A Comparative Analysis of Machine Learning Models for Corporate Default Forecasting

Alexander Seum

Dissertation written under the supervision of

Professor Eva Schliephake

Dissertation submitted in partial fulfilment of requirements for the MSc in Economics, at the Universidade Católica Portuguesa, 5.4.2023.

A Comparative Analysis of Machine Learning Models for Corporate Default Forecasting

Alexander Seum

Abstract

This study examines the potential benefits of utilizing machine learning models for default forecasting by comparing the discriminatory power of the random forest and XGBoost models with traditional statistical models. The results of the evaluation with out-of-time predictions show that the machine learning models exhibit a higher discriminatory power compared to the traditional models. The reduction in the sample size of the training dataset leads to a decrease in predictive power of the machine learning models, reducing the difference in performance between the two model types. While modifications in model dimensionality have a limited impact on the discriminatory power of the statistical models, the predictive power of machine learning models increases with the addition of further predictors. When employing a clustering approach, both traditional and machine learning models exhibit an improvement in discriminatory power in the small, medium, and large firm size clusters compared to the previous non-clustering specifications. Machine learning models exhibit a significantly higher ability to classify micro firms. The findings of this research indicate that the machine learning models exhibit superior discriminatory power compared to the traditional models across the different specifications. Machine learning models can be used to forecast the potential impact of corporate default of non-financial micro cooperations on the Portuguese labour market by estimating the number of jobs at risk.

Key words: credit risk, default forecasting, machine learning, random forest

Acknowledgements

I would like to express my sincere appreciation to Professor Eva Schliephake for serving as my supervisor during the completion of my master's thesis. Throughout the process, her guidance, expertise, and support were invaluable to me.

I am grateful to Banco de Portugal for providing me with the opportunity to conduct my research and for all of the resources and support they provided me. I would like to extend a special thanks to Miguel Portela and the whole BPLIM team, who went above and beyond to help me with my work.

Table of Contents

List of Figures

List of Tables

I. Introduction

The field of data analytics plays a key role in comprehending the economic context by providing insights into complex economic dynamics such as market trends or consumer behaviour. Given the importance of data analytics, the development of new technologies is an ongoing and essential process. The advancement in computer sciences along with higher computational power makes the analysis of big data and the use of new analytical methods on a broader basis applicable. New technologies in conjunction with more accurate analytical methods allow economic scientists to achieve a better understanding of economical behaviour and give better advice to the decision makers.

Over recent years, the discussion surrounding the use of machine learning techniques in financial services industry has gained momentum because of the availability of big data and the ease with which the machine learning algorithms can be applied.

One area which especially benefits from the new opportunities is risk management where data analytics historically plays a key role. The effective management of credit portfolios is a crucial aspect of bank management, banking supervisors and central banks. With the regulation of the Basel II Accord, the regulatory framework was established and became mandatory for the relevant institutions. In this context the prediction of defaults is essential. Traditionally, statistical methods were used to evaluate credit risk, however in recent years machine learning models have gained popularity as a tool for risk assessments.

The objective of the study is to explore the potential benefits of implementing machine learning algorithms for default forecasting. To assess their predictive power, the machine learning models random forest and XGBoost are compared to the traditional statistical models linear discriminant analysis, logistic regression and penalized logistic regression based on their ability to predict corporate default of non-financial cooperations.

To evaluate the various models, a comprehensive dataset containing firm-level variables and financial indicators for Portuguese non-financial cooperations is utilized. To quantify the performance of each model, the area under the receiver operating characteristics curve (AuROC) is calculated based on the observed default data and the out-of-sample probabilities of defaults in the following year obtained by the respective model from 2009 until 2020.

This analysis reveals the following main results:

i. When the various models are evaluated using out-of-time predictions, the machine learning models outperform traditional statistical models in terms of their discriminatory power.
ii. Decreasing the sample size of the training dataset does not significantly impact the discriminatory power of the statistical models, whereas the performance of the machine learning models declines as the sample gets smaller, reducing the performance gap.
iii. The ability of the machine learning models to distinguish default and non-default instances improves with addition of further variables. In contrast, the statistical models maintain their level of performance for different number of features.
iv. When implementing a clustering approach, the discriminatory power of both traditional and machine learning model improves in the small, medium, and large firm size clusters compared to the previous non-clustering specifications. However, in the micro cluster which contains for most of the firms in the dataset, machine learning models have a significantly higher predictive power.
v. Finally, this study estimates the number of jobs at risk in the next year due to corporate default for different discrimination thresholds and establishes a correlation between the estimated jobs at risk and the actual observed unemployment rate.

The results of this research contribute to the existing literature on default forecasting in a variety of ways. First, the study shows the effectiveness of machine learning algorithms to predict corporate default for non-financial cooperations using out-of-time predictions which simulate real-life predictions based on currently available information. In addition, this study estimates possible labour market implications due to the default non-financial micro cooperations by estimating the number of jobs at risk.

The rest of paper is organized as follows: chapter II offers an overview of the related literate on machine learning and default forecasting; chapter III describes the process of developing a predictive model including a description of the dataset and the data pre-processing steps; chapter IV includes the explanatory data analysis; chapter V describes the various models and the hyperparameter tuning process; chapter V presents the result across various

specifications; chapter VI discusses the limitations and implications; chapter VII concludes the main findings of the study.

II. Literature review

The use of machine learning models in the financial industry has become increasingly popular in recent years. As Chakraborty & Joseph, 2017 provide on introduction to the use of machine learning in the context of central banking as well as its implications for policy analysis. After presenting the fundamental concepts of machine learning and various models including random forest, neural networks, support vector machine, the authors provide several case studies of successful applications of machine learning models in the context of central banking. One of those case studies is about banking supervision under imperfect information in which various models are trained to identify problematic institutions based on balance sheet items. Their results show that machine learning models outperform conventional models in predicting advisory alerts and the discriminatory power of the random forest model is 11.1 percentage points higher compared to the logistic regression model.

Especially in the context of credit risk the benefits of implementing machine learning models have been well documented in the literature. For the classification of mortgage loans, Galindo & Tamayo, 2000 show that both the CART decision tree-based and the neural network models can achieve a lower error rates compared the standard Probit model. This aligns with the existing literature stating that neural networks are more accurate and robust when assessing credit risk compared to traditional models (Oreski et al., 2012).

Furthermore, machine learning techniques are used for the classification of corporate default. Moscatelli et al., 2020 compare the performance of the random forest and gradient boosting model to traditional statistical models in predicting corporate default of Italian non-financial cooperations. When the models are trained using publicly available information the machine learning models outperform the statistical models by approximately 2.6 percentage points. If high-quality information such as credit behavioural indicators are included in the training set, the performance gap between model types decreases.

Moreover, based on the findings of the literature boosting algorithm such as XGBoost are identified as the top-performing approach in credit scoring (Chang et al., 2018; Hamori et al., 2018). When compared to other machine learning models such as deep neural, networks, bagging or random forest, models based on boosting have a higher discriminatory power in credit scoring (Hamori et al., 2018).

As shown in the study of Frey & Osborne, 2017, machine learning models can also be used to estimate potential labour market implications. The authors examine how different occupations are affected by computerization and automatization in the United States. They predict the likelihood of a job being automated in the future using a machine learning algorithm and estimate that about 47% of the employment in the United States is at risk in the next decades.

III. Model Building

Predictive analysis is a data analysis technique that uses empirical models such as statistical or machine learning models to identify patterns and extract information from large and complex datasets to make accurate predictions of future outcomes. A predictive model also includes the evaluation of its predictive power (Shmueli & Koppius, 2011), which refers to its ability to correctly forecast future outcomes. The prediction of future default of Portuguese non-financial cooperations can be classified as a predictive learning problem because the training of the model includes the use of historical, past data to predict the likelihood of a future event to occur.

III.A. Research Objectives

The goal of this study is to predict the probability of corporate default in the next year and compare the discriminatory power of machine learning models and traditional statistical models. A firm is categorized as in default in a given year if the non-performing credit exceeds 5% of total credit drawn for at least one month in a year. A credit is classified as non-performing credit if the payment is past due for more than 90 days following the definition of default in Article 178 of the Regulation by the European Banking Authority (EBA, 2016). This study compares the predictive power of various models with several specifications including different types of forecasts, sample sizes, dimensionality, a clustering approach and estimates the impact of corporate default on the Portuguese labour market which will be explained in more detail in chapter V.

III.B. Data Collection and Study Design

This study uses the Central Balance Sheet Database provided by Banco de Portugal, which includes economic and financial information on all non-financial firms operating in Portugal for the period of 2008 until 2020. Most of the database is based on the information reported through Informação Empresarial Simplificada (IES). For companies with organized accounting, the IES is a mandatory annual declaration which allows them to fulfil several tax

obligations, including the delivery of annual accounting data to Banco de Portugal. This ensures that the information in the database is accurate, and the sample is representative. Since the IES is a mandatory declaration, the data is collected consistently and on a regular basis and there is no issue due to a sample selection bias. Furthermore, the dataset used in this study also provides a default dummy that indicates whether a firm defaulted in a given year following the definition described.

This database contains a variety of firm-level variables of Portuguese non-financial cooperations. Based on balance sheet variables in the dataset, various financial ratios and their corresponding growth rates are calculated. Growth rates are additionally calculated because they can provide a more informative perspective for the evaluation of the current state of a firm. The initial dataset contains a total of 109 variables including firm descriptives, balance sheet variables, and financial ratios with their corresponding growth rates.

III.C. Data Preparation

Data-pre-processing is a crucial step in the building of any predictive model because it has a significant impact on its performance. It describes the process of transforming the raw data into a better structured and efficient format.

Data cleaning is the process of identifying and correcting various error and issues within the dataset which otherwise can have a negative impact on the performance of the predictive model. This includes dealing with issues like missing data, infinite values, and outliers. Initially, the original dataset contains a total of 2,772,451 instances over the observation period. When addressing the potential issues of missing data and infinite values, it is important to understand the data and identify the origin. Since the dataset is based on a mandatory annual declaration, the accounting data provided by Banco de Portugal is complete and has no infinite values or missing data. However, the computation of several financial ratios and their respective growth rates results in the generation of infinite and missing values which can be attributed to the division by zero. To handle these occurrences, observations that include either missing or infinite values are dropped from the dataset which reduces the number of observations to 2,550,800.

The next step in the data cleaning process is the handling of outliers in the dataset. It is important to address them since outliers can have a significant negative impact on the performance of a predictive model leading to inaccurate forecasts. The computation of the financial ratios and their corresponding growth rates generates in several natural extreme

outliers that are significantly different from the remaining data. Therefore, a variable-specific trimming approach is employed that removes the top and bottom one percent of observations for these specific variables resulting in a final sample size of 2,469,758 observations.

Furthermore, feature normalization is implemented in this study which is as an important pre-processing step to rescale the predictors using Min-Max scaling. As a result, predictors range from 0 to 1 which helps can help to improve the model's performance and training speed.

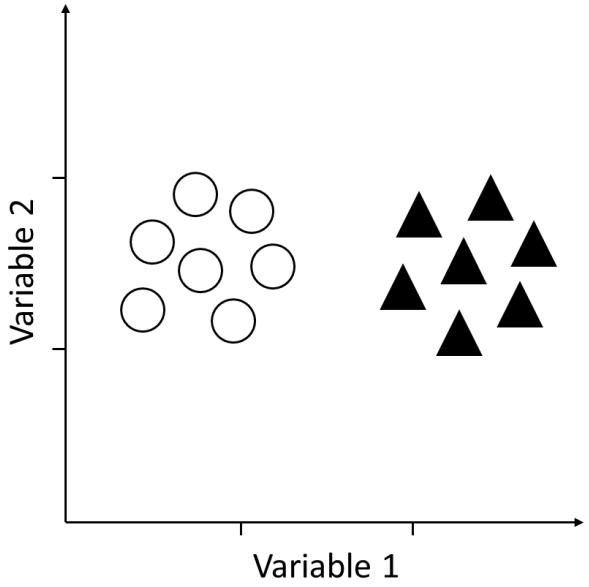
III.D. Selection of Variables

The selection of the most relevant features is an important step in the building of a predictive model as it involves the selection of the most relevant variables from the original dataset. As previously described, the initial dataset includes a total of 109 variables. However, it is likely that many of these variables are noisy and not informative and therefore have no or even a negative impact on the models' performance. Excluding irrelevant data is not only important to reduce the dimensionality of the dataset but also improves the performance and the generalization of the model. To select the most relevant variables for the classification prediction, there are two main techniques: supervised and unsupervised.

III.D.i. Supervised Feature Selection

Supervised feature selection techniques involve the use of the dependent variable, which is the variable that the model tries to predict, to identify and remove irrelevant features from the dataset. For instance, Variable 1 is able to distinguish between the two classes, while Variable 2 is similar for both classes and therefore does not enable the differentiation between the two classes as illustrated in FIGURE I which is based on Dy and Brodley (2004) and Haar et al. (2019). Consequently, Variable 2 is considered irrelevant and therefore removed from the dataset. There are numerous methods of a supervised feature selection, however, they can be classified into one of three groups: filter methods, wrapper methods, and embedded methods (García et al., 2015; Guyon et al., 2008).

FIGURE I IRRELEVANT FEATURES



Filter methods, which were implemented in this study, use statistical methods to assess the relationship between each input variable and the target variable. The obtained results of the statistical model are used to filter and choose the most relevant input variables that will be included in the final model (Kuhn & Johnson, 2013).

In contrast of evaluating the relationship between the input and target variable individually, wrapper methods evaluate the performance of various subsets of features. In these methods, features are added or removed to find the optimal combination to maximize the model's performance (Kuhn & Johnson, 2013). The most common wrapper methods are forward selection, backward elimination, and bi-directional elimination. However, compared to filter methods wrapper methods are more computationally intensive and have a tendency of overfitting (Kohavi & John, 1997).

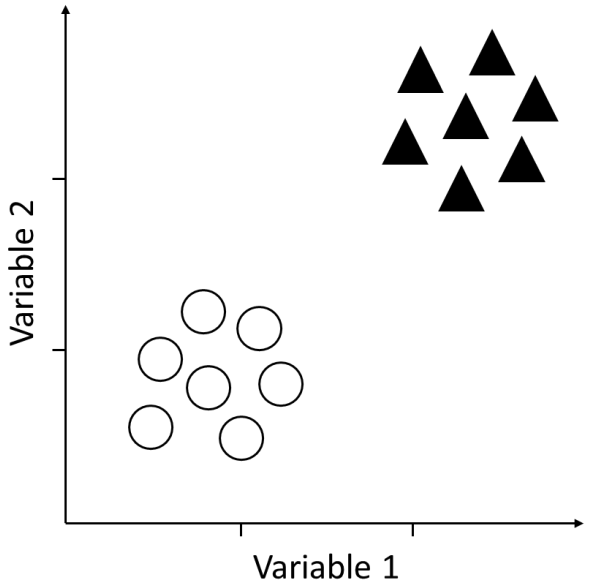
Finally, embedded methods are algorithms in which the feature selection process is integrated in the model's training process. However, not only the subset of the most important features is determined but also the optimal weights of each feature to maximize the accuracy of the model. Examples of embedded methods include tree-based algorithms like random forest or gradient boosting among others.

