Are you a Zombie Firm? An Early Warning System Based on Machine Learning Methods*

Angela De Martiis† Thomas L.A. Heil‡ Franziska J. Peter‡

November, 2022 · Link to most recent version

Abstract

This paper develops an early warning system based on machine learning methods and logistic regressions to predict zombie firms. We use feature selection methods on large datasets of listed firms from Europe and the US to find the most important variables that separate a zombie firm from a recovered zombie. We find that beyond debt and income, taxes and equity are recurring features. Altogether, we document that differently to standard preselected variables, an ensemble of features related to the firm capital, financial, and industry structure are needed to predict zombie firms and recovered zombies.

JEL codes: C55, C63, D22, G32, G33

*We thank Philip Valta, Julien Cujean, Marc Brunner, Ansgar Walther, Cláudia Custódio, Maria Guadalupe, Eileen Tipoe, Brent Glover, Bryan Routledge, James Albertus, Angela Maddaloni, Camelia Minoiu, and participants at Bern Data Science Day (University Bern), Big Data & Machine Learning in Finance Conference (Politecnico di Milano), Swiss Society of Economics and Statistics (University of Zurich), 28th Finance Forum (Nova SBE), EFMA Meeting (University of Leeds), CEBRA Annual Meeting (MIT Golub Center for Finance and Policy), EEA Women in Economics, IFABS Oxford Conference (University of Oxford), Tepper CMU Finance Brown Bag, the Philadelphia Fed, and ASSA/AEA poster session for valuable comments.

†Corresponding author. Institute for Financial Management, University of Bern, Engehaldenstrasse 4, 3012, Bern, Switzerland. I acknowledge WRDS Visiting Faculty access from Carnegie Mellon University Tepper School of Business during FS 2021. angela.demartiis@gmail.com.

‡Zeppelin University, Am Seemooser Horn 20, 88045 Friedrichshafen, Germany. thomas.heil@zu.de; franziska.peter@zu.de.

1 Introduction

Long after the Global Financial Crisis, the prevalence of insolvent borrowers kept alive by subsidized credit, also called zombie firms (Hoshi 2006; Caballero, Hoshi, and Kashyap 2008), remains a concern and a warning sign for financial regulators (see, Figure 1). This motivates us to develop an early warning system that classifies zombies and recovered zombie firms and predicts future zombies and recovered. To do so, we use machine learning methods on large data of listed firms from the U.S. and Europe. Proposing a mechanism that outdoes conventional measures can serve banks, regulators, and policy makers to predict the development of the zombie phenomenon in the future.

Given the high dimensionality of our datasets and the amount of explanatory variables that can be used to predict the zombie status, we use machine learning methods to exploit the ensemble of data and select the most important features (i.e., independent variables) classifying a firm as zombie or recovered. In a traditional approach, economists carry out the selection of variables. Such an approach becomes however undisciplined and intractable when there are many candidate explanatory variables. With the application of machine learning methods, we refrain from making a priori assumptions and let the data-driven algorithms select the most relevant features. The final model, combining the most informative features, can correctly predict zombies versus non-zombie firms for almost 80% and zombies versus recovered for roughly 88% of the firm sample.

To develop an early warning system, our empirical strategy follows a distinct selection process. First, we apply machine learning methods such as random forests, decision trees, and logistic-LASSO to find the most relevant features separating zombies from non-zombies and recovered zombie firms. Second, we fine-tune the machine learning algorithms' parameters to increase the power of the models. The prediction task takes place implicitly by matching the input space at time t with firms' zombie status two years later, i.e., at time $t + 2$. The actual early warning system is provided by a logistic model that includes the selected and generated features from the previous machine learning algorithms. We thereby propose an easy to implement but powerful mechanism which advantages lye in its practical application, usage, and interpretability. We focus on two early warning systems:

one to predict zombie firms from non-zombies and a second one predicting recovered zombies from zombie firms. The first is of particular interest for banks having to decide on whether, and under which conditions, to issue loans to their borrowers. The second carries additional information for policy makers seeking to target support measures to firms that are likely to recover from the zombie status.

Existing research examines the reasons why zombie firms remain alive and the consequences of their existence on the economy (Hoshi 2006; Caballero, Hoshi, and Kashyap 2008; McGowan, Andrews, and Millot 2018; Banerjee and Hofmann 2018; Acharya et al. 2019; Acharya et al. 2020), while it remains silent about the specific features that matter to capture future zombies and recovered zombie firms. Using firm level data from Compustat Global and Compustat North America, we develop a selection system that out of a wide range of explanatory variables returns a set of features predictive of the zombie and recovered zombie status.

