LUND UNIVERSITY

SCHOOL OF ECONOMICS AND MANAGEMENT

Corporate default prediction: a comparison between Merton model and random forest in an environment of data scarcity

A thesis submitted in June 2022 for the degree of

Master's in Finance

Aitor Díaz García
Matīss Mirošņikovs

Supervised by Anders Vilhelmsson

Lund University - School of Economics and Management

Department of Economics - Course code: NEKN02

Abstract

The aim of this paper is to compare the performance of the Merton model to a machine learning technique (random forest), in a context where the number of predictors is low or the dataset is quite small. Since random forest is a data-intensive method, the main goal is to find the minimum number of explanatory variables and observations that is needed for it to perform at least as well as the Merton model, an approach developed in the 70s that gives the probability of the firm defaulting. Results suggest that a minimum of 13 predictors is required for both models to have a similar performance and that the dataset should be formed by no less than 9,600 observations for random forest to be as accurate as the classic approach by Merton, providing that the proportion of defaults in the dataset is around 0.18%. Our study also suggests that the application of the over-sampling technique SMOTE combined with 100% under-sampling of the majority class leads to superior results for the random forest, with an accuracy higher than 0.8 as measured by area under curve (AUC). In general, our findings indicate that, although machine learning techniques perform better in absolute terms when predicting corporate defaults, it is important to consider the amount of data available and its quality before trying to apply any of them, since, in extreme data scarcity scenarios, traditional approaches as the Merton model perform better.

Keywords: Merton model, random forest, default prediction, SMOTE.

Acknowledgments

We would like to thankfully acknowledge all the inputs and dedication from our supervisor Anders Vilhelmsson during the writing process.

I, Aitor, would like to thank my family and especially my mum and dad for the never-ending love and support they have always given to me, especially in hard times when everything seemed to be dark. There is a part of you in every achievement I make. Also, I would like to thank my loyal group of friends for being there when I needed them the most. Lund has been a life-changing experience for me but without all of you having my back, nothing of this would have been possible.

Table of contents

List of figures

List of tables

1. Introduction

In this paper, we aim to compare one of the pioneering models for predicting the potential default of a company, the Merton model (1974) to more recent approaches based on machine learning algorithms, such as random forest. Since the Merton model has been the basis of all subsequent credit risk assessment models and it is still widely used among both practitioners and researchers, we intend to test to which extent methods such as random forest outperform it, especially in situations where data are scarce. Hence, a full assessment is to be made to provide some intuition on whether it is worth it to use machine learning algorithms instead of simpler approaches that also produce a highly reliable probability of default.

Corporate defaults have been a major issue over the last decades. The recent financial crises have pointed out several weaknesses among companies that led many of them to default, and have revived the debate on which are the appropriate mechanisms to deal with those situations. Hence, credit risk models that evaluate the default probability of a certain company are a key instrument for the whole financial system, especially when they shed light on the health of privately held firms, which are not that much controlled by regulators.

The development of such models started some decades ago, when Altman (1968) and Merton (1974) released their Z-score and Merton model, respectively. In contrast to the research by Altman that uses a regression-based approach, Merton conceives a model that gives importance to the economic mechanisms that produce default, therefore creating one of the first structural models for predicting default. His findings are the foundation for many subsequent studies on the matter, that focus on the improvement of predictive power, sensitivity, and specificity.

Later in time, a completely different set of techniques were developed. Those, which include classification trees, logistic regression, Least Absolute Shrinkage and Selection Operator (LASSO), and random forest among others, are known as machine learning (ML) techniques since they use a built-in algorithm to perform the calculations rather than an economic approach. The application of ML in finance has become widespread within the last years, and further developments that will increase the predictive power of the models are to be expected.

In this setting, many have used these approaches to try to outperform the classic techniques, including Merton model. One of the problems of this method is that the probability of default it yields is not accurate on its own, which implies that it is not a sufficient statistic for default prediction even though the ranking it provides among firms is correct (Bharath & Shumway, 2008). ML techniques use the available data to solve classification problems and recent studies have shown that, when an adequate dataset is employed, their performance is superior to those from previously existing methods (Barboza, Kimura & Altman, 2017; Moscatelli et al., 2020). However, a widely recognized drawback of these methods is the need for having certain previous knowledge to implement them correctly, apart from their "black box" nature (Kim, Cho & Ryu, 2020).

Although it is true that most of the literature that aims to compare the performance of ML techniques for default prediction agrees on the fact that these approaches outperform the traditional methods, it must be taken into consideration that generally, these studies include very wide and rich datasets where almost all the information (public and private) that is needed to predict a default is accounted for. In our study, we aim to make a comparison between Merton model and random forest in a context of data scarcity, that is, where only some publicly known financial variables are available and the number of observations is small. We consider this to be insightful since, in reality, many practitioners do not have enough time or resources to access the relevant data, and hence, they might have to consider different alternatives. To our knowledge, no previous literature has directly compared such different approaches in a data-scarce environment as we are doing in this paper.

In order to perform our study, we first use only the inputs from the original Merton model, and test how random forest performs with respect to Merton model. Then, we add a set of relevant financial variables obtained from publicly available datasets and compare both approaches again. Since the aim of the paper is to obtain the point where it would be more convenient to use a classic approach such as the Merton model instead of the more complex random forest, we drop some variables and reduce the number of observations in order to check how much the accuracy of the ML technique goes down.

Since datasets used in corporate default studies are severely imbalanced, we use the Synthetic Minority Over-sampling Technique (SMOTE) suggested by Chawla et al. (2002) to artificially increase the number of defaults. This method, when combined with

100% under-sampling of the majority class, significantly increases the level of accuracy of the model, another important output from our study.

The remainder of this paper is organized as follows. Section 2 reviews the main literature that has been written about default prediction models, starting from the first approaches by Altman (1968) and Merton (1974) and ending with the most recent techniques applied, that is, machine learning based models. Section 3 discusses the theoretical framework of the two main approaches applied in this study and gives further insight into their development over time and their current applications. Section 4 states the proposed methodology, which intends to give an insight into when machine learning techniques start outperforming the classic Merton model. Section 5 discusses the data collection process and the potential transformations that might have to be performed. Section 6 includes the main results and findings of the study, with the relevant analysis on them being discussed in section 7. The paper ends with a conclusion that aims to summarize the approach followed and the results obtained, and it also includes a reference section and an appendix.

2. Literature review

Over the last decades, a vast amount of literature has been written about corporate default prediction methods. After years of research, the development of new statistical and mathematical tools has increased the number of available options, albeit not all the techniques have performed as expected in this field and some older models have proved to be still fully valid. In this literature review, we aim to give insight into the main lines of corporate default study during the last fifty years, with an emphasis on the most relevant recent publications.

Among the pioneering research in this field, the work of Altman (1968) is noteworthy since he develops one of the first credit score models, known as the Z-score. Applying multiple discriminant analysis (MDA) and relying on historical data, the author produces a multivariate regression with financial ratios as independent variables that aims to categorize the company as either distressed or safe, including a gray "inconclusive" zone. However, the original model can only be applied to public firms since the market value of equity is an input in the regression, a fact that is fixed by Altman (1983) when he frames two variations of the earliest regression to account for private and non-manufacturing companies.

Shortly after Altman develops his Z-score model, Merton (1974) establishes the foundation of all the subsequent credit risk models with his proposed approach. Starting from the consideration that the market value of equity can be seen as a long call option on the asset market value (with debt value being the strike price), he conceives a model where the company defaults if the market value of the assets is below the market value of the debt, or in other words, if the value of the call option that equity-holders have is zero. The model relies on previous assumptions by Black and Scholes (1973) and develops formulae to calculate the probability of default (PD) and the distance to default (DD) with relatively few inputs.

Both models have proved to be valid to the present day and are still widely used among practitioners, as some authors like Altman et al. (2014) and Afik, Arad, and Galil (2016) point out. Altman et al. (2014) use a large dataset that includes a wide variety of countries to test the Z-score model proposed by Altman (1983) and show that it performs satisfactorily in an international context, although some variables in the regression may need to be slightly modified to account for country-specific factors. Meanwhile, Afik, Arad, and Galil (2016) test the Merton model along with some variants of it and alternative approaches and find that the simplest versions of it perform better than the alternative methods, which are more computationally intensive. According to the authors, these positive results when tested empirically and the relatively few inputs that it needs, explain why the Merton model is still one of the most applied techniques for credit risk assessment nowadays.

As time passed, some models which aimed to improve the one proposed by Merton (1974) arose. One of the most relevant ones is the KMV-Merton, which was first developed in 1989 by the company KMV and subsequently acquired by Moody's in 2002. Although this model is used daily by Moody's to assess credit risk of thousands of companies, some authors have also tested it independently with dissimilar results. Bharath and Shumway (2004) test a naïve alternative of the KMV-Merton model and conclude that it does not provide a sufficient statistic to predict corporate default, in contrast to the research by Yusof and Jaffar (2011), who consider the KMV-Merton model to be an adequate tool to forecast potential company defaults, by using a sample of Malaysian firms. Nevertheless, these results should be considered carefully since the actual version of the model is privately held by Moody's and hence, not known to most practitioners.

The introduction of the models by Altman (1968) and Merton (1974) for predicting corporate default entailed a revolution in the field, and new statistical techniques started to be applied to relevant research on the matter. Thus, Ohlson (1980) proposes the use of logistic regression analysis for predicting bankruptcies with his O-score model and after that, many other researchers applied novel techniques to improve the forecasting power of the models. Among those approaches, ML techniques have been given a remarkable focus in recent literature, due to their potential to outperform existing methods. According to Kim, Cho, and Ryu (2020), the main ML approaches include support vector machines, decision trees, random forest, and artificial neural networks. The authors argue that under ML, corporate defaults are considered to be classification problems with two possible outcomes: default or normal (i.e. no default). That is considered to be a limitation of the techniques since it implies the assumption that defaults occur independently, a fact often disproved by empirical data (Kim, Cho & Ryu, 2020). Besides, the lack of use of macroeconomic variables is pointed out by the researchers as one of the main limitations of the previous ML studies.

There are several papers that aim to compare different ML methods with traditional statistical approaches in order to check which one is the most accurate to predict corporate defaults. Along these lines, Barboza, Kimura, and Altman (2017) use a large dataset involving several decades to test many different models in order to find the most accurate one for forecasting defaults. The approaches considered include support vector machine, bagging, boosting, random forest, and artificial neural networks, and are compared to classic models such as linear discriminant analysis and logistic regression. Barboza, Kimura, and Altman (2017) show that the ML-based models outperform the traditional approaches, with random forest, boosting, and bagging being the most accurate ones, which is noteworthy considering that no transformation was done to the variables. In their article, they also consider different sets of variables and conclude that the one initially selected by Altman (1968) performs worse than some others that include relevant information regarding bankruptcy. The findings by Moscatelli et al. (2020), who test random forest and gradient boosted trees in a dataset formed by more than 300,000 non-financial Italian companies, point in the same direction. The authors find that ML techniques perform better than logistic regression and linear discriminant analysis when using only financial variables since they improve the forecasting accuracy by 10 percentage points. Conversely, when Moscatelli et al. (2020) also consider credit behavioral indicators, the difference in discriminant power between traditional and ML approaches becomes smaller, probably as an effect of having higher quality data, which improves the overall accuracy of the models.

Out of all ML techniques, random forest seems to be one of the most used and accurate approaches to predict corporate defaults. This method was introduced by Breiman (2001) and has been applied to many different fields and areas of study in order to deal with classification problems. For instance, Ali Khan et al. (2022) use random forest to model water surface salinity in Pakistan, Gurm et al. (2014) apply a random forest-based algorithm to predict the likelihood of needing a transfusion after certain coronary interventions, and Lock and Nettleton (2014) employ it to develop a model for estimating the winning probability of a certain team in an NFL game. Applications more related to finance have also been widely mentioned in the literature. As an illustration, we can cite the work by Tan, Yan, and Zhu (2019), who implement random forest to select stocks with the aim of forming a portfolio that outperforms a Chinese stock market index, with success in its application.