III.D.ii. Unsupervised Feature Selection

Unsupervised feature selection techniques do not require the use of the dependent variable. Their objective is to remove redundant variables from the feature vector. Two variables are considered redundant if they contain similar information regarding the

discrimination of the two classes. FIGURE II based on Dy and Brodley (2004) and Haar et al. (2019), illustrates Variable 1 and Variable 2 which demonstrate an example of two redundant variables. Hence, one of the variables can be removed without any loss of information.

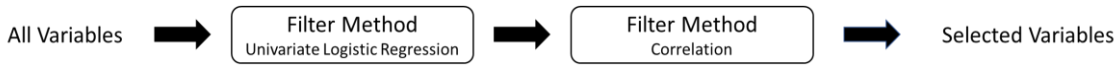
FIGURE II REDUNDANT FEATURES



III.D.iii. Feature Selection Process

This study implements a hybrid combination of both supervised and unsupervised filter methods to select the explanatory variables. From the initial dataset a total of 80 variables were considered as potential predictors of corporate default in the next year. The variable selection follows the steps illustrated in FIGURE III.

FIGURE III VARIABLE SELECTION PROCEDURE



To identify the most relevant features, univariate logistic regressions are utilized to estimate the out-of-sample probabilities of default in the next year. Based on the estimated probabilities and the observed default data the AuROC is calculated for each firm variable. If a variable does not exceed an AuROC of 60% it is considered as irrelevant because of its insufficient ability to correctly distinguish between the default and non-default class and is therefore removed from the dataset. In comparison, other studies use a threshold of 55% (Moscatelli et al., 2020). However, when working with high-dimensionality datasets, choosing a higher threshold, can be advantageous because it helps to reduce the dimensionality of the dataset and to choose the most relevant features. The complete list of all tested features with

their respective AuROCs are attached in the appendix TABLE A.1. Out of 80 variables 26 show a sufficient ability to distinguish the between default and non-defaults and exceed an AuROC of 60%.

When working with a dataset consisting mostly of balance sheet variables it is likely that features are selected that carry similar information to other features, and they are highly correlated with each other. These highly correlated variables do not add any additional information but cause collinearity problems (Kuhn & Johnson, 2013). To prevent collinearity issues and detect highly correlated variable pairs, the Pearson correlation coefficient is calculated. The Pearson correlation coefficient is a common method to measure the linear correlation between two variables. The coefficient ranges from -1 indicating a perfect negative correlation to 1 indicating a perfect positive correlation. In general, two variables are considered strongly correlated if their Pearson correlation coefficient exceeds 0.7. Consequently, out of the 26 relevant variables only variables that do not exceed a linear correlation of 0.7 are retained. After removing the redundant variables, the final dataset contains 15 variables which are displayed in TABLE I with their corresponding AuROC from the univariate logistic regression. In addition, the description of the final variables as well as the correlation matrix is provided in the appendix TABLE A.2 - 3.

TABLE I LIST OF FINAL VARIABLES

Variable	AuROC
Debt-to-Capital Ratio	0.695
Net Income	0.691
Operating Net Income	0.674
Equity	0.672
Income Tax	0.667
Retained Earnings	0.665
Current Ratio	0.659
Mismatch Ratio	0.653
Cash and Bank Deposits	0.646
Cash-to-Assets Ratio	0.641
Supplies and External Services	0.631
Insurance Schemes for Accidents at Work	0.620
Employee Expenses Except Salaries	0.612
Total Expenses	0.612
Obtained Funding	0.608

Note:
Own calculation based on data provided by Banco de Portugal.
The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the univariate logisitic regression.

III.E. Class Imbalance

A common scenario in many classification tasks that needs to be addressed is the imbalance of the dataset. An imbalanced dataset refers to a dataset in which the number of instances of one class is significantly higher compared to the other classes, resulting in an uneven class distribution. In a binary classification problem, an imbalanced dataset exhibits a significant difference in the number of observations that belong to each of the two classes (Kuhn & Johnson, 2013). Normally, most of the instances belong to the negative class which is also referred to as majority class while the remaining positive class instances belong to the minority class.

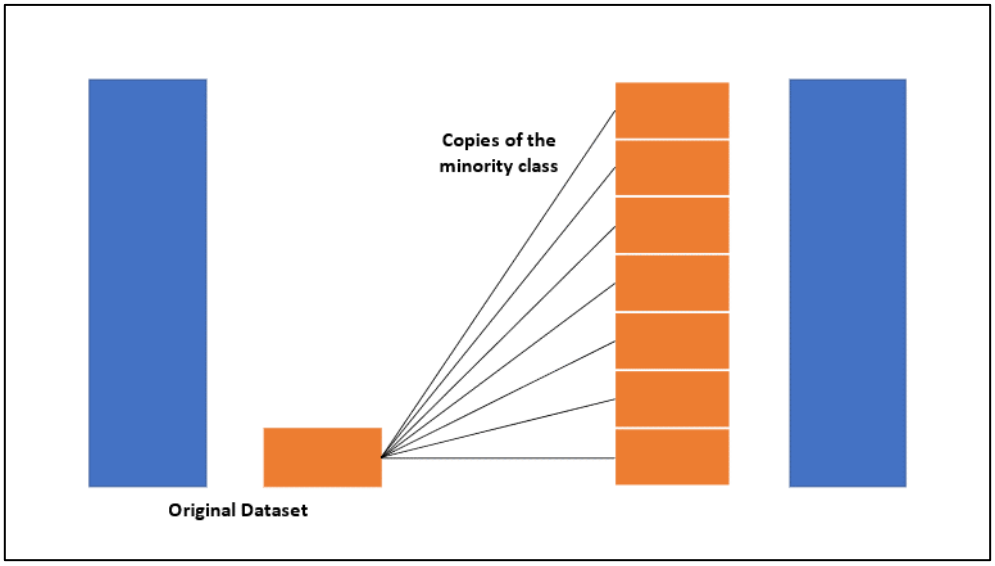
It is common to encounter datasets that are not perfectly balanced and minor disparities in the class representation can be considered negligible. However, there are many real-life classification problems in which the class distribution is severely imbalanced. Examples of imbalanced classification problems are the detection of credit card fraud (Dhankhad et al., 2018), medical diagnosis (Bach et al., 2017; Mena & Gonzalez, 2006), identification of spam messages (Liu et al., 2017) or in default forecasting (Khandani et al., 2010; Moscatelli et al., 2020). The forecasting of defaults poses a classification problem that is inherently imbalanced because the number healthy firms significantly outnumber the firms in default. Accordingly, this research is based on a severely imbalanced dataset in which over the observation period, approximately 92% of the instances belong to the non-default class, while the default instances account for only 8%.

However, severe class imbalance can cause several challenges during both the training and evaluation of a classification model. Due to the underlying imbalance of the class distribution, during the training process the model is not exposed to enough instances of the minority class compared to the majority class. Consequently, the model is more biased towards the majority class which can lead to severe misclassification errors for the minority class because the model is unable to capture and learn its underlying patterns.

To address the issue of the imbalanced datasets during the training process and improve the ability of the classification models to correctly identify cases of the minority class, various resampling techniques can be used to balance the class distribution. These methods can be broadly classified in upsampling/ oversampling and downsampling/ undersampling methods. To balance out the dataset, upsampling techniques increase the number of samples of the minority class until the ratio of the two classes reaches the desired level. In situations where

both classes are of equal significance, it can be advantageous to increase the number of the minority class until the distribution of both classes is completely balanced as displayed in FIGURE IV based on Kaur et al., 2019.

FIGURE IV OVERSAMPLING

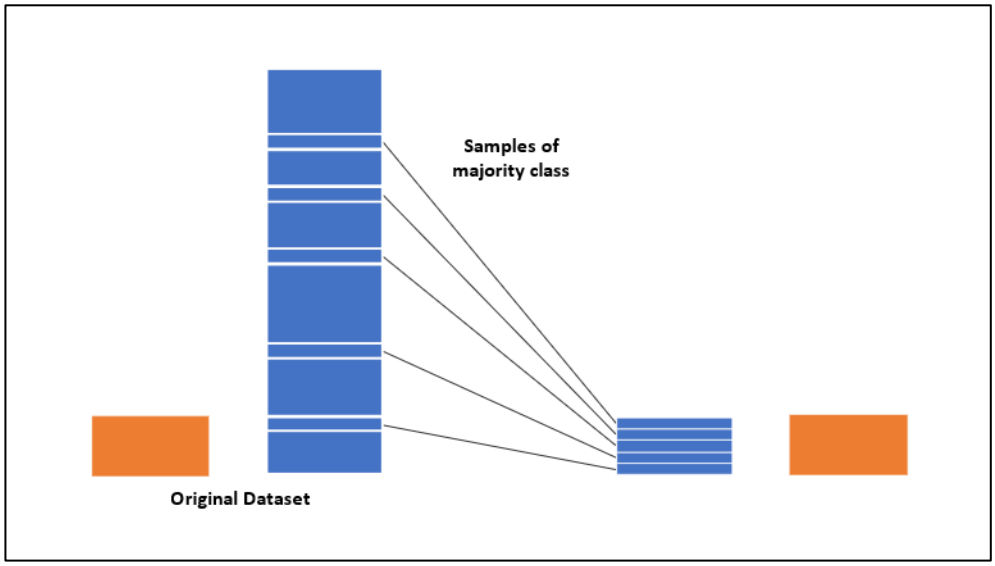


There are various upsampling techniques that can be used to achieve a balanced dataset including random oversampling in which random samples of the minority class are duplicated to increase its representation in the dataset. In cases with severe class imbalance where there only a small share of observations belongs the minority class, random oversampling can improve the performance of a classification model but it also increases the likelihood of overfitting which limits the models' generalization ability to new, unseen data (Batista et al., 2004). However, other techniques such as Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) have shown to be more effective in improving the performance of a classification models, including decision trees, support vector machines or neural networks, while also mitigating the likelihood of overfitting (Bach et al., 2017; Batista et al., 2004). SMOTE generates new, synthetic data points of the minority class with similar characteristics to the already existing observations based on the k-nearest neighbours algorithm (Chawla et al., 2002; Fix & Hodges, 1951).

Instead of increasing the sample size by creating new instances of the minority class, downsampling can be implemented to balance the class distribution by reducing the number of observations of the majority class. In comparison to upsampling techniques, downsampling methods reduce the sample size of the training dataset and thereby the computational complexity of the model as shown in the context of urban vegetation monitoring using random

forest (Feng et al., 2015) or spam filtering (Cormack et al., 2011). There are a various downsampling techniques available including random downsampling, cluster centroids or tomek links. To retain a balanced class distribution in the training datasets, this study employs random downsampling. In each year, instances from the majority class are randomly removed until a balanced class distribution of the majority and minority class is achieved.

FIGURE V DOWNSAMPLING



Especially in the presence of severe class imbalance the choice whether an upsampling or downsampling technique is implemented has a significant impact on the sample size of the training dataset. In such cases, upsamping techniques increase the computational complexity significantly to train the model. To highlight the enlargement of the sample, in the year 2009 the sample includes approximately 225,000 firms with a default rate of just under 8%. To equalize the number of instances of the default and non-default class, approximately a total of 189,000 new instances of the minority class needs to be created resulting in a sample size of 414,000. However, increasing the training datasets can be disadvantageous because large datasets pose significant challenges for classification models because of their computational complexity.

In comparison, implementing random downsampling in the training dataset implies reducing its sample size from approximately 225,000 to 36,000 observations. Consequently, implementing upsampling techniques results in a training dataset that is approximately 11.5 times larger compared to the training dataset obtained through downsampling methods. Therefore, utilizing downsampling can be particularly advantageous to reduce the computational complexity. After evaluating the performance of the various classification

models used in this study using both sampling techniques, the observed performance of the upsampling method SMOTE is comparable to the result obtained through random downsampling. Because there is no significant performance gain of using an upsampling method, random downsampling is implemented in the study due to its lower computational complexity.

IV. Explanatory Data Analysis

IV.A. Overview of Defaults

Over the observation period, the number of firms in the sample fluctuates between the minimum of 187,746 in 2015 and maximum of 225,118 non-financial cooperations in 2009. Subsequently, the yearly default rate is calculated based on the described definition of default. Table II displays the results, which indicate that the default rate peaked in 2013 at a rate of 10.71% and decreased over time to a minimum rate of 4.64% in 2020. From 2009 to 2020, the average observed default rate is 7.76%.

TABLE II YEARLY DEFAULT RATES

Year	N Firms	Default Rate
2009	225,118	7.86%
2010	222,387	7.74%
2011	216,683	8.41%
2012	208,002	10.57%
2013	196,595	10.71%
2014	188,429	9.42%
2015	187,746	8.15%
2016	190,971	7.63%
2017	194,658	6.81%
2018	205,651	5.99%
2019	210,816	5.22%
2020	222,702	4.64%

Note:
Own calculation based on data provided by Banco de Portugal. N Firms refers to the number of firms in the sample, while the Default Rate is the fraction of firms in default (according to the definition described in chapter IV.A.) relative to the total number of firms in a given year.

IV.B. Firm Classification and Distribution

To develop a better understanding of composition of default rates and the main influencing factors, this study follows the criteria outlined by the European Commission to distinguish between the four firm size classes: micro, small, medium, and large (European Commission, 2003). A firm is categorized as a micro firm if the number of employees does not

exceed 10 employees and a turnover or balance sheet total of €2 million. According to the definition of the European Commission, small firms do not exceed 50 employees and €10 million in turnover or balance sheet total. Finally, a medium-sized firm is defined as a firm that does employ less than 250 employees and a turnover or balance sheet total of €50 million and €43 million respectively. All remaining firms are classified as large firms.