In doing so, we contribute to two strands of the literature. The first, relates to studies applying machine learning methods to bankruptcy prediction (see, for instance, Gepp, Kumar, and Bhattacharya (2010), Chen (2011), Brezigar-Masten and Masten (2012), Liang, Tsai, and Wu (2015), and Bargagli Stoffi, Riccaboni, and Rungi (2020)). Brezigar-Masten and Masten (2012) apply decision trees to select and construct variables for bankruptcy prediction within a logit model and show how these additional variables yield superior prediction accuracy in standard models as well. Bargagli Stoffi, Riccaboni, and Rungi (2020), the paper closest to our, use machine learning methods to propose an alternative measure of zombie firms. The focus of this literature is however on methodological issues. In our paper, the interest is on two classes of firms, the zombie firms and the recovered zombies. Our contribution to this strand lies in the application of non-parametric methods, tree-based models, to find the most informative features categorizing a zombie firm from a recovered zombie, and ultimately predicting how a future zombie versus a recovered zombie looks like. Thereby, we rely on a statistical algorithmic outcome built on well-established empirical measures of zombie firms (Caballero, Hoshi, and Kashyap 2008; Acharya et al. 2020).

The second strand, relates to the literature examining the phenomenon of zombie firms, and more precisely their firm specific features. With this respect, the studies closest to our are those of Hoshi (2006) and Banerjee and Hofmann (2020).

We contribute to this literature by proposing an early warning system for zombie and recovered zombie firms which makes use of machine learning methods for feature selection. We take advantage of the granular classification of tree-based models to go beyond a standard categorization layer based on preselected variables and show that additional explanatory variables are decisive to categorize zombie firms and to understand the transition from zombie status to recovered.

We add to this literature by showing that capital and financial structure variables returned by tree-based models carry a higher weight for predicting zombies and recovered zombie firms and discriminating between the two classes of firms. The results furthermore indicate that the most informative features go beyond the firm income and debt structure and relate also to shareholders' equity and taxes, and divergences within the manufacturing industries point at the industry of operation of the firm as an additional feature to consider when modeling zombies and recovered zombie firms.

The rest of the paper continues as follows. Section 2 describes the data and the measure of zombie firms. Section 3 presents the empirical strategy, introducing tree-based models and logistic-LASSO. Section 4 presents the empirical results of the selected features for zombie firms and recovered zombies. Section 5 reports the underlying mechanism of the early warning system, its performance and results. Section 6 concludes.

Figure 1 about here

2 Data and Descriptive Statistics

2.1 Data

To develop an early warning system based on machine learning methods, we use high-dimensional firm level data on listed firms from the U.S. and Europe. We build two datasets using Compustat Global and Compustat North America Fundamentals Annual covering financial, balance sheet, and market data information.

We process the data by deleting observations with missing company unique identifier, *gvkey*, and missing information on fiscal year, *fyear*. We remove all

duplicates and drop year-company combinations that have less than 99% observations.[1] We drop variables that display missing values for more than 65% of their observations. We restrict the datasets to contain common variables that appear both in the Global and the North America dataset and construct three cross-sections, one for a pre-crisis year (2007), one for a post-crisis (2016), and another for a more recent period (2020).[2] In terms of standardization, we adopt a standard normal scaler and normalize all variables to total assets. This procedure leaves us with 120 input variables in both datasets per company-year.

Following existing studies (Acharya et al. 2020), we exclude firms belonging to the utilities, financial, and banking industries. We use 4-digit SIC codes and compute the 49 industry groups by Fama and French (1997). Next, we winsorize each variable at the 5% and 95% level, drop all observations below and above this threshold to reduce the effect of outliers, and delete rows with NaNs. We impute the remaining missing values with K-Nearest-Neighbors imputation, given its promising results in terms of predictive power.[3] Last, we lag all input variables with a two year lag for implicit prediction of zombies and recovered zombie firms.

Altogether, this data preparation process yields two well-stocked datasets consisting of 5,446 observations in year 2007, 5,476 in 2016, 4,966 in 2020 for the European sample, and 2,920 observations in 2007, 2,323 in 2016, 2,155 in 2020 for the U.S. sample.[4]

2.2 Measuring Zombie Firms

Following Caballero, Hoshi, and Kashyap (2008), Acharya et al. (2019) and Acharya et al. (2020), we define a zombie firm as a low-quality firm that receives subsidized credit at advantageous interest rates. In line with this definition, a firm is classified as zombie if it receives subsidized credit (Caballero, Hoshi, and Kashyap 2008). To classify firms, the actual interest payments made by the firms, R_{it}, is

[1] We take into account that these firms are likely listed dead firms.