Many articles have focused on using random forest for corporate default prediction, with relevant findings in most cases. Zhu et al. (2019) use several machine learning techniques to conduct a study on loan default by using data from Lending Club, a P2P lending platform that allows individuals to borrow and invest. The results show that random forest outperforms the other methods used (decision trees, support vector machines, and logistic regression) to predict loan defaults, and results suggest that the model could be generalized due to its accuracy. One step further go Kim, Cho, and Ryu (2021), who use geometric-lag variables in their analysis of ML methods (including random forest), and conclude that it boosts the discriminatory power of the models.

Although random forest models seem to perform well with large datasets, there can be differences in the outcomes depending on the data analyzed, especially when the study is performed in different countries (Behr & Weinblat, 2016). These authors employ ML techniques including random forest to analyze differences in corporate defaults in seven European countries (Germany, France, Finland, Spain, Portugal, Italy, and the United Kingdom), and find that the differences are significant, or in other words, that one model might not fit all countries. This opinion is also shared by Kim, Cho, and Ryu (2020), who also admit some limitations of ML techniques, such as the difficulty of its application or the absence of outputs apart from the probability of default. Nevertheless, Behr and Weinblat (2016) show that, despite country-specific differences, an international model for predicting credit default could be developed, with random forest outperforming the other techniques employed in their analysis.

The selection of variables has also been a broadly discussed topic among academics and researchers over the last few years. The main limitations of recent studies in the field include the omission of the stock price as an independent variable and the lack of use of multi-period models (Kim, Cho & Ryu, 2020). In addition to that, the work by Kohv and Lukason (2021) highlights the importance of non-financial variables, since they find that tax arrears have great predictive power for forecasting loan defaults, even more than ratios from classic financial variables. On the other hand, delays in handing in financial reports are not a significant variable in loan default prediction models, whereas it is important for models that consider bankruptcy. Kohv and Lukason (2021) focus on loan defaults (that are more unpredictable than bankruptcies, which tend to be more systematic) and employ a dataset of loans from a well-established commercial bank that includes mostly small and medium enterprises (SMEs).

We aim to finish this literature review by giving a short insight into the use of artificial neural networks to predict defaults. In this sense, we remark on the contribution of Kim, Cho, and Ryu (2021), who test two different algorithms to process sequential data and find that neural networks perform better than the other methods considered, besides that an ensemble model which combines the predicted probabilities of the logistic regression, support vector machine, random forest, recurrent neural networks, and long short-term memory models performs better than all the models separately. However, they also point out that neural networks have pitfalls, such as they cannot clearly indicate the importance of each individual explanatory variable for bankruptcy prediction due to their complexity.

3. Theoretical framework

Since the main purpose of this paper is to compare the performance of the Merton model to an ML technique (random forest) when data are scarce, it is essential to include a section in which the theory behind both methods is explained clearly. The abovementioned approaches were mainly chosen because of their dissimilar characteristics in terms of inputs needed and underlying mechanisms to predict defaults, as will be explained later in this part of the essay.

3.1. Merton model

This approach was developed by Merton (1974), and it was the first of a subsequent series of structural credit risk models. These methods consider the underlying economic conditions of the firm to forecast the likelihood of a default, so they take into account company-specific variables and are especially useful since they can spot flaws in the capital structure that can make the firm be in a situation of distress. Hence, they differ from reduced-form models which rely exclusively on statistical models.

Merton's idea is mainly based on the assumption that the firm defaults if and only if the value of assets is lower than the face value of debt (i.e. what must be paid to debtholders) at the time of maturity. This implies that the value of equity can be seen as a long call option written on assets, with the strike price equal to the face value of debt. This premise is key for the further development of the model since it means that an option pricing model can give some insight into the probability of default of the firm. In this sense, Merton continues with the previous work by Black and Scholes (1973).

In order for the model to work, some assumptions need to be made beforehand. Probably, the most relevant one is that debt is a zero-coupon bond with a face value of *K*, that matures at the time when the firm is terminated, *T*. This means that the firm defaults when the value of assets is below *K*, but since debt is only serviced once, that can only happen at *T*. Obviously, the firm cannot issue new debt or equity to pay the old debt. Consequently, there are two possible scenarios at time *T*: default or not default.

When the firm is terminated, both debt holders and equity holders get their share. It is common knowledge that debt holders have to be paid back before equity holders. Therefore, the value for them will be *K* unless the firm has defaulted. A default would mean that the value of assets at time *T* is lower than *K*, and then this would be the amount appropriated by the creditors. The following equation summarizes the outcome

$$B_T = \begin{cases} K \; if \; A_T \geq K \\ A_T \; if \; A_T < K \end{cases} \tag{1}$$

Following the previous discussion, equity holders do not get anything if the firm defaults, since debt holders receive the total asset value. Only if the firm does not default, they do receive the difference between the asset value and the face value of debt (i.e. the remaining value of the company after all debt holders have been serviced), as can be summarized in equation 2.

$$E_T = \begin{cases} A_T - K \; if \; A_T \geq K \\ \quad 0 \; if \; A_T < K \end{cases} \tag{2}$$

The most important outcome of the previous discussion is that equity can be seen as a long call option with the asset value as underlying and with strike price equal to the value of *K*, since the payoff to the equity holders is the difference between both if there is no default and zero otherwise. This is stated in

$$E_T = max[A_T - K, 0] \tag{3}$$

which implies that an option pricing model could give some information regarding the probability of the company defaulting, this being the reason why Merton continues with the previous work by Black and Scholes (1973). Concretely, he applies the assumption that the asset value follows a geometric Brownian motion (GBM), which is a stochastic process where the logarithm of the underlying variable follows a standard Brownian motion with a drift. Hence, this implies that the logarithm of A_T is normally distributed according to

$$lnA_T \sim \Phi\left(lnA_0 + \left(\mu_A - \frac{\sigma_A^2}{2}\right)T, \sigma_A\sqrt{T}\right) \tag{4}$$

where A_0 stands for the initial asset value, μ_A represents the return on assets (ROA) or the expected growth rate (and is usually proxied by the risk-free rate) and σ_A is asset volatility. By knowing this, it is straightforward to get the probability of default as the probability of the value of *K* being lower than the value of A_T at time *T*. Applying natural logarithms to both sides and standardizing the normally distributed variable, Merton found that the probability of default is

$$N(-DD) \tag{5}$$

where DD is a newly defined variable known as distance to default which measures the number of standard deviations that the firm is away from distress (see equation 6) and N is the cumulative normal distribution. Consequently, DD allows direct comparisons between firms, and although some studies have revealed that it is not a sufficient statistic for predicting defaults, it gives an adequate ranking of firms with respect to how far they are from financial distress (Bharath & Shumway, 2008).

$$DD = \frac{lnA_0 + \left(\mu_A - \frac{\sigma_A^2}{2}\right)T - lnK}{\sigma_A\sqrt{T}} \tag{6}$$

When implementing the model in practice, it is important to note that it is difficult to find accurate estimates for all the inputs in the model. As stated before, μ_A is usually proxied by the risk-free rate. This yields the risk-neutral probability of default, which is higher than the actual one since equity holders demand a return that is higher than the risk-free rate to compensate for risk, and this implies that PD goes down because DD is smaller. For σ_A and A_0, a system of two equations is generally required since they are not observable. The standard procedure is to use the initial equity value E_0 (calculated as market capitalization) and equity volatility σ_E to solve the following equations, obtained by employing the Black-Scholes formula for option pricing backwards (equation 7) and the Ito (1944) formula (equation 8), a mathematical approach that can be used since the underlying asset value follows a GBM:

$$E_0 = A_0 N\left(\frac{ln\frac{A_0}{K} + \left(r + \frac{\sigma_A^2}{2}\right)T}{\sigma_A\sqrt{T}}\right) - Ke^{-rT}N\left(\frac{ln\frac{A_0}{K} + \left(r - \frac{\sigma_A^2}{2}\right)T}{\sigma_A\sqrt{T}}\right) \tag{7}$$

$$\sigma_E E_0 = \sigma_A A_0 N\left(\frac{ln\frac{A_0}{K} + \left(r + \frac{\sigma_A^2}{2}\right)T}{\sigma_A\sqrt{T}}\right) \tag{8}$$

In the paper, we use this traditional approach, since more recent adaptations such as the KMV model are privately held and not all details are known. Therefore, we calculate the risk-neutral probability of default and get our estimates of asset volatility and asset value by using the market values of equity of the selected firms.

3.2. Random forest

Machine learning techniques have become extremely popular to deal with classification problems, i.e., those in which the dependent variable can only take discrete values. Within finance, one of the fields in which these kinds of situations arise the most is default prediction, since there are only two possible outcomes: default or not default. Hence, a dummy variable approach is necessary to create a relationship between a set of predictors and a binary independent variable.

Among the vast number of different techniques that are applied nowadays, tree-based methods have taken a step forward due to their strong performance in terms of accuracy. Trees can be used for both regression and classification problems and are especially useful when the number of predictors is large or when there are non-linear relationships between the dependent and the independent variables.

Random forest is ultimately based on regression trees. Those are simply splits of the data into different branches, according to a decision rule decided by the algorithm so that it splits the data on a certain variable at each node. A regression tree will have *n* nodes and *n+1* branches, where at each node a decision is taken by the tree. In figure 1, an example of a basic regression tree is provided for better illustration.

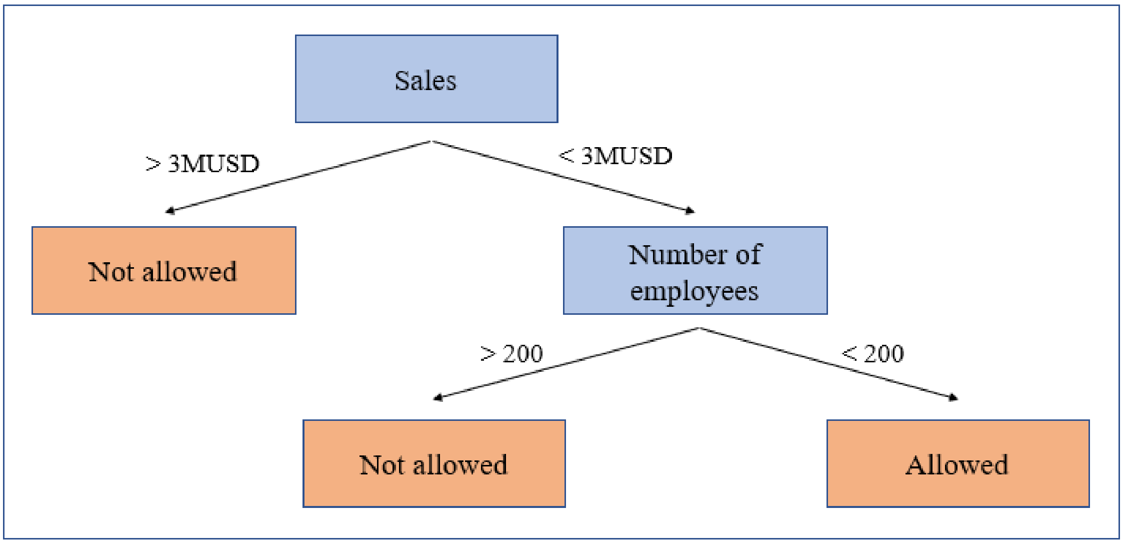

Figure 1. Example of regression tree. In this figure, the regression tree stands for the possibility of a certain company getting a subsidy from the government. Only firms with sales lower than 3 MUSD and less than 200 employees are allowed to get the subsidy, as shown in the picture.