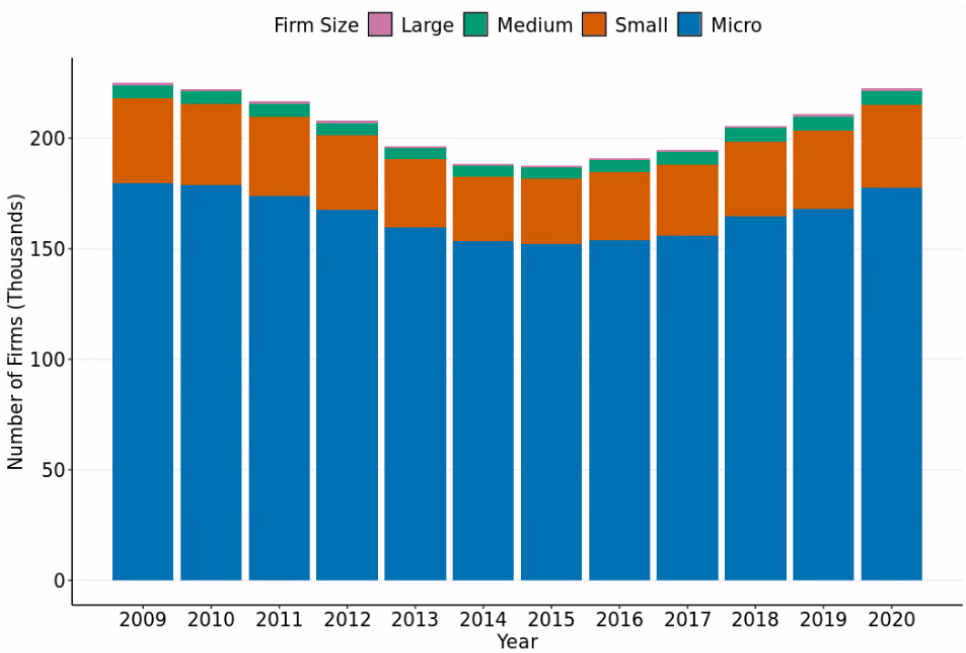
TABLE III FIRM SIZE CLASSIFICATION

Company Category	Staff Headcount	and	Turnover	or	Balance Sheet Total
Micro	< 10		≤ €2 m		≤ €2 m
Small	< 50		≤ €10 m		≤ €10 m
Medium	< 250		≤ €50 m		≤ €43 m
Large	≥250		> €50 m		>€43 m

Source: European Commission

The definition by the European Commission results in the following distribution of non-financial firms in Portugal displayed in FIGURE VI. Even though the number of firms in the sample fluctuates within the observation period the respective shares of the four classes remain approximately constant.

FIGURE VI NUMBER OF FIRMS BY SIZE

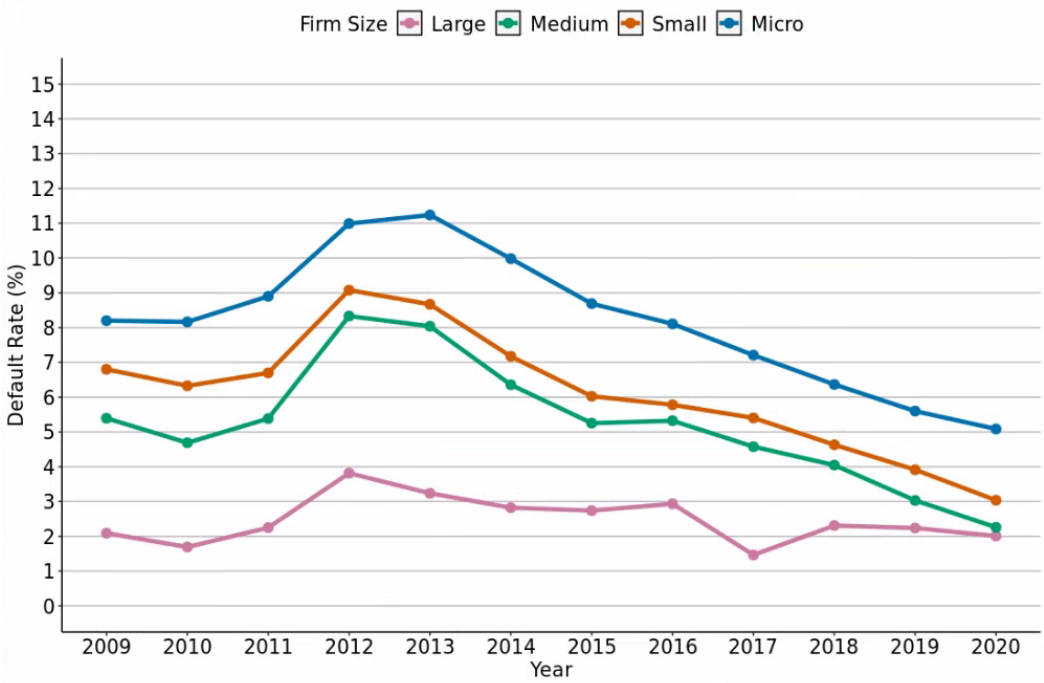


Note: Own calculation based on data provided by Banco de Portugal. The Number of Firms refers to the number of firms in the sample in a given year.

Most firms are classified as micro firms which account on average for 80.40% of the firms. More precisely, they represent between 79.7% and 81.4% which corresponds to 152,187 to 179,674 micro firms over the observations period. In comparison, the second largest group which is the group of small firms is significantly smaller and represents between 15.4% and 17.0% of Portuguese non-financial corporations. The two groups with the fewest observations are the groups of medium and large firms. On average these consistently account for 2.76% and 0.48% respectively.

Furthermore, the four classes differ not only in the number of firms but also in their default rates. As displayed in FIGURE VII, with increasing firm size the default rates decrease. Micro firms show highest default rate, peaking at 11.24% in 2013. The remaining groups small, medium, and large all reach their respective highest default rate in 2012. Afterwards, the default rates in the micro, small class decrease in every year and the same can be observed for the medium-sized class except for the year 2016 in which the default slightly increased compared to the previous year. In 2020 which is last year of the observation period, the lowest default rates can be observed in every class.

FIGURE VII DEFAULT RATE BY FIRM SIZE



Note: Own calculation based on data provided by Banco de Portugal. Based on the criteria defined by the European Commision, firm size classes are classified into four groups: micro, small, medium, and large. The Default Rate refers to the fraction of firms in default (according to the definition described in chapter IV.A.) relative to the total number of firms in a given year.

Due to the different default rates of each group, the proportion of defaults does not correspond to its representation in the sample. Micro firms exhibit the highest default rate

amongst all groups and as a result account for a higher share of defaults compared to their representation in the sample. In 2020 79.84% of the firms in the sample are micro firms but they account for 87.43% of the defaults. This can be observed over the whole sample period, the micro firms are on average responsible for 85.25% of all defaults each year.

V. Methods

V.A. Statistical Models

The study assesses the ability of various statistical and machine models to predict corporate default by identifying and learning complex patterns within the data.

Since Altman, 1968 first implemented the linear discriminant analysis (LDA) to classify corporate borrowers into default and non-defaults based on accounting ratios and other financial variables, it is also widely used method in the context of default forecasting and many other classification problems to separate two or more classes. The LDA is a feature reduction technique which aims to find a linear combination of features that best separates the classes of the dataset while retaining most of the relevant information.

The logistic regression (LOG) is another common model which is used in many binary classification problems. The LOG estimates the probability of a binary outcome based the explanatory variables which is then used to predict the class label. In the context of corporate default forecasting, the LOG model determines the state of a firm based on the estimated probability as either financially sound represented by 0 or by 1 if the firm is in default. As one of the first papers Ohlson, 1980 implements a logistic regression model to predict corporate bankruptcy. In this paper, the likelihood of bankruptcy is estimated based on financial ratios that reflect the current state of a firm.

The third statistical method used in this study is the penalized logistic regression (PLR). The PLR is a modification of the logistic regression which estimates the probability of binary outcome based on a set of explanatory variables. However, in the PLR a penalty term is added to avoid overfitting by reducing the impact of high variance explanatory variables. Furthermore, the literature has demonstrated that in certain prediction tasks, the PLR can outperform standard LOG (Zou & Hastie, 2005).

V.B. Machine Learning Models

The study implements the machine learning algorithms random forest (Breiman, 2001) and extreme gradient boosting (XGBoost) (Chen & Guestrin, 2016) to identify the complex

hidden patterns within the data and use them to build effective predictive models in the context of default forecasting.

Both models are tree-based algorithms which means that both utilize decision trees to build a predictive model. A decision tree is a supervised learning technique that can be applied in various fields to solve both regression and classification problems.

A decision tree has hierarchical tree-like structure, in which each of node of the tree represents a feature or variable of the dataset (Breiman et al., 1984). Starting from the first internal node, also referred to as root node, the data is split recursively into smaller subsets according to the decision rules which are derived from the features of the dataset. The decision rules are represented by the branches of the tree connecting two nodes. If a node cannot be split further, it becomes a leaf node which represents the outcome or class label (James et al., 2021). However, a limitation of using a single decision tree to solve a regression or classification problem is its tendency of overfitting on the training dataset which limits their ability to generalize to new data (Breiman et al., 1984).

Ensemble methods like random forest and XGBoost, address this limitation by aggregating many decision trees. Instead of relying on a single decision tree, these ensemble methods combine many individual models to obtain a single more powerful accurate (James et al., 2021). Moreover, both models differ in the way the individual decision trees are build and combined.

Bagging is another ensemble method that constructs decision trees using random samples (Breiman, 1996). However, when the trees are grown using the same features with different samples it, it is still likely that the bagged trees look similar to each other and as a result are highly correlated (James et al., 2021). In contrast, the random forest algorithm is also based on the bagging method. Moreover, each decision tree in a random forest is created using a random sample of observations as well as a random subset of features to decorrelate the trees (James et al., 2021). Since the decision trees are grown based on different samples and feature subsets, the prediction of the respective tree can differ. Finally, for a classification problem, the final prediction is determined by aggregating the predictions of individual decision trees and taking the majority vote (James et al., 2021).

The Gradient boosting model is based on the ensemble modelling technique boosting. Unlike in bagging, in boosting the individual trees are not grown parallel but successively. The idea behind this approach is to improve a weak model by combining several weak models in

series to build a strong model. This method builds sequence of models that attempts to correct the errors of the previous model. This is done by fitting a new model on the residuals of the previous model with the gradient descent algorithm to minimize the loss function (Friedman, 2001) (James et al., 2021). Gradient boosting is a specific implementation of boosting in which the models are constructed using decision trees whereas boosting can use any type of model. XGBoost is an extension of gradient boosting which includes several additional features that allow the model to handle complex datasets with a better accuracy and speed (Chen & Guestrin, 2016).

V.C. Hyperparameter Tuning

When building a machine learning model, it is important to distinguish between model parameters and hyperparameters. The model parameters are internal parameters that determine how input data is transformed into the desired output such as the coefficients in linear and logistic regression models. The model parameters are learned and updated during the training process by an optimization algorithm specific to the respective model including bagging in random forest (Breiman, 2001) or boosting in XGBoost (Chen & Guestrin, 2016).

On the other hand, hyperparameters are parameters that define the architecture of the learning algorithm and are set before the training process such as the number of trees in a random forest model (Breiman, 2001) or the learning rate in XGBoost (Chen & Guestrin, 2016). Hyperparameters are not learnt or changed during the training of the model but can have an significant impact on quality and speed of the training itself (Bergstra & Bengio, 2012). Unlike model parameters, hyperparameter control the learning process that determines the model parameters. Hyperparameters do not directly affect the predictions but can significantly affect the model's performance. The type and number of hyperparameters vary and are specific to the learning algorithm itself.

To maximize the performance of a model, it is essential to find the optimal combination of hyperparameters, which depends on the specific problem and dataset (Bergstra & Bengio, 2012). Moreover, the process of determining the optimal model architecture is more challenging because unlike the model parameters, hyperparameters cannot be directly learned or calculated from the data and therefore a trial-and-error approach is required.

The process of determining the optimal set of hyperparameter is called hyperparameter tuning or optimization (Feurer & Hutter, 2019). To assess and compare the performance of different combinations of hyperparameters, for each set it is necessary to train the model on the

training dataset, followed by testing on the testing dataset and finally evaluating it using a chosen validation metric, which is also referred to as hyperparameter metric (Bergstra et al., 2011). The choice of hyperparameter metric depends on the objective of the respective research question (Bergstra & Bengio, 2012). In this study, hyperparameters are chosen based on their ability to maximize the AuROC.

There are a variety of ways to tune hyperparameters (Bergstra & Bengio, 2012). First, they can be tuned manually but if the model is too complex with a large number of hyperparameters, handling this process manually becomes difficult and time consuming (Feurer & Hutter, 2019). Therefore, it is advantageous to use an automated hyperparameter tuning method such as grid search, random search, or Bayesian optimization to find the optimal hyperparameter values (Bergstra & Bengio, 2012; Feurer & Hutter, 2019).

Grid search is a common and simple automated algorithm for hyperparameter tuning (Feurer & Hutter, 2019). In this method a grid of potential values for each hyperparameter is defined. After the grid is specified, the algorithm builds a model for every possible combination of the specified hyperparameters and calculates its respective hyperparameter metric using cross-validation. Finally, the best performing set of hyperparameters in terms of the hyperparameter metric is selected as the optimal model architecture. While grid search is an effective method to identify the optimal hyperparameter set, iterating through every combination of hyperparameters in the grid is time-intensive and requires a high computation capacity (Feurer & Hutter, 2019).

A similar approach that addresses this limitation is random search (Bergstra & Bengio, 2012; Feurer & Hutter, 2019). Instead of evaluating all possible combinations of hyperparameters, random search only evaluates a randomly selected subset of hyperparameter combinations to find the optimal model architecture (Bergstra & Bengio, 2012). Although this method might not find the optimal combination of hyperparameters, it still achieves comparable results but requires less time compared to the grid search approach (Bergstra et al., 2011).

Nevertheless, both grid search and random search do not account for the results of previous iterations and each set of hyperparameters is evaluated individually. As a result, a large number of unsuitable hyperparameter sets is assessed resulting in an inefficient hyperparameter tuning process (Bergstra & Bengio, 2012).

To address this inefficiency and avoid the evaluation of unsuitable hyperparameter combinations, Bayesian optimization is a sequential model which choses the hyperparameter

that are evaluated next based on the results of previous evaluations (Bergstra et al., 2011). The main idea of this method is to focus on the more promising candidates to make fewer calls to the objective function. In this classification problem, the AuROC is used as the objective function to evaluate the performance of different hyperparameter combinations on the validation set, with the goal of maximizing this objective function. Compared to grid search and random search the Bayesian optimisation has proven its ability to obtain better results with fewer evaluations (Bergstra et al., 2011; Snoek et al., 2012).

However, to find the best possible hyperparameter combination in the fewest number of evaluations using the objective function, Bayesian optimization constructs a surrogate function which is a probabilistic model of the objective function based on the results of past evaluations (Snoek et al., 2012). Before the actual Bayesian optimisation process starts, various random hyperparameter combinations are evaluated to initialize the surrogate function.

The surrogate function can be considered as an application of the Bayes' Theorem (as shown in Equation 1) which relates hyperparameters to the probability of obtaining a certain score on the objective function. Equation 2 illustrates P(metric|hyper) which gives the probability of the given metric, like the AuROC, to be maximized given a combination of hyperparameters.

(1) $$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

(2) $$P(\text{metric}|\text{hyper}) = \frac{P(\text{hyper}|\text{metric})P(\text{metric})}{P(\text{hyper})}$$

This research utilizes the Gaussian Processes as the surrogate function to guide the hyperparameter search toward more promising candidates that are likely to yield an improvement to the objective function (Rasmussen & Williams, 2006).

After the surrogate model of the objective function is constructed, the model determines the hyperparameters that will be evaluated next by the objective function based on the criterion defined in the selection function. A common criterion of the selection function, which is utilized

in this study, is the expected improvement function (Jones et al., 1998). Consequently, the set of hyperparameters that maximizes the expected improvement criterion is evaluated next using the actual objective function.

