RISK MANAGEMENT MAGAZINE

Vol. 15, Number 3
September – December 2020

EXCERPT

https://www.aifirm.it/newsletter/progetto-editoriale/

Corporate Default Forecasting with Machine Learning

Mirko Moscatelli, Simone Narizzano, Fabio Parlapiano, Gianluca Viggiano

Corporate Default Forecasting with Machine Learning

Mirko Moscatelli, Simone Narizzano, Fabio Parlapiano and Gianluca Viggiano (Banca di Italia)

Sommario

Si presenta un raffronto tra modelli previsionali per il rischio di credito di tipo machine learning (ML) e modelli statistici standard quali la regressione logistica. Sfruttando un ampio dataset che include informazioni di bilancio e di Centrale dei Rischi per circa 300,000 imprese non finanziarie dal 2011 al 2017, si mostra come utilizzando soltanto informazioni di bilancio i modelli ML producono un incremento significativo in potere discriminate e precisione rispetto ai modelli statistici; il vantaggio si attenua quando si introducono anche le informazioni sul credito bancario.

Abstract

We compare statistical models usually employed in credit risk forecasting with machine learning algorithms (ML). Using a large dataset which includes financial ratios and credit behavioral indicators for about 300,000 Italian non-financial firms from 2011 to 2017, we show that training the models on financial statement data only, ML models record a significant improvement in discriminatory power and precision with respect to statistical models; however, this improvement is less pronounced when we enlarge the training dataset to include also credit behavioral data.

Keywords: credit scoring, machine learning, random forest, gradient boosting.

1 Introduction

Since the Basel II Accord, default forecasting methods based on a reduced-form regression approach have become popular in the banking industry. These methods consist of multivariate regression models which use firms' characteristics such as financial fundamentals to predict the credit quality of a firm. More recently, owing to the availability of large datasets and unstructured information, a growing strand of research suggests that models based on machine learning algorithms (henceforth ML) also constitute a suitable alternative for modelling default risk. ML refers to a class of models that can perform complex forecasting tasks when the relationship between predictors and the outcome is complex or unknown. As established in a number of works (Baesens et al., 2003; Brown and Mues, 2012; Barboza et al., 2017), ML models can perform highly accurate out-of-sample forecasts without imposing strong assumptions on the structure of the data generating process.

In this work we contrast statistical models with ensemble decision trees, a class of ML models which can handle both complex relationships across different variables and large datasets with correlated predictors. We use random forest (RDF) and gradient boosted tree (GBT) models, which combine a large number of predictions stemming from individual decision trees into a single (ensemble) forecast. The two models differ in the way individual trees are grown. In the RDF model, a random sample of the data and a random selection of variables are used for each tree in order to obtain less correlated individual predictions. The GBT model combines predictions obtained from trees that are tailored to deal sequentially with the forecasting errors of their predecessors.

We estimate the models using a large dataset covering financial and credit behavioral indicators for Italian non-financial firms. We test the out-of-sample performance for these competing models comparing one-year-ahead PD estimates and observed default data for the 2011-17 period.

Our analysis highlights the following main results:

(i) When using financial statement information usually available to external credit analysts, ML models outperform statistical models both in discriminatory power (the capacity to rank borrowers according to their riskiness) and precision (the ability to estimate PDs that deviate only marginally from actual default rates). When more information (namely from the Italian Credit Register) is added, gains from using ML are retained, albeit to a lesser extent.

(ii) We argue that improvements in forecasting performance from the use of ML are due to its capacity to exploit complex relationships between predictors and default outcomes.

2 Related literature

The use of the ML approach in credit risk modelling has gained momentum and has recently been applied in the field of early warning systems for banking crises, predictions of household mortgage or consumer credit default, and corporate default.

The most popular application of ML algorithms in default forecasting is in modelling consumer credit risk. Albanesi and Vamossy (2019) develop a model to predict consumer default based on deep learning (i.e. a combination of forecasts from deep neural network and gradient boosting) with high-dimensional data (over 200 variables). Deep learning models are shown to perform better than logistic regression and adapt to the behavior of the aggregate default rate quite closely. The model is able to capture the sharp rise in credit risk in the run-up to the 2007-09 crisis.

There are also numerous applications of ML to corporate default forecasting. Using a large dataset covering the North American corporate sector for the period 1987-2013, Barboza et al. (2017) show that ML provides an improvement of around 10 percentage points in AuROC over traditional models. In Bachman and Zhao (2017), ML models are compared to Moody's proprietary algorithm based on a regression model using corporate data for the United States; this exercise shows that ML models deliver an AuROC about 2-3 percentage points higher than a regression approach. However, their less transparent structure may lead to PDs that are difficult

to relate to firms' underlying characteristics. Furthermore, the inclusion of credit behavioral variables in the predictor set notably increases the AuROC of each model by over 10 percentage points.

3 Credit risk models at a glance

We compare two types of default forecasting models: statistical and ML models. Statistical models are especially fit for the purpose of inference, and typically rely on assumptions regarding the structural relationships between variables, the number of parameters that can be robustly estimated, and the distribution properties of the data generating process. ML models are mostly focused on prediction accuracy and make very weak assumptions on the data generating process. This feature allows the detection of data-driven interactions and non-linear or non-monotonic relationships between predictors and outcome variable. This is particularly relevant to credit risk applications, but it comes at a cost of less transparency compared with statistical models:[1] ML models do not provide estimates of the parameters that relate predictors to the outcome variable (the models are non-parametric) and this 'black box' feature can make their rationale and forecasts difficult to explain.