[2] We consider these specific cross-sections to account for economic downturns.

[3] We test mean-value imputation, but this yields general results that can bias the datasets.

[4] The European sample is composed of 30 countries (Germany, France, Italy, Netherlands, UK, Sweden, Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Latvia, Lithuania, Luxembourg, Malta, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Switzerland, Norway).

compared to an estimated benchmark, R^*_{it}, based on the firm's actual debt structure and an advantageous interest rate expected to be paid by the highest quality borrowers.[5] The difference between the effective interest expenses and the estimated benchmark, $R_{it} - R^*_{it}$, is referred to as the interest rate gap, x_{it}. The firms with a negative x_{it} are those expected to receive subsidized credit by their banking counterparts and are categorized as zombie firms by Caballero, Hoshi, and Kashyap (2008). We are however unable to use such an approach due to the lack of interest rate data for the U.S. sample. We therefore rely on the recent identification approach by Acharya et al. (2020), which uses the actual interest paid by high-quality (AAA-rated) borrowers and complements the resulting interest rate gap with firms' ratings and operating characteristics to distinguish zombies from non-zombie firms. A firm is categorized as zombie, or recovered zombie, based on two criteria: *(i)* it has an interest coverage ratio below median and a leverage ratio above median,[6] and *(ii)* it obtains credit at very low rates, at a rate below that paid by AAA-rated[7] borrowers with similar debt structure in any given year.

The final output variable that we use as input for the empirical strategy is equal to one if the firm is a zombie, to two if it is a recovered, and zero otherwise.

2.3 Descriptive Statistics

In Table 1, we report the summary statistics on a set of performance measures. The choice of variables follows the literature on zombie firms (see, for instance, Hoshi (2006), McGowan, Andrews, and Millot (2018), and Acharya et al. (2020)) and the statistics relate to the fraction of firms that we are interested in examining,

[5] As from Caballero, Hoshi, and Kashyap (2008), R^*_{it}, for firm i in year t, is defined as:

$$R^*_{it} = rs_{t-1}BS_{it-1} + \Big(\frac{1}{5}\sum_{j=1}^{5} rl_{t-j}\Big)BL_{it-1} + rcb_{\text{min over last 5 years, t}} \times Bonds_{it-1}, \tag{1}$$

where BS_{it}, BL_{it}, and $Bonds_{it}$ represent short-term loans, long-term bank loans, and total bonds outstanding (including convertible bonds and warrant-attached bonds), respectively, for firm i at end of year t; while rs_t, rl_t, and $rcb_{\text{min over the last 5 years, t}}$ represent the average short-term prime rate in year t, the average long-term prime rate in year t, and the minimum coupon rate on any convertible corporate bond issued in the last 5 years before t.

[6] Medians are computed at industry-year level.

[7] In addition to S&P ratings from Compustat daily updates, we use Aswath Damodaran synthetic ratings. We rely on S&P ratings whenever available, otherwise to the synthetic.

the class of zombie firms on the one hand and that of the recovered zombies on the other hand. Panel A, shows the results for the European sample, while panel B for the U.S. sample. As tree-based models do not account for time dynamics, we build three cross-sections across which we identify the most informative features separating zombie firms from recovered zombies.

In both samples and across cross-sections, we observe that zombie firms have a lower, if not negative, profitability compared to that of the recovered zombies and also higher leverage and tangibility, together with a lower investment capacity, cash availability, and a riskier financial position. We take these descriptive results as background features and we augment them with the application of tree-based methods to further understand zombie firms and recovered zombie firms.

Table 1 about here

3 Empirical Strategy

In this study, we develop an early warning system and we do so by considering two prediction tasks. On the one hand, we study zombies versus non-zombie firms, and on the other hand zombies versus recovered zombie firms. While the first task seeks to separate zombies from non-zombies, the second aims at separating zombies that have the potential to recover from those firms that remain in the zombie status. For both tasks, relevant variables are selected based on the high dimensional feature space fed into the tree-based models. At the same time, the large number of variables of the datasets that we use renders it cumbersome, if not impossible, to analyze the features of zombie firms and select important variables to predict the zombie status based on standard parametric models. Therefore, we resort to machine learning algorithms that are fit for the task (Breiman 2001). Figure 2 outlines our empirical strategy.

Figure 2 about here

We rely on variables selected by and derived from machine learning methods. The input space uses the whole set of balance sheet and market data information and the output constitutes the classification of a firm into zombie versus

non-zombie and recovered zombie versus non-recovered. The prediction task is performed implicitly by matching the input space at time t with the classification into the categories two years later, i.e., at time $t + 2$. The goal is to generate an early warning system that allows banks or policy makers to use current input data to predict the status of a firm two years ahead. Such a prediction mechanism delivers a system for future zombie firms and recovered zombie firms inspection.