After each evaluation, the surrogate function is updated incorporating the new results using the Bayes' rule. Consequently, the selection of the hyperparameters to be evaluated in the next iteration is more informed and the surrogate function becomes more accurate as it approximates the actual objective function with each iteration (Snoek et al., 2012).

Finally, the process of evaluating new hyperparameter combinations and updating the surrogate function is repeated until the optimal set of hyperparameters is found and no further improvement is found, or the predefined maximum number of iterations is reached.

This study implements Bayesian Optimization to find the optimal set of hyperparameters for the two machine learning models random forest and XGBoost model. The hyperparameters of the respective model are determined using all available observations of the year 2013. The resulting hyperparameter values are used for all remaining specifications discussed in this paper. To initialize the surrogate function, five random hyperparameter combinations are used that serve as initial estimates of the hyperparameters values which are updated during the optimization process. In addition, the maximum number of iterations is limited to 30. Alternatively, the optimization process is terminated before the 30^{th} iteration when no further improvement is achieved in 10 consecutive iterations.

To optimize the performance of the machine learning model random forest, there are three hyperparameter which can be optimized:

- mtry = number of predictors that are randomly sampled at each split
- trees = number of trees contained in the ensemble
- min_n = the minimum number of data points in a node that are required for the node to be split further

The results of the hyperparameter tuning of the random forest model using Bayesian optimization are presented in TABLE IV, showing the 10 best performing hyperparameter combinations in terms of the AuROC. The optimal set of hyperparameters is found during the 15^{th} iteration. As shown in the first row of TABLE IV, the optimal random forest model randomly based on the data used selects only one predictor at each split. In addition, a total of 1905 trees are included in model with a minimum of three data points required to split an internal node, otherwise there is no further split, and the internal node becomes a terminal node.

By utilizing the optimal combination of hyperparameters, the random forest model is able to achieve an average AuROC score of 86.31%. The score is presented as an average because it was obtained by utilizing the k-fold cross-validation method using 5 folds. The use of k-fold cross-validation allows for a more reliable and accurate evaluation of the performance of the model compared to other validation techniques. However, k-fold cross-validation will be discussed in more detail in chapter V.E. When considering all mean AuROC score of the 10 best hyperparameter combinations, it stands out that there are only minor variations between the observed results. Furthermore, it is noteworthy that already in the 4th iteration, as shown in line five of TABLE IV, a combination of hyperparameters is selected that achieves an AuROC score that is quite comparable to the best performing combination.

TABLE IV HYPERPARAMETER TUNING RANDOM FOREST

mtry	trees	min_n	mean	n	std_err	.iter
1	1905	3	0.8631416	5	0.0019078	15
1	998	2	0.8630706	5	0.0019256	19
1	1514	2	0.8630317	5	0.0019700	12
1	1061	2	0.8630277	5	0.0019443	20
1	1406	2	0.8630157	5	0.0019741	4
1	1954	2	0.8630107	5	0.0018933	8
1	1310	2	0.8630098	5	0.0018755	23
1	1757	2	0.8629885	5	0.0019472	21
1	1593	2	0.8629760	5	0.0019439	9
1	1829	3	0.8629724	5	0.0019418	13

Like random forest, XGBoost is a tree-based model and includes the hyperparamters mtry, min_n. In addition, three other hyperparameters are tuned:

- tree_depth: the maximum number of levels or nodes a tree can have
- learn_rate: determines the step size at each iteration when updating the weights of the decision trees
- loss_reduction: minimum reduction of the loss function required to split further

TABLE V presents the results of the 10 best performing hyperparameter combinations in terms of the AuROC obtained during the hyperparameter tuning of the machine learning model XGBoost using the Bayesian Optimization method. Although XGBoost is also a tree-

based machine learning model, the results table of the hyperparameter tuning do not include the number of trees. Since the XGBoost model already includes five hyperparameter with the potential addition of further parameters, the decision was made to limit the search space of hyperparameters during the tuning process by setting the numbers of trees to 1000. This number is still sufficient to learn enough from the data, while allowing for a faster and more efficient optimization process of the remaining hyperparameters. The first row of TABLE V shows the optimal hyperparameters values of the XGBoost model. The optimal XGBoost model based on this dataset considers four predictors at each split and requires a minimum of two data points to split an internal node with a maximum of 15 levels. Furthermore, the tuning process reveals an optimal learning rate of 0.0000013 which means that the weights of the decision trees are updated very slowly. Finally, the optimal loss reduction of 0.0001194 indicates the minimum reduction in the loss function that is required to split further. In addition, compared to the tuning process of the random forest model the optimal combination hyperparameters of the XGBoost model results achieves an average AuROC score of 85.69% during the second iteration of the Bayesian Optimization.

TABLE V HYPERPARAMETER TUNING OF XGBOOST

mtry	min_n	tree_depth	learn_rate	loss_reduction	mean	n	std_err	.iter
4	2	15	0.0000013	0.0001194	0.8569501	5	0.0023566	2
4	2	14	0.0000000	2.6216259	0.8553197	5	0.0024609	5
10	31	15	0.0000344	0.0000000	0.8551291	5	0.0023955	11
10	40	11	0.0000006	0.0000097	0.8523425	5	0.0025218	12
16	21	13	0.0009944	0.0000538	0.8520613	5	0.0021964	0
5	3	7	0.0353910	1.7822112	0.8496403	5	0.0015255	0
13	2	11	0.0000015	1.7023705	0.8491727	5	0.0029192	4
1	33	11	0.0000021	0.0000000	0.8480798	5	0.0017580	10
8	4	11	0.0000000	18.7156914	0.8458367	5	0.0023689	9
17	24	12	0.0980325	0.0000000	0.8457162	5	0.0014381	1

V.D. Model Evaluation

When evaluating the performance of classification model using a balanced testing dataset where both classes are equally important accuracy, which refers to the share of correct predictions, is a simple and sufficient metric (Kohavi & Provost, 1998). However, in

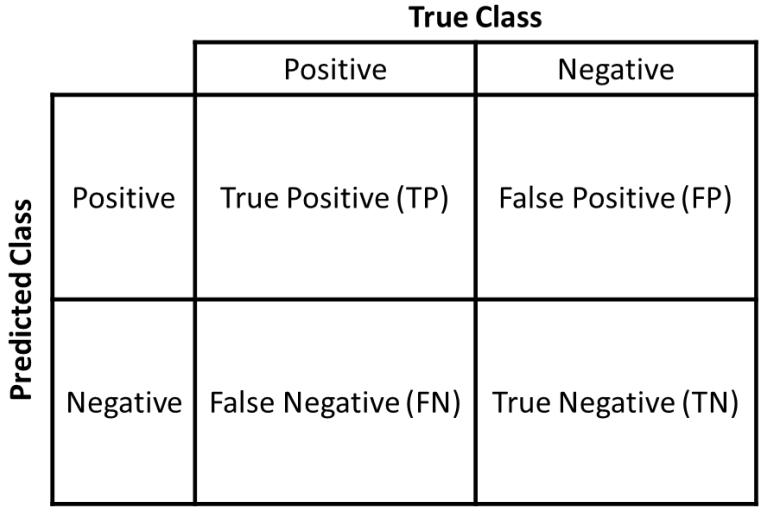
imbalanced datasets often a higher accuracy score is achieved due to the high number of correct predictions of the majority class while the minority group is neglected. Therefore, accuracy can be a misleading metric that does not reflect a models' ability distinguish between classes in the presence of class imbalance in the testing dataset. This phenomenon is also referred to as the accuracy paradox (Fawcett, 2001).

Therefore, when evaluating the overall performance of a binary classification model the use of alternative metrics that consider both the majority, and the minority class is advantageous. One commonly used metric is the AuROC (Fawcett, 2006; Powers, 2020; Tharwat, 2020). Furthermore, there are several other metrics available to evaluate the performance of classification models such as the area under precision recall, the F-score or the G-mean. However, it is important to note that these metrics capture different aspects of the model and therefore the choice of the respective metric depends on the context and objective. For imbalanced datasets where both classes are equally important AuROC is the most appropriate metric because it measures the ability of a model to correctly distinguish positive and negative classes regardless of the class distribution and is therefore implemented in this study.

The AuROC is derived from the confusion matrix which is a graphical representation of the actual and predicted classifications done by the classification model which is displayed in FIGURE VIII (Kohavi & Provost, 1998; Stehman, 1997). It helps to visualize performance of the model and provides a better understanding of its strengths and weaknesses, particularly where the model makes false predictions.

For a binary classification model, the confusion matrix consists of two rows and two columns. The rows of the confusion matrix represent the predicted class by the algorithm while the columns represent the true class.

FIGURE VIII CONFUSION MATRIX



The ROC curve is a plot of true positive rate (TPR) on the y-axis and the false positive rate (FPR) on the x-axis for all possible discrimination thresholds (Fawcett, 2006; James et al., 2021). A discrimination threshold usually refers to a probability and is used to determine whether specific case should be assigned to the positive or negative class. For a threshold of 0.5, all cases with a predicted probability greater than 0.5 are classified to the positive class. Any other case with a predicted probability lower than the respective discrimination threshold is classified as negative.

The TPR often also referred to as sensitivity measures the percentage of positives that are correctly identified. For the case of default forecasting, the sensitivity is the percentage of defaulting firm that are correctly identified as such (Kohavi & Provost, 1998).

(3) $$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$$

The FPR is defined is the percentage of negative cases that are incorrectly classified (Kohavi & Provost, 1998).

(4) $$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$$

The specificity is the proportion of negative class that is correctly classified. It gives the percentage of non-defaulting firms that are correctly identified (Kohavi & Provost, 1998).

(5) $$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{FP+TN}}$$

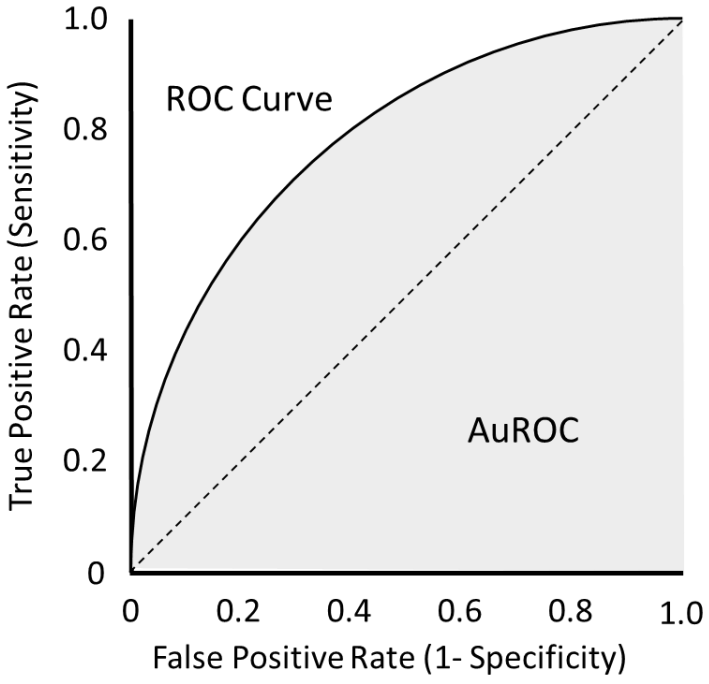
Consequently, the FPR can also be expressed as:

(6) $$\text{FPR} = 1 - \text{Specificity}$$

The AuROC is calculated as the area under ROC curve and is used to evaluate the performance of a binary classification model. The possible values of the AuROC range from 0 to 1 with higher AuROC values corresponding to a better performing model. A perfect classifier has an AuROC of 1 which will always have a TPR of 1, regardless of the FPR. This implies that the classifier does not make any mistakes in distinguishing the positive and negative classes.

For a binary classification model, an AuROC above 0.7 indicates a well-performing model. An AuROC of 0.5 indicates that the model has no predictive power and is as good as random guessing. Any model with an AuROC score of less than 0.5, which falls under the diagonal, is worse than random guessing. The AuROC is a good metric to evaluate a classifiers performance especially when one cares equally about the positive and negative class.

FIGURE IX ROC CURVE



V.E. Model Validation

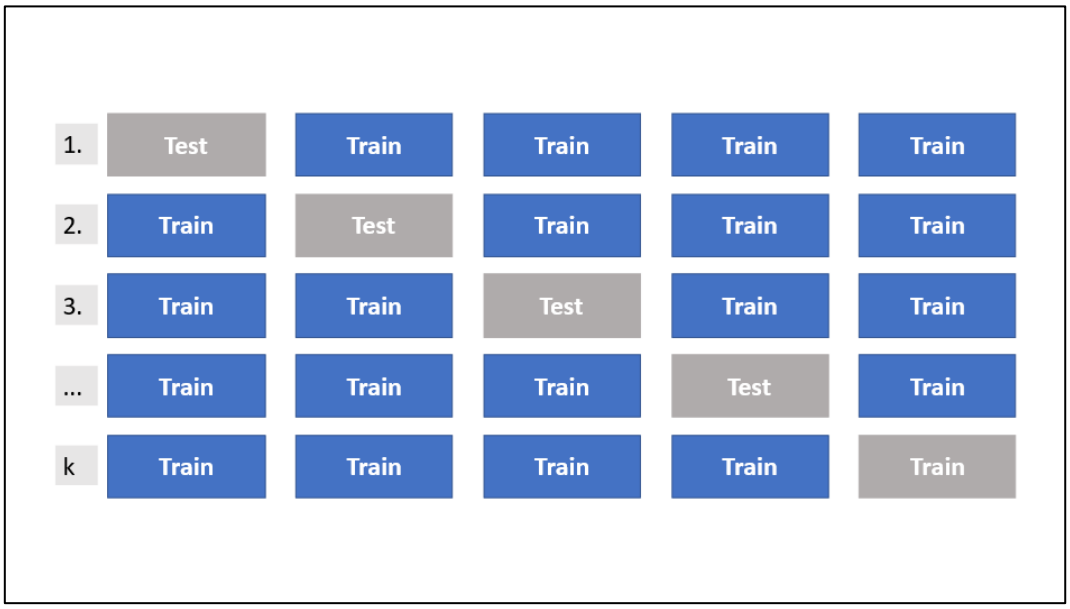
The following chapter will focus on the model validation which is used to evaluate the performance of a model and assess whether the trained model can be generalized. A models generalizability refers to its ability to accurately predicted outcomes for new, unseen data, based

on patterns learned from the training data. A common method is the validation set approach in which the data is split into a training set and a testing set which is not used for the training of the model (James et al., 2021). Most of the data is assigned to the training set and often a split 80/20 is used (James et al., 2021). After the model was fitted using the training set, the performance of fitted model is evaluated on the testing set.

The validation set approach is widely used, however it has some limitations. The performance of the model heavily depends on the one-time random split which can lead to a high variance of the results especially in small datasets. Also, a selection bias can occur when the testing set is randomly selected in a way that is no longer representative of the population and can result in an overestimation or underestimation of the models' true performance.

