10.1016/j.jbankfin.2004.04.008

Sun, J., Fujita, H., Chen, P., & Li, H. (2017). Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble. *Knowledge Based Systems*, *120*, 4–14. https://doi.org/10.1016/j.knosys.2016.12.019

Sun, J., Li, H., Huang, Q., & He, K. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge Based Systems*, *57*, 41–56. https://doi.org/10.1016/j.knosys.2013.12.006

Tafakori, L., Pourkhanali, A., & Rastelli, R. (2021). Measuring systemic risk and contagion in the European financial network. *Empirical Economics*, *63*(1), 345–389. https://doi.org/10.1007/s00181-021-02135-y

Touran, A., & Wiser, E. P. (1992). Monte Carlo Technique with Correlated Random Variables. *Journal of the Construction Division and Management*, *118*(2), 258–272. https://doi.org/10.1061/(asce)0733-9364(1992)118:2(258

Wilcox, J. W. (1971). A Simple Theory of Financial Ratios As Predictors of Failure. *Journal of Accounting Research*, *9*(2), 389–395. https://doi.org/10.2307/2489944

Wilcox, J. W. (1973). A Prediction of Business Failure Using Accounting Data. *Journal of Accounting Research*, *11*, 163–179. https://doi.org/10.2307/2490035

Xu, J., Lu, Z., & Xie, Y. (2021). Loan default prediction of Chinese P2P market: a machine learning methodology. *Scientific Reports*, *11*(1), 1–19. https://doi.org/10.1038/s41598-021-98361-6

Zhang, W. (2022). p-value based statistical significance tests: Concepts, misuses, critiques, solutions and beyond. *Computational Ecology and Software*, *12*(3), 80–122. https://www.researchgate.net/publication/360775123_p-value_based_statistical_significance_tests_Concepts_misuses_critiques_solutions_and_beyond

Zmijewski, M. E. (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research*, *22*, 59–82. https://doi.org/10.2307/2490859