We consider two machine learning methods: the first uses tree-based models, in particular random forests, to determine the most informative variables predicting a firm zombie status. The variables selected by the random forests are then fed into a final decision tree model that generates the split points of the variables. The split points are used to generate dummy variables as input for the final logistic regression model. Simultaneously, but independent from the random forests approach, we also use a second method, logistic-LASSO, to select important predictors. As both methods, random forests and logistic-LASSO, are based on different algorithms they can potentially select different variables. The variables selected by logistic-LASSO are thus also included into the final logistic regression model. The logistic regression model constitutes the early warning system used to select and construct the input variables for zombie firms and recovered zombie firms prediction.

3.1 Tree-Based Models

The advantage of a decision tree is its simplicity in combination with outstanding interpretability through elegant visualization. In contrast to classic statistic knowledge-based models, the tree finds the relevant characteristics directly from the data without the need for assumptions.

The idea behind decision trees is to split the input space X subsequently, into segments, where each segment decides for one class.[8] Accordingly, in each section, the outcome variable y is modeled with a different constant, e.g., a majority vote in classification problems. The algorithm does one binary split only for a single input feature at each iteration. After each iteration, the decision tree repeats the procedure in the new sub-samples until a stopping criteria is reached, thus preventing overfitting or underfitting, by setting constraints on tree size (e.g.,

[8] I.e., the whole dataset is split into subsets.

an upper bound on the depth of the tree) or performing tree pruning, another technique that reduces the size of the tree (Khandani, Kim, and Lo 2010).

Figure 3 shows an example of a decision tree procedure. In this paper, we focus on the CART algorithm by Breiman et al. (1984) to categorize two possible outputs, i.e., a zombie firm or a non-zombie, and to make prediction.

Figure 3 about here

For the task of predicting the zombie status, the algorithm underlying the decision tree searches repeatedly through the whole range of each firm financial and market information to select those variables that contribute the most to classify zombies versus non-zombies or zombie firms versus recovered, respectively. At each iteration, the algorithm searches for the best combination of split point and input variables to predict the output class. Such a chosen combination is called a node. At each node, the decision tree provides information on the selected variable, the chosen split point, and the percentage of observations used by the algorithm. The split point-variable combinations are chosen based on a loss function, such as the Cross-entropy or the Gini coefficient, which provide a measure of purity of the node, i.e., a measure that evaluates how well the algorithm predicts the output. For additional methodological details we refer to Appendix A.1.

While straightforward visualization and interpretability constitute an advantage of decision trees, they are unstable by nature, i.e., small changes in the dataset may change the resulting tree substantially. To account for this, and other, caveats Breiman (2001) introduced the so-called random forests.

Random forests introduce the concept of bootstrap aggregation (bagging) to decision trees. The random forests involve an ensemble (*forest*) of decision trees grown independently on a bootstrapped dataset sample in order to allow the averaging of the results. Additionally, at each node, the algorithm can only choose from a restricted set of randomly selected input features, which decreases the correlation among the ensemble of trees, thereby increasing out-of-sample performance.

As with many machine learning algorithms, random forests contain a certain amount of hyperparameters that require fine-tuning to increase the power of the model. The parameters of the random forest are, for example, the number of decision trees, the number of variables to choose from, the maximal number of nodes

in each tree (depth), or the amount of data required to make a split decision. It is a common approach to apply K-fold cross-validation to find the best hyperparameters (Khandani, Kim, and Lo 2010). K-fold cross-validation divides the dataset at hand, randomly, into K equally sized subsamples and uses K-1 subsets to train the model, and one subset to test the fitted model. This procedure is repeated until each of the K subsamples was the test sample once. In this study, we use 3-fold cross-validation (CV), in line with existing studies on default (Fuster et al. 2021), to ensure enough training data for the random forest while having a large enough test set for validation. The estimations are performed for all possible hyperparameters' combinations in a previously determined set of parameters.

We employ random forests with the aim of finding the most important and informative features categorizing zombie companies. We then use the features selected by the random forests to build binary decision trees and obtain an interpretable visualization of the decisive variables and their splitting points. Based on the selected variables and splitting points, we subsequently generate dummy variables to include in our final logistic regression model. In doing so, we are able to enrich the standard regression model by potentially novel non-linear information detected by the random forest and the decision tree algorithms.