In this section, we briefly review standard credit risk models (3.1) and introduce the ML models used in the empirical application, namely random forest and gradient boosted trees (3.2).

3.1 Statistical models

Statistical theory offers a variety of methods for estimating default probabilities, of which linear discriminant analysis and logistic regression are the most popular. Linear discriminant analysis (LDA) provides an assessment of corporates' credit quality using a linear discriminant function that classifies borrowers into groups (default and non-default) based on their characteristics. For more than three decades discriminant analysis was used extensively by practitioners and performed reasonably well in predicting bankruptcy and other types of distress of privately and publicly held non-financial firms in the international context (Altman, 1968; Altman, 1983; Altman et al., 2017). However, criticism of its underlying assumptions (firms belonging to two different populations, normal distribution of observables, and equal covariance matrices for the two populations) gradually opened the field to more flexible models such as logistic regression.

The logistic regression model (LOG) estimates default probability from firms' observable characteristics modeling the default event as a Bernoulli random variable, assuming the value 0 for financially sound firms and 1 for defaulted firms. The model assumes that firms belong to the same population and that a known structural relationship (additive and linear) exists between the observable characteristics of the firm and the credit score. Model parameters are usually estimated via maximum likelihood. LOG relaxes some of discriminant analysis' assumptions (multivariate normality and equal covariance matrices), and has the significant advantage that its results can be easily interpreted.

These features attracted strong interest and prompted its diffusion as a scoring model amongst practitioners.[2] Nevertheless, a number of limitations still apply: the lack of consideration of non-linear or complex interactions between observables and defaults, the sensitivity to outliers or missing data, and difficulties in fully exploiting large datasets. Penalized logistic regression model (PLR) has the same structural form as LOG, but it estimates parameters maximizing out-of-sample forecasting performance. Previous evidence shows that PLR can outperform standard logistic regression in some prediction tasks (Zou and Hastie, 2005).

Overall, statistical models are satisfactory forecasting devices that accommodate both accuracy and transparency requirements. This is owing to their plain functional form, which combines additively monotonic predictors of default into a probability with good out-of-sample performance.

However, the global financial crisis exposed some pitfalls in default forecasting based on commonly used rating systems approaches: i) their slow capacity to adapt to changes in the state of the economy, and ii) their limited ability to model complex non-linear interactions between economic, financial and credit variables. For example, credit ratings may overlook signals of a deteriorating economic and credit environment, such as a rapid increase in default rates or negative shocks to the supply of credit.

3.2 Machine learning models

We applied two ML algorithms used extensively in credit risk applications: random forest (Breiman, 2001) and gradient boosted trees (Friedman, 2000). The building blocks for these models are classification trees, which are partition algorithms that recursively split the dataset into smaller sets (or branches) that best separate defaulters from non-defaulters.[3] At each iteration, the decision tree algorithm chooses from the covariates space $\mathbf{X}$ a variable and a value for that variable, so as to minimize a measure of heterogeneity (impurity index) in the resulting subgroups with respect to the classification variable.[4] The process continues within each branch until a stopping condition is reached, such as too few observations in the branch or no significant reduction in impurity. The final branches - called leaves - contain a constant estimate for the probability of default of firms in each leaf, computed as the proportion of defaulted firms over the total number of firms in that leaf in the estimation (or training) dataset.

---

[1] A recent line of research has put forward methods to increase the interpretability of ML models (see Guidotti et al., 2019). Moreover, the use of ML models for the purpose of inference is also considered in a number of works (see Chernozhukov et al., 2018 and Joseph, 2019).

[2] For example, the Bank of Italy's In-house Credit Assessment System (BI-ICAS), which is used to assess credit claims posted as collateral in Eurosystem monetary policy operations, is based on a standard logit framework. BI-ICAS integrates two credit risk measures: the credit behaviour component that models monthly data sourced from the Italian Credit Register; and the financial component, based on yearly financial statement data reported in the company accounts data archive of the Bank of Italy. The statistical model of the BI-ICAS estimates at monthly frequency around 300,000 default probabilities for Italian non-financial limited companies. The forecasting performance of the model is assessed annually by the Eurosystem.

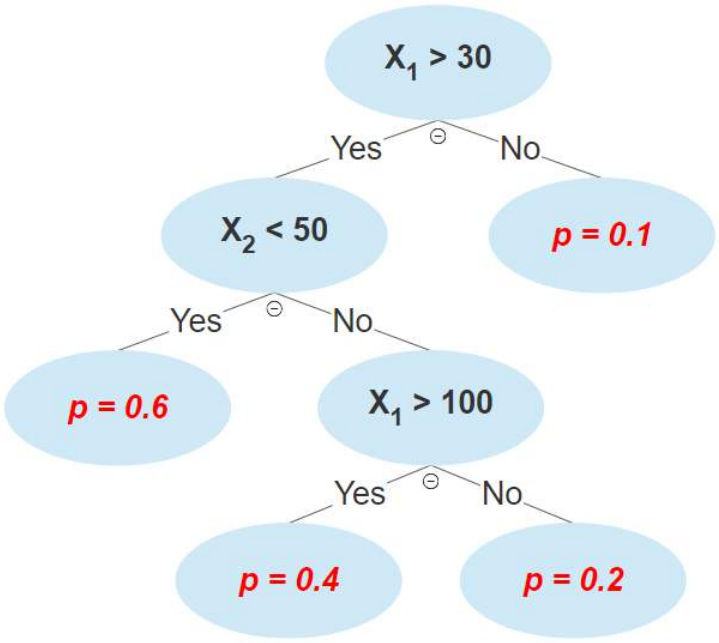
[3] A good introduction to trees, random forest and boosting algorithms can be found in Hastie, Tibshirani and Friedman (2001).

