Dynamic Balance Sheet Simulation and Credit Default Prediction: A Stress Test Model for Colombian Firms[*]

Diego Fernando Cuesta-Mora[†] Camilo Gómez[‡]

The opinions contained in this document are the sole responsibility of the authors and do not commit Banco de la República nor its Board of Directors.

Abstract

This paper presents a stress test model developed by the Financial Stability Department of the Banco de la República to assess the financial vulnerability of Colombian non-financial firms. The model supports the Central Bank's biannual Financial Stability Report and informs policy decisions by identifying firms that are exposed to credit risk under adverse economic conditions. The proposed model integrates three components: a dynamic balance sheet simulation framework; a suite of machine learning models to estimate credit default probabilities; and a final module that identifies firms at risk of default. This tool strengthens the Central Bank's capacity to monitor and evaluate risks in the corporate sector with a forward-looking perspective. The paper details each component and illustrates the model's results using a stress scenario.

Keywords: Stress Testing, Credit Risk, Credit Default, Machine Learning.
JEL Classification Numbers: G3, G21, G01, G17

[*] Felipe Clavijo, Santiago Cortés, Santiago Segovia, Carlos Quicazán, Juan Sebastián Mariño, Andrés Murcia, and Hernando Vargas provided valuable comments and discussions during the early stages of this project. We also thank the participants of the Banco de la República's internal seminar held in May 2025 for their insightful suggestions.

[†] Financial Stability Department, Banco de la República Colombia. E-mail: dcuestmo@banrep.gov.co
[‡] Financial Stability Department, Banco de la República Colombia. E-mail: agomezmo@banrep.gov.co

Simulación dinámica de balances y predicción del incumplimiento crediticio: Un modelo de prueba de estrés para firmas colombianas[*]

Diego Fernando Cuesta-Mora[†] Camilo Gómez[‡]

Las opiniones contenidas en este documento son responsabilidad exclusiva de los autores y no comprometen al Banco de la República ni a su Junta Directiva.

Abstract

Este documento presenta un modelo de prueba de estrés desarrollado por el Departamento de Estabilidad Financiera del Banco de la República para evaluar la vulnerabilidad financiera de las firmas no financieras colombianas. El modelo apoya el Reporte de Estabilidad Financiera semestral del Banco de la República y aporta al diseño de políticas al identificar firmas expuestas al riesgo crediticio en condiciones macroeconómicas adversas. El modelo propuesto integra tres componentes: un marco dinámico de simulación de balances; un conjunto de modelos de *machine learning* para estimar probabilidades de incumplimiento crediticio; y un módulo final que identifica firmas en riesgo de incumplimiento crediticio. Esta herramienta fortalece la capacidad del Banco de la República para monitorear y evaluar riesgos en el sector empresarial de forma prospectiva. El documento detalla cada componente e ilustra los resultados mediante un escenario de estrés.

Palabras clave: Prueba de estrés, Riesgo crediticio, Incumplimiento crediticio, *Machine learning*.
Clasificación JEL: G3, G21, G01, G17

[*] Felipe Clavijo, Santiago Cortés, Santiago Segovia, Carlos Quicazán, Juan Sebastián Mariño, Andrés Murcia y Hernando Vargas brindaron valiosos comentarios y aportes durante las primeras etapas de este proyecto. Agradecemos también a los participantes del seminario interno del Banco de la República, realizado en mayo de 2025.

[†] Departamento de Estabilidad Financiera, Banco de la República Colombia. E-mail: dcuestmo@banrep.gov.co

[‡] Departamento de Estabilidad Financiera, Banco de la República Colombia. E-mail: agomezmo@banrep.gov.co

1 Introduction

This paper presents a stress test model for Colombian firms, the analytical tool used by the Financial Stability Department of the Banco de la República (Central Bank of Colombia) to perform its financial vulnerability analysis of the Colombian non-financial corporate sector. The results of the firm's stress test model are typically presented in the Central Bank's biannual Financial Stability Report and serve as a key input for the financial stability assessment of the corporate sector, one of the main debtors of the financial system.

Following the 2008 global financial crisis, stress test models have been increasingly used in financial institutions to prospectively assess whether they have enough buffers to absorb severe shocks to the economy and financial system (Dent et al. 2016). This has enabled policymakers to design plans to accommodate buffers to such unexpected and extreme shocks. In this way, stress test models are not forecasting tools since their main aim is to quantify the financial system's resilience to extreme and improbable scenarios under restrictive behavioral assumptions.

The unprecedented COVID-19 crisis, characterized by supply and demand shocks and high levels of uncertainty (Guerrieri et al. 2022, Kalemli-Ozcan et al. 2020), showed the importance of complementing stress tests in financial institutions with more analytical tools to assess the financial soundness of the non-financial private sector (corporate sector hereafter) on a forward-looking basis. This necessity was twofold. First, the corporate sector is critical from a financial stability standpoint, given the firm's role in the credit market (direct channel) and the whole economy (indirect channel). Second, the financial information of the corporate sector tends to be less frequent than banks' balance sheets.

The model proposed in this paper has three building blocks. First, following the corporate stress test literature developed during the pandemic, we develop a dynamic balance sheet simulation framework based on accounting behavioral rules and econometric analysis. This block simulates the main accounts of firms' balance sheets, conditional on the firms' initial characteristics and a macroeconomic scenario (GDP growth, financial conditions, and the level of interest rates). The second block comprises a suite of machine learning models (ML) that enable us to predict the credit default probability of firms based on key financial and activity indicators. Finally, the third block combines the first two to define firms at credit risk. This final block is an input for banks' stress test models, where banks' idiosyncratic credit risk can be assessed. In fact, this final block is an input for the stress test model used by the Central Bank of Colombia (see Gamba et al. 2017).

The dynamic balance sheet simulation model provides an accounting micro and macro-consistent tool to evaluate the firms' exposure to the risks identified in a macroeconomic scenario and the most recent exposure of the financial statements to these firms. Moreover, the model provides prospective information about the financial situation of firms, which is key given the low frequency of this information. However, as explored in Section 4, the results must be read with caution since the model can be acid because it does not consider strategic behavior of the firms, such as prepaying debt, reducing size, renegotiating debt, etc. In fact, the model is based on the historical correlations observed in the data on a set of restrictive behavioral assumptions.

Estimated ML models indicate that extreme gradient boosting techniques are superior to logistic and random forest in out-of-sample performance metrics tailored to reduce default misclassification. However, simpler logistic models remain competitive. When ML models are applied to stress test scenarios based on the dynamic balance sheet simulation model, the identified defaulting firms are those whose operations are most likely to be affected according to the dynamic balance sheet simulation.

The rest of this paper proceeds as follows. Section 2 presents a review of the literature. Section 3 presents the data that we use for the model. Section 4 presents the dynamic balance sheet simulation framework and illustrates its results based on the stressed macroeconomic scenario presented in Banco de la República's Financial Stability Report from the second half of 2024. Section 5 elaborates on the ML set of models used for credit default predictions and presents their performance. Section 6 integrates the results of the last two sections with an application to the stress test scenario presented in the 2024-II Financial Stability Report. In particular, it describes the financial features of firms that would be predicted in default in a stressed scenario.

2 Literature Review

According to Borio et al. (2014), stress testing originated in engineering to assess the stability of an object while facing adverse conditions. In finance, top-down stress test models, where national supervisors and central banks model banks' financial resilience on a system-wide basis, can be traced back to the late 1990s with the launch of the Financial Sector Assessment Program by the IMF and the World Bank (Dent et al. 2016).[1] However,

[1] Before top-down stress test models, stress test models were carried out mainly from a bottom-up basis, where individual banks usually simulated the effect of market volatility in their investment holdings and their solvency (Dent et al. 2016). In contrast, top-down models are conducted by authorities (central banks, financial supervisors) and make use of the same behavioral model for all financial institutions and a common

as mentioned in Section 1, it was only until the 2008 global financial crisis that financial stability authorities around the world increased the use of stress test models to assess financial institutions' liquidity and solvency soundness to design plans to accommodate buffers to unexpected shocks prospectively. Examples of stress test modeling frameworks in different jurisdictions can be found in Anand et al. (2014), Burrows et al. (2012), Cabrera et al. (2012), Gamba et al. (2017), and Farmer et al. (2020).

In general, these models consist of macroeconomic adverse scenario design, the impact of such scenario on financial risks, and the impact of these risks on banks. Normally, stress test models rely on restrictive assumptions about the behavior of banks, such as mechanical rules of thumbs (Borio et al. 2014). In this way, as mentioned in Section 1, stress test models are not forecasting tools since their main aim is to quantify the financial system's resilience to extreme and improbable scenarios under restrictive behavioral assumptions.

As mentioned, the unprecedented COVID-19 crisis, characterized by supply and demand shocks and high levels of uncertainty (Guerrieri et al. 2022, Kalemli-Ozcan et al. 2020), highlighted the need to complement stress tests on financial institutions with more analytical tools to assess the financial soundness of the corporate sector in a forward-looking manner. This necessity was twofold. First, the corporate sector is critical from a financial stability standpoint, given the firm's role in the credit market (direct channel) and the whole economy (indirect channel). Second, the financial information of the corporate sector tends to be less frequent than banks' balance sheets.

In this way, several studies proposed analytical stress testing frameworks for the corporate sector during the pandemic. To simulate the financial variables of firms, works by Carletti et al. (2020), Demmou et al. (2021), Caceres et al. (2020) and Tressel & Ding (2021) elaborate on accounting-consistent frameworks for the main balance sheet accounts and profit and loss (P&L) items. These frameworks can depend on exogenous income shocks –e.g., at the sectoral level– (Carletti et al. 2020, Demmou et al. 2021) or can be complemented with firm-level regressions that relate macroeconomic variables to key financial and activity performance variables such as sales growth or leverage (Caceres et al. 2020, Tressel & Ding 2021). To build the financial simulation analysis of the corporate stress testing tool proposed in this paper, we closely follow the most comprehensive approach of Tressel & Ding (2021). However, we also consider sector heterogeneity components in regressions similar to Caceres et al. (2020).

This paper is also related to the default corporate finance literature. Since the seminal

macroeconomic scenario.

paper by Altman (1968), a prominent literature on firm failure has been developed. See, e.g., Bottazzi et al. (2011), Traczynski (2017), Cathcart et al. (2020), and Modina et al. (2023). In this branch of the literature, discriminant analysis and Logistic or Probit regressions were initially emphasized (for a review, see Ciampi & Gordini 2013 and Siggelkow & Fernandez 2024). Recently, studies have implemented ML methods to study corporate default. Altman et al. (1994) compare neural networks to linear discriminant analysis and find no high gains from the first approach. Ciampi & Gordini (2013) implement artificial neural networks to predict credit risk for small Italian companies and find that these models outperform traditional models. Xia et al. (2017) compare extreme gradient boosting (XGBoost) to other ML and traditional models. The authors find that the XGBoost-based sequential ensemble model with Bayesian hyper-parameter optimization improves credit scoring accuracy. Interestingly, the authors also find that simpler, traditional Logistic models remain competitive. Siggelkow & Fernandez (2024) show that random forests (RF) models enhance credit-risk evaluation for SMEs by improving prediction accuracy.

The models for firms' credit default we propose in the section 5 of the corporate stress test are inspired by these preview works. Given that some previous research has shown that traditional models as Logistic regressions remain competitive, we compare these simpler models with a set of data-driven ML approaches that use a wide range of covariates and specifications. In particular, we compare two families of models: i) Logit models and ii) tree-based methods such as RF and XGBoost. In total, we run seven models (five Logit and two tree-based models) and compare their performance metrics.

In summary, our paper builds on the recent literature on corporate stress tests developed during the pandemic. Moreover, it complements these models and their applications with ML models to predict credit default.

3 Data

We use data from three data sources to construct the dynamic balance sheet simulation framework and the battery of ML models. First, we employ annual financial information of firms that report their financial statements to the Colombian Superintendence of Companies and the Financial Superintendence. This dataset contains information from 1999 to 2023, comprising 517,850 observations and 66,166 firms. This set of information serves as the primary input for the dynamic balance sheet simulation model presented in Section 4. From these data, we utilize the main balance and P&L accounts (see Table 1) to construct various financial and activity performance indicators that measure dimensions such as profitability, leverage, debt burden, size, and size growth.

Table 1: Firm financial variables

Variables	Notes
Sector	
Total Assets	
Cash and equivalents	
Short-term financial liabilities	
Long-term financial liabilities	
Total liabilities	
Equity	
Operating income or sales	
Operating expense or costs	
Other operating income	
Other operating profits or losses	
Profit from operating activities	
Financial income	
Financial costs	Interest expenses before 2015
Profit before taxes	
Taxes	
Total profits	
Trade and other current receivables	
Other current financial assets	Short-term investments in cash definition (equation 5).
Current provisions for employee	Short-Term accrued payrolls in cash definition (equation 5)
Trade and other current payables	
Other current non-financial liabilities	
Other non-current financial assets	
Issued capital	

Notes: Balance and P&L accounts taken from annual financial information of firms.
Source: Authors' elaboration based on data from the Colombian Superintendence of Companies and Financial Superintendence.