Regression trees can also be seen as algorithms that split the data into a certain number of regions, with the aim of getting the smallest possible residual sum of squares (RSS) for all of them. Several methods exist to grow the trees, but avoiding overfitting is essential. Hence, pruning complex trees to make them simpler and more likely to predict accurately is a common procedure, and several methods for this are used in practice. Some advantages of regression trees include their easiness to be understood and to mirror human decisions and their ability to produce outcomes based on qualitative variables without the need for dummies (James et al., 2013).

For classification problems such as the one which involves default prediction, classification trees are needed. Their mechanism is essentially the same as those of regression trees, but instead of using RSS, they minimize a node purity measure. The Gini index is the most common one and it is defined as

$$G = \sum_{k=1}^{K} p_{mk}(1 - p_{mk}) \quad (9)$$

where p_{mk} is the fraction of training observations in the *mth* region that belong to the *kth* class (James et al., 2013). This means that the index will be lower the closer the observations are to zero or one, which is consistent with the intuition of the concept of node impurity. In contrast, the highest node impurity would be having half of the observations from one class and the other half from the other class.

One of the main disadvantages of classification trees is that they have a high variance. In order to solve this and to improve the forecasting power of the model, some approaches have been suggested. Breiman (1996) develops bootstrap aggregation (mostly known as bagging) to get an average prediction from many different trees rather than one single forecast. The underlying idea of this approach is to create different samples by bootstrapping. That means many samples will be obtained since bootstrapping works as random sampling with replacement. An unpruned tree will be fitted for each sample and the final tree will be the average of all the sample trees created. An advantage of this is that no validation dataset is required since the data that has not been used constitutes the out-of-bag (OOB) sample and is used to validate the data (Breiman, 1996). One main drawback of this procedure is that trees are highly correlated with each other, a contingency solved by the random forest approach.

Random forest was suggested by Breiman (2001) and is basically decorrelated bagged trees. According to the definition provided by the author in the original paper,

> a random forest is a classifier consisting of a collection of tree-structured classifiers *{h (x, k), k = 1,...}* where the *{k}* are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input *x* (Breiman, 2001, p.6).

This means that all trees grown vote for the dependent variable to belong to one class or another, and eventually it is assigned to the class that gets the most votes. To decorrelate the trees, the number of predictors taken into account at each node is reduced to *m*, an arbitrarily picked number that is lower than *P*, the number of predictors. The standard practice suggested by Breiman (2001) is to set $m = \sqrt{P}$. This incorporates a great advantage compared to bagging since not all the predictors are used in every tree and implies that all significant predictors will be considered at some point, because there will be some trees in which they are the main splitting variables. Hence, random forest uses all the available information, which leads to stronger predictive power. It is noteworthy that *m* should be lower if the predictors are highly correlated.

The number of trees to grow is an arbitrary decision, but it is standard to set it between 50 and 2,000. Intuitively, the computational time will increase if more trees are grown, which is one of the drawbacks of this approach. Another limitation is its data-intensive nature, which implies that predictions are generally poor in an environment of data scarcity, a fact that we intend to show in this paper by comparing this method to the less data-intensive Merton model in a corporate default framework.

4. Methodology

The aim of our study is to evaluate the performance of the Merton model compared to random forest in predicting corporate defaults. It is important to note that the Merton model is not exactly constructed to predict whether a default is to occur, as it rather returns a normalized distance to default which can then be transformed into a default probability. Hence, the strength of this approach lies in the proper ranking of firms according to their expected probability of default, since the actual numerical probability of default is not nearly as accurate on its own (Afik, Arad & Galil, 2016). Nevertheless, it is possible to introduce some threshold probability of default that separates high-risk and low-risk firms, where one could more accurately interpret a numerical value of the probability of default as a prediction of either solvency or insolvency. These values then can be directly compared to the actual condition of the firm later in time and also to predictions from other models.

Our research includes a comparison between the predictions of random forest and those of the Merton model. Initially, only the variables that are employed to calculate distance to default in Merton are used in both models, i.e., market value of equity, face value of debt, growth rate (this could be approached in different ways, but we stick to the standard assumption of setting it equal to the risk-free rate), and equity volatility. In this situation, it could be expected that the Merton model outperforms random forest, given that the initial number of variables is relatively small. However, we intend to find out at which moment the random forest algorithm starts outperforming the Merton model, and whether it is a consequence of expanding the sample size or adding more key variables.

The addition of variables is limited by variable preselection using a logit regression and checking correlation coefficients and multicollinearity between factors. Furthermore, for random forest to work properly it is needed to artificially expand the subset of defaults, since in most studies about defaults the dataset is quite imbalanced, with many more non-default observations than default ones. An oversampling method is preferable in this case in order to not lose valuable information.

To compare the methods, the receiver operating characteristic (ROC) curve is used, where the value to be compared is the area under curve (AUC). This measure mostly takes values between 0.5 and 1, where 1 is a perfect classifier and 0.5 is a completely random classifier. Therefore, the method with a higher AUC would be considered a better classifier. The curve is created by plotting false positive and true positive rates for all possible decision thresholds. We aim to analyze which approaches perform better in which circumstances and whether the added calculation time and data mining outweigh the potential gains from getting superior results over simpler calculations (i.e., those from the Merton model) without the need for other observations.

Another way to compare methods and show performance is by using a confusion matrix, where it can be directly seen how the observations are classified. However, to build a confusion matrix one has to decide on a decision threshold, which works well for random forest but not for the Merton model. In contrast, the ROC curve is acquired from results for all possible decision thresholds. This is why we have decided to include only ROC curves and the corresponding AUCs.

Since the dataset is severely imbalanced, a way to make it more proportionate is necessary for better results. That can be achieved by either under-sampling the majority class, over-sampling the minority class, or a mixture of both. In this setting, over-sampling is preferable because it does not lead to a loss of potentially valuable information, despite being slightly more challenging to implement. This is so since doing over-sampling with replacement is potentially not too effective, due to the fact that it does not introduce new data but rather has the same exact observations introduced as new ones, effectively not expanding the minority dataset. This is where Synthetic Minority Over-sampling Technique (SMOTE), developed by Chawla et al. (2002), is used. This method generates synthetic new observations in order to create a balanced dataset, which in our case leads to having the same number of default and non-default observations.

We choose SMOTE over some other over-sampling techniques such as Borderline-SMOTE (Han, Wang & Mao, 2005), Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN) (He et al., 2008), and Safe-Level SMOTE (Bunkhumpornpat, Sinapiromsaran & Lursinsap, 2009) because it yields more accurate results. Borderline-SMOTE only takes the borderline observations into account for the over-sampling process. ADASYN considers the characteristics of every class to assign it a weight according to its difficulty to learn. Finally, Safe-Level SMOTE uses a measure known as “safe level” to increase the quality of the process when many classes with different weights are involved. The reason why SMOTE works better on our dataset can be that we only consider two classes: default and non-default, and the remaining techniques are mainly designed for datasets with a larger number of classes.

First, the required rate of oversampling is calculated, which in our case is the one that provides a balanced dataset (for example, a 100% oversampling rate doubles the number of observations for the selected class). Then, the synthetic data are produced by usually using the 5 nearest neighbors. When taking one of the nearest neighbors, the difference between those two can be obtained in a two-dimensional form, which means that the difference between the observations is taken for all their variable values. This difference is then multiplied by a random number between 0 and 1, giving the new data point. In essence, a line is drawn between two observations and then a random point on it is chosen to be the new synthetic observation.

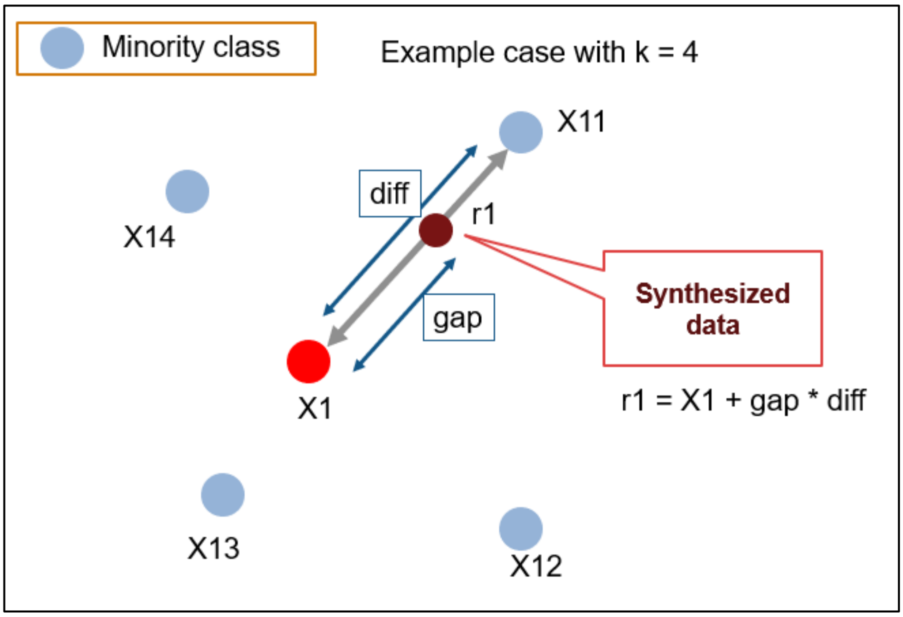

Figure 2. Illustration of the underlying mechanism of SMOTE. Source: GitHub.

To verify if random forest performs better than the Merton model in an environment of rather few observations but with all available information for them, the training sample is to be reduced in size. Since the performance of the Merton model is unchanged independently of the number of observations, a point can be found when the ML technique has enough information to outperform the Merton model.

The dataset is originally split into a training sample and a validation sample, with the respective weights being 75% and 25%. The AUC values are obtained by running random forest 10 times and averaging the result, with a different random sample being used each time. In this way, we simulate different situations in real life where some data might be the same, but the combination is different. After that, changes in the size of the training sample are made in order to find the point where random forest starts outperforming the Merton model, as developed further in the results section.

The software chosen to implement all the aforementioned models is MATLAB.

5. Data

5.1.Data description

It is common knowledge that a large dataset is generally required for studies that involve predicting corporate defaults, since defaults are generally rare events that do not happen often. For this study, we start from the same dataset as in Forssbæck and Vilhelmsson (2017), which includes data from 1980 to 2015 of all non-financial listed firms from the United States. This original dataset includes 164,803 observations and 1,018 defaults, which means the proportion of defaults in the total sample is 0.62%. In Forssbæck and Vilhelmsson (2017), a default is considered to happen when the firm is delisted due to the application of Chapter 7 or Chapter 11. Apart from the default dummy and the year of the observation, this dataset also includes the distance to default according to the Merton model, as obtained by the authors.

Since more variables are required, we gather data on the inputs of the Merton model (equity volatility, total equity, total debt, and risk-free rate) for the firms of the original dataset. This information was also included in the paper by Forssbæck and Vilhelmsson (2017). Since we could not find the values for all the observations, the dataset is reduced to 95,722 observations which include 254 defaults. This implies a default ratio of 0.27%, lower than before. In addition, we remove some observations that are copies of each other and non-numerical values. This leaves us with a final dataset of 76,656 observations and 140 defaults, which means that the default ratio goes down again to 0.18%. The imbalanced nature of the dataset was to be expected beforehand and is solved by applying the SMOTE technique, as explained in the methodology part.

In table 1 some descriptive statistics of the inputs of the Merton model are shown. The first row for each value shows the results for all the companies, the second one includes only the firms who did not default, and the last column involves the numbers of the firms who defaulted.