Another possible issue is overfitting, which occurs when a model is trained too well on specific characteristics in training set, causing it to lose its ability to generalize to new, unseen data. One way to mitigate these limitations is the use of k-fold cross validation. In this approach the data is randomly divided into k groups, also referred to as folds, of equal size. The first fold is treated as the test set and the remaining k-1 folds are used to train the model. Then, the model's performance is evaluated on the validation set (Berrar, 2019; James et al., 2021).

FIGURE X K-FOLD-CROSS-VALIDATION



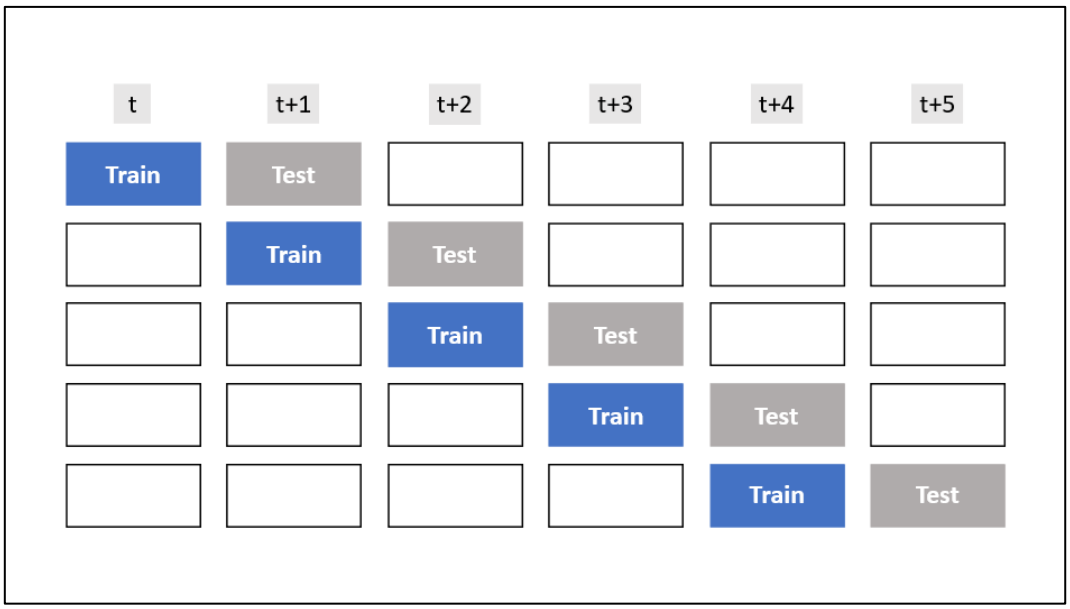
This process is repeated k times with a different fold selected as the validation set on each iteration, so each fold is used exactly once as the validation set. The final k-fold cross-validation estimate for the metrics like AuROC are obtained by averaging the results from all k iterations (Berrar, 2019; James et al., 2021).

$$\text{CV AUCROC}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \text{AUCROC}_i \quad (5)$$

Although k-fold cross-validation is a widely used and appropriate method, it has some limitations when applied to time-series modelling. The main objective in time series modelling is analyse past data and detect and capture patterns to make accurate predictions of future outcomes. One fundamental principal in the assessment of any time series modelling, is that the data used to evaluate a model is temporally distinct from the data used for training (Shumway et al., 2000). Even when applying k-fold validation repeatedly and splitting the data of each period into k folds, this does not capture the ability of the model predict future outcomes but rather how well the model can be trained to capture specific patterns at the respective particular point in time. In addition, k-fold cross validation does not represent a real life-scenario in which a model was trained based on past data and predictions are made using the current, available data at the point in time the prediction is made.

Therefore, to determine the performance and robustness of classification models in default forecasting and considering the temporal dynamics of the data, this study implements walk-forward validation which can considered as an extension of the k-fold cross validation. This approach is commonly used in financial modelling and algorithmic trading with the objective to determine the robustness of trading strategies and was originally introduced by Robert Pardo (Pardo, 1992). Similarly, to the k-fold cross-validation approach, the walk-forward validation involves splitting the dataset into a training and testing dataset. However, in this approach the model is fitted on the training dataset and tested on the testing dataset of the subsequent period. Furthermore, this process is repeated iteratively by training and testing the model on a rolling basis, while the training and testing subset move forward one period after each iteration as illustrated in FIGURE XI. The duration of each period of the training and the testing dataset varies and depends on the context of the specific question. For the purposes of this study, one period corresponds to one year for both the training and the testing dataset.

FIGURE XI WALK-FORWARD VALIDATION



In this study, a classification model is trained in a given year on all available observations of the specific year to predict corporate default in the next year using the selected predictors. Afterwards, the fitted model is applied on all available observations of the subsequent year to evaluate the performance of the fitted model. The detailed results correspond to the year in which the fitted model was tested and are presented in the appendix. The walk-forward validated results are calculated using the following formula:

$$\text{WFV AUCROC}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \text{AUCROC}_i \tag{6}$$

This formula shows the calculation of the AuROC using the walk-forward validation method where n represents the number of testing datasets in the data and AUCROC_i the AuROC score when tested on the i^{th} testing dataset.

VI. Results

In this section, the described statistical and machine learning models are assessed in terms of their ability to predict corporate default in the next year. To compare the forecasting ability of the various models, this study employs several specifications to ensure the validity of the results. These specifications include the type of forecast, the sample size of the training dataset, the dimensionality of the model and a clustering approach. Furthermore, based on the findings, the implications of corporate default of non-financial cooperations on the labour market are estimated.

VI.A. Type of Forecast

When discussing the type of forecast this refers to the data that was used to train and test the model. If forecasts are made for observations that were also included in the training sample, it is an in-sample forecast. Consequently, if the observations in the test sample were not included in the training process, it is an out-of-sample prediction.

Furthermore, the timing of the training and testing data is important to consider when making forecasts. If the tested data is in the same period as the data that was used to fit the model, it can be described as an in-time forecast. On the other hand, out-of-time predictions utilize data from different periods.

In the first step, the predictive power of the chosen models is evaluated through out-of-sample predictions using the validation set approach. In each year 80% of the data is used for training of the models and the remaining 20 percent of the same year are reserved for testing. This is the only specification of this study that implements the validation set approach while all following specifications utilize the walk-forward validation.

The validation set approach is repeated for every year in the observation period using the full sample and the top 10 variables with the highest AuROC score in the univariate logistic regression as discussed in chapter III.D.iii.. This serves an initial evaluation of the models' performance when they are tested on data of the same period as they are trained. TABLE VI illustrates the AuROC of the respective model in every year from 2008 until 2019 and shows that the machine learning models outperform the statistical models over the entire observation period. The linear discriminant model achieves an average AuROC of 66.36%. In comparison, the predictive performance of the logistic regression and the penalized logistic regression model is quite similar and they achieve an average AuROC score of 68.60% and 68.53% respectively. However, the average AuROC of the penalized logistic regression model which is the best performing statistical model is 9.70 percentage points lower compared to the random forest model and 9.54 percentage points compared to the XGBoost.

TABLE VI DISCRIMINATORY POWER OF OUT-OF-SAMPLE PREDICTIONS

	Statistical Models			ML Models	
Year	Linear Discriminant Analysis	Logistic Regression	Penalized Logistic Regression	Random Forest	Extreme Gradient Boosting
2008	61.83%	63.46%	63.43%	74.15%	74.07%
2009	65.34%	68.48%	68.44%	75.19%	75.10%
2010	66.48%	69.28%	69.25%	76.32%	76.35%
2011	65.21%	67.46%	67.42%	76.77%	76.67%
2012	66.33%	67.40%	67.31%	78.46%	78.25%
2013	67.00%	69.30%	69.24%	79.28%	79.17%
2014	66.89%	68.99%	68.85%	79.26%	79.15%
2015	67.45%	69.93%	69.85%	79.45%	79.08%
2016	68.80%	70.97%	70.94%	79.56%	79.46%
2017	68.64%	69.69%	69.65%	80.44%	80.23%
2018	66.70%	69.19%	69.09%	79.58%	79.40%
2019	65.67%	68.99%	68.94%	80.33%	80.01%

Note:
Own calculation based on data provided by Banco de Portugal. The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the respective model.

VI.B. Out-of-Time Predictions

After the initial evaluation of the models using data from the same year, this specification evaluates the performance of the models based on out-of-time predictions. In contrast to the previous specification, out-of-time predictions represent a real-life scenario because the predictions of a specific year are made based on the available information of the previous year.

For this specification the full sample with the top 10 variables is used. The AuROC for the respective models are displayed in TABLE VIII, but this time for the period of 2009 until 2019 because the year 2008 was used to fit the models for the year 2009.

Similarly, to the previous specification, TABLE VII illustrates that both the logistic and penalized logistic achieve comparable walk-forward validated AuROC scores of 69.05% and 69.03% respectively. They outperform the linear model by approximately by 2.40 percentage points, but their predictive power is still significantly lower compared to the machine learning models. The random forest model is the best performing classifier and which an average AuROC score of 79.03%, followed by XGBoost with an average score of 78.66%.

The results of the out-of-time and in-time predictions are quite comparable which confirms the validity of the models and indicates that they were not overfitted in the previous specification. Consequently, they can be generalized on new out-of-time data and that they are not only valid for the same period.

TABLE VII DISCRIMINATORY POWER OF OUT-OF-TIME PREDICTIONS

	Statistical Models			ML Models	
	Linear Discriminant Analysis	Logistic Regression	Penalized Logistic Regression	Random Forest	Extreme Gradient Boosting
2009	64.77%	69.12%	69.09%	75.93%	75.61%
2010	66.06%	69.30%	69.51%	76.20%	76.04%
2011	66.32%	68.91%	68.90%	76.64%	76.40%
2012	65.30%	67.52%	67.48%	78.09%	77.81%
2013	66.45%	68.38%	68.32%	79.20%	78.89%
2014	67.12%	68.82%	68.73%	80.46%	80.06%
2015	68.16%	70.18%	70.13%	80.41%	79.94%
2016	68.26%	70.12%	70.09%	80.73%	80.31%
2017	68.53%	70.66%	70.62%	80.74%	80.28%
2018	66.56%	68.83%	68.79%	79.92%	79.41%
2019	65.65%	67.74%	67.66%	80.97%	80.55%

Note:
Own calculation based on data provided by Banco de Portugal. The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the respective model.

VI.C. Sample Size of the Training Dataset

Another important aspect to consider is the sample size used to train the models. The following section discusses the forecasting ability of the various models when only a limited share of the available observations is used in the training process. Four different variations were implemented which use 100%, 50%, 10% and 5% of all Portuguese non-financial cooperations in each year. The used observations in the smaller datasets are randomly selected and then the models were tested using the full samples. This examination can be particularly advantageous as simulates a scenario in which only a limited amount of data is available.

The impact of the sample size on the model performance is evaluated by comparing the model performance using the top 10 variables with four different sample sizes. In TABLE VIII the walk-forward validated AuROCs of the four variations for the respective models are illustrated. In addition, the detailed tables of all the AuROCs including the predictions for every year are provided in the appendix (Table A.4 - A.7). The results in TABLE VIII and the

corresponding FIGURE XII suggest that discriminatory power of the statistical models remains constant as the sample size is reduced.

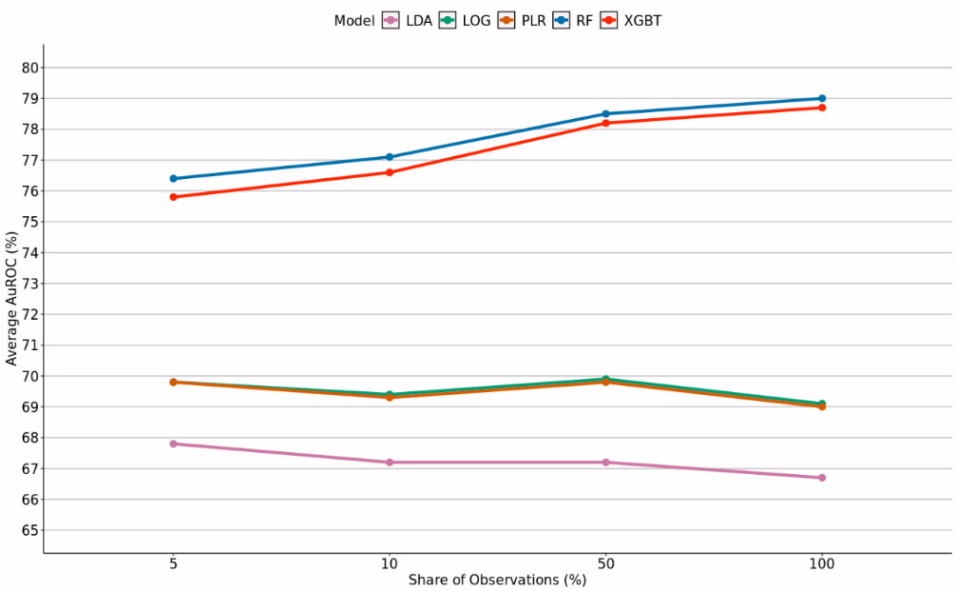
In contrast, the performance of machine learning models deteriorates when the sample size gets smaller. The AuROC of the machine learning model random forest drops from 79.00% with the full sample to 76.40% when only 5% of the available observations were used to train the models. The same trend can be observed for the XGBoost in which the AuROC decreases from 78.70% to 75.80%. The deterioration in performance of the machine learning can be attributed to several factors. Firstly, with decreasing sample size the chance of a sampling bias increases. This means that the model is trained on data that is not representative of the sample. Consequently, the performance deteriorates when tested with the actual data. In addition, with decreasing sample size the probability of overfitting increases, and the models are more likely capture noise in the data rather than the actual underlying patterns. Finally, to learn more complex patterns in the data a smaller sample size might not be sufficient to train the model effectively.

TABLE VIII DISCRIMINATORY POWER SAMPLE SIZE

	Statistical Models			ML Models	
	Linear Discriminant Analysis	Logistic Regression	Penalized Logistic Regression	Random Forest	Extreme Gradient Boosting
5%	67.80%	69.80%	69.80%	76.40%	75.80%
10%	67.20%	69.40%	69.30%	77.10%	76.60%
50%	67.20%	69.90%	69.80%	78.50%	78.20%
100%	66.70%	69.10%	69.00%	79.00%	78.70%

Note:
Own calculation based on data provided by Banco de Portugal. The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the respective model.

FIGURE XII DISCRIMINATORY POWER SAMPLE SIZE



VI.D. Dimensionality

Furthermore, the forecasting ability of the models is affected by the dimensionality of the model. The dimensionality refers to the number of variable or features used in a model to make predictions. Although it may seem beneficial to add more variables to improve the performance of the model, increasing the number of variables does not lead to a continuous improvement in the predictive power of a model (Hughes, 1968). After the optimal number of variables is exceeded, which depends on the specific question and model, the predictive power starts to decline (Hughes, 1968). This is commonly referred to as the curse of dimensionality or the Hughes phenomenon. With the addition of further variables, the model becomes more complex, which increases the chances of overfitting and therefore deteriorates the performance of the model and limits its ability to be generalized.