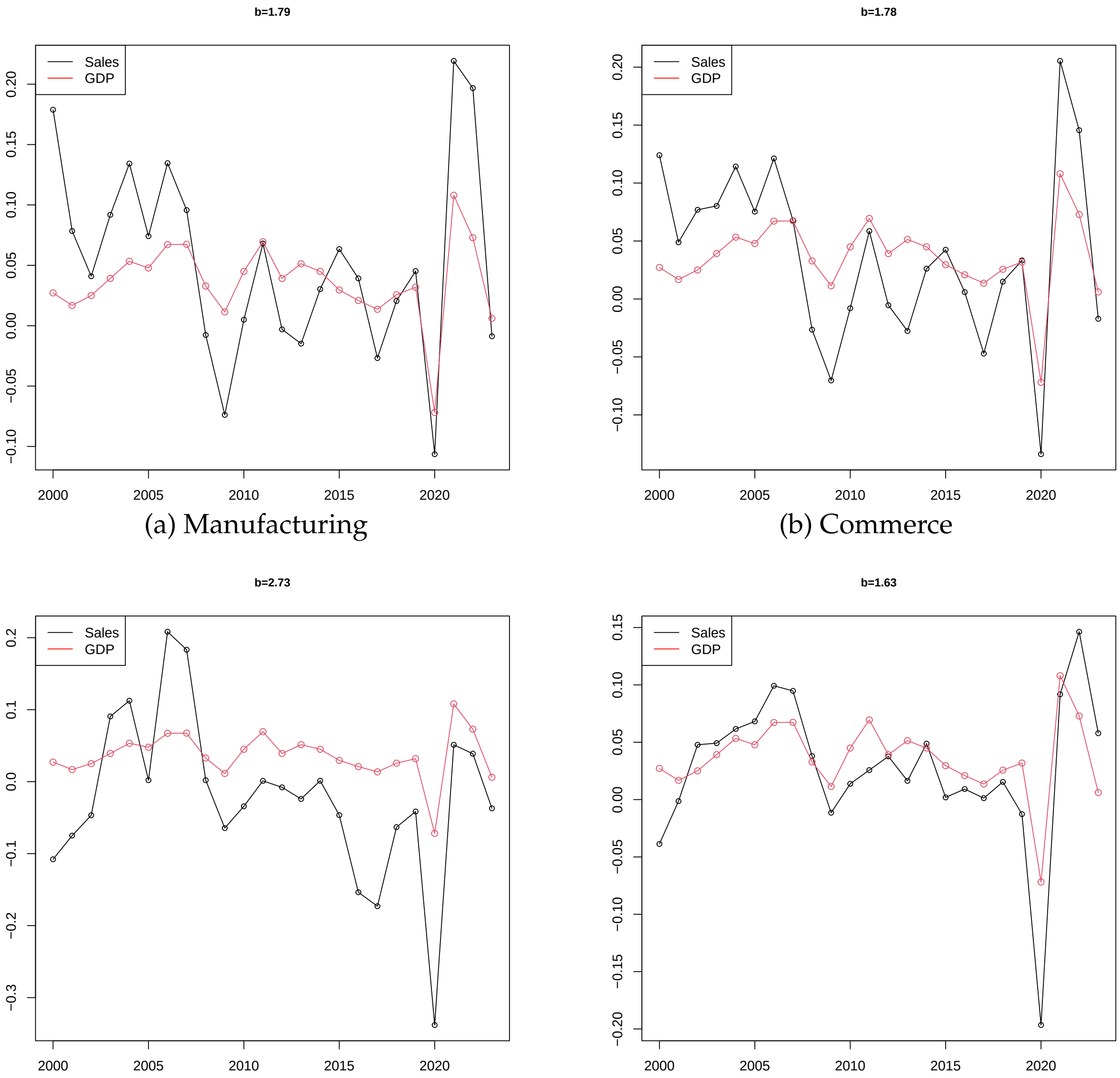
Second, we merge the firm's annual data with the Colombian credit registry, which is reported by credit institutions to the Colombian Financial Superintendence. This dataset provides information about firms' delinquency days in a given year from 2005, allowing us to define credit default status as being one or more months past due. This merged dataset is the main input for the set of ML models developed in Section 5. Finally, we utilize macroeconomic information provided by the National Statistics Office and an aggregate boom credit measure calculated by the Financial Stability Department of the Central Bank of Colombia.

Figure 1 presents, for selected sectors, time series of average sales growth for firms with available annual financial information and aggregate GDP growth. It also displays the results of linear regressions of sectoral average growth of sales on aggregate GDP (b values in each panel).[2] According to this figure, there is a heterogeneous relationship between GDP growth and average sales by sector. On the other hand, Figure 2 presents the yearly average of firms' financial leverage and the aggregate credit boom indicator. This figure displays a positive relationship between financial leverage and the credit boom indicator. These data movements motivate the regression set-up used for dynamic balance sheet simulation developed in Section 4, where sectoral heterogeneities are taken into account, and the relationship between micro leverage and macro credit dynamics is modeled.

The merged dataset used for the set of ML models to predict default one period ahead includes lagged financial information from various indicators. Consequently, the dataset employed in the models begins in 2006 to account for the one-period lag. Figure 3 displays the annual percentage of firms in default since 2006 for the dataset employed to develop ML models. Interestingly, defaulting firms increased in 2008 and 2020, periods characterized by economic distress. Moreover, for the whole period, the proportion of firms in default relative to the total sample is 13.4%, indicating a highly imbalanced dataset. Section 5 discusses the implications of this imbalance and the strategies implemented to address it.

[2] Formally, for each sector s, the time series of average log change in sales is regressed on aggregate GDP growth following the formula $\overline{\Delta \ln \text{Sales}}_{st} = a_s + b_s \cdot \text{GDP}_t + v_{st}$, where $\overline{\Delta \ln \text{Sales}}_{st}$ refers to average log change in sales in sector s and year t, and GDP_t to aggregate GDP growth, and v_{st} to the error term.

Figure 1: Average sales growth for selected sectors and aggregate GDP growth



(a) Manufacturing

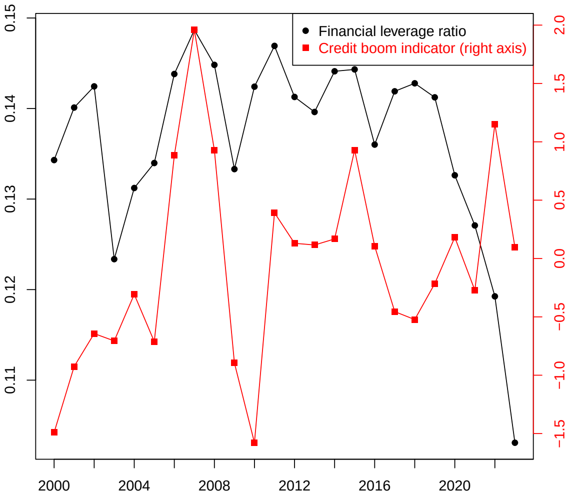
(b) Commerce

(c) Construction

(d) Real estate

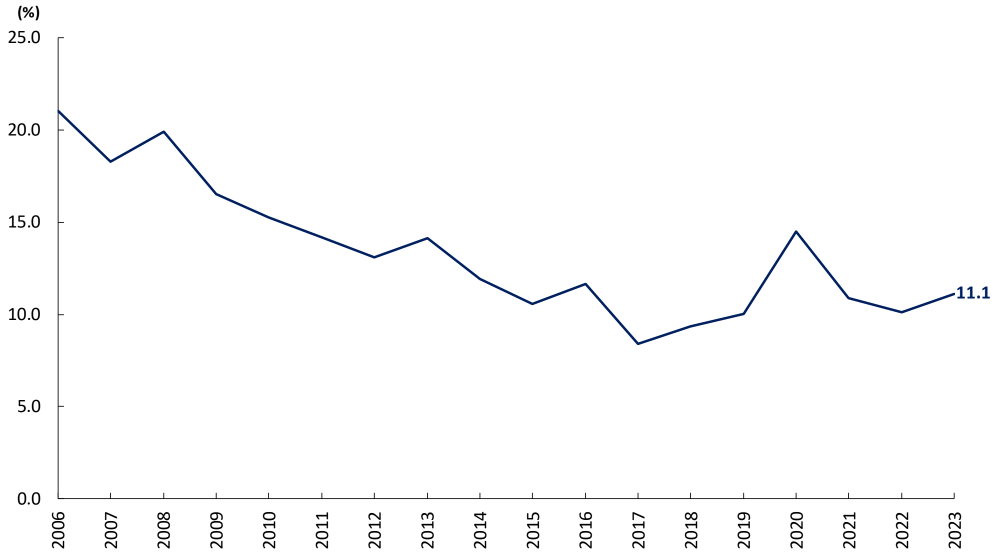
Notes: Average sales growth for selected sectors and aggregate GDP growth. b is the estimated coefficient, for each sector s, of regression $\overline{\Delta \ln \text{Sales}}_{st} = a_s + b_s \cdot \text{GDP}_t + v_{st}$, where $\overline{\Delta \ln \text{Sales}}_{st}$ refers to average log change in sales in sector s and year t, and GDP_t to aggregate GDP growth, and v_{st} to the error term. Source: Authors' elaboration based on data from the Colombian Superintendence of Companies and Financial Superintendence.

Figure 2: Average financial leverage and aggregate credit boom indicator



Notes: Yearly average of firms' financial leverage and aggregate credit boom indicator. Source: Authors' elaboration based on data from the Colombian Superintendence of Companies and Financial Superintendence.

Figure 3: Percentage of firms in default by year



Notes: Yearly number of firms in default (past-due days higher than 30) as a percentage of firms. Source: Authors' elaboration based on data from the Colombian Superintendence of Companies and Financial Superintendence.

4 Dynamic balance sheet simulation framework

Based on the data described in section 3, this part presents the balance sheet simulation framework. First, we describe the accounting rules used to simulate the main financial accounts and P&L items of firms. Secondly, we outline the regression analysis aimed at connecting macroeconomic aggregates to key financial indicators. This regression analysis is an input for the firm accounting model. Finally, we illustrate dynamic balance sheet simulation results based on the stressed macroeconomic scenario presented in Banco de la República's Financial Stability Report from the second half of 2024. The framework described in this section closely follows the work of Tressel & Ding (2021).[3]

4.1 Accounting consistent behavioral rules

To produce the main firms' financial variables in a consistent way, we use the following accounting behavioral rules proposed by Tressel & Ding (2021). These rules are based on a macroeconomic scenario and simulated sales and costs, which are constructed with the regression analysis explained in the next section.

Define $h = 0, ..., H$ as the balance sheet simulation periods, where $h = 0$ refers to the initial point and H to the final point. To simulate total profits, financial costs are modeled with the following equation:

$$\text{FinancialCosts}_{ih} = \text{FinancialDebt}_{ih-1} \cdot [i_{i0}^{eff} + a \cdot \Delta\text{MPR}_h], \tag{1}$$

where a is the share of variable interest rate commercial loans (close to 80% for the case of the Colombian banking sector), MPR_h refers to the monetary policy rate in the macroeconomic scenario, and i_{i0}^{eff} to the initial effective rate of financial debt, measured as financial costs over financial liabilities.

With equation (1) and given values of sales and costs, profits before taxes can be computed[4] and used in tax calculation according to the following equation:

$$\text{Taxes}_{ih} = \text{ProfitsBeforeTaxes}_{ih} \cdot \tau \cdot 1[\text{ProfitsBeforeTaxes}_{ih} > 0], \tag{2}$$

where $1[\cdot]$ is the indicator function, and τ the estatutory income tax rate for corporates.

With the above value of taxes, profits and –assuming that dividends are not distributed–

[3] Similar frameworks are developed by Caceres et al. (2020), Carletti et al. (2020), Demmou et al. (2021). We opt for the framework proposed by Tressel & Ding (2021) since it connects a regression analysis with a high-level, detailed accounting simulation.

[4] Interest income is assumed to be constant.

equity can be computed as follows:

$$\text{Profits}_{ih} = \text{ProfitsBeforeTaxes}_{ih} - \text{Taxes}_{ih}, \tag{3}$$

$$\text{Equity}_{ih} = \text{Equity}_{ih-1} + \text{Profits}_{ih}. \tag{4}$$

Following Tressel & Ding (2021), we define the initial net cash of a firm as:

$$\begin{aligned}\text{Cash}_{i0} = \text{Cash\&Equivalents}_{i0} + \text{Short-TermInvestments}_{i0} + \text{AccountReceivables}_{i0} \\ -(\text{Short-TermAccruedPayrolls}_{i0} + \text{AccountPayables}_{i0} + \\ \text{OtherShort-TermNon-FinancialLiabilites}_{i0}).\end{aligned} \tag{5}$$

With this definition, profits are accumulated in net cash. Moreover, if cash needs arise, financial debt increases. The above is summarized in the following expressions:

$$\text{Cash}_{ih} = \text{Cash}_{ih-1} + \text{Profits}_{ih}, \tag{6}$$

$$\text{FinancialDebt}_{ih} = \text{FinancialDebt}_{ih-1} - \text{Cash}_{ih} \cdot 1[\text{Cash}_{ih} < 0]. \tag{7}$$

Finally, liabilities and assets are given by:

$$\text{Liabilities}_{ih} = \text{Liabilities}_{ih-1} + \Delta\text{FinancialDebt}_{ih}, \tag{8}$$

$$\text{Assets}_{ih} = \text{Liabilities}_{ih} + \text{Equity}_{ih}. \tag{9}$$

4.2 Regression analysis

The regression analysis aims to capture, at the firm level, the statistical relationship between key macroeconomic and financial aggregates and firms' financial indicators, conditional on a rich set of initial firm characteristics. Moreover, the proposed models capture a differential effect of the macroeconomic variables on firm financial indicators. In this way, the regression analysis is an input for the previously described accounting rules. Specifically, this regression analysis conditions the following variables to a macroeconomic scenario: sales and cost growth and financial leverage. The above affects profits, equity, and financial debt in the accounting model (see equations 3, 4, and 7).

More formally, following Tressel & Ding (2021), we run the following firm-level dynamic OLS regressions

$$Y_{it} = \alpha \cdot Y_{it-1} + \delta \cdot \text{CharFirm}_{it-1} + \Psi_s^{\text{Macro}} \cdot \text{Macro}_t + \Psi^{\text{Fin}} \cdot \text{Fin}_t + d_s + v_{it}, \tag{10}$$

where i and t index the firm and year, Y_{it} refers to log change in sales or financial leverage, CharFirm_{it-1} to a vector of firms' characteristics, Macro_t to a set of macroeconomic variables, Fin_t to the credit boom indicator, d_s to a sector fixed effect, and v_{it} to the error term.

The initial firms' characteristics (CharFirm_{it-1}) vector includes a measure of profitability (ROA), indebtedness (financial leverage ratio), size (log of assets), turnover (sales-to-assets ratio), and growth (log change in sales).[5] Moreover, the model controls for a dynamic factor given by the lagged dependent variable (Y_{it-1}) to consider persistence or mean-reversion if the outcome variable is a ratio or a growth rate, respectively. Finally, we allow the coefficients of the macroeconomic variables Ψ_s^{Macro} to vary across economic sectors to capture heterogeneities in the relationship between sectors and the aggregate economic cycle.

The dependent variables considered in the model are the log change in sales and the financial leverage ratio.[6] For the case of the macroeconomic and financial variables, GDP growth and a credit boom indicator were selected after a careful regression analysis of different candidates. In the end, the models used are i) the log change in sales as a function of the credit boom and as a sectorally varying function of GDP growth, and ii) financial leverage as a function of the credit boom. The selection of variables and models are in line with those proposed by Tressel & Ding (2021)[7]. However, in contrast to the models used by Tressel & Ding (2021), we allow the effect of GDP on sales to vary depending on the economic sector.

Finally, the following regression is used to measure the elasticity of costs with respect to sales at the firm-sector level:

$$\Delta \ln \text{Costs}_{it} = \alpha \cdot \Delta \ln \text{Costs}_{it-1} + \delta \cdot \text{CharFirm}_{it-1} + \Psi_s \cdot \Delta \ln \text{Sales}_{it} + v_{it}. \tag{11}$$

[5] In the regression analysis, ratios are measured with lagged denominators, e.g., $\text{ROA}_{it} = \frac{\text{Profits}_{it}}{\text{Assets}_{it-1}}$.

[6] Notice from equation (7) that financial debt also has a simulated value based on accounting behavioral rules. Therefore, in simulations, the average of projected values from regression and equation (7) are used.