3.2 Logistic-LASSO

The logistic-LASSO combines the well-known logistic regression with the least absolute shrinkage and selection operator (LASSO) introduced by Tibshirani (1996). LASSO, in general, adds a penalty term to the loss function that introduces the ability to decrease the value of coefficients to zero, and it can therefore act as a feature selection algorithm. The LASSO minimizes the following quantity:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|. \tag{2}$$

where RSS is the residual sum of squares denoting the loss function from classic linear regression and p gives the number of independent variables (features), whose paramaters β are chosen in a way that minimizes the loss function. The penalty parameter λ is chosen by cross validation.

The Logistic-LASSO adds the well-known logistic link function and constitutes the appropriate method when dealing with a binary dependent variable, e.g., zombie and non-zombie. Logistic-LASSO allows us to select variables from the overall feature space, which are subsequently included into the final logit model together with the tree selected variables and the tree generated dummies.

Since LASSO and its variants are designed for variable selection, but based on different algorithms, we include it as a complement to the decision trees and random forests to increase the power of our final early warning prediction system.

The final early warning system is a logit model which includes the variables selected and generated (i.e., via the splitting points) by the decision trees and the logistic-LASSO models. The logit model is used to deliver a prediction about a firm's zombie status in two years time, based on current observations of the selected variables. In the following section we describe the selected features, before assessing the prediction accuracy of the final logit model (Section 5).

4 Evidence on Feature Selection

4.1 Zombie Firms

Figures 4 and 5 show the results of the random forests and document the most important and informative features separating zombie firms from non-zombies in the U.S. and Europe. The three panels relate to the three cross-sections, a pre-crisis (2007), a post-crisis (2016), and a more recent (2020) time period.

We document nine variables that are repeatedly selected across the cross-sections for Europe and the U.S., with minor variations for the latter sample. The selected variables relate to a firm's income, debt, and earnings. The most informative income-related variable selected by the random forests is pretax income, also known as earnings before tax. As a profitability measure signaling the firm cash-generating ability, pretax income is a decisive feature across all time periods and samples. Historical attention to income variables values them as financial distress predictors (Hofer 1980; Whitaker 1999; Platt and Platt 2002).

In terms of debt, the most informative variables are long-term debt total, liabilities total and debt in current liabilities, with long-term debt showing especially

high coefficients for all cross-sections of the U.S. sample. Although the variables report high values in both samples, long-term debt is the amount of debt that must be retired, while debt in current liabilities, or short-term debt, is bank debt often rolled over (Platt and Platt 2002). As a zombie is a firm kept alive by bank rolled over loans (Caballero, Hoshi, and Kashyap 2008), this finding points at dissimilar debt agreements between zombies and non-zombies in the U.S. and Europe, and at differences in terms of capital structure (Rajan and Zingales 1995), financial structure (Allen, Chui, and Maddaloni 2004), debt maturity (Custódio, Ferreira, and Laureano 2013), and institutions (Altman, Dai, and Wang 2022). This result might also explain the lower prevalence of zombies in the U.S., as also documented by Favara, Minoiu, and Perez-Orive (2022).

The difference between what the firm owns and what it owes is a recurring variable separating a zombie firm from a non-zombie. Shareholders' equity reports high coefficients during the post-crisis and in 2020, for the U.S. sample, and during the post-crisis period for Europe. Even tough it is a variable often used to predict financial distress (Fama and French 1992; Dichev 1998), it has received little attention by the literature studying zombie firms. Its high informativeness suggests that looking into specific types of investors matter to understand zombie firms, given that shareholders have decision rights regarding the firm's financial choices, especially in terms of debt-related decisions (Valta 2016).

Among the selected variables, we also find retained earnings and income taxes total. The results indicate that retained earnings, the firm's profit after dividends, costs, and income taxes, a component of shareholders' equity total, can be considered to distinguish a zombie firm from a non-zombie. This profitability measure, often observed to predict bankruptcy, signals the ability of the firm to use internal funds versus outside funds for growth opportunities. A zombie firm, however, is more likely to continue operating using outside generated capital, which entails higher risk. Income taxes total relates to all taxes that are imposed on the firm's income by local, provincial/state, national, and foreign governments. It is very uncommon to find taxes-related variables in studies examining zombie firms. Our results can therefore hint at country or state-specific taxes as possible determinant of a firm zombie status as well. As our analysis however focuses on the meta perspective, we leave country-specific features for future research.

Figures 4 and 5 about here

Complementing the random forests, Figures 6 and 7 show the zombie-predicting variables returned by LASSO feature selection algorithm. While the results report a set of debt-related variables, they also show relevant industries. In terms of debt structure, long-term debt, short-term debt, and notes payable are the variables with the highest coefficient in the European sample, where notes payable is the total amount of short-term notes and borrowings and is a component of debt in current liabilities,[9] while in the U.S. we find long-term debt followed by short-term debt. The results underline the importance of firm debt structure to discriminate between a zombie firm and a non-zombie.