To test the robustness of the forecasting models with respect to the dimensionality, this study examines three variations using the top 5, top 10 and top 15 variables respectively with the highest AuROC score in the univariate logistic regression as discussed in chapter III.D.iii. using all available observations for training and testing for each year.

In TABLE IX and FIGURE XIII, the walk-forward validated AuROCs for the different variations are presented with the full tables being provided in the appendix (A.8 -A.10). The discriminatory power of the statistical models does not benefit from higher dimensionality. The performance of the logistic regression, and the penalized logistic regression remain constant while the performance of the linear discriminant analysis slightly declines. Compared to that,

the discriminatory power of the machine learning models improves significantly as the dimensionality increases. The AuROC of the random forest and XGBoost models increase from 74.90% and 74.10% when the top 5 variables used to 82.50% and 82.10% respectively when the top 15 variables are utilized.

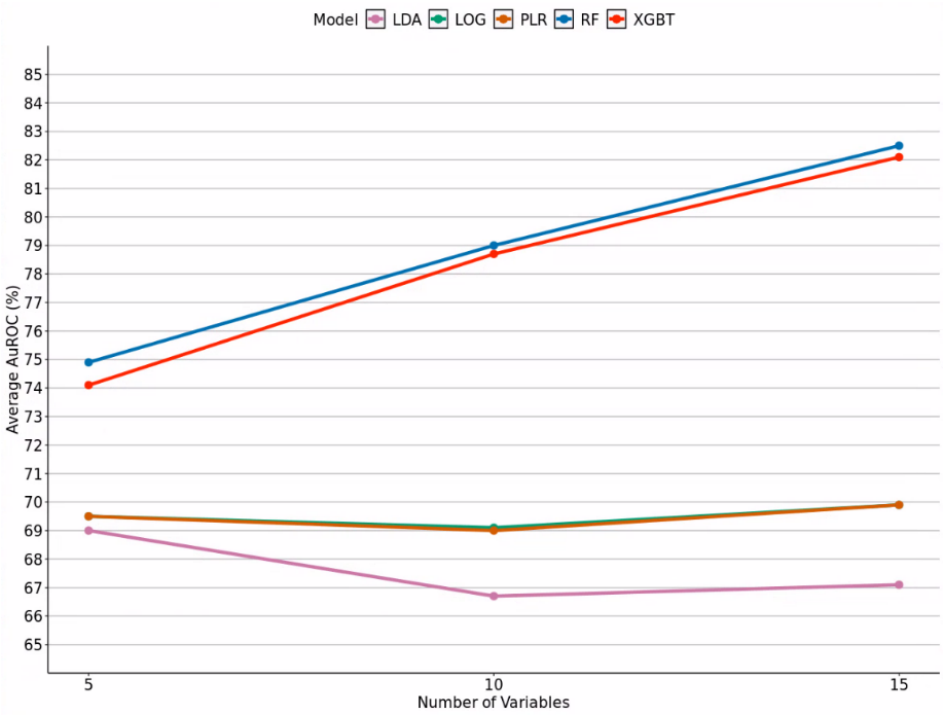
The performance of the machine learning models keeps improving when the number of features is increased until 15 features are used. Consequently, it is likely the optimal number of features has not been reached yet and potentially more features could be added to further enhance the predictive power of the two models.

TABLE IX DISCRIMINATORY POWER DIMENSIONALITY

	Statistical Models			ML Models	
	Linear Discriminant Analysis	Logistic Regression	Penalized Logistic Regression	Random Forest	Extreme Gradient Boosting
5	69.00%	69.50%	69.50%	74.90%	74.10%
10	66.70%	69.10%	69.00%	79.00%	78.70%
15	67.10%	69.90%	69.90%	82.50%	82.10%

Note:
Own calculation based on data provided by Banco de Portugal. The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the respective model.

FIGURE XIII DISCRIMINATORY DIMENSIONALITY



VI.E. Firm Size Clustering Approach

Finally, a clustering approach is implemented in which the data are divided into four groups based on respective their firm size following the definition of the European Commission. Accordingly, the firms are classified as micro, small, medium, or large firms. Clustering can be particularly useful to identify cluster specific patterns within the respective cluster which otherwise would not have been discovered and thereby improve the performance of a classification model.

When firms are divided into clusters based on their size and the models are trained using the top 10 variables, the analysis reveals than the machine learning models consistently outperform the statistical models across clusters. The ability to distinguish default and non-default is the highest among all models in the medium cluster, especially the random forest model achieves an AuROC of 85.90%.

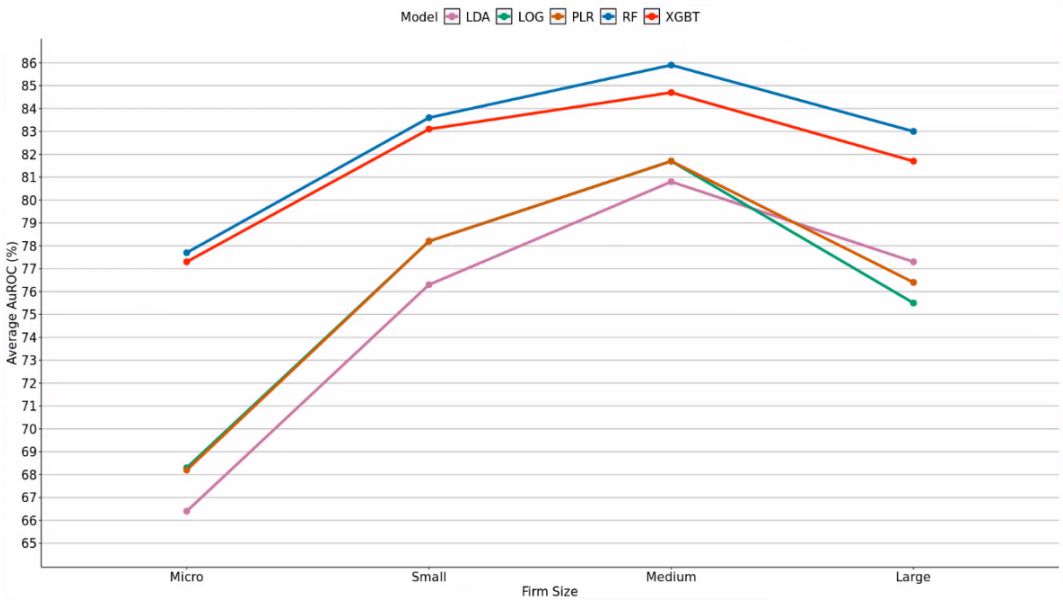
It stands out that when comparing the results from TABLE X with previous results in TABLE IX using the top 10 variables, the predictive power of all statistical and machine learning models improves. Furthermore, the difference in performance between the two model types diminishes significantly in all but the micro cluster. In the micro cluster, the AuROC deteriorates slightly but is still comparable to the performance presented in TABLE IX using the top 10 variables. Nevertheless, the performance gap observed in pervious specifications remains. This result can be attributed to the fact that most of the firms in the sample belong to the micro cluster. Consequently, they have a significant influence on the overall results when no cluster approach is implemented.

TABLE X DISCRIMINATORY POWER FIRM SIZE CLUSTERS

	Statistical Models			ML Models	
	Linear Discriminant Analysis	Logistic Regression	Penalized Logistic Regression	Random Forest	Extreme Gradient Boosting
Micro	66.40%	68.30%	68.20%	77.70%	77.30%
Small	76.30%	78.20%	78.20%	83.60%	83.10%
Medium	80.80%	81.70%	81.70%	85.90%	84.70%
Large	77.30%	75.50%	76.40%	83.00%	81.70%

Note:
Own calculation based on data provided by Banco de Portugal. The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the respective model.

FIGURE XIV DISCRIMINATORY POWER FIRM SIZE CLUSTERS



VI.F. Labour Market Implications of Corporate Default

Finally, the following chapter combines the findings of this study to build a model that examines the labour market implications of corporate default in Portugal. It is important to note that in the context of this study default does not imply bankruptcy because most firms that are classified as to be in default appear again in the sample in following years. However, being classified as default indicates a certain financial instability for a firm, which can put the jobs associated with the firm at risk.

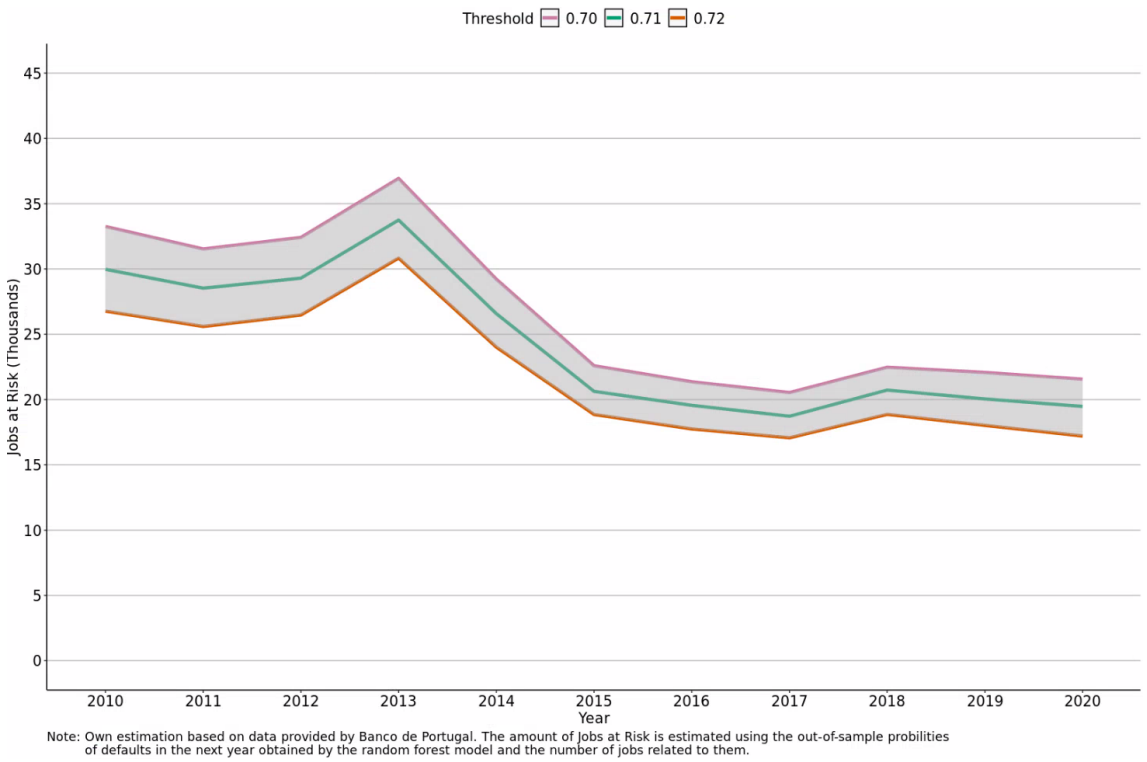
The majority of all Portuguese non-financial cooperations are micro firms which are also responsible for most of the corporate defaults. Approximately 20% of all jobs in the sample throughout the observation period are provided by micro firms. The following chapter examines corporate default among micro firms and combining the findings of this research.

As illustrated in TABLE X of the previous chapter, both traditional and machine learning models exhibit the most difficulties in correctly distinguishing defaults and non-defaults among micro firms. However, out of the five models, random forest exhibits the highest predictive power to correctly distinguish default and non-default firm in the micro cluster and is therefore selected for the subsequent evaluation. Furthermore, the results of TABLE IX indicate that by increasing the number of predictors from 10 to 15, the performance of the random forest classifier can be improved when all available observations are used.

Based on the sample using the top 15 predictors, the out-of-sample probabilities of default in the next year are obtained by the random forest model. Whether a firm is then

classified as in default depends on the specific discrimination threshold chosen. For the further evaluation, the discrimination thresholds 0.70, 0.71, 0.72 are considered. These specific discrimination thresholds are derived from the actual observed default rate among micro firms. Over the observation period, on average 13,525 micro firms default every year. When applying the threshold of 0.71, the classifier predicts an average of 13,296 micro firms to default yearly. Compared to that, a discrimination threshold of 0.70 overestimates the average number of defaults by predicting a total of 14,441 yearly defaults of micro firms. Increasing the discrimination threshold 0.72 leads to an underestimation of the average number of defaults predicting 12,173 yearly defaults. For each of the three thresholds, the defaulting firms, and the corresponding jobs at risk due to the default are calculated. The total jobs risk over the observation period for the respective discrimination threshold are illustrated in FIGURE XV. The grey area between the lines of the 0.70 and 0.72 discrimination thresholds represents an estimation of the number endangered jobs for the respective year. In 2013, the jobs at risk due to corporate default reached their highest level when approximately 30,800 to 36,900 jobs are endangered. After 2013 the number of jobs at risk decreases and since 2015 levels out between 18,000 and 22,500 approximately.

FIGURE XV JOBS AT RISK

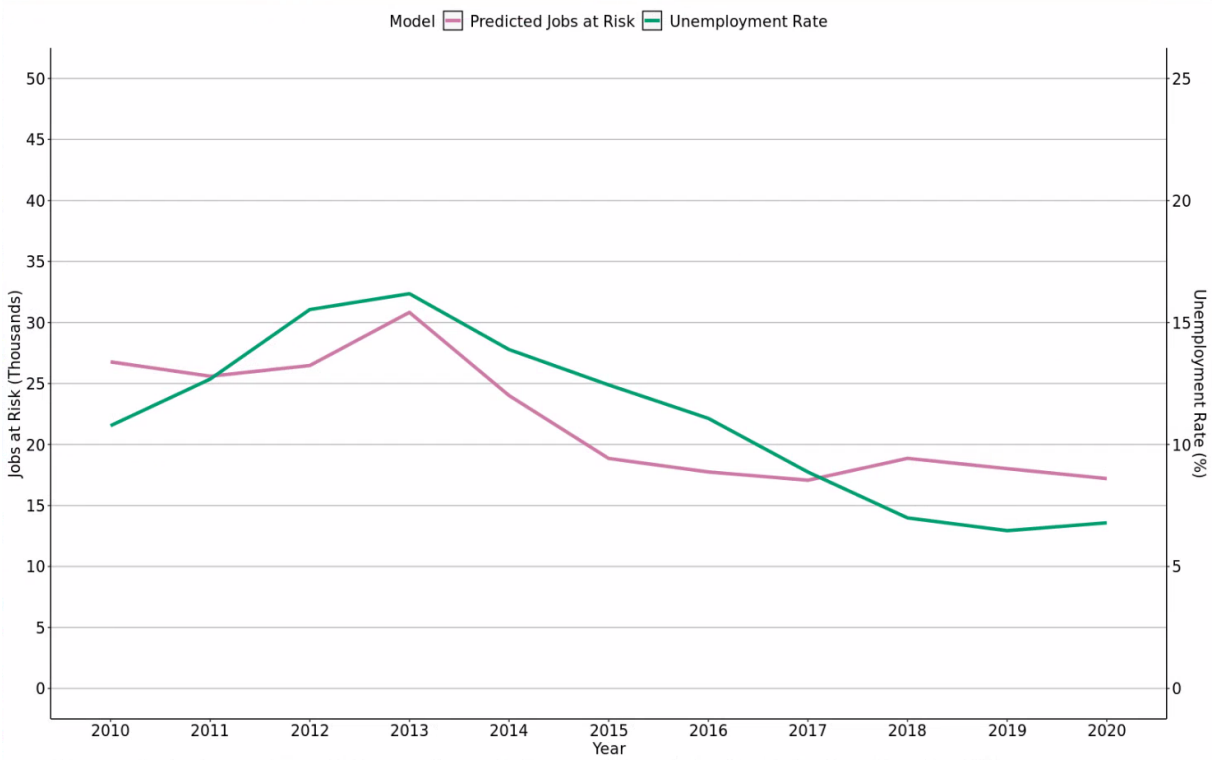


Note: Own estimation based on data provided by Banco de Portugal. The amount of Jobs at Risk is estimated using the out-of-sample probilities of defaults in the next year obtained by the random forest model and the number of jobs related to them.