[7] Similar to Tressel & Ding (2021), although dynamic OLS estimation may be subject to bias, the exclusion of firm-level fixed effects from the specification (while including sector fixed effects) mitigates the potential bias arising from the dynamic panel correlation between the lagged dependent variable and the fixed effect.

In sum, the dynamic balance sheet simulation model offers a tool to evaluate the firms' exposure to the risks identified in a macroeconomic scenario and the most recent exposure of the financial statements to these firms. However, since the model is based on some restrictive accounting behavioral rules and correlations observed in the data, its results must be carefully read. In particular, the model does not account for strategic behaviors such as prepaying debt, reducing debt size, or renegotiating debt, etc. When we present the results, we discuss, based on an out-of-sample comparison exercise presented in Appendix B, how the mentioned assumptions can affect results.

4.3 Results

We now present the results of the framework described above for dynamic balance sheet simulation. This simulation is based on the accounting-consistent behavioral rules and the results from the regression analysis presented earlier. To illustrate the balance sheet simulation of firms, the results are based on the stressed macroeconomic scenario presented in Chapter 3 of the Financial Stability Report published in the second half of 2024 by Banco de la República. The macroeconomic scenario presented in this report examines a hypothetical adverse macroeconomic scenario characterized by high-risk perception and fiscal uncertainty. In this scenario, the Colombian peso would depreciate against the U.S. dollar, leading to inflationary pressures and unanchored inflation expectations. As a result, the monetary policy interest rate would rise, increasing the cost of debt, and investment, consumption, and GDP would contract, negatively impacting credit demand and employment. The 2023 observed financial data of firms serves as the starting point of the exercise. From this point on, dynamic balance sheet simulation is conducted in a two-year horizon.

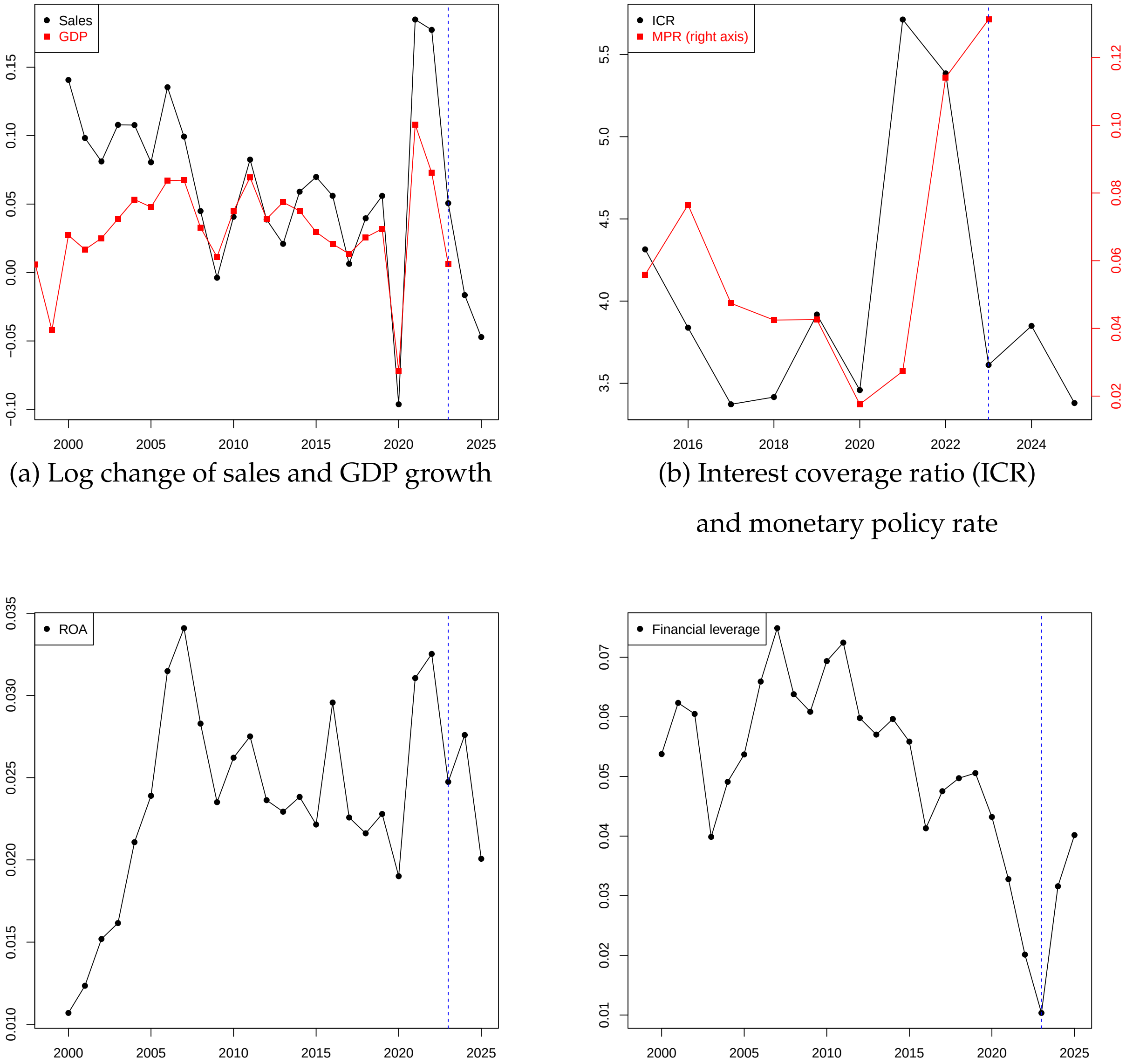
Figure 4 displays the median simulated values of key financial variables under the described stressed macroeconomic scenario. Specifically, this figure presents the observed median values of firms up to 2023 (vertical blue dashed lines) and the simulated values two years ahead, based on the methodology previously presented. Panels a, b, c, and d present, respectively, the median log change of sales, interest-coverage-ratio (ICR) —defined as the ratio between operating profits and financial costs—, ROA, and financial leverage. Some plots include, for reference, observed key macroeconomic variables in red. According to the results, and consistent with the macroeconomic scenario, firms would experience financial pressures during the stress horizon. In particular, the median firm would exhibit sales contractions (Figure 4, panel a). In line with this sales drop and the upward pressures on the monetary policy interest rate in the stressed macroeconomic scenario, the ICR would decrease in the stress horizon (Figure 4, panel b). Moreover, ROA and financial leverage would decrease and increase, respectively (Figure 4, panels c and d).

Appendix A shows the results of the regression analysis on which the dynamic simulations are based. In general, regressions point to mean reversion and persistence effects for growth variables and the financial leverage variable, respectively (lagged dependent variable). When it comes to estimates of macroeconomic and financial variables, we find that the credit boom indicator has a positive effect on sales growth and financial debt. Regressions also capture sectoral heterogeneous cost-to-sales elasticities and correlations between average sales growth and aggregate GDP growth.

As mentioned earlier, the dynamic balance sheet simulation model is designed as a macro-micro consistent stress test tool to assess the financial resilience of corporations in extreme and improbable scenarios, as well as under restrictive behavioral assumptions. To assess the extent to which these assumptions affect the conclusions drawn from the model, Appendix B presents an out-of-sample comparison exercise. In particular, dynamic balance sheet simulations starting from 2021 through periods 2022 and 2023 are compared with observed data. According to the results, the model tends to accurately estimate the distribution of operational profits and ROA. However, the model seems to overestimate financial leverage ratios, as well as financial obligations and costs. As a result, the proportion of firms with an ICR lower than one is also likely to be overestimated.[8]

[8] In the context of this study, the conservative bias observed in the out-of-sample results aligns with the Basel IRB credit risk modeling guidelines. Moreover, it is consistent with best practices for evaluating the impacts of stress scenario.

Figure 4: Median simulated values of key financial variables under the stressed macroeconomic scenario



(a) Log change of sales and GDP growth

(b) Interest coverage ratio (ICR) and monetary policy rate

(c) Return on assets (ROA)

(d) Financial leverage

Notes: Balance sheet simulation results of median key firms' financial variables based on the stressed macroeconomic scenario presented in Chapter 3 of the Banco's de la República Financial Stability Report of the second half of 2024, and the dynamic balance sheet simulation framework presented in accounting rules (1)-(9) and regressions (10)-(11). Some plots include, for reference, observed key macroeconomic variables. Source: Authors' elaboration based on data from the Colombian Superintendence of Companies and Financial Superintendence.

5 Machine learning models for credit default prediction

In this section, we present the set of ML models used to predict corporate credit default. We begin by defining the classification problem, outlining the performance metrics considered, and detailing the methods employed to optimize them in the context of highly imbalanced data. Subsequently, we introduce the different sets of models used and compare their performance based on the reference metric. Our primary focus is on optimizing model performance rather than emphasizing economic interpretability. However, in Section 6, we provide insights into the characteristics of firms classified as defaulting based on their key financial variables.

5.1 Classification problem and model tuning

A classification problem aims to assign qualitative responses to an individual observation based on its observable characteristics. In our case study, the objective is to determine whether a firm is in credit default ($Y = 1$) or not ($Y = 0$), based on its financial variables.

The variable of interest of firm i in year t is defined as:

$$Y_{it} = \text{Default}_{it},$$

where Default_{it} is equal to one if firm f is in default (30 or more past-due days) in any of its credits during the year t (0 otherwise).

The goal of the model is to find a function $f(\cdot)$ that predicts the default variable given a set of observable characteristics. In particular, our models are based on a set of lagged firm-level financial indicators (X_{it-1}), the lagged default variable (to capture default persistence), and a set of firm sectoral dummies ($\mathbf{S}_{it}$):

$$\hat{Y}_{it} = f(X_{it-1}, Y_{it-1}, \mathbf{S}_{it}).$$

All variables included in the models, except those related to the economic sector, are lagged by one period. Using lagged variables in the estimation ensures that the models can predict default one period ahead. This is also the practice commonly followed in the corporate finance literature (see the works cited in Section 2). The specific set of financial indicators included in X_{it-1} are presented in Table 3 below.

The classification process embedded in f involves two interconnected steps. First, an estimation of the conditional probability must be performed. The probability of default is calculated conditional on the observed characteristics previously described:

$$P(Y_{it} = 1 \mid X_{it-1}, Y_{it-1}, \mathbf{S}_{it}).$$

Second, a classification threshold T must be selected to assign to each observation a predicted default category. Based on the estimated probability, a label is assigned to the observation:

$$\hat{Y}_{it} = \begin{cases} 1 & \text{if } P(Y_{it} = 1 \mid X_{it}, Y_{it-1}, \mathbf{S_{it}}) \geq T \\ 0 & \text{if } P(Y_{it} = 1 \mid X_{it}, Y_{it-1}, \mathbf{S_{it}}) < T \end{cases}.$$

This approach systematically classifies firms into default and non-default states, providing a systematic assessment of credit risk.

Confusion matrix and performance metrics

To evaluate the performance of a classification model like the one described above, the confusion matrix is used. This tool compares the model's predictions with the actual observed values in a dataset. The columns of this matrix represent the predicted values: positives ($\hat{Y}_{it} = 1$) and negatives ($\hat{Y}_{it} = 0$), while the rows correspond to the actual observed data (Table 2). The usefulness of the confusion matrix lies in its ability to summarize the model's correct and incorrect classifications concisely. Specifically, it provides insights into how well the model predicts positive values (true positives, TP) and negative values (true negatives, TN). Similarly, the confusion matrix highlights the extent to which the model misclassifies observations, indicating the number of false positives (FP) and false negatives (FN). FP and FN are also referred to as Type I and Type II errors, respectively.

Table 2: Confusion matrix

		Forecast	
		Negatives ($\hat{Y}_{it} = 0$)	Positives ($\hat{Y}_{it} = 1$)
Observed	Negatives ($Y_{it} = 0$)	True negatives (TN)	False positives (FP)
	Positives ($Y_{it} = 1$)	False negatives (FN)	True positives (TP)

Based on the confusion matrix, the following performance metrics are usually used in classification models:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}}, \tag{12}$$

$$\text{Precision} = \frac{\text{Correct Positive Predictions}}{\text{Total Positive Predictions}} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{13}$$

$$\text{Recall} = \frac{\text{Correct Positive Predictions}}{\text{Total Positives}} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{14}$$

The most well-known metric for evaluating a classification model is accuracy (equation 12), which measures the proportion of correctly classified observations. However, in contexts where the focus is on evaluating the model's performance in a specific and minority class, accuracy may not be a good measure of model performance. This is particularly relevant in our case of credit default, as we face a highly imbalanced class problem. In fact, as shown in Section 3, only 13.4% of the observations in the whole sample correspond to firms in credit default.

In our study, we aim for a model that reduces the error of misclassifying a high-risk firm as a non-defaulting firm. Therefore, we are more interested in precision (equation 13) and recall (equation 14) metrics. Precision measures the proportion of correctly classified defaulting firms over the total positive predictions, that is, the percentage of correctly predicted defaulters. Recall, on the other hand, measures the proportion of true positives over the total number of defaulting firms, meaning the percentage of correctly identified defaulters within the total defaulting population. Precision and recall are class-specific metrics, as they focus on quantifying model performance for a particular class that may be of greater interest to researchers. It is important to note that maximizing precision is equivalent to minimizing false positives, while maximizing recall is equivalent to minimizing false negatives.

Based on the above, in each proposed model, our primary objective is to maximize the F_2-score, which combines precision and recall, giving greater weight to recall. In other words, we place more importance on identifying truly defaulting firms while minimizing false negatives. Formally, the F_2-score is defined as:

$$F_2 = (1 + 2^2)\frac{\text{Precision} \cdot \text{Recall}}{2^2 \cdot \text{Precision} + \text{Recall}} = \frac{(1 + 2^2) \cdot \text{TP}}{(1 + 2^2)\text{TP} + 2^2 \cdot \text{FN} + \text{FP}}. \tag{15}$$

Strategies to address class imbalance

As mentioned above, we are facing a classification problem with highly imbalanced classes (13.4% of the observations correspond to firms in default). Class imbalance occurs when the relative frequency of one class in the sample is significantly lower compared to the remaining categories. This characteristic can reduce the effectiveness of classification models, especially when the focus is on the minority class. The problem arises because the model's predictions may exhibit good overall performance metrics while being biased toward the majority class, lacking the ability to correctly identify the minority class of interest, such as the default class.

To address this issue, several complementary solutions can be employed by researchers (Kuhn et al. 2013). In this research, we implement three main strategies. First, we follow sampling methods to train models with a *synthetic* balanced sample. When dealing with an imbalanced dataset, up-sampling and down-sampling techniques can be applied. The first technique involves simulating or imputing additional observations in the sample to improve balance between classes, whereas the second reduces the number of observations to achieve such balance. In this paper, we utilize the Synthetic Minority Over-Sampling Technique (SMOTE) to oversample the minority class, ensuring the training sample is completely balanced. The SMOTE algorithm randomly selects an observation from the minority class and, using the k-nearest neighbors (five in our application), creates a synthetic point by randomly combining features between the selected point and its neighbors. When categorical features are present, the technique imputes the most common category among the neighbors.[9] It is important to highlight that SMOTE should only be applied to the training sample, as the evaluation and test subsets must maintain the real data distribution (see below the data partitioning strategy). Otherwise, model performance evaluation would be overly optimistic and could reduce the model's ability to classify new data.