Table 1. Descriptive statistics for the inputs of the Merton model for the whole sample, the firms who did not default and the firms who defaulted. Equity volatility is measured as the annual standard deviation of equity, total equity and total debt are expressed in MUSD and the risk-free rate is the monthly 3-month T-Bill in percent. Asset value and asset volatility are obtained as outputs of the system of equations presented in equations (7) and (8). The sample corresponds to all listed non-financial companies in the United States between 1980 and 2015.

		Equity volatility	Total equity	Total debt	Risk-free rate	Asset value	Asset volatility
Median	**All**	0.3566	425.9598	126.1538	0.2450	577.3849	0.2590
	Non-defaults	0.3568	426.5343	126.3443	0.2450	578.6780	0.2591
	Defaults	0.2720	111.9758	73.0820	0.3817	174.6872	0.1772
Average	**All**	0.4403	4828.1653	2761.6902	0.2557	7185.9939	0.3349
	Non-defaults	0.4405	4831.0857	2766.0814	0.2555	7192.6377	0.3351
	Defaults	0.3411	3232.0552	361.7331	0.3918	3554.8725	0.2548
Standard deviation	**All**	0.3155	19845.9764	20504.3693	0.2382	32004.8536	0.2769
	Non-defaults	0.3156	19861.5398	20522.8404	0.2378	32032.0402	0.2769
	Defaults	0.2645	7344.8026	698.1752	0.3493	7748.6392	0.2447
Q 1	**All**	0.2329	87.1700	23.1288	0.0075	122.9512	0.1634
	Non-defaults	0.2330	87.4046	23.1398	0.0075	123.2154	0.1635
	Defaults	0.1752	36.7740	21.0625	0.0467	52.0925	0.1093
Q 3	**All**	0.5528	2015.8606	689.8573	0.3958	2741.7898	0.4198
	Non-defaults	0.5530	2016.3682	690.5833	0.3958	2742.4280	0.4200
	Defaults	0.4168	1129.0049	293.2278	0.6263	1831.4091	0.3310

In the table, it can be clearly seen that both the median and the average values of total equity, debt, and assets are lower for the defaulted companies. This means that, on average, the firms who experienced financial distress are smaller in size, which goes in line with the general intuition that smaller companies default more often or, in other words, that size is a good default predictor. The standard deviation is lower for the firms who defaulted in all variables but the risk-free rate, but that was something to be expected since the sample of not defaults is way bigger, so more values mean more dispersion around the mean. Is it noteworthy though that the standard deviation of asset volatility is really close for both subsamples, which implies that the firms who defaulted were quite volatile before, another stylized fact mentioned in previous literature. The fact that having a smaller dataset for defaulted firms distorts the final results can be checked while looking at the first and third quantiles of the distribution of each variable, also stated in table 1, since the values for the defaults are always smaller except for the risk-free rate. To conclude, is it noteworthy that the risk-free rate has a larger median, average, and standard deviation for the defaulted firms.

5.2. Variable preselection

Since our paper aims to detect the point at which random forest starts outperforming the Merton model, more variables are needed. We collect several data and financial ratios from Refinitiv Eikon for all the firms and years of our dataset. We initially select 50 variables, with the aim of including all the relevant publicly available information. It is important to note that we do not consider private data because the goal of our study is to determine which model performs better when data are scarce and it is impossible to access undisclosed information. In order not to forget any relevant financial variable, we base our selection on the previous work by Behr and Weinblat (2016), Barboza, Kimura, and Altman (2017), and Moscatelli et al. (2020), since all of them consider financial ratios in their studies. In the following table, the initial 50 variables considered are stated:

Table 2. List of initial variables considered. All values correspond to the relevant year except the changes in variables, which stand for the one-year difference as measured in relative terms. The selection of variables is based on previous research by Behr and Weinblat (2016), Barboza, Kimura, and Altman (2017) and Moscatelli et al. (2020). Source: Refinitiv Eikon.

Variable	Code	Variable	Code	Variable	Code
Revenue	REV	**Total assets**	AST	**EBIT**	EBIT
EBITDA	EBITDA	**Net debt**	N_DBT	**Total equity**	EQT
Cash	CASH	**Total debt**	DBT	**Cash flow**	CFL
Fixed assets	FIX_AST	**Current assets**	CRR_AST	**Short-term debt**	ST_DBT
Working capital	W_CAP	**Retained earnings**	RET_EAR	**Shares outstanding**	N_SHA
Share price	SHA_P	**Earnings per share**	EPS	**ROE (pre-tax)**	ROE
Return on assets	ROA	**Enterprise value / Market capitalization**	EV/MCAP	**Net profit margin**	NPM
Total debt / Total assets	DBT/AST	**Fixed assets / Total assets**	FIX_AST/AST	**Short-term assets / Short term debt**	ST_AST/ST_DBT
Equity / Fixed assets	EQT/FIX_AST	**Total debt / Cash flow**	DBT/CFL	**Short-term debt / Total debt**	ST_DBT/DBT
Working capital / Total assets	W_CAP/AST	**Retained earnings / Total assets**	RET_EAR/AST	**EBIT / Total assets**	EBIT/AST

Sales / Total assets	REV/AST	**EBIT / Sales**	EBIT/REV	**EBITDA / Sales**	EBITDA/REV
Net debt / Equity	N_DBT/EQT	**Net debt / EBITDA**	N_DBT/EBITDA	**Cash / Total assets**	CASH/AST
Cash flow / Total debt	CFL/DBT	**Market capitalization / Total debt**	MCAP/DBT	**Change in assets**	ΔAST
Book value per share	BVPS	**Share price / Book value per share**	SHA_P/BVPS	**Change in sales**	ΔREV
Change in employees	ΔEMP	**Change in ROE**	ΔROE	**Log revenue**	LOG_REV
Log assets	LOG_AST	**Log cash**	LOG_CASH	**Log total debt**	LOG_DBT
Log current assets	LOG_CA	**Log short-term debt**	LOG_STD		

Variable preselection is done in the following way (including both the inputs of the Merton model and the 50 variables previously discussed). First, univariate logit regressions are run for each potential explanatory variable. The AUC is acquired for every regression, where the ones with a value higher than 0.55 are kept, that is, those who provide some explanatory or discriminatory power. After doing that, 28 variables remain: volatility (Merton), total equity (Merton), risk-free rate (Merton), EBIT, EBITDA, EQT, CFL, ST_DBT, RET_EAR, SHA_P, ROE, ROA, EV/MCAP, NPM, DBT/AST, DBT/CFL, ST_DBT/DBT, RET_EAR/AST, EBIT/AST, REV/AST, EBIT/REV, CFL/DBT, MCAP/DBT, ΔAST, BVPS, SHA_P/BVPS, ΔROE and LOG_STD. The AUC of the logit regression for each of the variables can be consulted in the appendix. It is noteworthy that some look-ahead bias might be present, since we use the full sample for the preselection with logit.

The next step consists of creating a correlation matrix, in which pairs of variables with a strong correlation (equal or higher than 0.7) are identified. Then, only one of them is kept, the one with the higher AUC (calculated in the previous point). This is done to avoid multicollinearity and to improve the performance of the model. The correlation matrix is also located in the appendix.

The first two steps of the preselection return the following 20 variables, which are the final variables for our study: volatility (Merton), total equity (Merton), risk-free rate (Merton), ST_DBT, RET_EAR, SHA_P, ROE, ROA, EV/MCAP, NPM, DBT/AST, DBT/CFL, ST_DBT/DBT, RET_EAR/AST, CFL/DBT, MCAP/DBT, ΔAST, SHA_P/BVPS, ΔROE, and LOG_STD.

It is important to remark that some selection of the variables happens also after acquiring the results of the random forest algorithm, where the aim is to identify the most important predictors and check the point where the random forest approach gives the same results as the Merton model. After running the random forest model with these variables as the predictors their importance is acquired, and even further selection of variables can be done to identify the difference in performance between the two models.

6. Results

In this section, we present the results of the study. We first compare the two approaches when using only the inputs of the Merton model, and after that, we include the key financial variables obtained in the previous step.

6.1. With Merton model inputs only

First, the power of the Merton model and random forest is compared only with the data that is necessary to obtain the distance to default measure within the Merton model, i.e. value and volatility of equity, risk-free rate (as a proxy of expected growth) and total debt.

We test four different approaches. Firstly, we use the original imbalanced dataset without any kind of over-sampling. The second procedure uses over-sampling with replacement, this is, by just repeating the same default observations until the dataset is balanced. Then, we use SMOTE as explained in the methodology section. In order to do that, we use the 5 closest neighbors after checking that increasing the number of closest neighbors does not improve the final result. Eventually, we combine the SMOTE technique with 100% under-sampling of the majority class (i.e., we reduce the majority class to half), as suggested in Chawla et al. (2002) to get a synthetic dataset where the former minority class is now the most frequent.

In table 3, the results are shown for each of the approaches, and for both the Merton model and random forest. It is important to highlight that the AUC of the Merton model changes slightly among methods because every time we run the models, a different validation sample is used.

Table 3. Area Under Curve (AUC) of Merton model and random forest when using only the inputs of the Merton model as explanatory variables.

	Original dataset (no change)	**Over-sampling with replacement**	**SMOTE**	**SMOTE + under-sampling**
Merton model	0.7307	0.7280	0.7328	0.7359
Random forest	0.5747	0.5594	0.6654	0.7152

The expectation of random forest performing poorly in this setting is confirmed, as Merton model outperforms random forest in terms of AUC. This is to be expected, due to the data-intensive nature of the ML technique. However, the over-sampling method can increase the accuracy of the AUC of random forest. This can be seen in figures 3, 4, 5 and 6, which represent the ROC curve of Merton model and random forest for each of the scenarios.

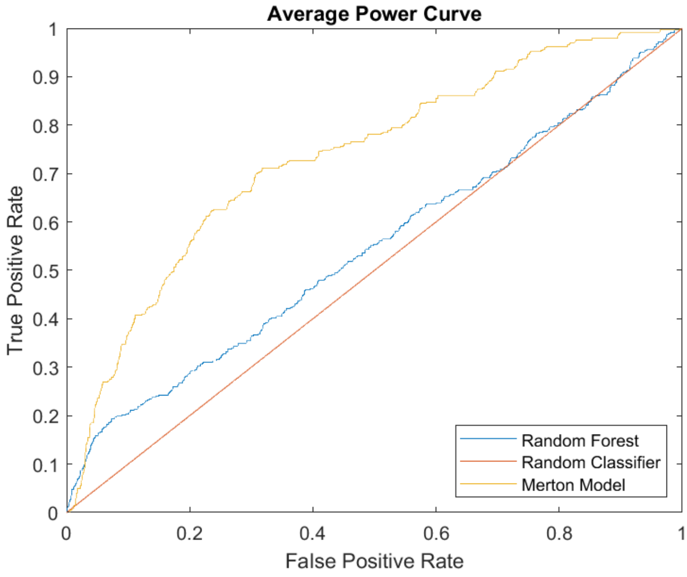

Figure 3. Receiving Operating Characteristic (ROC) curve for Merton model and random forest when using the original dataset (with Merton model inputs only).

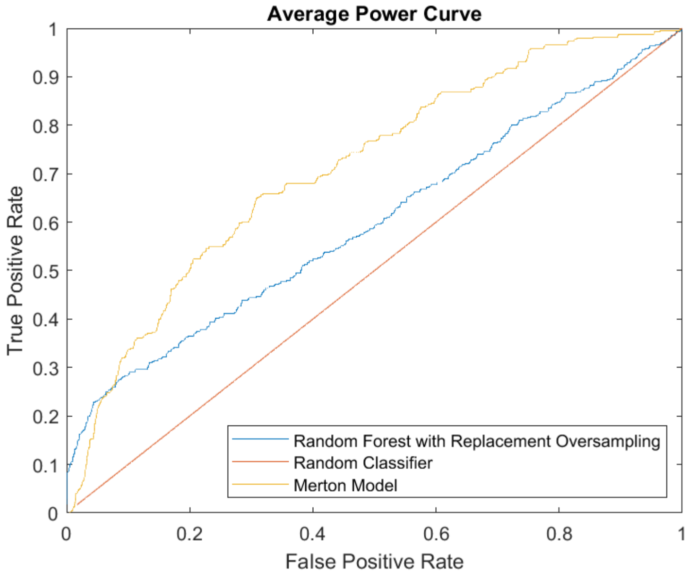

Figure 4. Receiving Operating Characteristic (ROC) curve for Merton model and random forest when using over-sampling with replacement (with Merton model inputs only).