In terms of industries, European firms operating within the computer software or liquor industry (Figure 6) appear more likely to be classified as zombies compared to firms in healthcare, retail or wholesale. In the U.S., electronic equipment, measuring and control equipment, and textiles are returned as relevant industries capturing zombie firms (see, Figure 7). These results contrast with the narrative of zombie firms mostly populating the retail industry,[10] point at the existence of disparities within manufacturing, at potential differences between zombie status and bankruptcy status, and at the necessity to consider the industry of operation when modeling zombie firms as the presence of zombies in an industry can put non-zombies in that same industry at risk (Caballero, Hoshi, and Kashyap 2008).

The results suggest that firm capital and financial structure variables composed of debt, income, but also shareholders' equity and taxes, jointly with industry features, can discriminate between a zombie firm and a non-zombie.

Figures 6 and 7 about here

4.2 Recovered Firms

The second classification task considers the most important and informative features defining a recovered zombie firm and separating it from a zombie. The results

[9] As from Compustat guide, the item of notes payable includes not only all long-term debt in default, but also, for example, current bank loans, loans payable to parent companies, and loans payable to shareholders.

[10] February 5, 2020, Financial Times article: https://on.ft.com/3t4EUEt.

of the random forests, presented in Figures 8 and 9, show unanimously across time and country, that long-term debt total and pretax income are the variables with the highest coefficient. The remaining selected variables differ only slightly from those selected within the first classification task (zombies versus non-zombies) and refer to the firm composition in terms of debt structure and profitability, but also taxes and equity. Although these features have a lower weight, as indicated by their coefficients, their ensemble together with long-term debt and pretax income is what discriminates between a recovered zombie firm and a zombie firm.

Figures 8 and 9 about here

Figures 10 and 11 report the selected features returned by the logistic-LASSO. The results show the importance of industries such as pharmaceutical products in the U.S. and computer software and retail in Europe, together with firm capital structure characteristics such as pretax income, income taxes, as well as long-term and short-term debt. The variables that collectively display the highest weight in determining the recovery status of a zombie firm are debt, income, and taxes, while secondary relevance is assigned to sales and short-term investments for U.S. firms, and to intangibles and inventories for firms in Europe.

In addition to signaling the importance of firm specific capital structure features, and country specific differences that are beyond the scope of this paper, the results indicate that there is a set of informative variables, beyond debt and income, that contribute to identifying and separating zombies from recovered zombie firms. The random forests, as well as logistic-LASSO, also indicate that the predictors of zombie firms and recovered zombies present additional information to those used in the zombie literature. This suggests that existing models might be incomplete. Also, the transition from zombie status to recovered zombie status is composed of diverse facets that can hardly be captured by standard parametric models. In the next section, we propose an early warning system that discriminates between a zombie company and a non-zombie or a recovered zombie firm.

Figures 10 and 11 about here

5 Early Warning System

After selecting the most informative features, we combine them in a logit model. We include the selected variables from the random forests and logistic-LASSO, and add the information from the final decision trees by generating dummy variables based on the selected split points. The aim is to provide an early warning system for zombie firms on the one hand and predict the zombie firms that will recover on the other hand. For this purpose, a logit model is fit to the selected and generated variables and subsequently used for prediction on unseen (test) data.

We start by evaluating the performance of the logistic models on a 15% out-of-sample test set in different ways. We start by presenting the *Accuracy* of the prediction, which shows the difference between the actual observed values and the predicted values, followed by *Precision*, *Recall*, and *F1 score*.

The *Precision*, is the ratio of the correctly predicted positive (e.g., the zombies) observations to the total predicted positive observations. It tells us how precise the model is out of those predicted positive, meaning how many of them are actually positive. The final values are expressed in percentage terms and result from dividing the true positives (TP) by the TP added to the false positives (FP). We also compute the *Recall* to know how many of the actual positives the model can capture through labeling them as TP. This measure tells us what proportion of the actual positives was correctly identified. The outcome is a percentage value derived from dividing the TP by the TP summed to the false negatives (FN).

Both *Precision* and *Recall* have however drawbacks in terms of informativeness, especially when, for example, the cost of FP is high, i.e., if a zombie firm is predicted as non-zombie the consequence can be negative for the bank extending the loan to that firm. For this reason, we additionally compute the *F1 score* which gives us a balance between *Precision* and *Recall* and functions especially well when there is an uneven class distribution, meaning a large number of actual negatives. The *F1 score* can be read as an harmonic mean of the *Precision* and *Recall*.[11]

Tables 2 and 3 about here

[11] Following Python's scikit-learn metrics and scoring to evaluate a model prediction, we compute the *Precision* as: $\frac{TP}{(TP+FP)}$, the *Recall* as: $\frac{TP}{(TP+FN)}$, and the *F1 score* with the following formula: $2 \times \frac{(Precision \times Recall)}{(Precision+Recall)}$.