As a second strategy to address class imbalance, we make specific choices in the model's hyperparameter tuning strategy. In imbalanced contexts, the traditional accuracy metric may fail to predict the minority class. Instead, an alternative approach to improve model performance is to optimize different metrics, such as precision, recall, and the F_2-scores, which assign greater weight to the minority class. As mentioned before, we focus on the F_2-score as the performance metric over which hyperparameters are estimated through a five-fold cross-validation procedure.

[9] If the categorical feature of the k neighbors is bimodal, one mode is randomly selected.

As a final strategy, we implement classification threshold tuning. As previously explained, classifying an observation into a given class depends not only on the estimated conditional probability but also on the threshold at which classification into one category or another occurs. By default, the threshold is set at 0.5; however, it can be adjusted to enhance the model's performance for the minority class and, consequently, improve the F_2-score. Under this approach, it is essential that tuning is performed using a dataset independent of the training and test sets. Using training predictions could introduce an optimistic bias in the model's performance, whereas using the test sample would prevent an impartial evaluation of model performance and limit the comparison between different models. For this reason, an independent sample is used for the cutoff calibration. The data partitioning strategy is explained below.

Data partitioning strategy

To address the classification problem using different models aimed at maximizing the F_2-score under the strategies to address the class imbalance previously described, it is essential to randomly divide the sample into three mutually exclusive subsets while maintaining approximately the same proportion of defaulting firms. This partitioning enables model evaluation at different stages, ensuring its generalization ability and reducing the risk of overfitting the training data. The original sample is divided as follows.

- **Training set (75% of the data)**: This subset is used to estimate the model parameters after applying the SMOTE technique. With this set, optimal hyperparameters through a 5-fold cross-validation procedure are found with the goal of maximizing the F_2-score metric.

- **Evaluation set (10% of the data)**: This subset is used to determine the optimal decision threshold that maximizes the F_2-score after model training with the training set.

- **Test set (15% of the data)**: This subset is used to compare the performance of different models based on the selected performance metric using data that was not used during training or evaluation. This ensures a fair comparison across models based on the F_2-score.

5.2 Set of machine learning models

Based on the practices for addressing class imbalance and the data partitioning strategy, this section provides a brief description of the variables and the family of models used. All models were estimated using five-fold cross-validation on the training sample, which was adjusted using the SMOTE methodology. Once the hyperparameters maximizing the metric of interest were identified, the optimal threshold was determined using the evaluation sample. Finally, the performance of each model was compared by applying it to the test sample.

Table 3: Firm-level variables used to predict default

Type of variables	Variables
Default	Dummy variable that takes the value of 1 if the firm is in default and 0 otherwise.
Economic sector	Dummy variable that takes the value of 1 for the economic sector to which the firm belongs. The 14 economic sectors considered are: 1) professional activities, 2) agriculture, 3) commerce, 4) construction, 5) electricity, 6) financial, 7) hospitality, 8) information and communications, 9) real estate, 10) manufacturing, 11) mining, 12) restaurants, 13) transportation, 14) others.
Operational profitability	Ratio of operating expenses to operating income Ratio of operating income to assets Operational ROA Operating margin Operational ROE
Profitability	ROA Net margin ROE
Leverage	Ratio of assets to equity Ratio of financial obligations to assets
Financial burden	Interest coverage ratio Ratio of financial obligations to operating income
Size	Log of assets Log of sales
Growth	Annual growth of operating expenses Annual growth of operating income Annual growth of financial obligations Annual growth of equity Annual growth of assets

Notes: Default and financial indicators are used with one lag. In ML models, ratios are measured with numerator and denominator in the same period, e.g., $\text{ROA}_{it} = \frac{\text{Profits}_{it}}{\text{Assets}_{it}}$. Growth variables were calculated using the arcsinh transformation, which allows handling values close to or equal to zero without definitional issues.
Source: Authors' elaboration based on data from the Colombian Superintendence of Companies and Financial Superintendence.

As explained above, a set of lagged financial indicators summarizing each firm's economic performance is used to predict default. In addition, the set of predictors includes lagged default and sectoral dummies. Table 3 presents the set of predictors used, grouped

Table 4: Description of ML models used to predict default

Model	Features used to predict default	Hyperparameters grid	Selected hyperparameters
Logit 1	Lagged default Sector dummies Lagged financial variables from Table 3		
Logit 2	Variables from Logit 1 Squared terms of lagged financial variables from Table 3		
Logit 3	Variables from Logit 2 Interactions between lagged financial variables from Table 3		
Logit Lasso	Variables from Logit 3	$\lambda \in \{0, 0.00001, 0.00002, \ldots, 0.0001\} \cup$ $\{0.0001, 0.0003, 0.0005, \ldots, 0.001\} \cup$ $\{0.001, 0.002, 0.003, \ldots, 1\}$	0.00001
Logit Ridge	Variables from Logit 3	$\lambda \in \{0, 0.00001, 0.00002, \ldots, 0.0001\} \cup$ $\{0.0001, 0.0003, 0.0005, \ldots, 0.001\} \cup$ $\{0.001, 0.002, 0.003, \ldots, 1\} \cup$ $\{1, 1.5, 2, \ldots, 100\}$	0.684
RF	Lagged default Sector dummies Lagged financial variables from Table 3	min.node.size $\in \{5, 10, 15\}$	min.node.size = 10

XGBoost	Lagged default
	Sector dummies
	Lagged financial variables from Table 3
		max_depth $\in \{2, 4, 6\}$
		eta $\in \{0.1, 0.15\}$
		gamma $\in \{0, 0.025, 0.005, 0.0075, 0.01, 0.015\}$
		min_child_weight $\in \{5, 10\}$
			max_depth = 4
			eta = 0.1
			gamma = 0.01
			min_child_weight = 10

Source: Authors' elaboration based on data from the Colombian Superintendence of Companies and Financial Superintendence.

by the type of information they provide. In total, a set of 35 features is considered (14 sectorial dummies, 19 lagged financial indicators, and the lagged default status). Based on these variables, two families of models are estimated. These models are summarized in Table 4, and we will describe them in what follows. Specifically, logistic regression models are presented first, followed by classification tree-based methods. In total, we run seven models (five logit and two tree-based methods) and compare their performance metrics.

Logit models

Logistic regression is widely used due to its simplicity, interpretability, computational efficiency and robust performance, which can sometimes rival more complex models (Xia et al. 2017).[10] For these reasons, this study adopts that approach and estimates five logit-type models (Logit 1 to Logit Ridge in Table 4). Logit 1 to Logit 3 models are simple logistic regressions where variables are included sequentially, resulting in more complex and flexible models at each step. In particular, Logit 1 includes lagged default, economic sector dummies and lagged financial variables from Table 3. Logit 2 extends this specification by incorporating the squared terms of lagged financial variables. Logit 3 adds to Logit 2 the interactions between lagged financial variables.

To mitigate overfitting in the presence of a large number of explanatory variables in the previous models, Logit Lasso and Logit Ridge are employed based on the variables selected for the most complex model from previous steps, i.e., Logit 3.[11] These techniques constrain excessive model fitting to the training data and enhance generalization to new data, effectively balancing the bias-variance trade-off. The Lasso penalization reduces the magnitude of the coefficients and, depending on the degree of penalization, can shrink some of them to zero, effectively performing variable selection.[12] On the other hand, Ridge penalization also reduces the magnitude of the coefficients. Still, it does not shrink them exactly to zero, retaining all variables in the model.[13] Since regularization is sensitive to the scale of variables, all explanatory variables were standardized prior to analysis. Based on this regularization, two additional models are defined.

For Lasso and Ridge models, the optimal degree of penalization is determined using five-fold cross-validation, maximizing the F_2-score as the performance metric. Table 4 presents the grid search performed in the training sample over the penalization parameters and the optimal values obtained in this process.

Tree-based classification models

Classification trees are methods in which the predictor space is segmented into non-overlapping regions through a recursive binary splitting algorithm. That is, at each step, one split is performed based on the most relevant variable in that step without considering what the best tree would be. Specifically, starting from the previous divisions, additional

[10] In the logistic model, the inverse of the log-likelihood function $J(\beta)$ is minimized using the logistic function $h_\beta(z_1, ..., z_k) = \frac{e^{\beta_1 z_1 + \ldots + \beta_k z_k}}{1+e^{\beta_1 z_1 + \ldots + \beta_k z_k}}$ as a parameterization for the probability of default, where β is the parameters vector.

[11] The Elastic-Net model, which incorporates both types of regularization, was also estimated. However, it is excluded from the final set of models since, based on the hyperparameters that maximized the F_2-score, the model converged towards a Ridge regularization.

[12] In this case, a penalization parameter $\lambda \sum_{j=1}^{p} |\beta_j|$ is introduced in the optimization program, where λ refers to the penalty parameter, β_j to the coefficients of the logistic function and p to the number of predictors.

[13] In Ridge, penalization is conducted with the term $\lambda \sum_{j=1}^{p} \beta_j^2$.

non-overlapping regions are constructed iteratively based on the variable and the variable cutoff that minimizes the cost function. In classification problems, the Gini criterion is normally employed as the cost function. This process is recursively repeated, generating a series of partitions in the predictor space. Ultimately, predictions are made based on the most representative class within each region. Decision trees are appealing due to their high interpretability, ease of explanation, and intuitive nature. However, due to their high flexibility, they suffer from high variance, which can lead to overfitting and poor performance on the test sample (James et al. 2013). To address these limitations, RF and XGBoost methods are used. These models enhance performance by introducing randomness into tree construction and combining multiple trees rather than relying solely on a single one.

RF is based on constructing an ensemble of decision trees generated simultaneously. Specifically, bootstrap sampling with replacement is used to create pseudo-samples for estimating individual trees. In the estimation, we use 200 threes. In each tree, a random subset of predictors is selected as potential candidates for determining the optimal data partition at each iteration. Unlike bagging, which utilizes all explanatory variables, RF randomly selects a subset of predictors at each split. This strategy aims to mitigate overfitting by reducing the excessive influence of certain variables on the prediction. In our estimations, we use a random subset of predictors at each split, equal to the square root of the total number of predictors. For the estimation of the RF model for firm default, we consider the following variables: lagged default, sectoral dummies, and lagged financial variables (see Table 4).

To control the complexity of the model and mitigate overfitting, some hyper-parameters can be adjusted, each affecting the structure of individual trees. The most relevant hyper-parameters include: i) the rule used to select the best split at each node (splitrule) and ii) the minimum terminal node size (min.node.size), which sets the minimum number of observations required for a node to remain unsplit. For splitrule, the Gini index was used as the decision criterion. The value of min.node.size was determined via five-fold cross-validation, optimizing performance based on the F_2-score metric (see Table 4 for the hyperparameter search space and the results obtained).

Boosting is a tree-based learnizing algorithm that follows a slow, sequential learning process, where trees are fitted iteratively such that each new tree learns from the prediction errors of the previous step. In each iteration, a new decision tree is added to reduce residuals, gradually improving the model's performance in areas where it previously failed to generalize well. Unlike bagging and RF, boosting is sequential, meaning that the construction of each tree depends on the previously grown trees. Among the algorithms that utilize boosting, XGBoost stands out as it combines features from RF with the sequential tree-building concept and then integrates the results of each tree. This method is widely used in classification problems due to its computational efficiency. For its estimation, we consider the same variables as those used in RF.

To control model complexity and mitigate overfitting in XGBoost, several hyperparameters can be tuned, each influencing different aspects of the learning process. The main hyperparameters considered in this study include: i) the maximum tree depth (max_depth), which sets the maximum number of partitions within each tree, defining its complexity; ii) the learning rate (eta), which controls the step size in each iteration, where smaller values require more trees to converge, while larger values may lead to overfitting; iii) the minimum reduction in the loss function (gamma), which imposes an additional penalty to prevent the creation of multiple irrelevant terminal nodes and; iv) the minimum size of terminal nodes (min_child_weight), which controls the minimum number of observations required in a node before performing a split, directly affecting tree complexity. To select the optimal values for these hyperparameters, we use five-fold cross-validation, optimizing performance according to the F_2-score metric (see Table 4). In addition, we set the number of boosting iterations (nrounds) to 200. The fraction of predictors randomly sampled at each tree level (colsample_bytree) was set to 0.9, and the proportion of the training data used in each iteration (subsample) was set to 0.8.

5.3 Results

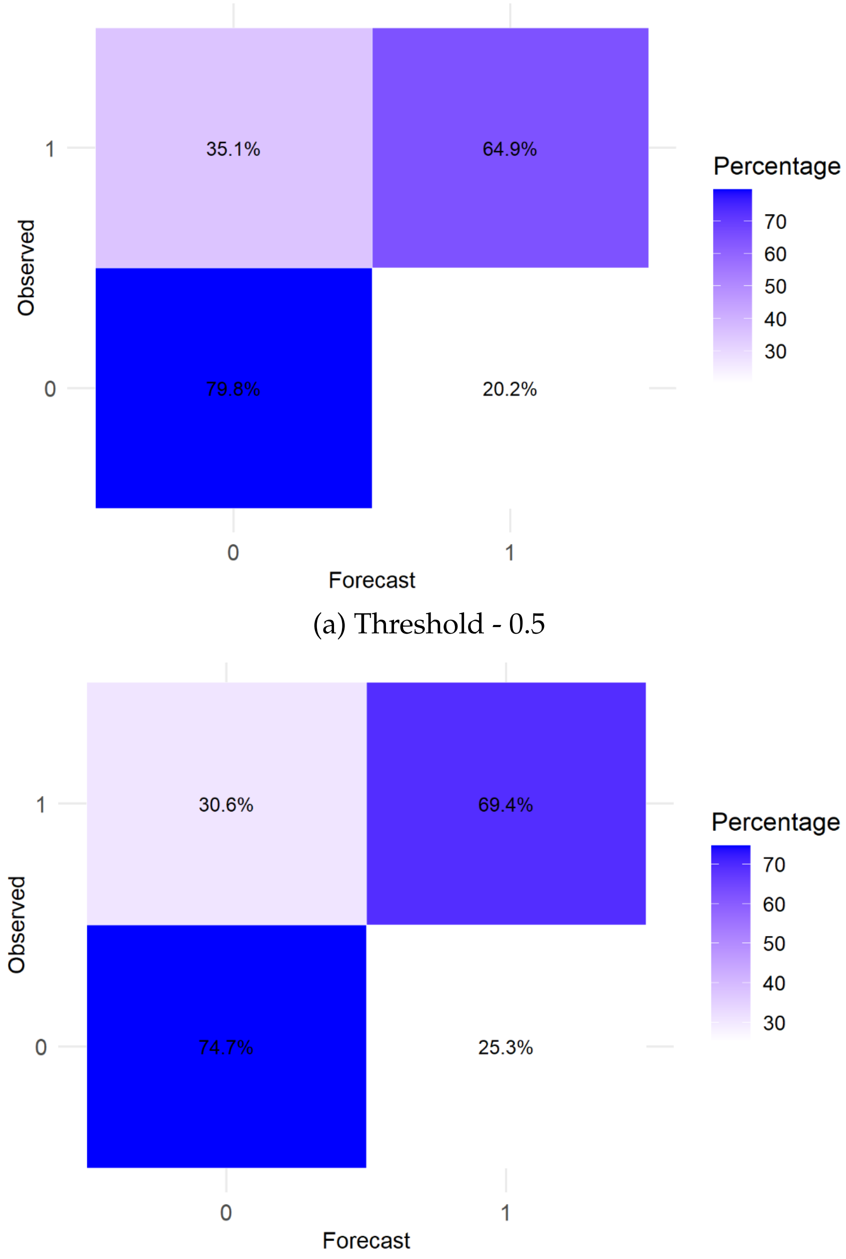
This section presents the performance assessment of the classification models on the test sample, employing the four previously introduced performance metrics, with particular emphasis on the F_2-score. The first part of Table 5 summarizes the results obtained for the test sample when the classification threshold for predicting default is set at 0.5. Conversely, the second part of Table 5 reports the results when the classification threshold is optimized based on the evaluation sample.