[4] An example of a measure of impurity, as well as the one that we use, is the Gini coefficient. Given a set of observations S with a binary value for each observation $y_i \in \{0,1\}$, the Gini impurity coefficient is defined as $GI(S) := 2 \cdot \sum_i p_i (1-p_i)$, where p_i is the percentage of observations in S such that $y_i=1$. The Gini impurity coefficient ranges between 0, when $p_i = 0$ or $p_i = 1$, and 0.5 when $p_i = 0.5$.

A simplified classification tree containing only two splitting variables X_1 and X_2 and four leaves is shown in Figure 1. The first level splits the sample into two branches depending on the value of X_1. Firms with X_1 greater than 30 end up in a leaf and receive a default probability equal to 0.1, whereas firms with X_1 less than or equal to 30 are classified in a middle branch and are divided further according to the value of X_2. Firms with X_2 smaller than 50 end up in a leaf and receive a default probability equal to 0.6, while firms with X_2 greater or equal to 50 undergo a final split according to the value of X_1: if X_1 is greater than 100 their default probability equals 0.4; otherwise, they receive a default probability of 0.2.

Notice that the classification tree model: i) captures interactions between variables: the effect of X_2 on the default probability strongly depends on the value of X_1,for instance, if X_1 is greater than 30 the value of X_2 is irrelevant for the estimated default probability; ii) captures non-linear relationships, since the same variable can be used more than once with different values to split the tree, for example, X_1 is used twice to split the sample; and iii) the default probabilities are not a continuous function of the variables: a small increase from 100 to 101 in the value of X_1 results in the default probability of the firm going from 0.2 to 0.4.

Figure 1: Decision tree



Notes: A decision tree for predicting firms' default probabilities grown using two variables, namely X_1 and X_2. Decision rules according to which branches are split are reported in black, while estimated default probabilities are reported in red.

Classification trees have the desirable property of being low-bias, meaning that the leaves, defined using multiple variables simultaneously, can fit the data extremely well. However, this flexibility results in an undesirable high-variance property, meaning that out-of-sample predictions are highly sensitive to small changes in the estimation dataset, which often leads to low forecasting power.

This shortcoming is addressed by ensemble classifiers, such as those provided by the random forest and gradient boosted tree models. Instead of using one single classification tree, these models grow a number of trees, and the final prediction is obtained as the average of the predictions stemming from the individual trees. In particular, the predictive function F takes the form of $F(x)=\frac{1}{k}\sum_{i=1}^{k} T_i(x)$, where $\{T_1, \ldots, T_k\}$ are the k different classification trees and $T_i(x)$ is the probability of default that the tree T_i associates to a borrower with covariates x. The two models differ in how the different set of trees is constructed.

Random forest (RDF) grows the set of trees: i) using a different bootstrapped sample of the original dataset for each tree (i.e. a sample with replacement having the same number of observations of the original dataset); and ii) selecting at each branch the best split using only a randomly selected subset of the covariates. This procedure implies that the trees differ from one other, since the underlying information set is different.

Gradient boosted tree (GBT) grows the set of trees recursively: at each step classification errors from the previous trees are used as the dependent variable to grow the next tree. Namely, the first tree T_1 is a standard classification tree as described above; from the second tree onwards, trees are grown using the same x covariate set, but a different dependent variable, computed as the difference between the 0/1 default outcome and the estimations of the previous trees.[5] In other words, the first tree T_1 will be trained on the model $y = T_1(x)$, the second tree T_2 will be trained on the model $y - T_1(x) = T_2(x)$, the third tree T_3 on the model $y - T_1(x) - T_2(x) = T_3(x)$, and so on, each time trying to predict the forecasting errors of the previous trees.[6] Iterated learning from previous

[5] Regression trees differ from classification trees because the output variables are continuous rather than numerical. They are generated in the same way, the only difference being that the impurity function is the variance of the outcome variable instead of the Gini impurity coefficient.

[6] Residuals can be interpreted as negative gradients in F of the quadratic loss function $\frac{1}{2}(y\text{-}F(x))^2$ from which the name "gradient boosting" is derived.

forecasting errors can achieve very accurate predictions, but can also lead to overfitting, thus the number of trees used is a very important hyperparameter to be chosen via cross-validation.[7]

4 The training dataset

4.1 Corporate default

We use an extensive dataset of financial and credit behavioral indicators for Italian non-financial firms for the period 2011-17. Our dependent variable, namely financial default, is sourced from the Italian Credit Register and reflects a system-wide definition of a borrower's non-performing status. A firm is classified in default on a given year if the ratio of non-performing credit to total credit drawn from the banking system is greater than 5 per cent for at least one month.[8] The default rate, i.e. the ratio of borrowers classified as non-performing in a given year to the total number of borrowers not in default at the beginning of the year, gives an aggregate measure of credit risk which we aim to model at firm level (Table 1).