In addition, there is a strong correlation between the predicted jobs at risk due to corporate default of Portuguese non-financial cooperations and the actual observed

unemployment rate in Portugal which is based on data provided by Statista. The predicted jobs at risk using the 0.72 discrimination threshold and the actual unemployment rate have a correlation coefficient of 0.835 which is illustrated in FIGURE XVI. Therefore, this strong correlation shows that these predictions of endangered jobs can be used as a leading indicator for future unemployment.

FIGURE XVI JOBS AT RISK COMPARED TO UNEMPLOYMENT RATE



Note: Own estimation based on data provided by Banco de Portugal. The amount of Jobs at Risk is estimated using the out-of-sample probilities of defaults in the next year obtained by the random forest model and the number of jobs related to them. The unemployment rate obtained from Statista.

Based on the findings of this research, the machine learning models random forest and XGBoost demonstrate superior discriminatory ability compared to traditional statistical models. The gap of discriminatory power remains significant across specifications which shows that it is beneficial to implement machine learning models for default forecasting. Especially when considering the micro cluster, the random forest model outperforms the best performing statistical model by 9.40 percentage points. Moreover, this approach can used to forecast the potential impact on the Portuguese labour market by estimating the number of jobs at risk due corporate default of non-financial micro cooperations.

VII. Discussion

The results of this research suggest that it is beneficial to implement machine learning models such as random forest or XGBoost for default forecasting because of their superior discriminatory power compared to traditional models. However, it is important to note that this study has some limitations.

First, when comparing the predictive power of statistical and machine learning models the quality of information used to train them has a substantial impact on the performance. As shown by Moscatelli et al., 2020, the addition of high-quality information such as credit behavioural indicators can lead to significant improvements in the discriminatory power for both statistical and machine learning models; however, statistical models benefit more from high-quality information.

Another limitation is the frequency of the data since this study is based on yearly data. However, the use of higher frequency such as quarterly or monthly data can potentially improve the accuracy of the predictive models.

Furthermore, the computational complexity of the hyperparameter tuning on a large dataset presents another limitation because as a result the hyperparameters are only tuned once. To optimize the model's performance a more frequent hyperparameter tuning can be advantageous. Furthermore, this study discusses the implementation of a clustering approach based on the firm size and its implications on the labour market. While this research focuses on micro firms, future research could explore the other firms size cluster as well as alternative clustering approaches based on other characteristics such as geographic location or industry.

Finally, while this approach focuses on the labour market implications there are many other possible macroeconomic consequences of corporate default that can be evaluated using a similar approach.

VIII. Conclusion

The study examines the potential benefits of implementing machine learning models for default forecasting by comparing the discriminatory power of the random forest and XGBoost models to that of the conventional statistical models linear discriminant analysis, logistic regression and penalized logistic regression. To evaluate the predictive power of the various models, their AuROC scores are compared across different specifications using a comprehensive dataset containing of firm-level variables and financial indicators for approximately 200,000 Portuguese non-financial cooperations from 2009 until 2020.

When the models are evaluated using out-of-time predictions, the machine learning models exhibit a significantly higher discriminatory power compared to the traditional models. With the reduction in sample size, the difference in predictive power between the two model types decreases. However, machine learning models maintain a higher level of performance. While modification in model dimensionality have a limited impact on the ability of statistical models to distinguish default and non-default instances, the predictive power of machine learning models increases with the addition of further predictors.

The study applies a clustering approach which divides firms based on their size. Both traditional and machine learning models exhibit an improvement in discriminatory power in all but the micro cluster in comparison to non-clustering specifications. Regarding the classification of micro firms, machine learning models demonstrate a significantly higher discriminatory ability, while traditional models exhibit limited performance. Furthermore, this research aims to evaluate the macroeconomic impact on the labour market by estimating the number of endangered jobs due to corporate default of non-financial cooperations.

The study can serve as a foundation for future research to evaluate the macroeconomic impact of corporate default using machine learning techniques.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609.

Bach, M., Werner, A., Żywiec, J., & Pluskiewicz, W. (2017). The study of under-and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis. *Information Sciences*, *384*, 174–190.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*, *24*.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2).

Berrar, D. (2019). *Cross-Validation*. 542–545.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L., Friedman, J. H., Olshen, R., & Stone, C. J. (1984). *Classification and Regression Trees*.

Chakraborty, C., & Joseph, A. (2017). *Machine learning at central banks*.

Chang, Y.-C., Chang, K.-H., & Wu, G.-J. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*, *73*, 914–920.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Cormack, G. V., Smucker, M. D., & Clarke, C. L. (2011). Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval*, *14*, 441–465.

Dhankhad, S., Mohammed, E., & Far, B. (2018). Supervised machine learning algorithms for credit card fraudulent transaction detection: A comparative study. *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, 122–125.

Dy, J. G., & Brodley, C. E. (2004). Feature selection for unsupervised learning. *Journal of Machine Learning Research*, *5*(Aug), 845–889.

EBA. (2016). *Final Report on the Application of the Definition of Default*. European Banking Authority. https://www.eba.europa.eu/eba-harmonises-the-definition-of-default-across-the-eu

European Commission. (2003). Commision Recommendation of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises. *Official Journal of the European Union*.

Fawcett, T. (2001). Using rule sets to maximize ROC performance. *Proceedings 2001 IEEE International Conference on Data Mining*, 131–138.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Feng, Q., Liu, J., & Gong, J. (2015). UAV remote sensing for urban vegetation mapping using random forest and texture analysis. *Remote Sensing*, *7*(1), 1074–1094.

Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. *Automated Machine Learning: Methods, Systems, Challenges*, 3–33.

Fix, E., & Hodges, J. L. (1951). *Discriminatory analysis: Nonparametric discrimination: Consistency properties*.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics*, *15*, 107–143.

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining*. Springer.

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (2008). *Feature extraction: Foundations and applications* (Vol. 207). Springer.

Haar, L., Anding, K., Trambitckii, K., & Notni, G. (2019). Comparison between Supervised and Unsupervised Feature Selection Methods. *ICPRAM*, 582–589.

Hamori, S., Kawai, M., Kume, T., Murakami, Y., & Watanabe, C. (2018). Ensemble learning or deep learning? Application to default risk analysis. *Journal of Risk and Financial Management*, *11*(1), 12.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, *14*(1), 55–63.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R* (Second Edition).

Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, *13*(4), 455.

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR)*, *52*(4), 1–36.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767–2787.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1–2), 273–324.

Kohavi, R., & Provost, F. (1998). *Glossary of Terms Special Issue on Applications of Machine Learning and the Knowledge Discovery Process*.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Liu, S., Wang, Y., Zhang, J., Chen, C., & Xiang, Y. (2017). Addressing the class imbalance problem in twitter spam detection using ensemble learning. *Computers & Security*, *69*, 35–49.

Mena, L. J., & Gonzalez, J. A. (2006). Machine Learning for Imbalanced Datasets: Application in Medical Diagnostic. *Flairs Conference*, 574–579.

Moscatelli, M., Narizzano, S., Parlapiano, F., & Viggiano, G. (2020). *Corporate default forecasting with machine learning*.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 109–131.

Oreski, S., Oreski, D., & Oreski, G. (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. *Expert Systems with Applications*, *39*(16), 12605–12617.

Pardo, R. (1992). *Design, testing, and optimization of trading systems* (Vol. 2). John Wiley & Sons.

Powers, D. M. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *ArXiv Preprint ArXiv:2010.16061*.

Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian processes for machine learning* (Vol. 1). Springer.

Shmueli, G., & Koppius, O. R. (2011). Predictive Analytics in Information Systems Research. *MIS Quarterly*, 553–572.

Shumway, R. H., Stoffer, D. S., & Stoffer, D. S. (2000). *Time series analysis and its applications* (Vol. 3). Springer.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, *25*.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62.1*, 77–89.

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

Appendix

TABLE A. 1 1 LIST OF TESTED VARIABLES

Variable	AuROC	Variable	AuROC	Variable	AuROC	Variable	AuROC
Debt-to-Capital Ratio	0.695	Employee Expenses	0.612	Fixed Tangible Assets	0.559	Net-Debt-to-Equity Ratio (Growth Rate)	0.504
Debt-to-Assets Ratio	0.695	Total Expenses	0.612	Fixed Tangible Assets and Intangible Assets	0.558	Capitalized Production	0.502
Equity-to-Assets Ratio	0.695	Employee Expenses - Salaries	0.611	Sales	0.554	Debt-to-Ebita Ratio	0.502
EBT	0.695	Social Security Expenses	0.608	Total non-current Assets	0.549	Interim Dividends	0.500
Net Income	0.691	Obtained Funding	0.608	Return on Assets (Growth Rate)	0.545	Firm Age	0.500
EBITDA	0.681	Expenses of Depreciations and Amortizations	0.603	Legal Reserves	0.541	Equity-to-Assets Ratio (Growth Rate)	0.500
EBIT	0.678	Current Liabilities - State and Other Public Entities	0.600	Operating Subsidies	0.538	Remaining non-current Assets	0.498
Operating Net Income	0.674	Salaries of Corporate Bodies	0.596	Financial Investments	0.537	Remaining Expenses	0.498
Equity	0.672	Curren Liabilities	0.594	Cash Discounts Granted	0.527	Non-Current Liabilities -Obtained Funding	0.496
Income Tax	0.667	Salaries	0.589	Impairment Losses	0.525	Remaining non-current Liabilities	0.494
Retained Earnings	0.665	Indirect Taxes	0.580	Current Assets - State and other Public Entities	0.523	Non-Current Liabilities	0.493
Current Ratio	0.659	Services	0.579	Variation in Production	0.523	Debt-to-Ebita Ratio (Growth Rate)	0.491
Mismatch Ratio	0.653	Remaining Income	0.578	Supplementary Income	0.513	Interest Expenses	0.489
Cash and Bank Deposits	0.646	Liabilities	0.576	Total Assets	0.513	Return on Assets	0.488
Total Income	0.643	Remaining Current Liabilities	0.576	Natural Logarithm of Total Assets	0.513	Debt-to-Equity Ratio (Growth Rate)	0.481
Cash-to-Assets Ratio	0.641	Donations	0.570	Equity and Liabilities	0.513	Inventories and Biological Assets	0.477
Turnover	0.638	Current Ratio (Growth Rate)	0.567	Mismatch Ratio (Growth Rate)	0.511	Subscribed Capital	0.470
Supplies and External Services	0.631	Costs of Goods Sold and Material Consumed	0.566	Remaining Current Assets	0.509	Cash-to-Assets Ratio (Growth Rate)	0.465
Insurance Schemes for Accidents at Work	0.620	Income from Financial Assets	0.560	Current Assets -Customers	0.508	Current Liabilities - Suppliers	0.452
Employee Expenses Except Salaries	0.612	Asset Turnover Ratio	0.560	Total Current Assets	0.508	Debt-to-Equity Ratio	0.414

Note:
Own calculation based on data provided by Banco de Portugal. The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the univariate logisitic regression.

TABLE A. 2 VARIABLE DESCRIPTION

Variables	Description
Debt-to-Capital Ratio	Ratio measures the proportion of a company's debt in its capital structure and is used to assess the financial leverage of a company and the level of risk associated with its debt.
Net Income	Total amount of a company's revenue or earnings after all expenses, including taxes, interest, and depreciation, have been deducted.
Operating Net Income	The amount of profit a company earns from its main or core business operations, after deducting all operating expenses.
Equity	Portion of a company's total assets that is owned by its shareholders.
Income Tax	Liability on the balance sheet representing the amount of tax that has been assessed by the government but has not yet been paid.
Retained Earnings	Portion of a company's net income that has been kept by the company instead of being paid out as dividends to shareholders.
Current Ratio	Ratio that is a measure of a firm's short-term liquidity and ability to pay its current liabilities.
Mismatch Ratio	Ratio that measures the extent to which a company's short-term liabilities and short-term assets are mismatched.
Cash and Bank Deposits	Current assets held by a company in the form of cash or cash equivalents such as bank accounts, money market accounts, and short-term investments
Cash-to-Assets Ratio	Ratio that measures a firm's liquidity and how easily it can service debt and short-term liabilities if the need arises.
Supplies and External Services	Supplies and external services are expenses incurred by a company for goods or services purchased from external suppliers.
Insurance Schemes for Accidents at Work	Insurance schemes for work accidents at work and occupational diseases
Employee Expenses Except Salaries	Costs incurred by a company in relation to its employees, such as benefits, bonuses, payroll taxes, and other related expenses, that are not included in the salaries of the employees
Total Expenses	Sum of all the costs incurred by a company in a given period, including operating expenses, interest expense, depreciation, and taxes
Obtained Funding	Amount of money that a company has raised from external sources such as loans, bonds, or other types of financing