Without adjusting the classification threshold, the XGBoost model achieves the best performance, as indicated by the F_2-score. The second-best performing model is Logit 2, which exhibits a performance similar to that of the Random Forest model, followed by Logit 1. In contrast, the model with the lowest predictive capacity is Logit 3. This lower performance suggests that, given the model's complexity, it may be overfitting the training data, thereby significantly limiting its generalization ability to new data. This observation aligns with the high accuracy value and, simultaneously, the low recall value, indicating that a large proportion of firms are classified as non-defaulting when, in reality, they defaulted. As a result, accuracy increases due to the class imbalance in the test sample, which favors non-defaulting firms, while recall decreases due to a higher number of FN. The behavior of the Logit Lasso model in these metrics is similar, which is consistent with the near-zero value of the penalty hyperparameter. However, both Lasso and Ridge exhibit improved performance, suggesting that the Logit 3 model may indeed have been overfitted.

When employing the tuned classification threshold, the XGBoost model remains the best-performing model according to the F_2-score, closely followed by Logit 2. Although the performance of Logit 3 and Lasso improved, their predictive capacity remains inferior to that of the aforementioned models. Notably, the simplest model, Logit 1, achieves an F_2-score comparable to that of Logit 2. This result suggests that, despite incorporating fewer explanatory variables, Logit 1 maintains a similar ability to generalize and predict defaults. Given its superior performance in the test sample, the subsequent analysis focuses on the detailed results of the XGBoost model.

Based on the confusion matrix, under the untuned threshold, the proportion of true positives (TP) for the XGBoost model is 64.9%, while the true negative (TN) rate reaches 79.8% (Figure 5, panel a). When analyzing the confusion matrix with the tuned threshold, the TP rate improves to 69.4%. This improvement is achieved because a lower threshold results in a greater number of firms being classified as defaulting. However, this adjustment also leads to a reduction in the TN rate (Figure 5, panel b). Conversely, as the threshold increases beyond the tuned value, the proportion of true positives progressively declines until no firm is classified as defaulting, causing the F_2-score to converge to zero (Figure 6).

Figure 5: Confusion Matrix of XGBoost



(a) Threshold - 0.5

(b) Tuned threshold

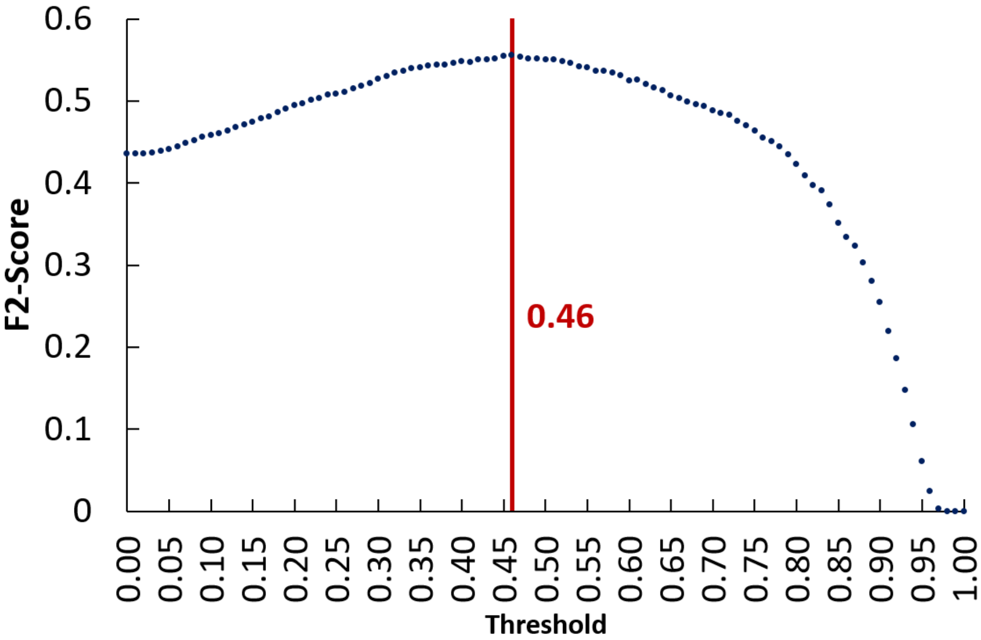
Source: Authors' elaboration based on data from the Colombian Superintendence of Companies and Financial Superintendence.

Table 5: Out-of-sample model performance

Model	**Threshold - 0.5**				**Tuned threshold**				
	Accuracy	Recall	Precision	F_2-score	Threshold	Accuracy	Recall	Precision	F_2-score
Logit 1	0.800	0.612	0.355	0.535	0.480	0.783	0.641	0.337	0.543
Logit 2	0.783	0.635	0.336	0.539	0.460	0.740	0.688	0.297	0.545
Logit 3	0.832	0.531	0.403	0.499	0.360	0.704	0.710	0.270	0.535
Logit Lasso	0.835	0.555	0.413	0.520	0.390	0.733	0.692	0.290	0.542
Logit Ridge	0.715	0.675	0.272	0.521	0.490	0.691	0.700	0.258	0.522
RF	0.739	0.674	0.293	0.535	0.500	0.739	0.674	0.293	0.535
XGBoost	0.778	0.649	0.331	0.544	0.460	0.740	0.694	0.298	0.548

Source: Authors' elaboration based on data from the Colombian Superintendence of Companies and Financial Superintendence.

Figure 6: F_2-score in the evaluation subsample vs. threshold



Source: Authors' elaboration based on data from the Colombian Superintendence of Companies and Financial Superintendence.

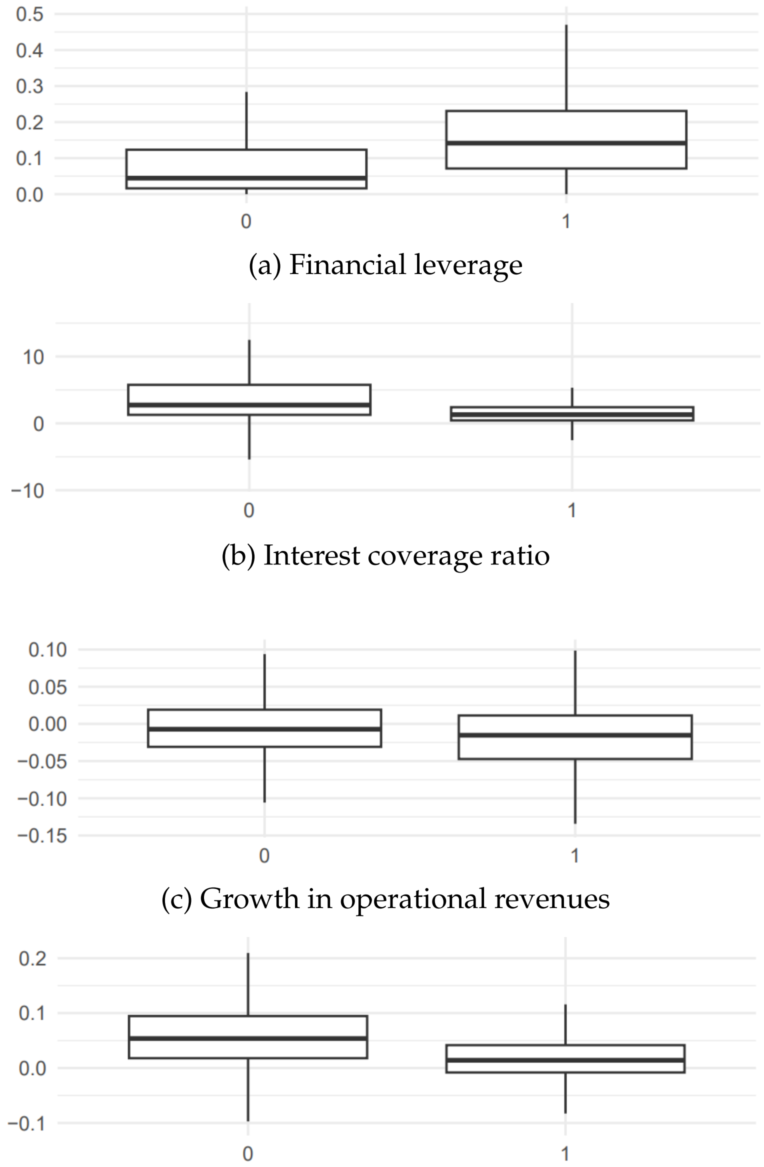
6 Classification under a stressed scenario

This section integrates the results from sections 4 and 5. Specifically, it utilizes the financial indicators of firms simulated under the methodology described in section 4 and applies the ML models presented in section 5 to identify firms that may be classified as in default under the given stress scenario. These identified firms are those whose operations would be most affected according to the dynamic balance sheet simulation. Therefore, the individual identification of each defaulting firm becomes a valuable input for stress-testing exercises such as the stress test model used by the Central Bank of Colombia (Gamba et al. 2017), where idiosyncratic shocks to borrowers are considered to assess the resilience of credit institutions.

The results presented correspond to those obtained for the year 2025 using the XGBoost model. In particular, this section presents boxplots comparing firms classified as in default (1) versus those classified as not in default (0). When analyzing the distribution of firms in terms of their financial leverage (measured as the ratio of financial obligations to total assets), it is observed that firms classified as in default generally exhibit higher levels of this indicator in the preceding period (Figure 7, panel a). Similarly, when examining the interest coverage ratio (measured as the ratio of earnings before taxes to interest expenses), firms classified as in default tend to display lower coverage ratios in the previous period compared to those that were not classified as in default (Figure 7, panel b). Regarding firms' revenues, those with a lower probability of being classified as in default are generally those that experienced higher growth in operational revenues in the previous period (Figure 7, panel c). Finally, in terms of return on assets (ROA), firms that exhibited lower ROA in the preceding period have a higher probability of being classified as in default (Figure 7, panel d).

It is essential to note that while each variable offers valuable insights into the distribution of firms in default, the XGBoost model's classification process considers all variables jointly. Therefore, a firm with a negative or near-zero ROA is not necessarily classified as in default. This is because, despite potentially having low profitability due to the nature of its business, the firm may simultaneously exhibit robust financial leverage or interest coverage indicators. Consequently, these results should be interpreted as illustrative and analyzed holistically in conjunction with the additional variables.

Figure 7: Financial indicators for firms classified in default (1) and not in default (0)



(a) Financial leverage

(b) Interest coverage ratio

(c) Growth in operational revenues

(d) Return on assets (ROA)

Source: Authors' elaboration based on data from the Colombian Superintendence of Companies and Financial Superintendence.

References

Altman, E. I. (1968), 'Financial ratios, discriminant analysis and the prediction of corporate bankruptcy', *The Journal of Finance* **23**(4), 589–609.
URL: *http://www.jstor.org/stable/2978933*

Altman, E. I., Marco, G. & Varetto, F. (1994), 'Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the italian experience)', *Journal of Banking and Finance* **18**(3), 505–529.
URL: *https://www.sciencedirect.com/science/article/pii/0378426694900078*

Anand, K., Bédard-Pagé, G. & Traclet, V. (2014), Stress testing the Canadian banking system: A system-wide approach, Financial system review., Bank of Canada.

Borio, C., Drehmann, M. & Tsatsaronis, K. (2014), 'Stress-testing macro stress testing: Does it live up to expectations?', *Journal of Financial Stability* **12**, 3–15. Reforming finance.
URL: *https://www.sciencedirect.com/science/article/pii/S1572308913000454*

Bottazzi, G., Grazzi, M., Secchi, A. & Tamagni, F. (2011), 'Financial and economic determinants of firm default', *J Evol Econ* **21**, 373–406.

Burrows, O., Learmonth, D. & McKeown, J. (2012), RAMSI: a top-down stress-testing model, Bank of England Financial Stability Papers 17, Bank of England.

Cabrera, W., Rueda, J. G. & Mendoza, J. C. (2012), Credit risk stress testing: An exercise for Colombian banks, Temas de Estabilidad financiera 073, Banco de la República de Colombia.

Caceres, C., Cerdeiro, D., Pan, D. & Tambunlertchai, S. (2020), Stress testing u.s. leveraged corporates in a covid-19 world, IMF Working Papers 2020/238, International Monetary Fund.

Carletti, E., Oliviero, T., Pagano, M., Pelizzon, L. & Subrahmanyam, M. G. (2020), 'The COVID-19 Shock and Equity Shortfall: Firm-Level Evidence from Italy', *The Review of Corporate Finance Studies* **9**(3), 534–568.
URL: *https://doi.org/10.1093/rcfs/cfaa014*

Cathcart, L., Dufour, A., Rossi, L. & Varotto, S. (2020), 'The differential impact of leverage on the default risk of small and large firms', *Journal of Corporate Finance* **60**, 101541.
URL: *https://www.sciencedirect.com/science/article/pii/S0929119918305443*

Ciampi, F. & Gordini, N. (2013), 'Small Enterprise Default Prediction Modeling through Artificial Neural Networks: An Empirical Analysis of Italian Small Enterprises', *Journal of Small Business Management* **51**(1), 23–45.
URL: *https://ideas.repec.org/a/taf/ujbmxx/v51y2013i1p23-45.html*

Demmou, L., Calligaris, S., Franco, G., Dlugosch, D., McGowan, M. A. & Sakha, S. (2021), Insolvency and debt overhang following the covid-19 outbreak: Assessment of risks and policy responses, OECD Economics Department Working Papers 1651, OECD Publishing.
URL: *https://www.oecd-ilibrary.org/content/paper/747a8226-en*

Dent, K., Westwood, B. & Segoviano, M. (2016), Stress testing of banks: an introduction, Quarterly Bulletin 2016Q3, Bank of England.