Within the 2011-17 time period, corporate sector credit risk peaked in 2014, in the aftermath of the European sovereign debt crisis and the associated slowdown of the Italian economy. Following the monetary policy measures adopted by the European Central Bank, the gradual improvement in the business cycle and the exit of vulnerable firms from the market, the aggregate default risk also decreased, with default rate levels approaching about 2.5 per cent in 2017.

The sudden increase in the number of defaults in both 2012 and 2014, which more than doubled compared to the previous year, may pose a challenge to slow-adapting credit risk models. From a qualitative point of view, the ability of models to forecast defaults in different stages of the credit cycle is a desirable property.

Table 1: Default rate

Year	N Firms	Default Rate
2011	222,879	1.20%
2012	233,157	2.45%
2013	259,166	2.23%
2014	249,566	4.76%
2015	252,059	4.12%
2016	260,156	3.31%
2017	269,657	2.53%

Notes: Own calculation based on National Credit Register data. N Firms refers to the number of firms included in our sample, while Default rate is the proportion of firms in default in a given year over the number of firms not in default at the beginning of the year.

4.2 Financial and credit behavioral indicators

The majority of academic works on credit risk modelling use economic and financial ratios as potential indicators of corporate defaults. We add to these predictors a set of credit behavioral indicators on the firm-bank relationship.

Our dataset, drawn from the Company Accounts Data system (provided by Cerved Group) and the Italian Credit Register, contains a wide array of firm-level variables for Italian non-financial companies. Starting from financial ratios (time lag of one year), we compute 24 indicators covering: profitability, financing structure, debt sustainability and asset types. Credit behavioral indicators (with a time lag of two months) include eight variables related to a firm's financial flexibility, that is the proportion of drawn to granted bank credit for different facilities, and the occurrence of delinquencies within a firm-bank credit relationship. After including firms' descriptive indicators, such as economic sector and geographical area, our set of default predictors contains 38 variables. Variables were then selected using the following criteria:

1. using univariate logit regressions for the probability of default, variables with an AuROC lower than 55 per cent were dropped;
2. using the Kolmogorov-Smirnov test, variables with insignificant differences in the distributions between the default and non-default groups were dropped;
3. from the list satisfying 1) and 2), only less correlated variables were retained (linear correlation < 0.7).

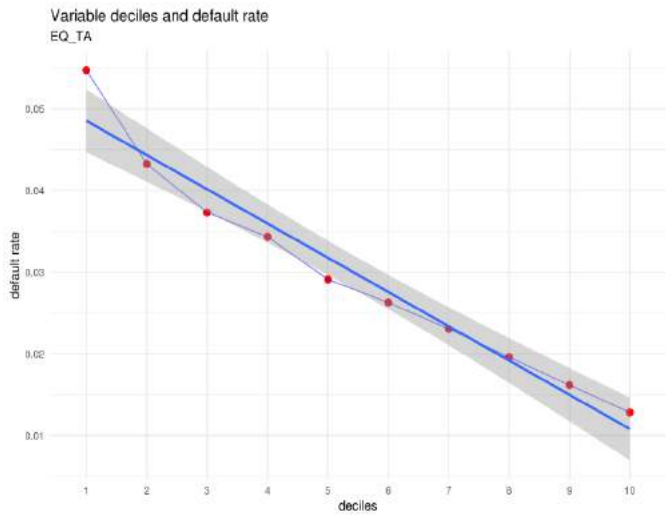
[7] Hyperparameters define the general characteristics of a model, such as its complexity, and can either be set a priori or learned from the data through optimizers such as grid search.

[8] The status of non-performing loans includes different stages of impairment: past-due 90 days, unlikely to pay and bad loans.

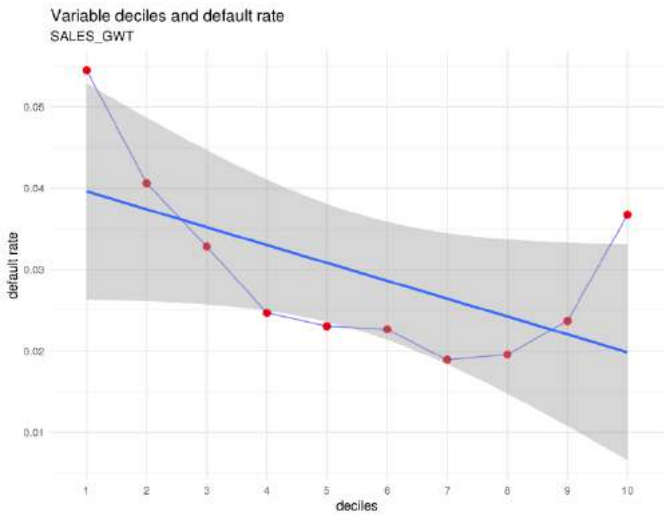
We ended up with 26 predictors, and the estimation dataset includes about 250,000 yearly observations for defaulted and non-defaulted firms. A complete description of predictors can be found in the Appendix.
Variables displaying a linear relationship with the default rate are suitable candidates to be default predictors in all models. For example, financial leverage (equity over total assets) correlates linearly with the default rate (Figure 2, Panel A). ML models are expected to benefit more than statistical models from non-linear or non-monotonic indicators. For example, the sales growth rate displays a non-monotonic relationship with default risk, with low- and high- sales-growth firms being riskier (Figure 2, Panel B).