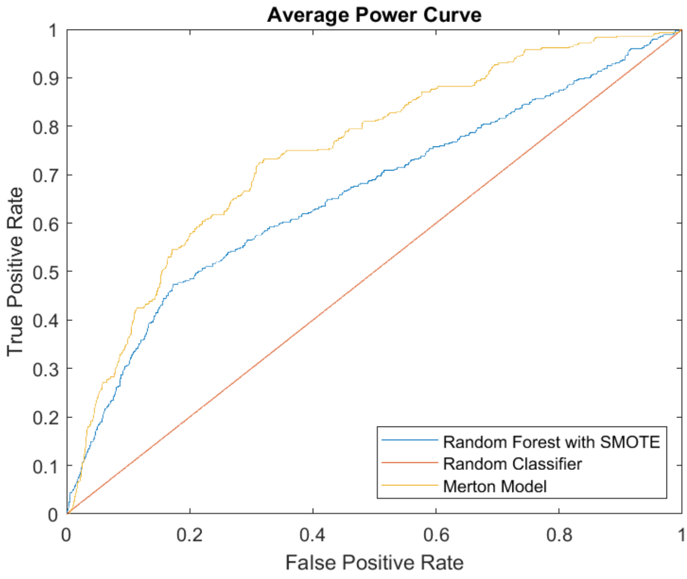

Figure 5. Receiving Operating Characteristic (ROC) curve for Merton model and random forest when using SMOTE (with Merton model inputs only).

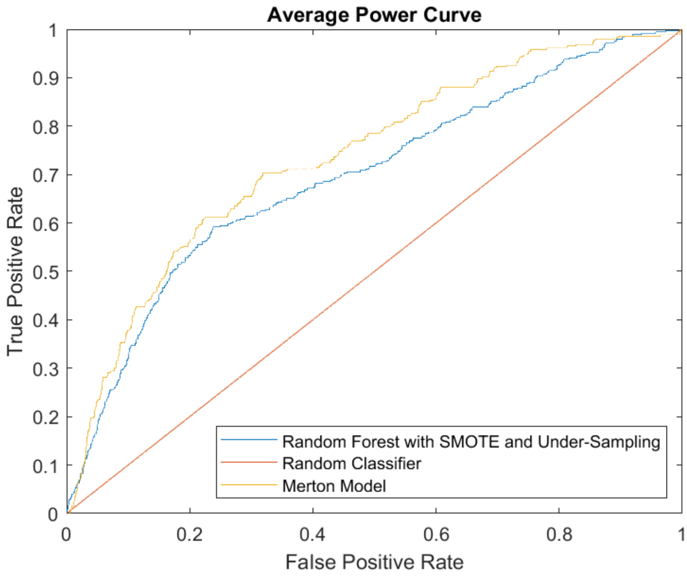

Figure 6. Receiving Operating Characteristic (ROC) curve for Merton model and random forest when using SMOTE and 100% under-sampling (with Merton model inputs only).

In the previous figures, it can be seen how the ROC of random forest becomes slightly better when the over-sampling technique used is more sophisticated, a fact that goes in line with both intuition and the AUC results. Therefore, the closest results to the Merton model ROC are achieved when SMOTE is used, especially if combined with under-sampling of the majority class as suggested in previous literature. This finding is important because it gives insight into how to obtain better results when using an ML technique if the number of predictors is low.

6.2. Using all key financial variables

The next step of the study includes all the financial variables considered after the preselection process discussed in section 5. Since ML algorithms perform better when a great amount of data are available, it is to be expected that random forest outperforms Merton model in this setting, especially when applying the over-sampling techniques to balance the dataset.

In table 4, the AUCs of the Merton model and random forest are stated when all the key financial variables are used. It can be seen that, in this situation, random forest outperforms the Merton model in every scenario, with great improvement when the over-sampling techniques are applied (an AUC of 0.8404 means that the underlying approach has a decent predicting power). The clear conclusion at this point is that, when using imbalanced datasets, it is essential to apply some kind of technique to balance them in order to improve the accuracy of the ML-based algorithm used.

Table 4. Area Under Curve (AUC) of Merton model and random forest when using all the key financial variables as explanatory variables.

	Original dataset (no change)	Over-sampling with replacement	SMOTE	SMOTE + under-sampling
Merton model	0.7400	0.7254	0.7269	0.7439
Random forest	0.7567	0.7556	0.7916	0.8404

As in the previous subsection, we also plot the ROC for each of the models in all scenarios. Results are displayed in figures 7, 8, 9, and 10 and show how the accuracy of random forest improves when more variables are taken into account, always outperforming Merton even when the dataset is not balanced. The brilliant performance of random forest when SMOTE and 100% under-sampling is applied is especially noteworthy, since it confirms the findings by Chawla et al. (2002) that the application of both methods yields more accurate results as measured by AUC.

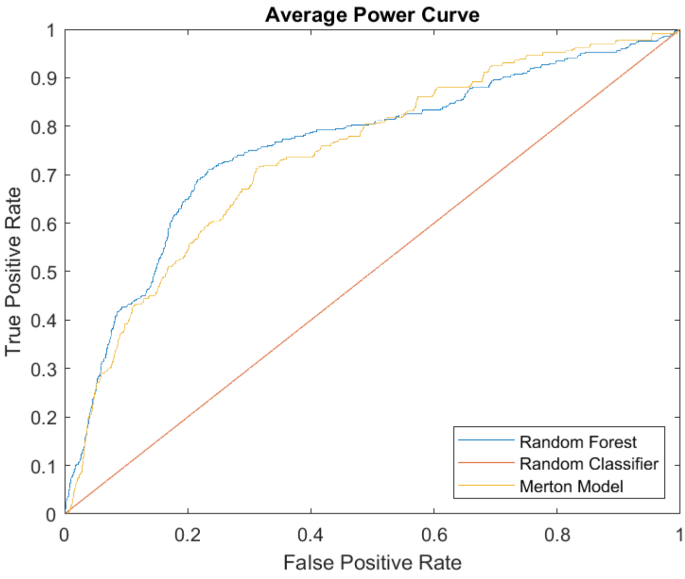

Figure 7. Receiving Operating Characteristic (ROC) curve for Merton model and random forest when using the original dataset (with all key financial variables).

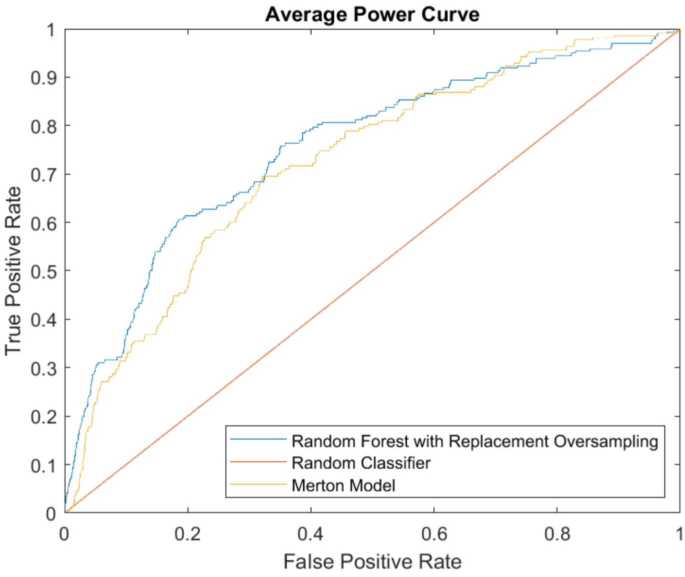

Figure 8. Receiving Operating Characteristic (ROC) curve for Merton model and random forest when using over-sampling with replacement (with all key financial variables).

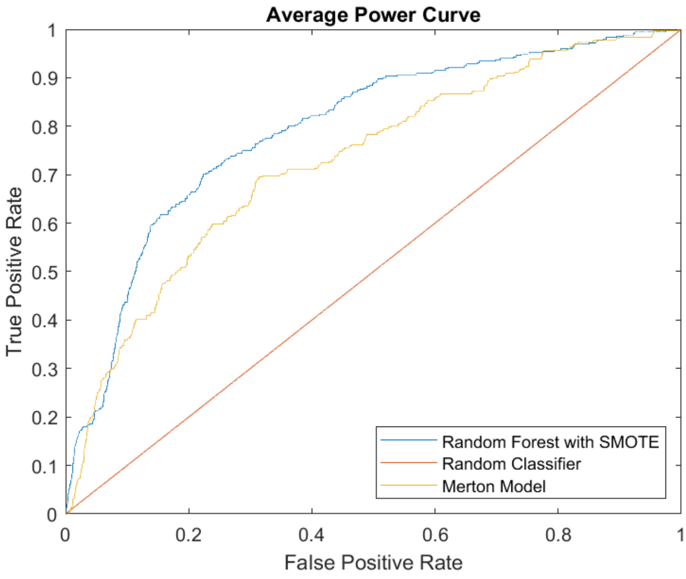

Figure 9. Receiving Operating Characteristic (ROC) curve for Merton model and random forest when using SMOTE (with all key financial variables).

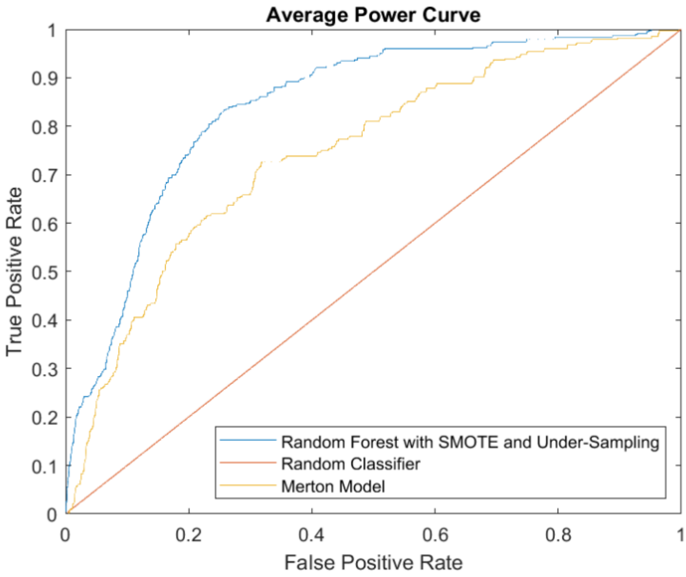

Figure 10. Receiving Operating Characteristic (ROC) curve for Merton model and random forest when using SMOTE and 100% under-sampling (with all key financial variables).

6.3. Number of predictors

At this point, we have confirmed the expectations of Merton performing better only when its four inputs are considered, and random forest being more accurate when financial information is available. Hence, we aim to find the point when random forest starts performing better in terms of accuracy. In order to do that, we perform two separate studies. First, we test how many financial variables are needed for the ML technique to be better than the classic approach by Merton. To perform this test, we obtain the relative importance of each of the variables for the random forest algorithm. This way we can rank the variables according to their predictive power. This is shown in figures 11 and 12.