Farmer, J. D., Kleinnijenhuis, A. M., Nahai-Williamson, P. & Wetzer, T. (2020), Foundations of system-wide financial stress testing with heterogeneous institutions, Bank of England working papers 861, Bank of England.
URL: *https://ideas.repec.org/p/boe/boeewp/0861.html*

Gamba, S., Óscar Jaulin, Lizarazo, A., Mendoza, J. C., Morales, P., Osorio, D. & Yanquen, E. (2017), SYSMO I: A systemic stress model for the Colombian financial system, Borradores de Economia 1028, Banco de la Republica de Colombia.

Guerrieri, V., Lorenzoni, G., Straub, L. & Werning, I. (2022), 'Macroeconomic implications of covid-19: Can negative supply shocks cause demand shortages?', *American Economic Review* **112**(5), 1437–1474.

James, G., Witten, D., Hastie, T., Tibshirani, R. et al. (2013), *An introduction to statistical learning*, Vol. 112, Springer.

Kalemli-Ozcan, S., Gourinchas, P.-O., Penciakova, V. & Sander, N. (2020), 'Covid-19 and sme failures', *IMF Working Papers* **2020**(207).

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

Modina, M., Pietrovito, F., Gallucci, C. & Formisano, V. (2023), 'Predicting smes' default risk: Evidence from bank-firm relationship data', *The Quarterly Review of Economics and Finance* **89**, 254–268.
URL: *https://www.sciencedirect.com/science/article/pii/S1062976923000558*

Siggelkow, C. & Fernandez, R. M. (2024), ‘Sme default prediction using random forest including nonfinancial features: An empiricial analysis of german enterprises’, *Journal of the International Council for Small Business* **5**(2), 129–147.
URL: *https://doi.org/10.1080/26437015.2023.2224108*

Traczynski, J. (2017), ‘Firm default prediction: A bayesian model-averaging approach’, *Journal of Financial and Quantitative Analysis* **52**(3), 1211–1245.

Tressel, T. & Ding, X. (2021), Global corporate stress tests—impact of the covid-19 pandemic and policy responses, IMF Working Papers 2021/212, International Monetary Fund.

Xia, Y., Liu, C., Li, Y. & Liu, N. (2017), ‘A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring’, *Expert Systems with Applications* **78**, 225–241.
URL: *https://www.sciencedirect.com/science/article/pii/S0957417417301008*

A Regression Analysis

Table A1 presents the results of the regressions (10) and (11). Specifications include the lagged outcome variable and a vector of lagged firm characteristics. Column 1 presents the regression of sales growth as a function of the credit boom and sector×GDP interactions. Financial leverage is modeled in column 2 as a function of the credit boom indicator. Column 3 presents regressions of log change in costs as a function of sector$\times\Delta \ln \text{Sales}_{it}$ interactions (see equation 11). Robust standard errors are in parenthesis.

Table A1: Regression analysis

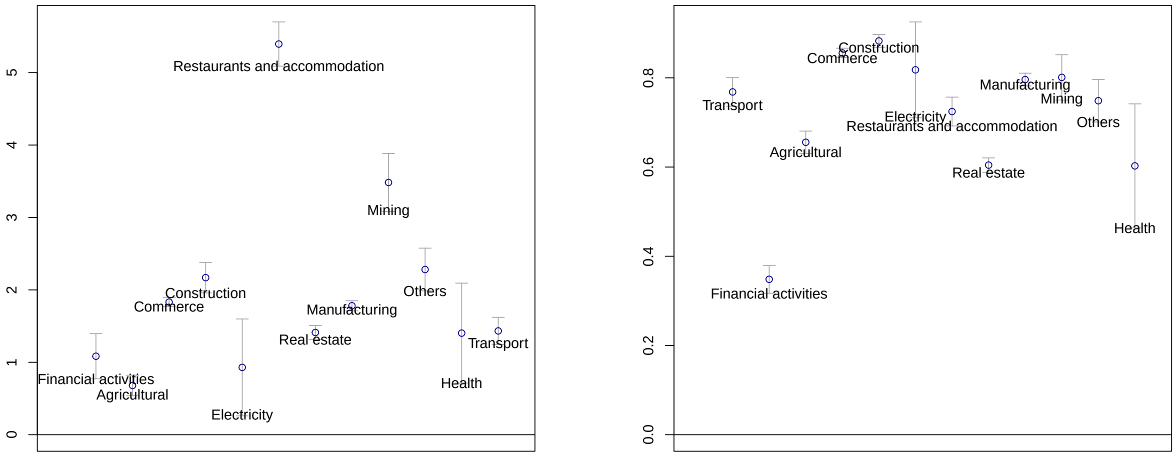
VARIABLES	(1) Log change of sales	(2) Financial leverage	(3) Log change of costs
Lagged dependent variable	-0.07 ***	0.73 ***	-0.49 ***
	(0.00)	(0.00)	(0.01)
Lagged firm characteristics			
Log change of sales		-0.003 ***	0.38 ***
		(0.00)	(0.01)
Financial leverage	0.07 ***		-0.02 *
	(0.00)		(0.01)
Sales-to-assets ratio	0.003 ***	0.005 ***	-0.02 ***
	(0.00)	(0.00)	(0.00)
ROA	0.08 ***	-0.05 ***	0.29 ***
	(0.01)	(0.00)	(0.02)
Log of assets	0.02 ***	0.002 ***	0.001 ***
	(0.00)	(0.00)	(0.00)
Credit boom indicator	0.02 ***	0.002 ***	
	(0.00)	(0.00)	
Sector fixed effects	Yes	Yes	No
Sector x GDP growth	Yes	No	No
Sector x Log change of sales	No	No	Yes
N	301,689	304,609	299,135
R2	0.03	0.59	0.31

Regression results from equation (10) and (11). Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Results point to mean reversion and persistence effects for growth variables and financial leverage variable, respectively (lagged dependent variable). Regarding lagged firm characteristics, we find that firms with larger assets, higher profitability, and higher financial leverage and sales-to-assets ratio tend to grow more (Column 1). Sales growth and ROA, on the other hand, correlate negatively with financial leverage (Column 2). Finally, we find that sales-to-assets rate correlate negatively with the log change of costs (Column 3).

When it comes to estimates of macro and financial variables, we find that there is a positive effect of the credit boom indicator on sales growth and financial debt. In particular, an increase of 1SD in the boom indicator implies an increase of 0.14 percentage points in the financial debt ratio. Figure A1 depicts sector×GDP-growth effects on sales growth (panel a) and sector×$\Delta \ln \text{Sales}_{it}$ (panel b). In general, regressions capture a sectoral heterogeneous correlation between average sales growth and aggregate GDP growth which is consistent with the co-movements observed in the historical data (see Section 3). In particular, restaurants and accommodation, mining, construction, and manufacturing sectors display a strong sales elasticity with respect to aggregate GDP. Regressions also capture a heterogeneous cost-to-sales elasticity among sectors, where construction and commerce show the highest elasticities.

Figure A1: Effects of GDP growth and sales growth accross sectors



(a) Sector×GDP-growth effects on log change of sales

(b) Sector×$\Delta \ln$ Sales effects on log change of costs

Authors' calculations of Sector×GDP-growth effects (Panel a, Ψ_s^{Macro} in equation 10) and Sector×$\Delta \ln$ Sales effects (Panel b, Ψ_s in equation 11). Robust confidence intervals calculated at the 10% significance level.

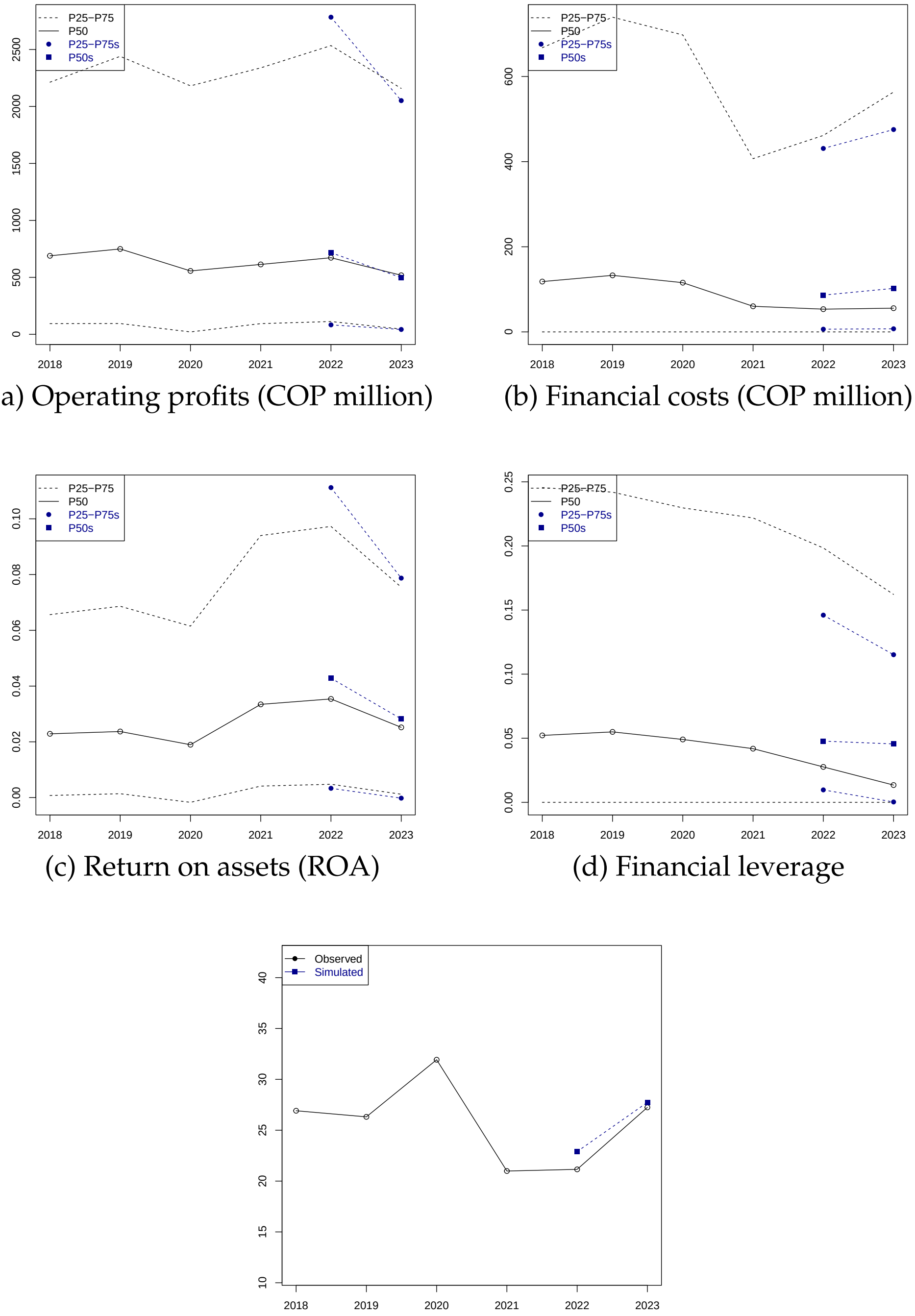
B Out-of-sample performance of the dynamic balance sheet simulation

This appendix presents an out-of-sample exercise to provide some insights about possible biases presented in the dynamic balance sheet simulation model. In particular, we run the dynamic balance sheet simulation starting in 2021 through periods 2022 and 2023 based on regression results until 2021 and macroeconomic variables observed in 2022 and 2023. The two-year horizon is chosen based on the commonly stress test horizons used by Banco de la República. Based on these results, we compare the simulated distribution of key financial variables with the actual 2022 and 2023 observed data.[14]

Figure B1 shows the results. According to them, the model tends to correctly estimate the distribution of operational profits and ROA (Figure B1, panel a and c). However, the model seems to overestimate financial leverage ratios, as well as financial obligations and costs for half firms (percentiles 25 and 50, Figure B1, panels b and d). As a result, the proportion of firms with an ICR lower than one also tends to be overestimated, particularly in 2022 (Figure B1, panel e).

[14] For sound comparisons, results are only presented for firms observed in 2021 and with a complete vector of lagged firm characteristics, CharFirm_{it-1}, at the starting point of the exercise.

Figure B1: Out-of-sample results of the dynamic balance sheet simulation (2023-2024)



(a) Operating profits (COP million)

(b) Financial costs (COP million)

(c) Return on assets (ROA)

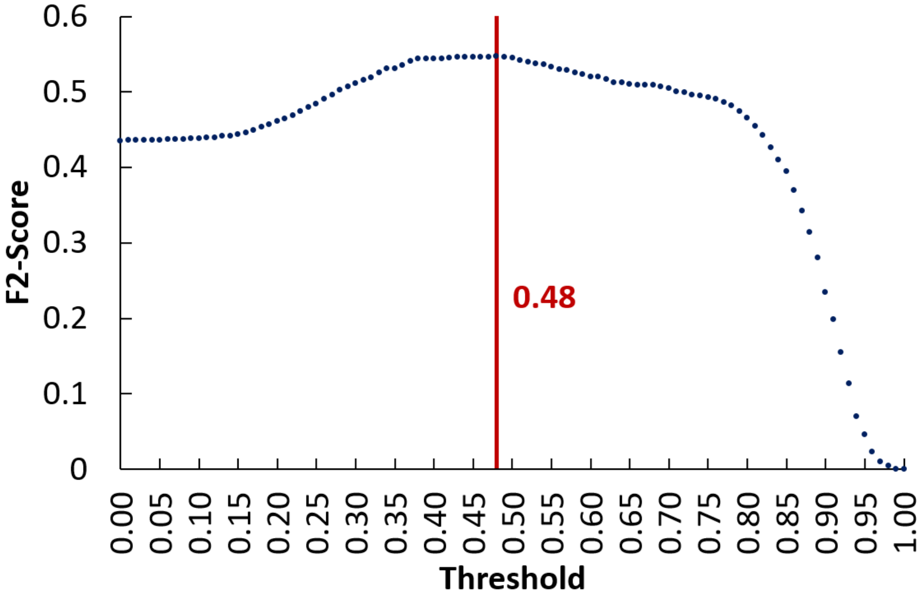
(d) Financial leverage

(e) Percentage of firms with ICR<1

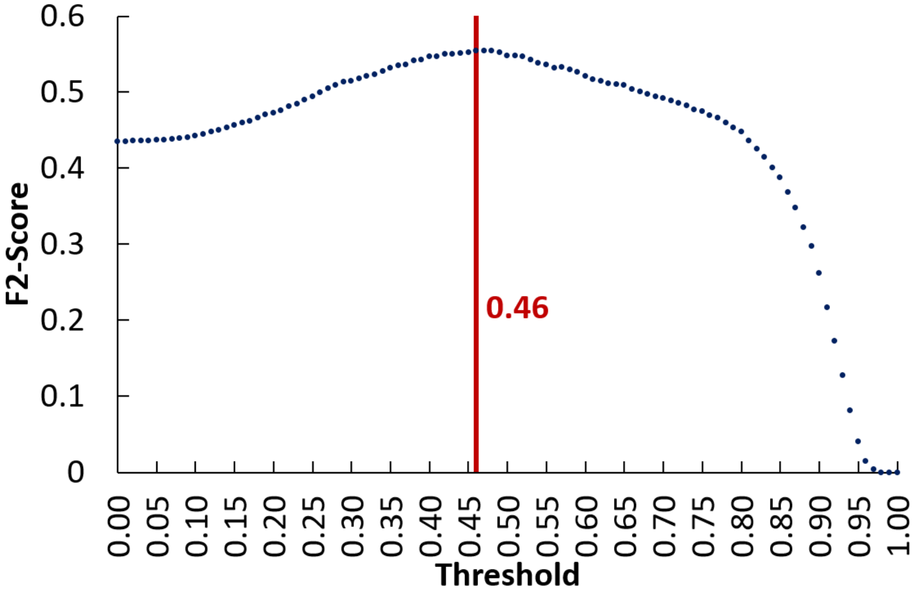
Observed (black lines) and balance sheet simulation (blue lines) of key firms' financial variables based on observed macroeconomic data for 2022 and 2023. Px refers to the x percentile of the corresponding variable.