In Tables 2 and 3, we report the results of the *Accuracy* (A), *Precision* (P), *Recall* (R), and *F1 score* (F1) for both samples over the years 2007, 2016, and 2020. The prediction results for the class of zombies versus non-zombies are shown in Table 2, while those for zombies versus recovered in Table 3. Both tables display a *Confusion matrix* (CM), which represents a contingency table showing the difference between the correctly and incorrectly estimated, where the class of the non-zombies is displayed in the top line and that of the zombies in the bottom, while in the diagonal we find the correctly estimated versus the incorrectly estimated (recovered) zombies.[12]

The column *Accuracy* shows how often the classifier is correct. In both tables, the results indicate a high prediction accuracy that falls between 75% and 77% for zombies versus non-zombie firms (Table 2) and between 85% and 88% for zombies versus recovered (Table 3) with respect to the European sample. We record similar values also for the U.S. sample with an accuracy between 75% and 78% for zombies versus non-zombies (Table 2) and between 86% and 88% for zombie firms versus recovered (Table 3). The column *Precision* shows, when predicting zombie firms, how often this prediction is correct. For zombies versus non-zombie firms we find that the values are between 60 and 75%, for the representative years, with similar results for both country samples. Comparing the class of zombie firms with that of the recovered, the precision results are between 84% and 91%.

Recall shows instead how often a zombie firm is predicted as zombie. For the class of zombies versus non-zombies, we observe *Recall* values lower than *Precision*, between 14% and 22% for the European sample and between 4% and 49% for the U.S. sample, respectively. With respect to the *F1 score*, column *F1*, the score reaches its best value at 1 (i.e., perfect *Precision* and *Recall*) and worst at 0. The results are particularly high for predicting zombies versus recovered firms, thus reinforcing our understanding of future zombies and their transition into the recovered zombie status, while more modest for zombies versus non-zombie firms.

In Tables B1 and B2 in the Appendix we show the estimated logit parameters for the task of zombies versus non-zombie firms, while in Tables B3 and B4 those for predicting the recovered from the zombie firms. Altogether, we can observe

[12] In Appendix B.2, we also report prediction results on the whole sample.

that the input variables examined above are statistically highly significant in most of the cross-sections and that the split variables generated from the decision trees (indicated by the addition *Split*) add an additional layer of information with respect to the firm debt structure, profitability, and taxes. This indicates, that the relationships between these variables and the future zombie status of a firm are more complex than what can be implied by a standard logit regression analysis, thus highlighting the usefulness of using tree-based models to understand future zombies and recovered zombie firms.

6 Conclusion

In light of the COVID-19-induced crisis, there is agreement within the literature that the prevalence of zombie firms is a renewed concern for financial stability. This study contributes by providing empirical evidence on the firm specific features that matter to understand future zombies and recovered zombie firms. Using machine learning methods, precisely tree-based models, on two high-dimensional datasets of listed firms from Europe and the U.S., we develop an early warning system that returns a set of informative and recurring features (i.e., independent variables) predicting zombie firms and recovered zombies.

The main findings of this paper show that to examine tomorrow's zombies and recovered zombie firms all aspects of a firm capital, financial, and industry structure are relevant. A sorting mechanism is fit for such a data demanding task that discriminates between a zombie firm and a non-zombie or a recovered zombie. The results further show an additional layer of informative variables pointing at the firm shareholders' equity and taxes, beyond income and debt, as features to consider when examining future zombie firms and recovered zombies.

In carrying out this exercise, this paper has however limitations. We are, on the one hand, limited to listed firms and we acknowledge that this could underestimate the number of zombie firms. On the other hand, the large number of selected features does not allow us to make comparisons across countries and provide an interpretation in the context of finance theory. We therefore encourage future research on the set of variables returned by tree-based models, including private firms, to improve the identification of future zombie and recovered zombie firms.