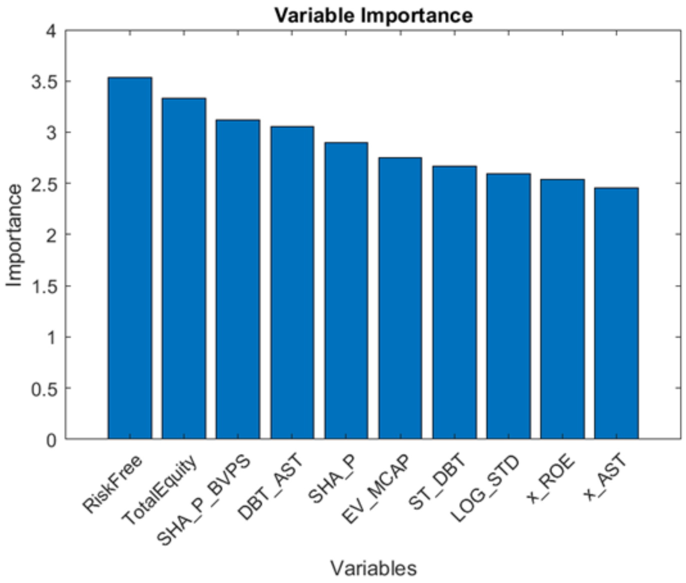

Figure 11. Variable importance (I) for the random forest approach.

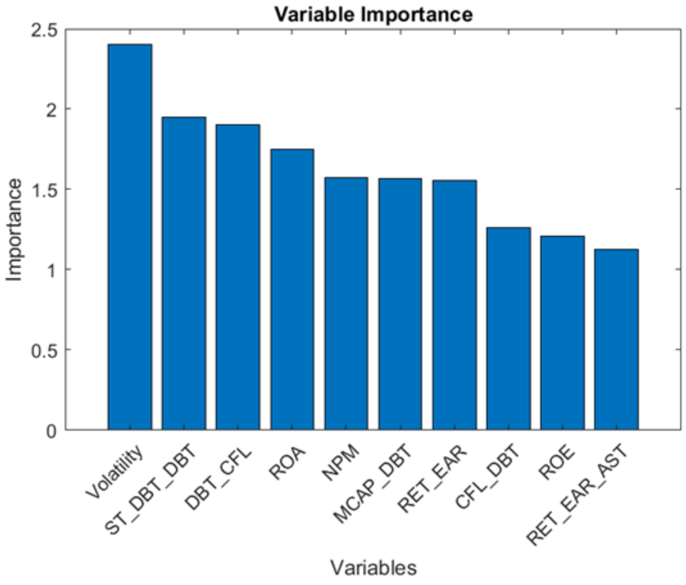

Figure 12. Variable importance (II) for the random forest approach. Please note that the y-axis has a different scale than in the previous figure.

It can be seen that the variables with the highest predictive power are the risk-free rate and total equity, both inputs of the original Merton model. This gives credit to Merton because even using a completely different approach, those values seem to be relevant for default prediction, which means his selection of variables was fairly good. On the other hand, we observe that the variables with the lower predictive power are CFL_DBT, ROE, and RET_EAR/AST.

The next step consists of gradually removing the least important variables and running the model until we find the point where the AUC of both approaches is the same. This way we can have an idea of how many financial variables are needed for the random forest to be an adequate model to predict default, or in other words, when is it worth it to employ this computationally intensive technique instead of the classic Merton model. The results from this study are shown in table 5, where the AUC of taking the *x* most important variables is shown and compared with those from the Merton model. Again, the differences in AUC in the Merton model are due to different random samples used every time.

Table 5. AUC comparison between Merton model and random forest with a different number of variables.

Number of predictors	Merton model	Random forest
10	0.7339	0.7070
11	0.7332	0.7228
12	0.7356	0.7313
13	0.7338	0.7415

From these results, it seems that the 13 most important financial variables are needed for random forest to outperform the Merton model. This is valuable information because it gives an idea of the minimum number of predictors that are needed to start considering random forest as an actual option to predict defaults. This can be more clearly seen in figures 13 and 14, where the ROC curve of random forest when using 10 variables and 13 variables respectively is plotted against the same values for the Merton model. It is important to mention that these values for random forest are obtained by applying SMOTE with 100% under-sampling of the minority class, since that approach yields better results as shown in the previous subsection.

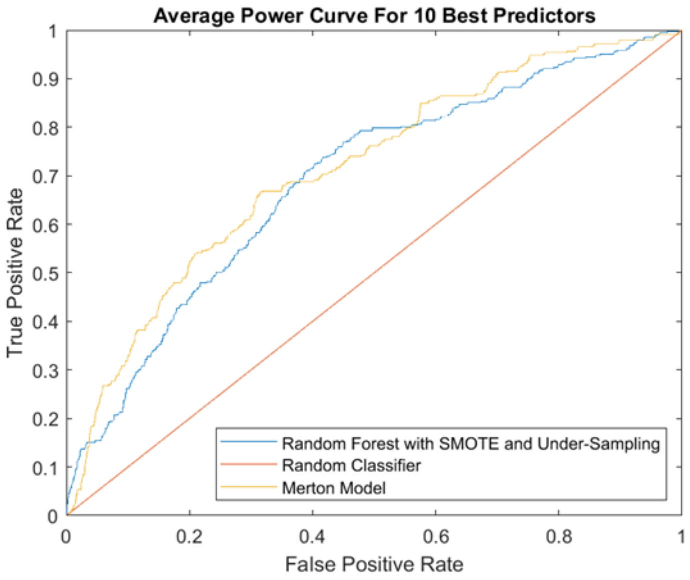

Figure 13. Receiving Operating Characteristic (ROC) curve for Merton model and random forest when using only the 10 most important variables for random forest.

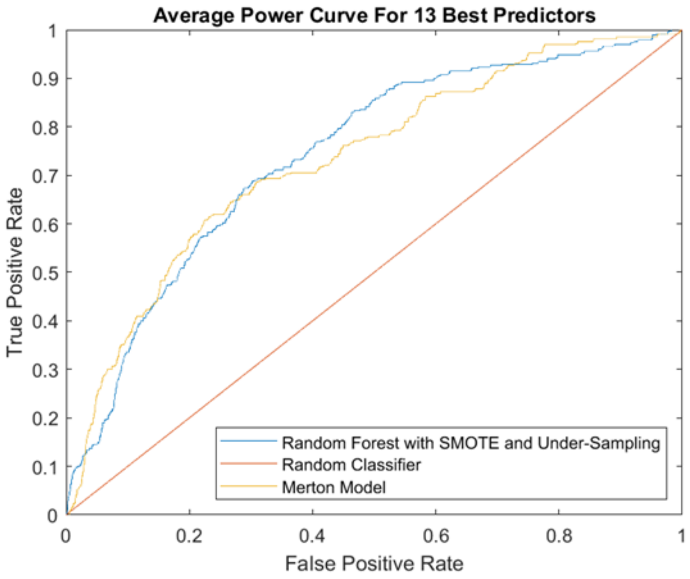

Figure 14. Receiving Operating Characteristic (ROC) curve for Merton model and random forest when using only the 13 most important variables for random forest.

6.4. Data size

The study considered in this subsection consists of deliberately reducing the training sample until we reach the point where the AUC of both methods is similar. Since random forest is a data-intensive approach, we aim to test which is the minimum number of observations that are needed for it to be considered a suitable alternative for predicting corporate defaults. As expected, when reducing the training sample the AUC gets lower, as can be seen in table 6.

Table 6. AUC comparison between Merton model and random forest with different training sample sizes. In this study, we use the 20 selected predictors as in subsection 6.2.

Size of training sample	Merton model	Random forest
50%	0.7391	0.7784
45%	0.7462	0.7823
40%	0.7387	0.7788
35%	0.7294	0.7686
30%	0.7358	0.7729
25%	0.7334	0.7759
20%	0.7332	0.7575
15%	0.7311	0.7403
12.5%	0.7295	0.7293
10%	0.7284	0.7232
5%	0.7297	0.7191
3%	0.7297	0.6897
2.5%	0.7309	0.6740

Results in table 6 show that, when using 12.5% of the total dataset as the training sample, the results in terms of AUC are similar to those of the Merton model. Since the dataset is formed by 76,656 observations, an approximate number of 9,600 observations is then needed for random forest to be a good method for predicting corporate defaults. It is noteworthy that, again, we apply SMOTE and 100% under-sampling of the majority class to test the accuracy of random forest.

For better insight, figure 15 can be consulted. It shows the AUC for different training sample sizes. It can be clearly seen that the Merton model is stable in terms of AUC (what is to be expected since its accuracy does not depend on the sample size), while random forest performs worse when the training sample is smaller. From 25% and larger training samples, both approaches show a similar behavior, as they move up and down together due to the randomness of the samples. At the 25% mark random forest stabilizes and only a slight increase in performance is observed, since from previous results we know its AUC reaches 0.8404 when the 75% of the dataset is used (see table 4).

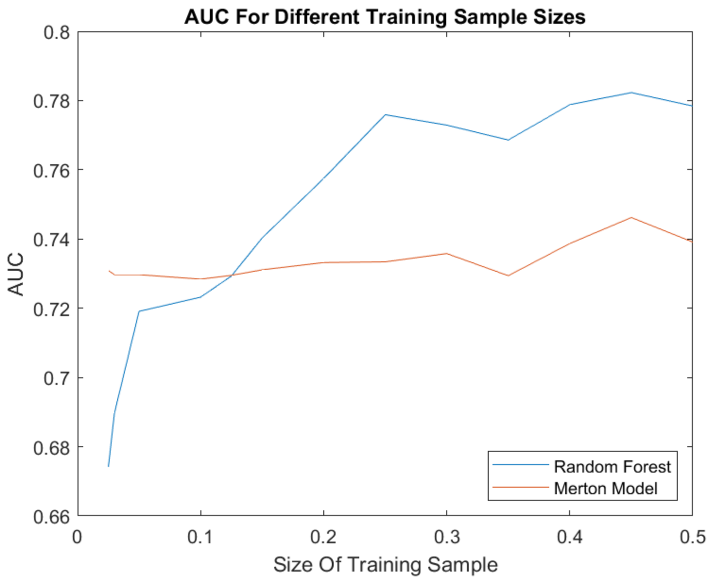

Figure 15. AUC for different training sample sizes for both the Merton model and random forest.

A further illustration is provided in figures 16 and 17. They show the difference between the ROC curve of the random forest when using 50% of the total dataset as the training sample in contrast to using only 10% of it. As can be spotted, in the first case the ML technique clearly outperforms Merton model, while in the second case both approaches seem to have a similar predictive power.

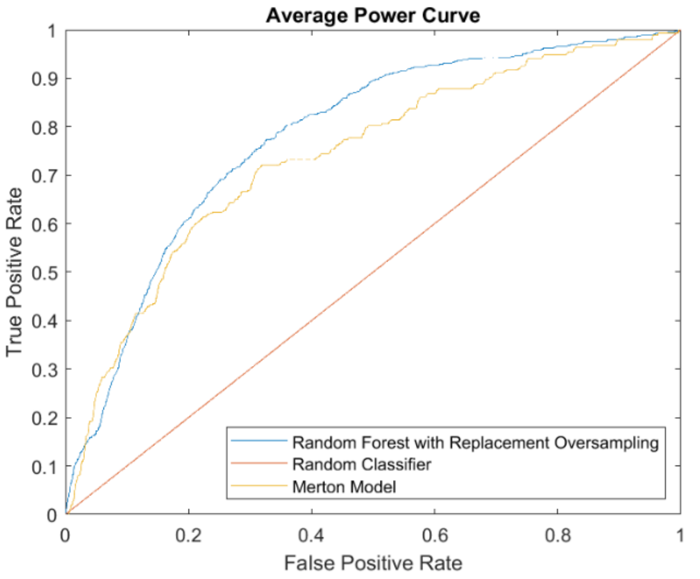

Figure 16. Receiving Operating Characteristic (ROC) curve for Merton model and random forest when using 50% of the original dataset as training sample.