Figure 2: Default rate and predicting variables deciles

Panel A: Equity over total assets



Panel B: Sales growth



Notes: Own calculation based on National Credit Register and Cerved data.

Figure 2 plots (dots) the default rates associated with different deciles of two of the variables employed to predict defaults.

5 Calibration

The estimation of credit risk models is subject to the rare-event problem, i.e. the dataset usually only contains a small proportion of default observations compared to non-default observations, causing a weak discriminatory power: on the one hand, more importance is attributed to variables that identify sound firms (which are strongly represented in the estimation sample) rather than to variables that help distinguish distressed firms; on the other hand, the model tends to give very low PDs to all the firms.

Undersampling is commonly used to overcome this issue. In a first stage, the model is estimated on a balanced sample, namely a sample with an equal number of default and non-default observations, with the latter randomly sampled from the original dataset.[9] Since this balanced sample does not reflect the real level of credit risk, during the second stage, a recalibration is performed via algebraic manipulation using Bayes correction (see Sugiyama et al., 2017; Dal Pozzolo et al., 2015).[10]

A second issue related to the estimation of ML models is hyperparameter calibration, which defines the general structure of the models. In RDF, we need to calibrate the number of variables selected at each split, whereas in GBT, we need to set the number of trees and leaves on each tree.

Using *k*-fold cross-validation, we choose the hyperparameters that maximize the out-of-sample performance.[11] First, we randomly divide the training set into *k* disjoint subsets (folds) of equal size; then, we train the model *k* times, each time on a dataset composed by the union of different *k-1* folds, and we use the remaining fold to compute an out-of-sample AuROC for the model. Finally, this gives us *k* out-of-sample accuracy measures for each combination of the parameters, and we select the parameters presenting the highest average AuROC.

6 Results

In this section, we compare the credit risk models described in Section 2 based on discriminatory power and precision. First, we evaluate their performance using a limited training dataset, which includes financial indicators and firms' characteristics only. This dataset contains only publicly available financial information. Then we use the whole dataset, which covers both financial and credit behavioral information from the Italian Credit Register, with the latter set of indicators usually available only to lenders or supervisors.

[9] See Wallace and Dahabreh (2014).
[10] See Appendix 2 for a detailed description of the calibration procedure.
[11] For an introduction to cross-validation, see for example Geisser (1993).

6.1 Discriminatory power

We assess the discriminatory power of the one-year default probabilities estimated by the different models with area under the curve (AuROC).[12] This is a measure of the ability of the model to assign higher default probabilities to firms that will default compared with financially sound firms: a random model that does not discriminate between sound and distressed firms has a 0.5 AuROC, while a perfect model has an AuROC of 1.

We first report the out-of-sample AuROC for the different models using only financial ratios and firm characteristics (such as geographic area, economic sector and firm size) that can be collected from publicly available sources (Table 2). The AuROC ranges from 72 to 77 per cent, a level of accuracy comparable to Wang and Dwyer (2011), Bacham and Zhao (2017), and Barboza et al., (2017).

We find that tree-based models outperform statistical models over the entire time span, with an average increase in discriminatory power over the LOG model of about 2.6 per cent. Linear discriminant analysis (LDA) and penalized logistic regression (PLR) display results very close to the LOG model.

Table 2: Discriminatory power with restricted dataset (financial indicators)

Year	Linear discriminant analysis	Logistic regression	Penalized logistic regression	Random forest	Gradient boosted trees
	LDA	LOG	PLR	RDF	GBT
2012	73,7%	73,9%	73,9%	76,6%	76,3%
2013	73,7%	73,9%	73,9%	77,2%	77,3%
2014	72,2%	72,3%	72,4%	74,4%	73,9%
2015	73,7%	73,7%	73,7%	76,1%	76,0%
2016	72,6%	72,6%	72,6%	75,3%	75,3%
2017	73,0%	73,0%	73,0%	75,7%	75,4%

Notes: Own calculation based on Cerved data. The AuROC score is computed using out-of-sample Probabilities of defaults obtained from the various models and observed default data.

We then expand the training set to include also credit behavioral indicators. The AuROCs for these models are reported in Table 3. We observe an increase in overall discriminatory power when credit behavioral indicators are included, amounting to about 10 percentage points in AuROC. This finding is consistent with Bacham and Zhao (2017), where credit behavioral indicators lead to similar increase in accuracy over a model based exclusively on financial ratios and firm characteristics. While still outperforming statistical models, tree-based models now provide a smaller increase in discriminatory power over the logistic model. We interpret this finding as the effect of the different information set: with high quality data, the logistic regression already provides very good forecasting, approaching the upper bound in discriminatory power. As a result, the improvement in default forecasting owing to the use of ML is necessarily less pronounced.

[12] See Fawcett (2004), Chawla (2009) and Xu-Ying et al. (2009).