TABLE A. 3 CORRELATION MATRIX

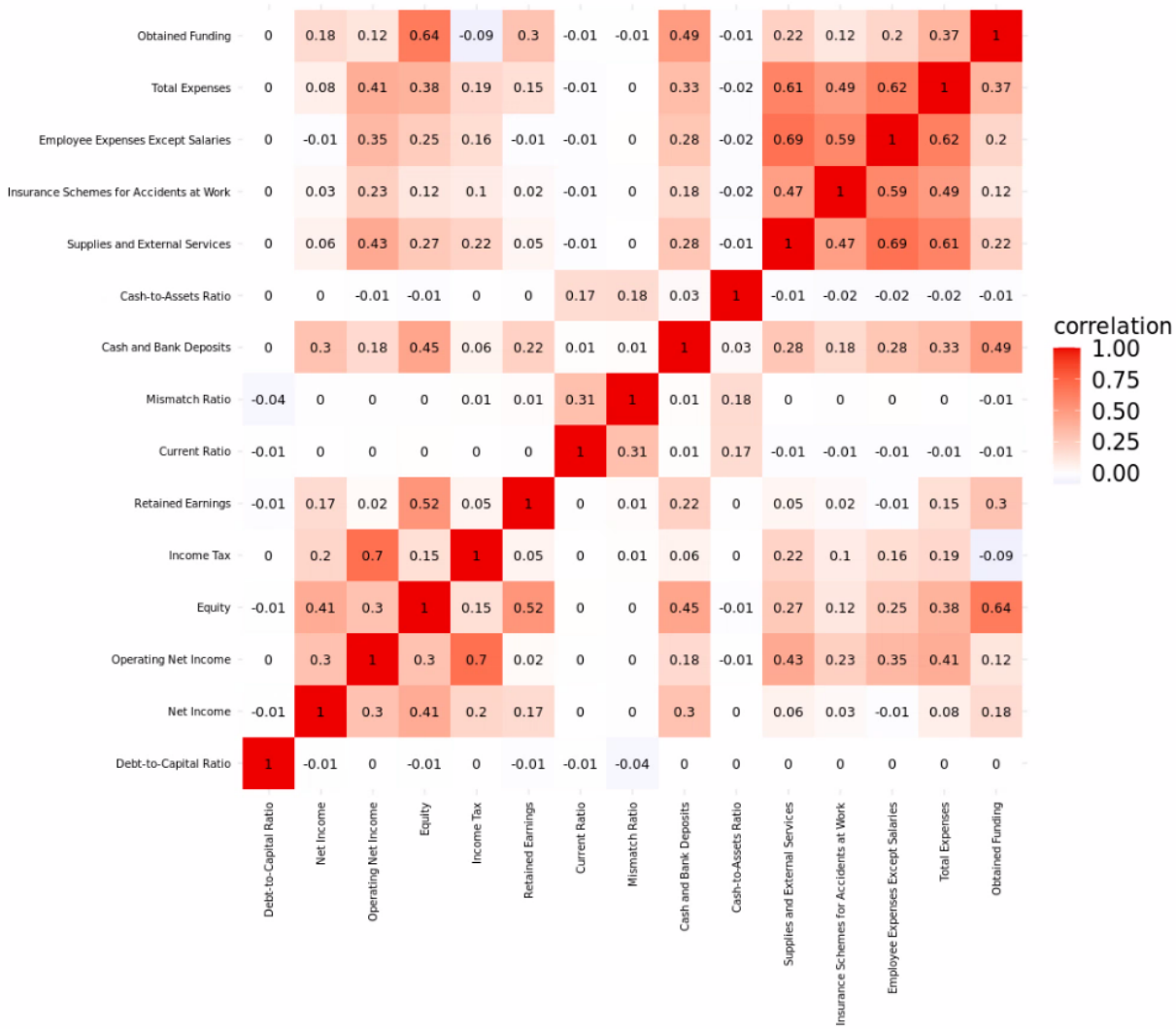


TABLE A. 4 DISCRIMINATORY POWER OF OUT-OF-TIME PREDICTIONS WITH COMPLETE DATASET

	Statistical Models			ML Models	
	Linear Discriminant Analysis	Logistic Regression	Penalized Logistic Regression	Random Forest	Extreme Gradient Boosting
2009	64.77%	69.12%	69.09%	75.93%	75.61%
2010	66.06%	69.30%	69.51%	76.20%	76.04%
2011	66.32%	68.91%	68.90%	76.64%	76.40%
2012	65.30%	67.52%	67.48%	78.09%	77.81%
2013	66.45%	68.38%	68.32%	79.20%	78.89%
2014	67.12%	68.82%	68.73%	80.46%	80.06%
2015	68.16%	70.18%	70.13%	80.41%	79.94%
2016	68.26%	70.12%	70.09%	80.73%	80.31%
2017	68.53%	70.66%	70.62%	80.74%	80.28%
2018	66.56%	68.83%	68.79%	79.92%	79.41%
2019	65.65%	67.74%	67.66%	80.97%	80.55%

Note:
Own calculation based on data provided by Banco de Portugal. The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the respective model.

TABLE A. 5 DISCRIMINATORY POWER OF OUT-OF-TIME PREDICTIONS WITH 50% DATASET

	Statistical Models			ML Models	
	Linear Discriminant Analysis	Logistic Regression	Penalized Logistic Regression	Random Forest	Extreme Gradient Boosting
2009	66.27%	69.30%	69.28%	75.53%	75.14%
2010	66.28%	69.62%	69.59%	75.76%	75.57%
2011	67.05%	69.50%	69.46%	76.29%	76.06%
2012	65.69%	69.56%	69.52%	77.49%	77.24%
2013	66.73%	68.24%	68.13%	78.70%	78.37%
2014	67.97%	70.35%	70.30%	79.89%	79.55%
2015	68.06%	70.48%	70.40%	79.68%	79.36%
2016	68.69%	70.06%	69.98%	80.02%	79.66%
2017	69.68%	71.39%	71.38%	79.96%	79.66%
2018	66.82%	70.14%	70.07%	79.57%	79.23%
2019	66.18%	69.71%	69.62%	80.17%	79.82%

Note:
Own calculation based on data provided by Banco de Portugal. The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the respective model.

TABLE A. 6 DISCRIMINATORY POWER OF OUT-OF-TIME PREDICTIONS WITH 10% DATASET

	Statistical Models			ML Models	
	Linear Discriminant Analysis	Logistic Regression	Penalized Logistic Regression	Random Forest	Extreme Gradient Boosting
2009	64.09%	65.68%	65.59%	74.35%	73.83%
2010	65.14%	69.33%	69.29%	74.38%	73.96%
2011	68.61%	69.64%	69.67%	74.91%	74.27%
2012	66.72%	68.56%	68.52%	76.25%	75.95%
2013	66.08%	66.37%	66.36%	77.16%	76.88%
2014	68.14%	69.97%	69.89%	78.55%	78.08%
2015	68.11%	70.11%	69.99%	77.90%	77.54%
2016	68.71%	70.80%	70.72%	78.73%	78.31%
2017	69.64%	72.16%	72.11%	78.26%	77.83%
2018	67.64%	69.53%	69.47%	78.20%	77.56%
2019	66.68%	71.30%	71.23%	78.89%	78.42%

Note:
Own calculation based on data provided by Banco de Portugal. The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the respective model.

TABLE A. 7 DISCRIMINATORY POWER OF OUT-OF-TIME PREDICTIONS WITH 5% DATASET

	Statistical Models			ML Models	
	Linear Discriminant Analysis	Logistic Regression	Penalized Logistic Regression	Random Forest	Extreme Gradient Boosting
2009	65.04%	68.37%	68.34%	73.71%	73.26%
2010	65.68%	67.85%	67.80%	73.82%	73.46%
2011	68.83%	70.11%	70.03%	74.21%	73.62%
2012	65.98%	67.91%	67.92%	75.57%	74.99%
2013	68.56%	70.53%	70.52%	76.62%	76.07%
2014	68.93%	70.69%	70.72%	77.84%	76.90%
2015	68.22%	70.85%	70.85%	77.82%	77.31%
2016	69.10%	70.90%	70.81%	77.89%	77.25%
2017	70.55%	71.92%	71.88%	77.95%	77.27%
2018	67.26%	68.70%	68.70%	76.70%	76.23%
2019	67.16%	69.70%	69.83%	78.04%	77.36%

Note:
Own calculation based on data provided by Banco de Portugal. The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the respective model.

TABLE A. 8 DISCRIMINATORY POWER OF OUT-OF-TIME PREDICTIONS WITH TOP 5 VARIABLES

	Statistical Models			ML Models	
	Linear Discriminant Analysis	Logistic Regression	Penalized Logistic Regression	Random Forest	Extreme Gradient Boosting
2009	68.71%	69.20%	69.19%	72.47%	71.88%
2010	68.07%	69.19%	69.17%	72.67%	71.78%
2011	68.38%	68.79%	68.77%	73.00%	72.47%
2012	69.36%	69.52%	69.52%	74.54%	73.88%
2013	69.45%	69.54%	69.52%	75.58%	74.99%
2014	66.78%	68.23%	68.22%	76.45%	75.55%
2015	69.78%	70.58%	70.57%	76.33%	75.46%
2016	69.46%	70.27%	70.26%	76.07%	75.23%
2017	69.16%	69.69%	69.67%	75.52%	74.56%
2018	69.66%	69.84%	69.81%	75.31%	74.57%
2019	69.99%	69.55%	69.55%	75.99%	75.04%

Note:
Own calculation based on data provided by Banco de Portugal. The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the respective model.

TABLE A. 9 DISCRIMINATORY POWER OF OUT-OF-TIME PREDICTIONS WITH TOP 10 VARIABLES

	Statistical Models			ML Models	
	Linear Discriminant Analysis	Logistic Regression	Penalized Logistic Regression	Random Forest	Extreme Gradient Boosting
2009	64.77%	69.12%	69.09%	75.93%	75.61%
2010	66.06%	69.30%	69.51%	76.20%	76.04%
2011	66.32%	68.91%	68.90%	76.64%	76.40%
2012	65.30%	67.52%	67.48%	78.09%	77.81%
2013	66.45%	68.38%	68.32%	79.20%	78.89%
2014	67.12%	68.82%	68.73%	80.46%	80.06%
2015	68.16%	70.18%	70.13%	80.41%	79.94%
2016	68.26%	70.12%	70.09%	80.73%	80.31%
2017	68.53%	70.66%	70.62%	80.74%	80.28%
2018	66.56%	68.83%	68.79%	79.92%	79.41%
2019	65.65%	67.74%	67.66%	80.97%	80.55%

Note:
Own calculation based on data provided by Banco de Portugal. The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the respective model.

TABLE A. 10 DISCRIMINATORY POWER OF OUT-OF-TIME PREDICTIONS WITH TOP 15 VARIABLES

	Statistical Models			ML Models	
	Linear Discriminant Analysis	Logistic Regression	Penalized Logistic Regression	Random Forest	Extreme Gradient Boosting
2009	65.02%	69.33%	69.32%	79.18%	78.79%
2010	66.66%	70.35%	70.43%	78.73%	78.55%
2011	66.66%	69.46%	69.44%	80.41%	80.06%
2012	65.98%	68.87%	68.81%	82.12%	81.74%
2013	66.83%	69.64%	69.58%	83.39%	83.02%
2014	67.87%	70.21%	70.17%	84.63%	84.32%
2015	68.47%	71.56%	71.61%	84.43%	84.15%
2016	68.75%	70.97%	70.89%	84.41%	83.97%
2017	68.77%	71.20%	71.17%	84.44%	83.93%
2018	66.92%	69.19%	69.14%	82.54%	81.95%
2019	65.90%	68.35%	68.27%	82.99%	82.51%

Note:
Own calculation based on data provided by Banco de Portugal. The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the respective model.

TABLE A. 11 DISCRIMINATORY POWER OF OUT-OF-TIME PREDICTIONS WITH MICRO FIRM DATASET

	Statistical Models			ML Models	
	Linear Discriminant Analysis	Logistic Regression	Penalized Logistic Regression	Random Forest	Extreme Gradient Boosting
2009	63.48%	66.54%	66.50%	74.52%	74.28%
2010	66.36%	68.56%	68.50%	74.88%	74.60%
2011	66.20%	67.79%	67.76%	75.44%	75.13%
2012	64.65%	66.51%	66.46%	76.85%	76.52%
2013	65.46%	66.72%	66.57%	77.81%	77.42%
2014	67.02%	68.68%	68.61%	79.13%	78.67%
2015	67.39%	69.17%	69.13%	79.15%	78.65%
2016	68.15%	69.75%	69.67%	79.37%	78.93%
2017	68.72%	70.66%	70.64%	79.45%	79.09%
2018	66.16%	67.51%	67.37%	78.68%	78.30%
2019	66.54%	69.48%	69.39%	79.32%	78.98%

Note:
Own calculation based on data provided by Banco de Portugal. The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the respective model.

TABLE A. 12 DISCRIMINATORY POWER OF OUT-OF-TIME PREDICTIONS WITH SMALL FIRM DATASET

	Statistical Models			ML Models	
	Linear Discriminant Analysis	Logistic Regression	Penalized Logistic Regression	Random Forest	Extreme Gradient Boosting
2009	75.51%	76.21%	76.25%	80.69%	80.23%
2010	74.09%	76.24%	76.15%	80.46%	80.35%
2011	74.23%	75.80%	75.79%	79.99%	79.39%
2012	76.05%	77.79%	77.73%	83.01%	82.41%
2013	77.32%	77.70%	77.71%	84.41%	83.67%
2014	78.19%	78.91%	78.92%	85.84%	85.29%
2015	77.00%	78.51%	78.51%	84.65%	84.27%
2016	77.06%	80.16%	80.13%	85.47%	84.63%
2017	77.69%	80.23%	80.24%	85.48%	84.99%
2018	75.45%	79.03%	79.01%	83.83%	83.22%
2019	76.68%	79.72%	79.71%	86.01%	85.50%

Note:
Own calculation based on data provided by Banco de Portugal. The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the respective model.

TABLE A. 13 DISCRIMINATORY POWER OF OUT-OF-TIME PREDICTIONS WITH MEDIUM FIRM DATASET

	Statistical Models			ML Models	
	Linear Discriminant Analysis	Logistic Regression	Penalized Logistic Regression	Random Forest	Extreme Gradient Boosting
2009	82.87%	83.08%	83.04%	86.19%	85.01%
2010	78.73%	78.09%	78.09%	82.51%	82.24%
2011	80.15%	78.38%	78.49%	83.78%	82.25%
2012	81.55%	82.15%	82.17%	85.96%	84.60%
2013	81.34%	83.59%	83.58%	87.46%	86.56%
2014	80.01%	81.50%	81.51%	87.98%	87.17%
2015	78.17%	83.73%	83.78%	86.61%	86.09%
2016	81.31%	81.53%	81.54%	85.31%	84.07%
2017	78.80%	79.47%	79.56%	84.99%	84.28%
2018	79.80%	80.66%	80.69%	84.48%	82.74%
2019	86.57%	86.38%	86.45%	89.09%	87.23%

Note:
Own calculation based on data provided by Banco de Portugal. The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the respective model.

TABLE A. 14 DISCRIMINATORY POWER OF OUT-OF-TIME PREDICTIONS WITH LARGE FIRM DATASET

	Statistical Models			ML Models	
	Linear Discriminant Analysis	Logistic Regression	Penalized Logistic Regression	Random Forest	Extreme Gradient Boosting
2009	72.86%	61.79%	62.10%	74.58%	75.69%
2010	76.89%	70.11%	80.18%	86.22%	86.84%
2011	76.64%	78.60%	78.88%	81.40%	79.19%
2012	75.16%	78.24%	78.35%	84.31%	82.86%
2013	80.55%	79.16%	79.11%	84.95%	81.60%
2014	82.47%	84.74%	84.83%	85.50%	83.98%
2015	76.41%	78.34%	78.24%	81.55%	79.29%
2016	91.07%	91.23%	91.09%	83.25%	81.76%
2017	77.07%	66.68%	66.64%	85.93%	82.34%
2018	72.46%	71.63%	71.67%	78.51%	81.60%
2019	68.58%	69.53%	69.47%	86.49%	83.01%

Note:
Own calculation based on data provided by Banco de Portugal. The AuROC score is calculated using the observed default data and the out-of-sample probabilities of defaults in the next year obtained by the respective model.