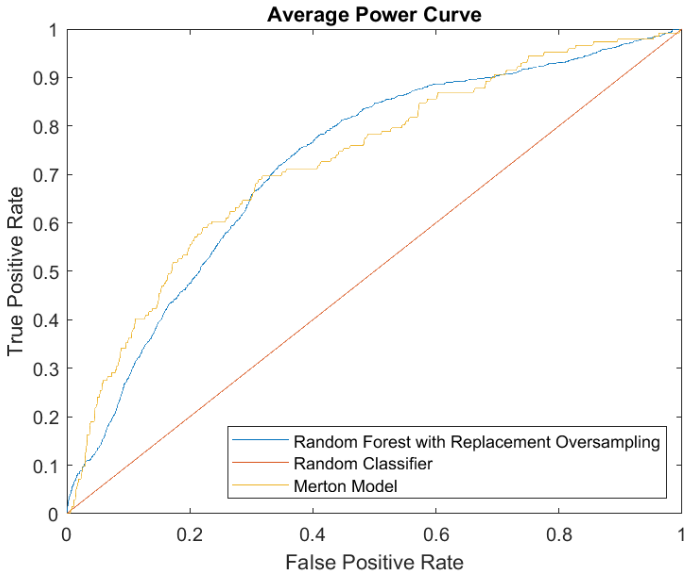

Figure 17. Receiving Operating Characteristic (ROC) curve for Merton model and random forest when using 10% of the original dataset as training sample.

As is developed in the following section, our findings confirm that ML techniques are not suitable for every situation if the goal is to predict corporate defaults accurately. Hence, according to our study, random forest requires a dataset containing at least 9,600 observations, plus 13 or more relevant financial variables. Besides, the application of an over-sampling technique is key to gaining predictive power, with SMOTE with 100% under-sampling of the majority class being the best option.

7. Discussion and analysis

In this paper, we review two completely different methods to predict corporate defaults. While the classic Merton model has been used for decades, newer techniques based on algorithms are becoming more and more popular. In recent literature, there is a trend of highlighting the accuracy of ML techniques for default prediction. However, many of those studies do not consider the fact that, in practice, the access to relevant data can be very limited, so their findings can be misleading for practitioners. Our results show in which cases is it better to avoid the use of those data-intensive techniques and stick to a more conventional approach.

7.1. Discussion of the main findings

According to our results, there are several conditions that should be fulfilled to get a decent accuracy from random forest. First of all, a sufficient number of predictors is necessary. Using the full dataset, at least 13 explanatory variables have to be used to match the performance of Merton model. If considering all the predictors, around 9,600 observations should be taken into account, where there should be approximately a minimum of 20 default observations. It has to be added that neither of those mentioned prerequisites would be enough without SMOTE and under-sampling of the majority class.

If the previous requirements are not fulfilled it is better to use Merton model, since its performance does not depend on the sample size and only requires four input variables. Thus, in a real-life situation, given the case that one has to evaluate whether some companies are going to default, it is more convenient to apply Merton model for cases of few observations. It also takes much less time to perform, which is a clear advantage when time constraints occur. Hence, even if there is some mathematical optimization going on to solve the two simultaneous equations, it is yet much quicker than going the random forest route. Merton model does also not require any variable preselection or any additional manipulations with the data. Additionally, using ML techniques requires some advanced knowledge, whereas calculations with the Merton model do not need deeper technical skills, since it can almost be simplified down to plugging in numbers in an equation.

The dataset we use is not ideal, and shuffling up the validation sample for each iteration makes it even more imperfect. However, this is supposed to represent the randomness of data that could be encountered in a practical setting, in addition to the luck factor of what data goes to the training and validation sample, with the associated increase or decrease in performance. The necessity for practicality would mean that, in any scenario of subpar data, lack of time, or absence of specific knowledge, Merton model should be used, which explains why it is still widely used to estimate default probabilities.

In contrast, even though the classifications are not perfect every time, from a regulator perspective the most important observations are the default ones, that is, misclassifying high-risk firms as low risk. The opposite ones are not that destructive to get wrong, since a regulator would usually rather overestimate risk than underestimate it. That is an argument against the use of Merton model, since it has been empirically proved that it underestimates the probability of default (Bharath & Shumway, 2008).

Finally, it is important to remark that, in previous literature, models are usually tested in an environment that maximizes their performance. Contrary to that, our research has more to do with cases when the perfect environment is not reached and the circumstances are not constructed to create a certain outcome.

7.2. Limitations

When it comes to limitations, the hardware available to us imposes some restrictions, since running machine learning algorithms takes a relatively long time and is resource-intensive. Theoretically, random forest gives the best results when an infinite number of trees are grown. Hence, one is supposed to attempt to grow as many trees as possible for better results. In our case, due to time and resource constraints, the number of trees grown for each iteration is limited to 50, and despite the out-of-bag classification error stabilizing even before that number of trees, growing a larger forest would have probably been more insightful. However, we are of the opinion that the number is adequate for our research purpose.

In terms of variable preselection, the use of LASSO was to be optimal. Nonetheless, it takes an excessive amount of time to run on ordinary devices which are not tailored to do heavy computing work. Therefore, given the time frame for writing this essay, it was not a suitable option to apply.

From a data perspective, a larger dataset is always preferable, both in terms of observations and variables. However, data on defaults and the associated financial variables can be difficult to find, since quite many defaulted companies later go on to be delisted from public stock trading. However, gaining more information would lead to the problem mentioned before, i.e., excessive computational time. Furthermore, it makes sense not to have a huge dataset since the study focuses on data scarcity scenarios.

Additionally, one might argue that a direct comparison between the two methods is not completely possible and does not give a clear picture of what the takeaway from this should be. Merton model is not created to have a clear decision threshold to classify firms as defaults or non-defaults, whereas random forest is based on classification trees, which are built just for that purpose. Therefore, one might have to critically judge what the comparison means exactly: if one method ranks observations better than the other without giving an accurate classification or if the methods can be compared by performance of classification at every decision threshold. If it is the latter, then it does not help in a real-life setting, since then a decision threshold must be decided to distinguish defaults from non-defaults and high risk from low risk.

Another fact to think about is that for models such as random forest there is no statistical inference, meaning that it is not clear whether the acquired results are statistically significant or just a fluke. Running the model 10 times with different samples might not be enough to say how the model performs exactly on average every time, while Merton model always performs in the same specific way. In essence, random forest still is a black box while Merton model provides clear intuition for the outcomes produced.

8. Conclusion

In this paper, we compare two different methods for predicting corporate defaults: the Merton model, developed by Merton in 1974, and random forest, created by Breiman in 2001. Both are different in terms of the underlying mechanism they use to predict defaults. Hence, while the Merton model is a structural model formed by various equations that take the economic situation of the firm into account, random forest is an ML technique that generates decorrelated classification trees to predict whether a company will default or not.

Although previous literature shows that ML techniques generally outperform classic approaches when predicting situations of distress in firms, it is important to consider that those studies usually include large datasets, non-financial variables, or non-public information that is difficult to acquire in practice. Our study aims to obtain the minimum required number of observations and financial variables that are needed to consider random forest as a suitable option when predicting corporate defaults, since in reality there can be constraints in terms of time or resources that make the application of ML techniques unsuitable or, at least, not recommendable. Therefore, we only use publicly available financial variables and a dataset with a small number of defaults, with the purpose of simulating a real-life scenario of data scarcity.

We first run random forest using only the four inputs of the Merton model as explanatory variables: debt value, growth rate (proxied by the risk-free rate), equity value, and equity volatility. As expected, its accuracy as measured by AUC is lower than the one of the Merton model.

Then, we add more financial variables to the study in order to check how much the accuracy improves. Initially, 50 variables are obtained, but after a preselection process that involves running a logit regression and checking for multicollinearity, we get a final number of 20 variables that are added to the dataset. Besides, due to the imbalanced nature of the data (only 0.18% of all the observations correspond to defaulted companies) we apply SMOTE, an over-sampling technique developed by Chawla et al. (2002). Results show that, when using the aforementioned variables and SMOTE with 100% under-sampling of the majority class, the predicting power of random forest visibly outperforms that from the Merton model, again as expected.

The key point of the whole paper is to find both the minimum number of predictors and the minimum number of observations required for random forest to be a better option than Merton model in predicting corporate defaults. We find that the 13 best predictors are needed to have a similar AUC between both models and that at least 9,600 observations are required for random forest to perform acceptably, providing that the percentage of non-default observations holds.

According to our findings, random forest should only be used when the previous conditions are met. Since it is a data-intensive technique, it is important to know the point where it actually starts working, especially for empirical studies where data are scarce. Besides, we also show that the application of an over-sampling technique is indispensable when dealing with such an imbalanced dataset and that for default prediction and random forest, SMOTE with 100% under-sampling of the minority class seems to be the preferable option.

This paper also aims to give some food for thought for future research. Along these lines, it could be optimal to test different ML techniques to check their performance when data are scarce and compare it with the one from the Merton model. Concretely, we consider artificial neural networks to be an interesting method that should be tested in the future in this kind of setting. Besides, adding non-financial variables and focusing the data scarcity situation only on a reduced number of observations would be insightful since it is to be expected that those variables have a higher predicting power, which could imply that the sample size might be reduced without losing accuracy.

Corporate default prediction has been a major topic in finance for decades and nothing seems to indicate that situation will change in the future. Therefore, it is essential to keep applying the most innovative techniques in order to obtain models that can accurately predict whether a firm is to be in distress. However, keeping in mind the importance of not losing touch with the most classic approaches is also key, since there will always be situations where they have something to say.

References

Afik, Z., Arad, O., & Galil, K. (2016). Using Merton Model for Default Prediction: An Empirical Assessment of Selected Alternatives, *Journal of Empirical Finance*, vol. 35, pp. 43-67.

Ali Khan, M., Izhar Shah, M., Faisal Javed, M., Ijaz Khan, M., Rasheed, S., El-Shorbagy, M., Roshdy El-Zahar, E. & Malik, M. (2022). Application of Random Forest for Modelling of Surface Water Salinity, *Ain Shams Engineering Journal*, vol. 13, no. 4, p. 101635.

Altman, E.I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *The Journal of Finance*, vol. 23, no. 4, pp. 589-609.

Altman, E.I. (1983). Corporate Financial Distress. A Complete Guide to Predicting, Avoiding, and Dealing with Bankruptcy, New York: John Wiley & Sons.

Altman, E.I., Iwanicz-Drozdowska, M., Laitinen, E. & Suvas, A. (2014). Distressed Firm and Bankruptcy Prediction in an International Context: A Review and Empirical Analysis of Altman's Z-Score Model, *SSRN Electronic Journal*.

Barboza, F., Kimura, H. & Altman, E.I. (2017). Machine Learning Models and Bankruptcy Prediction, *Expert Systems with Applications*, vol. 83, pp. 405-417.

Behr, A. & Weinblat, J. (2016). Default Patterns in Seven EU Countries: A Random Forest Approach, *International Journal of the Economics of Business*, vol. 24, no. 2, pp. 181-222.

Bharath, S. & Shumway, T. (2004). Forecasting Default with the KMV-Merton Model, *SSRN Electronic Journal*.

Bharath, S. & Shumway, T. (2008). Forecasting Default with the Merton Distance to Default Model, *Review of Financial Studies*, vol. 21, no. 3, pp. 1339-1369.

Black, F. & Scholes, M. (1973). The Pricing of Options and Corporate Liabilities, *Journal of Political Economy*, vol. 81, no. 3, pp. 637-654.

Breiman, L. (1996). Bagging Predictors, *Machine Learning*, vol. 24, no. 2, pp. 123-140.