Table 3: Discriminatory power with complete dataset (financial and credit behavioral indicators)

Year	Linear discriminant analysis	Logistic regression	Penalized logistic regression	Random forest	Gradient boosted trees
	LDA	LOG	PLR	RDF	GBT
2012	83,8%	84,0%	84,0%	84,6%	84,7%
2013	83,2%	83,3%	83,3%	84,2%	84,4%
2014	81,1%	81,6%	81,6%	82,5%	82,7%
2015	82,8%	82,9%	82,9%	84,4%	84,6%
2016	82,9%	83,0%	83,0%	84,1%	84,0%
2017	82,9%	83,1%	83,1%	84,1%	84,2%

Notes: Own calculation based on Cerved and National Credit Register data. The AuROC score is computed using out-of-sample probabilities of defaults obtained from the various models and observed default data.

6.2 Backtesting

To assess the degree to which default probabilities match realized defaults, we perform a binomial-style test for different credit quality buckets, using the Credit Quality Steps (CQS) defined by the Eurosystem for validation and annual monitoring of credit rating systems. In particular, we use a backtesting strategy outlined in Coppens et al. (2017) as the 'traffic light approach': for each bucket, we test how often realized default rates are compatible with forecasted PDs and in the range of usual statistical deviations. A colour is assigned based on the p-value of the test, with green indicating that realized defaults are below the expected threshold, and yellow and red indicating that there is a positive or strongly positive discrepancy between expected and realized defaults respectively.

In Table 4 we report the results of the backtesting exercise for ML and statistical models trained on the full information set of financial and credit behavioral data: each cell reports the realized default rate within a certain PD bucket, and the PD thresholds represent the upper limit for PD in each interval.

For the years 2013 and 2015-17, characterized by declining default rates (Table 1), all rating systems report satisfactory performances. Structural models (LDA, LOG and PLR), however, show a lower capacity to correctly classify high credit quality borrowers, recording several red warnings in the CQSs 1-2 and 3. Tree-based models, instead, do not report significant discrepancies between expected and realized default for these years.

In the years 2012 and 2014, characterized by a strong increase in default rates, the estimated PDs often do not match realized defaults. Structural models tend to record weaker performances, presenting red warnings in all of the credit quality buckets.

Overall, these results show that the assessment of high credit quality borrowers and the adaptation to rapid deterioration in aggregate default risk is a common problem for structural models, while ML models can partially mitigate these issues. In particular, RDF predictions tend to be more precise compared to other models both in economic upturns and downturns.

7 Conclusions

We compared statistical models usually employed in credit risk modelling with ML models, namely random forest and gradient boosted tree models. We used a large dataset which includes financial ratios and credit behavioral indicators for about 300,000 Italian non-financial firms for the years 2011-17.

When the models are trained using only publicly available information, ML models have a more accurate forecasting performance, both in terms of discriminatory power and precision, compared with statistical models. This gain is reduced when high quality information, such as credit behavioral indicators, is added to the training set.

We argue that the better forecasting performance of ML models is due to their ability to capture more precisely the complex relationship between the available firms' indicators and the default outcome. Our results suggest that the joint use of statistical and ML models by lenders or credit analysts may be beneficial for the accurate assessment of potential borrowers. For example, ML models, which are relatively non-transparent, may be used as a benchmark for more transparent statistical models.

Table 4: Backtesting

		2012					2013				
CQS	*Threshold*	*LDA*	*LOG*	*PLR*	*RDF*	*GBM*	*LDA*	*LOG*	*PLR*	*RDF*	*GBM*
CQS1-2	*0,1%*	0,4%	0,5%	0,5%	0,0%	0,0%	1,0%	0,6%	0,6%	0,0%	0,0%
CQS3	*0,4%*	0,6%	0,7%	0,7%	0,4%	0,6%	0,4%	0,4%	0,5%	0,2%	0,2%
CQS4	*1%*	1,3%	1,4%	1,4%	1,1%	1,6%	0,6%	0,7%	0,7%	0,5%	0,6%
CQS5	*1,5%*	2,3%	2,5%	2,5%	2,3%	3,1%	0,9%	1,1%	1,0%	0,8%	1,2%
CQS6	*3%*	4,4%	4,5%	4,5%	4,5%	4,9%	1,7%	1,9%	1,8%	1,5%	2,0%
CQS7	*5%*	9,0%	9,0%	9,0%	8,9%	8,1%	3,4%	3,5%	3,5%	3,1%	3,6%
		2014					2015				
CQS	*Threshold*	*LDA*	*LOG*	*PLR*	*RDF*	*GBM*	*LDA*	*LOG*	*PLR*	*RDF*	*GBM*
CQS1-2	*0,1%*	0,5%	0,7%	0,7%	0,0%	0,0%	4,8%	2,7%	2,0%	0,5%	NA
CQS3	*0,4%*	1,0%	1,1%	1,0%	0,2%	0,5%	0,9%	0,8%	0,8%	0,1%	0,1%
CQS4	*1%*	1,4%	1,6%	1,6%	0,7%	1,4%	0,9%	0,9%	0,9%	0,3%	0,5%
CQS5	*1,5%*	2,1%	2,3%	2,3%	1,5%	2,7%	0,9%	1,1%	1,1%	0,6%	1,0%
CQS6	*3%*	3,3%	3,4%	3,3%	3,2%	4,1%	1,7%	1,8%	1,8%	1,4%	1,9%
CQS7	*5%*	5,4%	5,6%	5,6%	6,3%	6,4%	2,8%	2,8%	2,8%	3,1%	3,5%
		2016					2017				
CQS	*Threshold*	*LDA*	*LOG*	*PLR*	*RDF*	*GBM*	*LDA*	*LOG*	*PLR*	*RDF*	*GBM*
CQS1-2	*0,1%*	0,0%	6,4%	7,9%	0,0%	NA	12,5%	2,0%	1,8%	0,4%	0,0%
CQS3	*0,4%*	1,1%	1,0%	1,0%	0,0%	0,1%	0,5%	0,5%	0,4%	0,1%	0,2%
CQS4	*1%*	0,8%	0,8%	0,8%	0,3%	0,6%	0,7%	0,7%	0,7%	0,3%	0,6%
CQS5	*1,5%*	1,0%	1,1%	1,1%	0,8%	1,1%	1,0%	1,0%	1,1%	0,7%	1,2%
CQS6	*3%*	1,8%	1,9%	1,8%	1,6%	2,2%	1,7%	1,7%	1,6%	1,7%	2,2%
CQS7	*5%*	3,6%	3,5%	3,5%	3,5%	3,9%	3,7%	3,7%	3,7%	3,6%	3,7%