C Threshold tuning in the evaluation sample

Figure C1: Threshold tuning

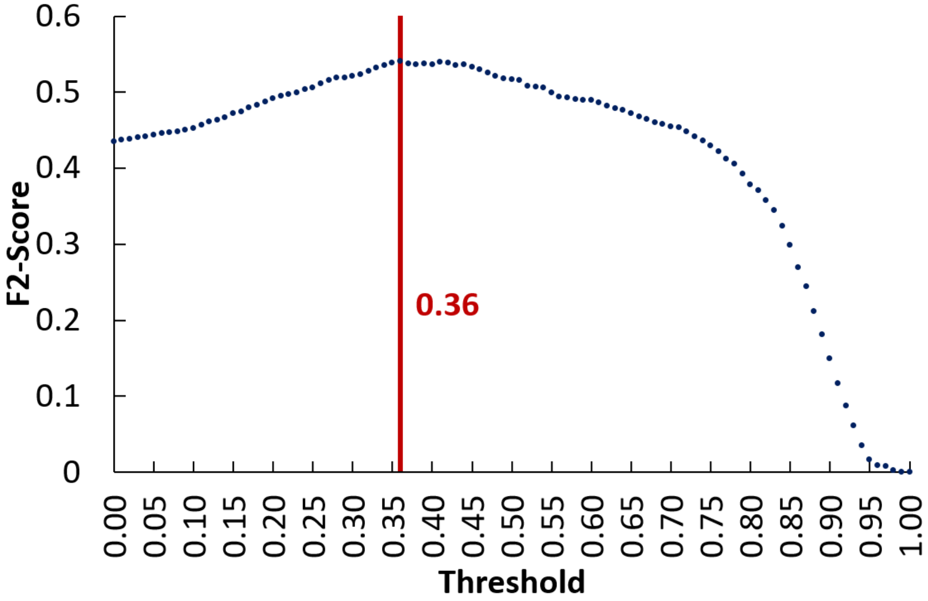


(a) Logit 1

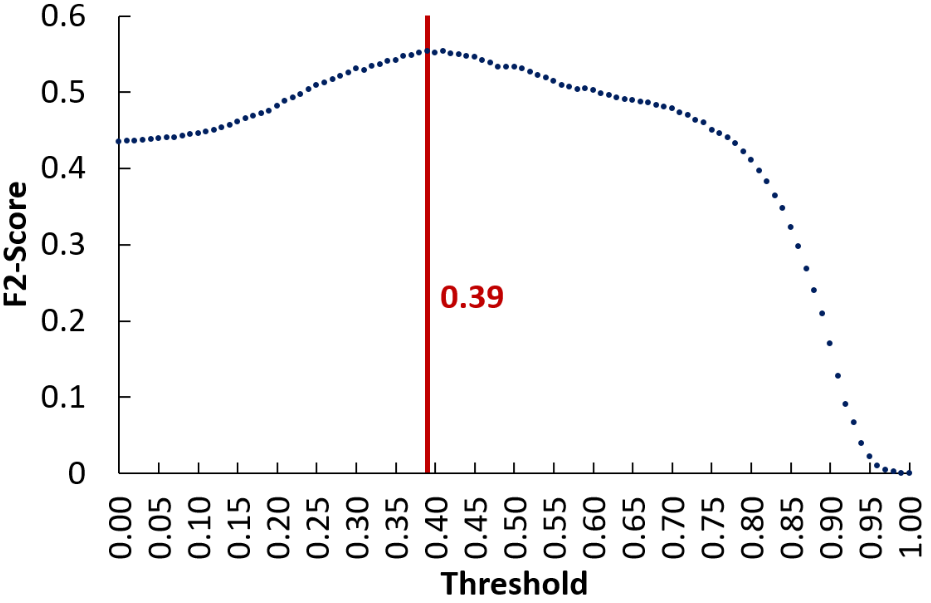


(b) Logit 2

Figure C1: Threshold tuning (continued)

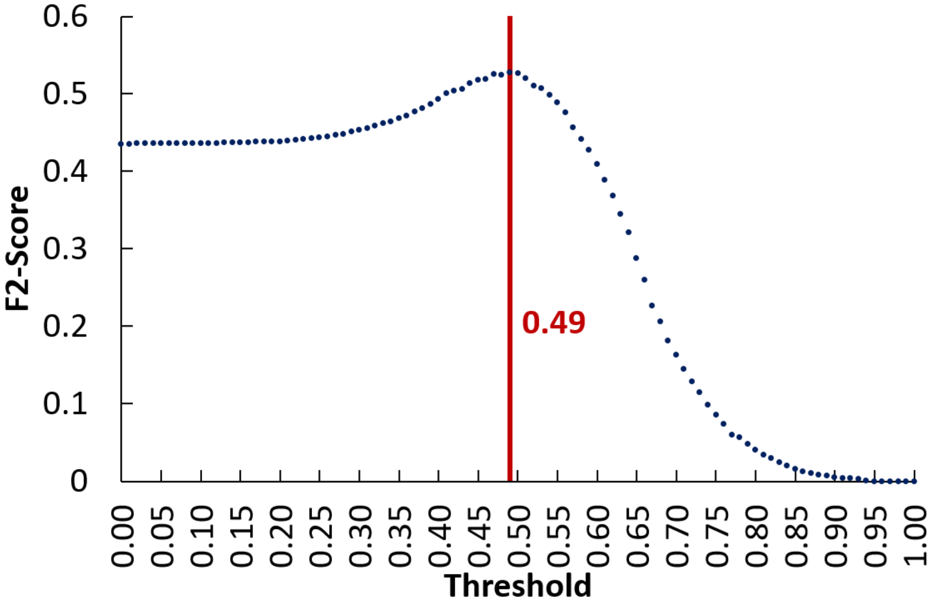


(c) Logit 3

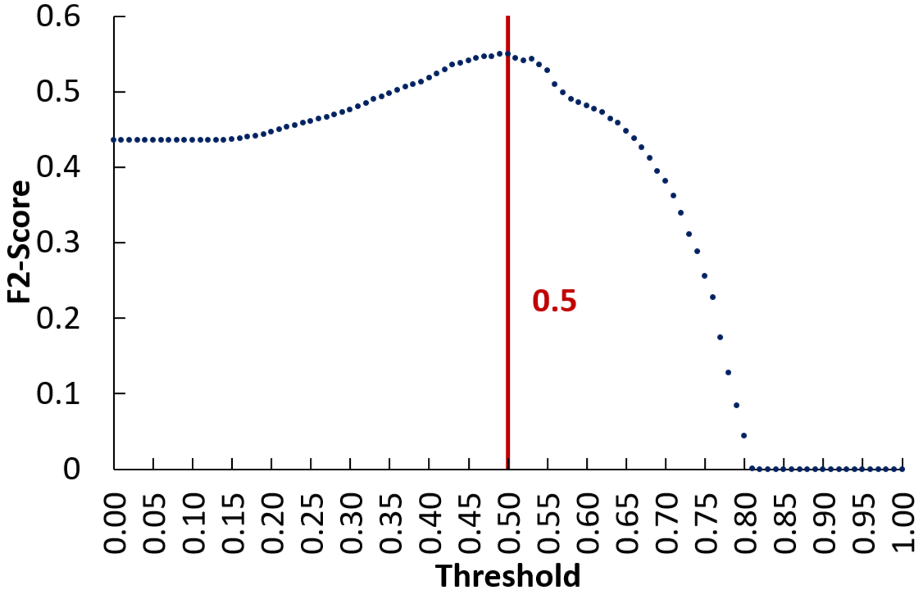


(d) Logit Lasso

Figure C1: Threshold tuning (continued)