Breiman, L. (2001). Random Forests, *Machine Learning*, vol. 45, pp. 5-32.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-Level-Smote: Safe-Level-Synthetic Minority Over-sampling Technique for Handling the Class Imbalanced Problem, Advances in Knowledge Discovery and Data Mining, Berlin: Springer, pp. 475-482.

Chawla, N., Bowyer, K., Hall, L. & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique, *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357.

Forssbæck, J. & Vilhelmsson, A. (2017). Predicting Default - Merton vs. Leland, *SSRN Electronic Journal*.

GitHub. (2020). Oversampling Imbalanced Data, Available online: https://bit.ly/3auljaU [Accessed 7 June 2022]

Gurm, H., Kooiman, J., LaLonde, T., Grines, C., Share, D. & Seth, M. (2014). A Random Forest Based Risk Model for Reliable and Accurate Prediction of Receipt of Transfusion in Patients Undergoing Percutaneous Coronary Intervention, *PLoS ONE*, vol. 9, no. 5, p. e96385.

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: a New Over-sampling Method in Imbalanced Data Sets Learning, Advances in Intelligent Computing, Berlin: Springer, pp. 878-887.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. 2008 IEEE International Joint Conference on Neural Networks, pp. 1322-1328. Available online: https://bit.ly/39AGnvO [Accessed 7 June 2022]

Itô, K. (1944). Stochastic integral, *Proceedings of the Japan Academy, Series A, Mathematical Sciences*, vol. 20, no. 8, pp. 519-524.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R, New York: Springer.

Kim, H., Cho, H. & Ryu, D. (2020). Corporate Default Predictions Using Machine Learning: Literature Review, *Sustainability*, vol. 12, no. 16, p. 6325.

Kim, H., Cho, H. & Ryu, D. (2021). Corporate Bankruptcy Prediction Using Machine Learning Methodologies with a Focus on Sequential Data, *Computational Economics*, vol. 59, no. 3, pp. 1231-1249.

Kim, H., Cho, H. & Ryu, D. (2021). Predicting Corporate Defaults Using Machine Learning with Geometric-lag Variables, *Investment Analysts Journal*, vol. 50, no. 3, pp. 161-175.

Kohv, K. & Lukason, O. (2021). What Best Predicts Corporate Bank Loan Defaults? An Analysis of Three Different Variable Domains, *Risks*, vol. 9, no. 2, p. 29.

Lock, D. & Nettleton, D. (2014). Using Random Forests to Estimate Win Probability Before Each Play of an NFL Game, *Journal of Quantitative Analysis in Sports*, vol. 10, no. 2.

Merton, R. (1974). On the Pricing of Corporate Debt: The Risk Structure of Interest Rates, *The Journal of Finance*, vol. 29, no. 2, pp. 449-470.

Moscatelli, M., Parlapiano, F., Narizzano, S. & Viggiano, G. (2020). Corporate Default Forecasting with Machine Learning, *Risk Management Magazine*, vol. 15, no. 3, pp. 4-14.

Ohlson, J. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy, *Journal of Accounting Research*, vol. 18, no. 1, pp. 109-131.

Tan, Z., Yan, Z. & Zhu, G. (2019). Stock Selection with Random Forest: An Exploitation of Excess Return in the Chinese Stock Market, *Heliyon*, vol. 5, no. 8, p. e02310.

Yusof, N.M. & Jaffar, M.M. (2011). The Analysis of KMV-Merton Model in Forecasting Default Probability. 2012 IEEE Symposium on Humanities, Science and Engineering Research, pp. 93-97. Available online: https://bit.ly/36bFWXI [Accessed 7 June 2022]

Zhu, L., Qiu, D., Ergu, D., Ying, C. & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm, *Procedia Computer Science*, vol. 162, pp. 503-513.

Appendix

Table A-1. Area Under Curve (AUC) of all predictors after running a logit regression.

Volatility	**Total Equity**	**Total debt**	**Risk-free**	**REV**
0.6853	0.6015	0.4961	0.7572	0.5096
AST	**EBIT**	**EBITDA**	**N_DBT**	**EQT**
0.5085	0.6736	0.6210	0.4324	0.5721
CFL/DBT	**DBT**	**CFL**	**FIX_AST**	**CRR_AST**
0.4742	0.4610	0.6656	0.4207	0.4669
ST_DBT	**W_CAP**	**RET_EAR**	**N_SHA**	**SHA_P**
0.5692	0.4975	0.7720	0.5027	0.6337
EPS	**ROE**	**ROA**	**EV/MCAP**	**NPM**
0.2986	0.7122	0.6961	0.6304	0.7158
DBT/AST	**FIX_AST/AST**	**ST_AST/ST_DB**	**EQT/FIX_AST**	**DBT/CFL**
0.6194	0.5326	0.5069	0.3244	0.6092
ST_DBT/DBT	**W_CAP/AST**	**RET_EAR/AST**	**EBIT/AST**	**REV/AST**
0.6076	0.5082	0.7551	0.6872	0.5524
EBIT/REV	**EBITDA/REV**	**N_DBT/EQT**	**N_DBT/EBITDA**	**CASH/AST**
0.5879	0.5412	0.4550	0.4391	0.4794
CFL/DBT	**MCAP/DBT**	**ΔAST**	**BVPS**	**SHA_P/BVPS**
0.6370	0.6321	0.5888	0.5807	0.6312
ΔREV	**ΔEMP**	**ΔROE**	**LOG_REV**	**LOG_AST**
0.4223	0.4768	0.5929	0.5179	0.5227
LOG_CASH	**LOG_DBT**	**LOG_CA**	**Log STD**	
0.4841	0.5287	0.4805	0.5706	

Table A-2. Correlation matrix of the 28 remaining variables after running the logit regression.

DBT/AST	DBT/CFL	ST_DBT/DBT	RET_EAR/AST	EBIT/AST	REV/AST	EBIT/REV	CFL/DBT	MCAP/DBT	ΔAST	BVPS	SHA_P/BVPS	ΔROE	LOG_STD
1.0000													
0.0157	1.0000												
-0.0848	0.0036	1.0000											
-0.0425	0.0024	-0.0143	1.0000										
-0.0147	-0.0002	-0.0186	0.0583	1.0000									
-0.0052	-0.0021	-0.0078	-0.0023	0.8939	1.0000								
-0.0099	0.0006	-0.0079	0.0982	0.0198	0.0116	1.0000							
-0.0089	-0.0002	0.0049	0.0166	0.0055	0.0077	0.0096	1.0000						
0.0035	0.0000	0.0000	-0.0015	-0.0003	-0.0023	-0.0001	0.0753	1.0000					
-0.0136	-0.0004	0.0043	0.1190	0.0027	-0.0412	0.0072	0.0069	0.0026	1.0000				
-0.0079	-0.0002	0.0190	-0.0099	-0.0037	-0.0046	0.0003	-0.0063	0.0006	0.0044	1.0000			
-0.0045	0.0001	0.0013	0.0009	0.0000	0.0000	0.0000	0.0000	0.3447	0.0033	0.0000	1.0000		
0.0244	-0.0054	-0.0042	-0.0157	0.0027	0.0118	-0.0029	-0.0056	-0.0093	-0.0746	-0.0002	0.0041	1.0000	
0.3198	0.0105	0.5103	0.0752	-0.0092	0.0256	0.0136	-0.0039	-0.0084	0.0091	0.0099	0.0005	0.0021	1.0000

	Volatility	Equity	Risk-free	EBIT	EBITDA	EQT	CFL	ST_DBT	RET_EAR	SHA_P	ROE	ROA	EV/MCAP	NPM
Volatility	1.0000													
Equity	-0.0969	1.0000												
Risk-free	0.0199	-0.0797	1.0000											
EBIT	-0.0148	0.0407	-0.0159	1.0000										
EBITDA	-0.0161	0.0458	-0.0173	0.9768	1.0000									
EQT	-0.0200	0.0676	-0.0245	0.7481	0.8055	1.0000								
CFL	-0.0158	0.0441	-0.0164	0.9575	0.9926	0.7838	1.0000							
ST_DBT	-0.0139	0.0711	-0.0203	0.2654	0.2480	0.3802	0.2076	1.0000						
RET_EAR	-0.0199	0.0642	-0.0235	0.7946	0.8505	0.9538	0.8431	0.2835	1.0000					
SHA_P	0.0195	-0.0012	0.0030	0.0189	0.0194	0.0169	0.0189	0.0047	0.0180	1.0000				
ROE	-0.0329	0.0032	-0.0021	0.0006	0.0005	-0.0001	0.0005	-0.0003	0.0001	-0.0013	1.0000			
ROA	-0.0431	0.0083	-0.0027	0.0016	0.0015	0.0008	0.0015	-0.0003	0.0011	-0.0043	0.2780	1.0000		
EV/MCAP	0.0000	-0.0007	-0.0069	0.0001	0.0001	0.0003	0.0000	0.0026	0.0000	-0.0001	-0.0004	-0.0002	1.0000	
NPM	-0.0277	0.0052	0.0053	0.0009	0.0009	0.0012	0.0009	0.0009	0.0011	0.0002	0.0075	0.0198	0.0002	1.0000
DBT/AST	-0.0029	0.0293	-0.0755	0.0076	0.0104	0.0159	0.0096	0.0231	0.0141	-0.0066	-0.0271	-0.0235	0.0204	-0.0108
DBT/CFL	-0.0074	0.0003	0.0016	0.0003	0.0001	0.0007	0.0002	0.0029	0.0000	-0.0002	-0.0002	0.0000	-0.0001	0.0005
ST_DBT/DBT	0.0368	0.0084	-0.0081	0.0071	0.0065	0.0109	0.0056	0.0373	0.0108	0.0154	-0.0152	-0.0174	-0.0008	-0.0080
RET_EAR/AST	-0.1765	0.0396	0.0552	0.0064	0.0070	0.0089	0.0068	0.0043	0.0091	-0.0098	0.0235	0.0423	0.0000	0.0898
EBIT/AST	-0.0393	0.0071	-0.0050	0.0013	0.0012	0.0004	0.0012	-0.0007	0.0007	-0.0038	0.2753	0.8328	-0.0003	0.0166
REV/AST	0.0192	-0.0183	-0.0084	-0.0064	-0.0062	-0.0094	-0.0052	-0.0125	-0.0065	-0.0049	0.2407	0.7422	-0.0027	0.0099
EBIT/REV	-0.0298	0.0061	0.0062	0.0010	0.0011	0.0014	0.0010	0.0010	0.0012	0.0002	0.0070	0.0190	0.0003	0.9011
CFL/DBT	-0.0168	0.0058	-0.0037	-0.0002	-0.0002	-0.0003	-0.0002	-0.0004	-0.0002	-0.0062	0.0036	0.0068	-0.0001	0.0080
MCAP/DBT	0.0037	-0.0005	-0.0010	-0.0001	-0.0002	-0.0003	-0.0002	-0.0004	-0.0003	0.2760	-0.0007	-0.0010	-0.0001	-0.0002
ΔAST	-0.0053	0.0121	0.0220	-0.0010	-0.0022	-0.0034	-0.0018	-0.0042	-0.0030	0.0019	0.0107	0.0143	-0.0024	0.0085
BVPS	0.0220	0.0036	0.0015	0.0661	0.0676	0.0699	0.0644	0.0228	0.0703	0.8633	-0.0012	-0.0041	-0.0001	0.0003
SHA_P/BVPS	-0.0018	-0.0004	0.0024	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001	0.0907	0.0010	-0.0001	0.0000	0.0000
ΔROE	0.0151	0.0012	0.0002	0.0003	-0.0002	-0.0024	0.0004	-0.0028	-0.0025	-0.0026	-0.0134	-0.0064	0.0014	-0.0024
LOG_STD	-0.0784	0.1957	-0.0719	0.0714	0.0773	0.1012	0.0724	0.0848	0.0927	0.0013	-0.0201	-0.0114	0.0080	0.0109