Notes: Own calculation. The Threshold column reports the upper limit of the CQS interval identified by the Euro Credit Assessment Framework (ECAF) scale. For instance, a firm is classified in CQS3 if the default probability is between 0.1 and 0.4 per cent. The percentages in the colored cells represent the realized default rate for each CQS in each year. The green, yellow and red colors denote the p-value of the "traffic light approach" test; where the H0 hypothesis is that true probabilities of defaults are less than or equal to the thresholds. The green shading indicates a p-value greater than 20 per cent, the yellow shading between 1 and 20 per cent, and the red shading less than 1 per cent.

Mirko Moscatelli, Simone Narizzano, Fabio Parlapiano and Gianluca Viggiano

References

- Albanesi, S. & D.F. Vamossy (2019). Predicting Consumer Default: A Deep Learning Approach. *NBER Working Papers*, No. 26165.
- Altman, E.I. (1968, September). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance*, *23*, 589-609.
- Altman, E.I. (1983). *Corporate Financial Distress: A Complete Guide to Predicting, Avoiding and Dealing With Bankruptcy.* New York: John Wiley & Sons.
- Altman, E.I., M. Iwanicz-Drozdowska, E.K. Laitinen & A. Suvas (2017, June). Financial Distress Prediction in an International Context: A Review and Empirical Analysis of Altman's Z-Score Model. *Journal of International Financial Management and Accounting*, *28*, 31-171.
- Bacham D. & J. Zhao (2017, July). Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modeling. *Moody's Analytics Risk Perspectives*, *9*, 30-35.
- Baesens, B., T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens & J. Vanthienen (2003, June). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*, 627-635.
- Barboza, F., H. Kimura & E. Altman (2017, October). Machine learning models and bankruptcy prediction. *Expert Systems with Applications: An International Journal*, *83*, 405-417.
- Breiman, L. (2001, October). Random Forests. *Machine Learning*, *45*, 5-32.
- Brown, I. & C. Mues (2012, February). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*, 3446-3453.
- Chawla, N.V. (2009). Data Mining for Imbalanced Datasets: An Overview. In O. Maimon and L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 875-886). Boston: Springer.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey & J. Robins (2018, February). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*, C1-C68.
- Coppens, F., F. González & G. Winkler (2017). The performance of credit rating systems in the assessment of collateral used in Eurosystem monetary policy operations. *European Central Bank Occasional Paper Series*, No. 65.
- Dal Pozzolo, A., O. Caelen & G. Bontempi (2015). When is Undersampling Effective in Unbalanced Classification Tasks?. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 200-215). Cham: Springer.
- Fawcett, T. (2004, September). ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, *31*, 1-38.
- Friedman, J.H. (2000, October). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*, 1189-1232.
- Friedman, J., T. Hastie & R. Tibshirani (2001). *The Elements of Statistical Learning: Data Mining, Interference, and Prediction.* New York: Springer.
- Geisser, S. (1993). Predictive Inference: An Introduction. *Monographs on Statistics and Applied Probability*, 55.
- Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi & F. Giannotti (2019, January). A Survey of Methods for Explaining Black Box Models. *ACM computing surveys*, *51*, 93.
- Joseph, A. (2019). Shapley regressions: A framework for statistical inference on machine learning models, *Bank of England Working Papers*, No. 784.
- Sugiyama, Masashi, Neil D. Lawrence & Anton Schwaighofer (2018). *Dataset shift in machine learning*. Cambridge: The MIT Press.
- Wallace, B.C., and I.J. Dahabreh (2014, October). Improving class probability estimates for imbalanced data. *Knowledge and Information Systems*, *41*, 33-52.
- Wang, J. & D.W. Dwyer (2011). Moody's Analytics RiskCalc™ v3.1 Italy, *Moody's Analytics*.
- Xu-Ying, L., Wu, J. & Zhou, Z.H. (2009, October). Exploratory Undersampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*, 539-550.

Appendix

We describe the financial and credit behavioral indicators used to predict default of non-financial firms. By means of graphical inspection, those variables presenting a non-linear or non-monotonous relationship with respect to the default outcome are labelled "NL".