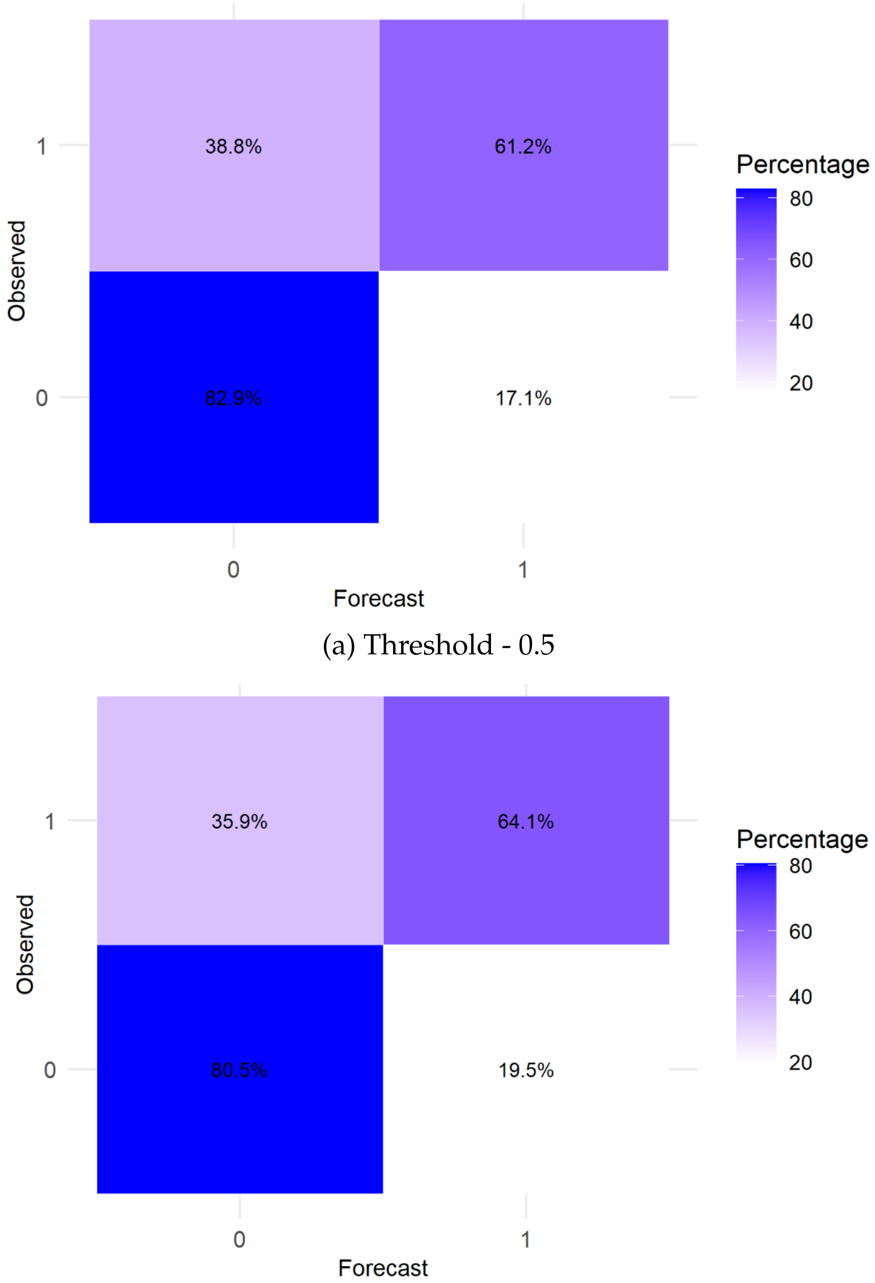
(e) Logit Ridge



(f) Random Forest

D Confusion Matrix per model

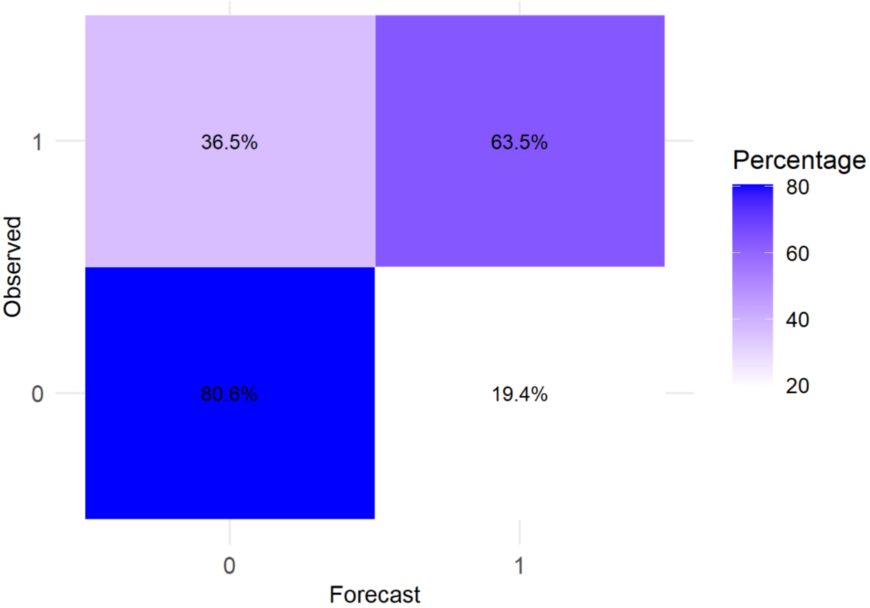
Figure D1: Logit 1



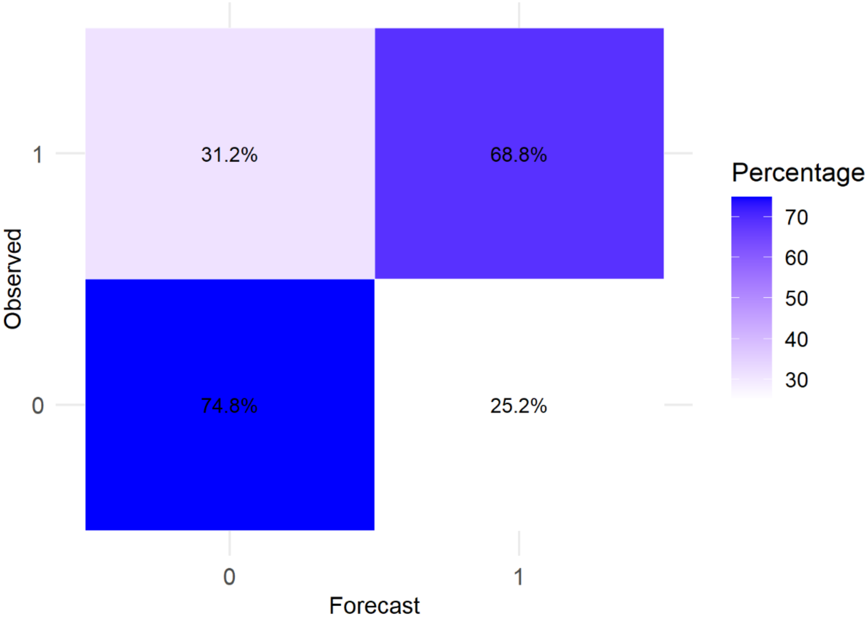
(a) Threshold - 0.5

(b) Tuned Threshold

Figure D2: Logit 2

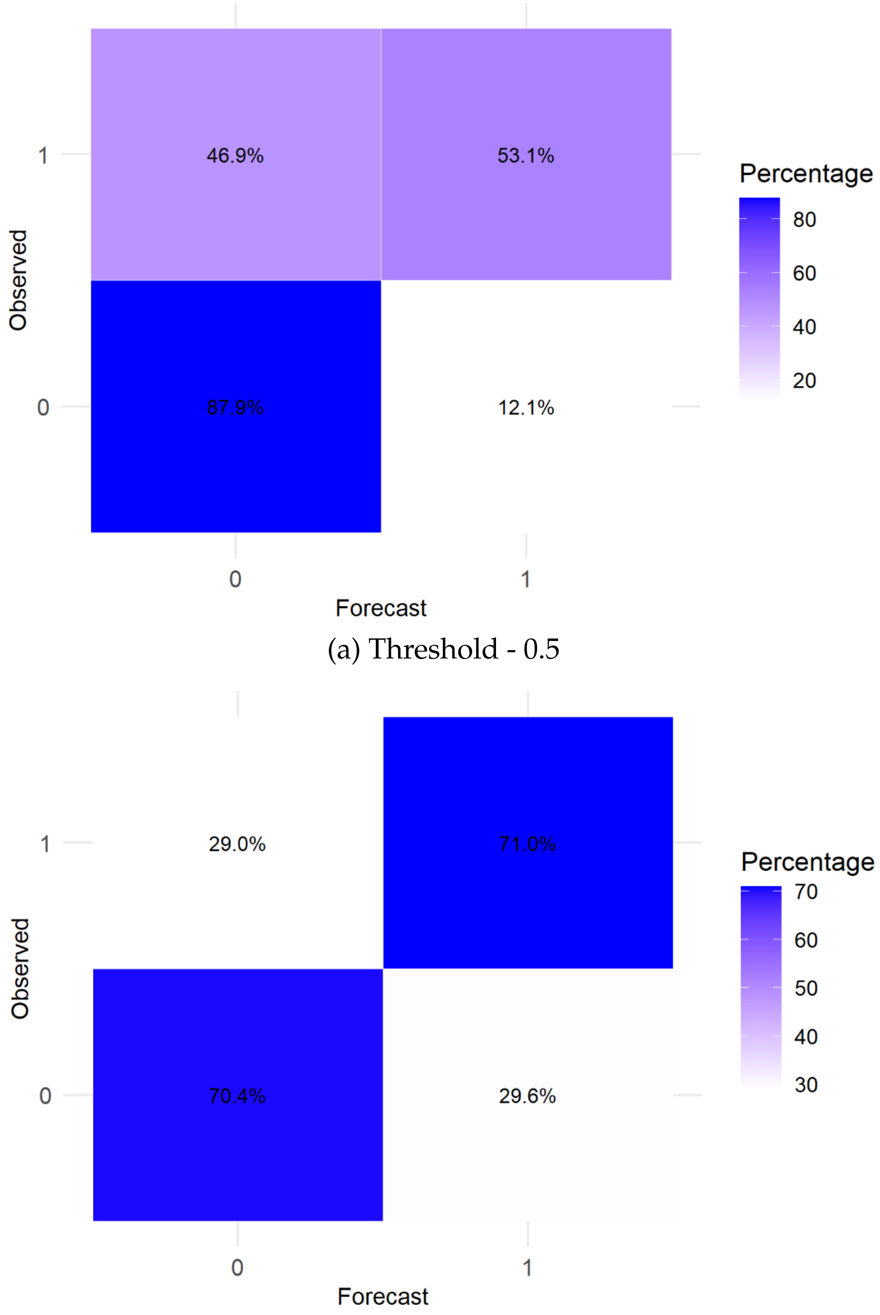


(a) Threshold - 0.5



(b) Tuned Threshold

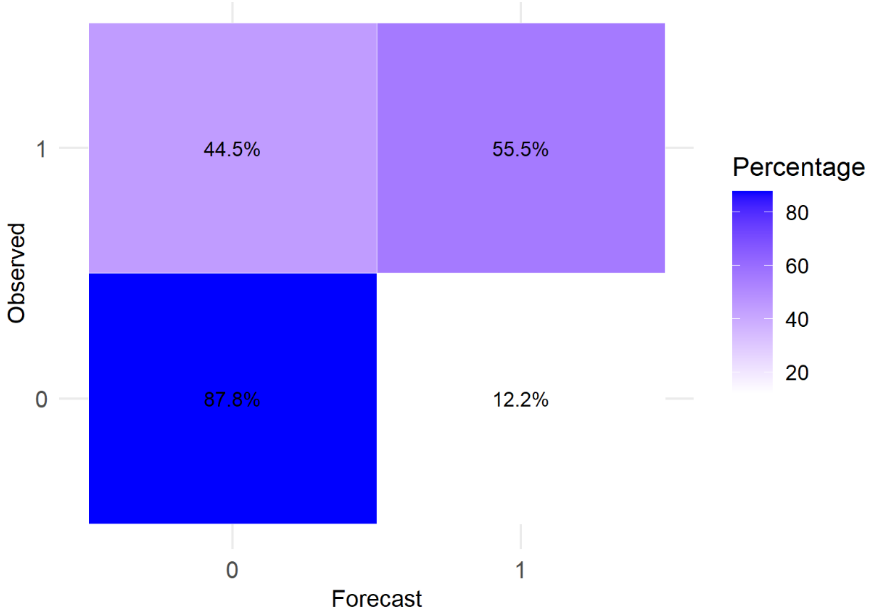
Figure D3: Logit 3



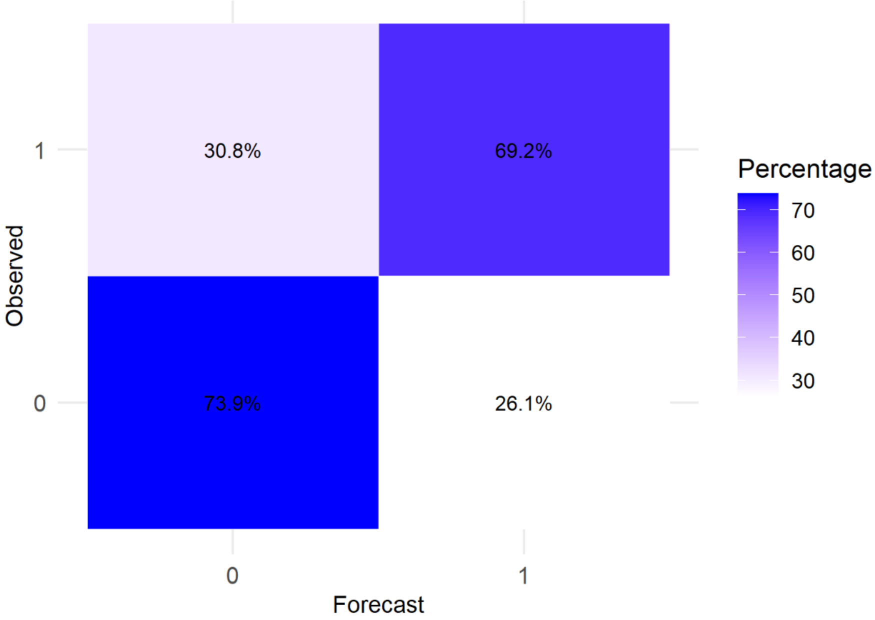
(a) Threshold - 0.5

(b) Tuned Threshold

Figure D4: Logit 4

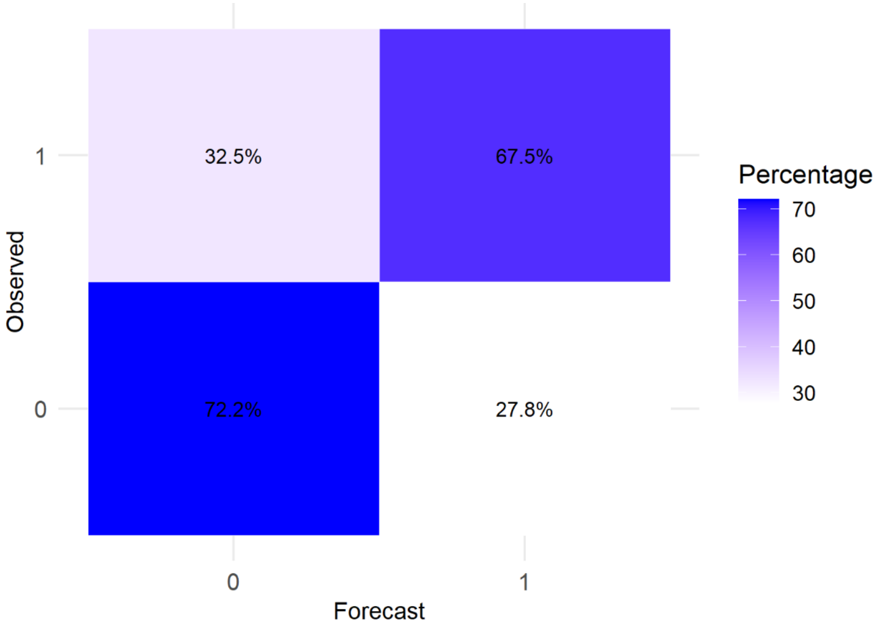


(a) Threshold - 0.5

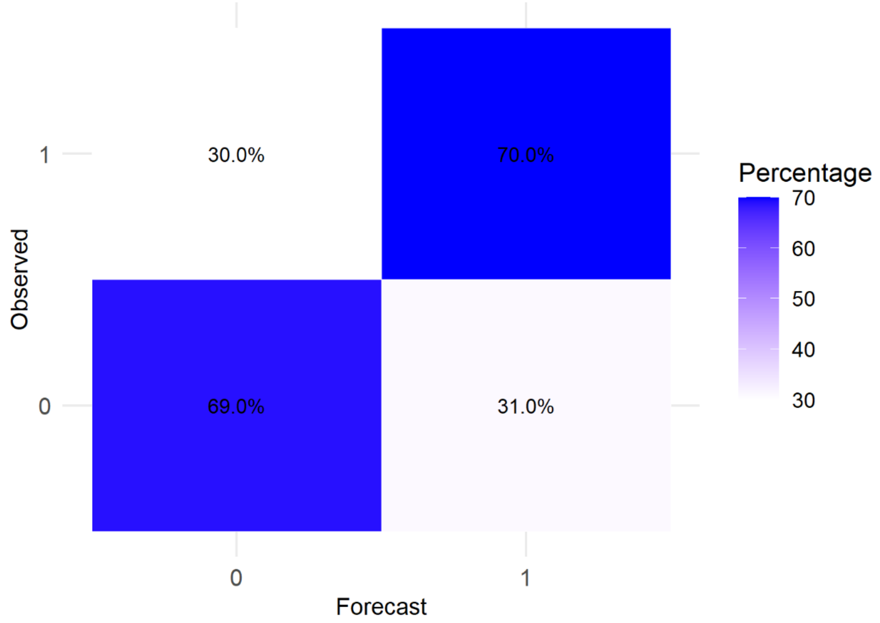


(b) Tuned Threshold

Figure D5: Logit 5

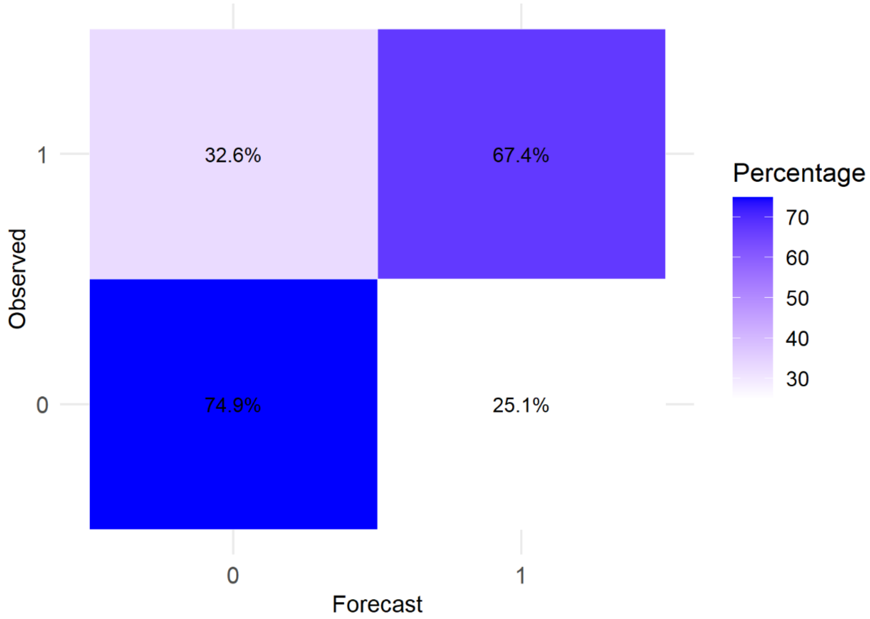


(a) Threshold - 0.5



(b) Tuned Threshold

Figure D6: Random Forest



(a) Threshold - 0.5