Variable	Description
TURNOVER (Asset Turnover Ratio)	Ratio between net sales and total assets. The asset turnover ratio is an efficiency ratio that measures a firm's ability to generate sales from its assets.
VA_TA (Value Added to Total Assets)	Ratio between economic value added and total assets. Operating profitability ratio that measures the firm's ability to generate value from its assets.
EBITDA_MARGIN (EBITDA to Net Sales) - ***NL***	Operating profitability ratio that measures how much earnings the firm is generating before interest, taxes, depreciation, and amortization, as a percentage of revenue.
PFN/PN (Net Debt to Equity)	Measure of a firm's financial leverage, calculated by dividing its net liabilities by stockholders' equity.
EQ_TA (Equity to Total Assets)	Ratio between equity and total assets. Used to assess a company's financial leverage
PFN/EBITDA (Net Debt to EBITDA)	Debt sustainability ratio gives an indication as to how long a firm would need to operate at its current level to pay off all its financial debt.
IE_CASHFLOW (Interest Expenses to Cash Flow)	Ratio that indicates the enterprise's ability to pay interest from generated cash flow.
DSCR (Debt Service Coverage Ratio)	Ratio of debt sustainability that refers to the amount of cash flow available to pay interest expenses and annual principal payments on financial debt.
FIN_MISMATCH (financial mismatch)	Ratio of the mismatch (difference) between short-term liabilities and short-term assets and total assets. Negative value of the ration (short-term liabilities > short-term assets) indicates that the firm has enough short-term assets to meet its short-term liabilities.
CASH_ST_DEBT_S (Current Assets to Short Term Debt)	Liquidity ratio that measures a firm's ability to pay off short-term debt obligations with cash and cash equivalents.
CASH_TA (Cash to Total Assets)	Ratio between cash and liquid assets to total assets. It measures a firm's liquidity and how easily it can service debt and short-term liabilities if the need arises.
RECEIVABLES_TURNOVER (Receivable Turnover Ratio) - ***NL***	Efficiency ratio that measures how efficiently a firm is using its assets. It measures the number of times over a given period (usually a year) that a firm collects its average accounts receivable.
PAYABLES_TURNOVER (Payable Turnover Ratio) - ***NL***	Efficiency and liquidity ratio that measures how many times a firm pays its creditors over an accounting period.
LOG_ASSETS (Natural Logarithm of Total Assets)	Measures the size of the firm.
SALES_GWT (Net Sales Growth) - ***NL***	Measures a firm's growth in a specific year. It also measures the stability of a firm's performance.
Variable	**Description**
DG_CR_TOT (Drawn amount to Granted Amount) - ***NL***	Financial flexibility ratio. It measures the percentage of available credit that the firm is

	actually using. It refers to all the different types of loans.
DG_CR_REV (Drawn Amount to Granted Amount of uncommitted short term loans) - ***NL***	Financial flexibility ratio. It measures the percentage of uncommitted short-term loans that the firm is actually using.
DG_CR_AUT (Drawn Amount/Granted Amount, short term loans) - ***NL***	Financial flexibility ratio. It measures the percentage of self-liquidating short-term loans that the firm is actually using.
DUMMY_SCONF (Overdrawns)	Dummy equal to 1 if the firm has an overdrawn amount greater than the granted amount, and 0 otherwise.
DEF_STORIA_CRED (Deteriorated loans)	Dummy equal to 1 if the firm has deteriorated loans, and 0 otherwise.
MORTGAGE (Mortgage)	Dummy variable equal to 1 if long-term loans are more than 90 per cent of total loans. It is used to mitigate the impact on PD estimation of a high drawn/granted ratio which is physiological for mortgages.
DUMMY_REV	Dummy equal to 1 if the firm has uncommitted short-term loans, and 0 otherwise.
DUMMY_AUT	Dummy equal to 1 if the firm has short-term loans, and 0 otherwise.
Variable	**Description**
AREA_CVD (geographical area)	Dummy variables identifying the geographical region where the firm operates (North-East, North-West, Center, South and Islands).
ATECO_CVD (economic sector)	Dummy variables identifying firms' economic sector.
DIM_CVD (size)	Dummy variables identifying firm size as defined by the European Commission.

RISK MANAGEMENT MAGAZINE

Anno 15, numero 3
Settembre – Dicembre 2020

In collaborazione con

prometeia

Poste Italiane - Spedizione in abbonamento postale – 70% aut. DCB / Genova nr. 569 anno 2005
TESTATA INDIPENDENTE CHE NON PERCEPISCE CONTRIBUTI PUBBLICI (legge 250/1990)

IN QUESTO NUMERO

Rivista scientifica riconosciuta da ANVUR e AIDEA

Risk Management Magazine
Anno 15 n° 3 Settembre – Dicembre 2020

Direttore Responsabile (Chief Managing Editor)
Maurizio Vallino

Condirettore (Deputy Managing Editor)
Corrado Meglio

Proprietà, Redazione e Segreteria:
Associazione Italiana Financial Industry Risk Managers (AIFIRM), Via Sile 18, 20139 Milano

Registrazione del Tribunale di Milano n° 629 del 10/9/2004

ISSN Print 2612-3665 – **ISSN Online** 2724-2153
DOI 10.47473/2016rrm

E-mail: risk.management.magazine@aifirm.it; Tel. +39 389 6946315

Stampa
Algraphy S.n.c. - Passo Ponte Carrega 62-62r 16141 Genova

SPEDIZIONE IN ABBONAMENTO POSTALE AI SOCI AIFIRM RESIDENTI IN ITALIA, IN REGOLA CON L'ISCRIZIONE

Rivista in stampa: 22 dicembre 2020