A Appendix

A.1 Decision Trees

We follow CART algorithm[13] to construct input space regions and find exclusive regions $R_1, \ldots, R_j$ with rectangular shape. In a sample with input and output (y, X), where y is a discrete variable with classes K and $X = (x_1, x_2, \ldots, x_p)$ input variables, we require the algorithm to find the best input variable and split point s at each iteration. The proportion of y for each region R_j, is:

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k), \tag{1}$$

where I is the indicator function. The Cross-entropy is given by:

$$L(p) = -\sum_{k=1}^{K} \hat{p}_{mk} log(\hat{p}_{mk}), \tag{2}$$

where p_k is class k probability. The impurity reaches its minimum if all observations are correctly classified. A direct computation of the regions, minimizing Eq. 3, is not feasible as the input space can be split in infinite combinations of sub-rectangles. We thus start with a top-down approach of binary splitting. Assuming a splitting variable l and splitting point s, we choose the first pair of regions as:

$$R_1(l, s) = \{X | X_l \leq s\} \text{ and } R_2(l, s) = \{X | X_l > s\}. \tag{3}$$

Last, we find splitting variable l and split point s by solving $arg \max_k \hat{p}_{lk}$. After partitioning the input space in two regions, based on the best splitting variable and split point, the process is repeated within each region. Below, we provide a decision tree model example applied to our setting.

[13] Classification and Regression Trees (CART) algorithm.

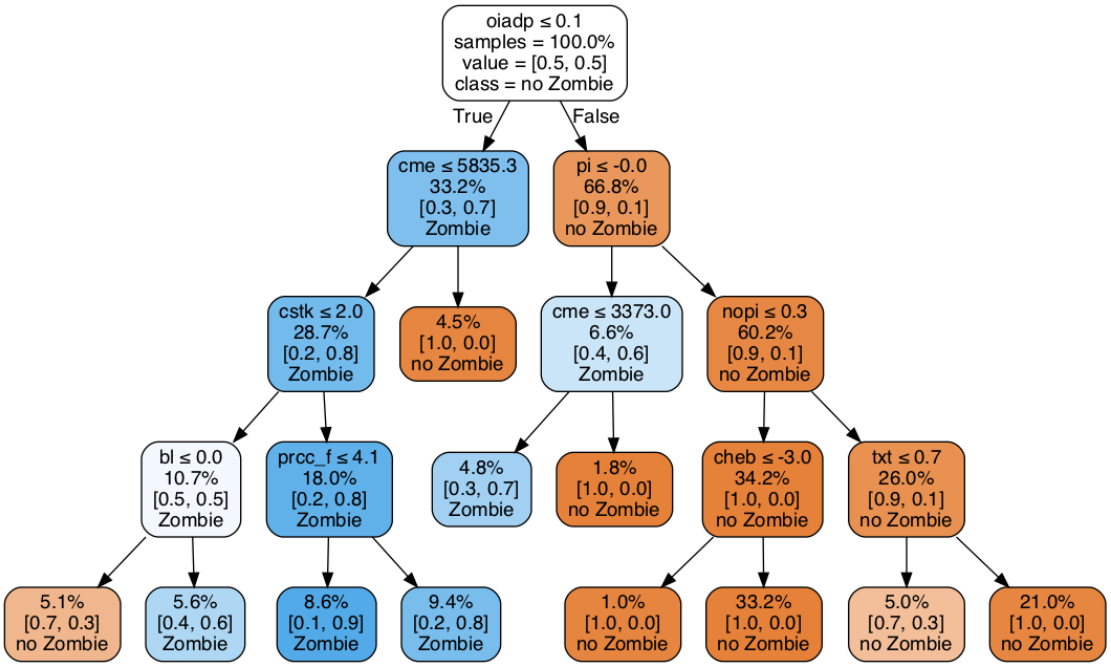

Figure A1: Tree Example Zombie versus Non-Zombie Firms. This figure reports a decision tree zombie versus non-zombie firms in Europe, for the 2016 cross-section. Higher splits provide higher importance for the decision. The nodes are where a decision is taken. Every node represents a feature, an individual independent variable acting as input. The decision iteration of CART algorithm is provided at the top of each node. The purity of the nodes is given by a higher entropy, and a darker color. The samples refer to the number of observations in the node. Value is the number of samples in each class (Zombie or no Zombie). Zombie firms are defined following Acharya et al. (2020), additional information on the identification of zombie firms is provided in Section 3. The final variable takes value 0 for the class no Zombie and 1 for Zombie. The figure is for illustration, it does not correspond to a tree used in the paper.

Legend: *oiadp* Operating Income After Depreciation, *pi* Pretax Income, *cstk* Common Stock, *nopi* Nonoperating Income (Expense) Total, $prcc_f$ Stock Price Closing Fiscal, *cheb* Cash and Cash Equivalents at Beginning of Year, *txt* Income Taxes Total.

B Additional Results

B.1 Logistic Regressions

	2007	2016	2020
	(1)	**(2)**	**(3)**
Income taxes other	−4.4059**		
	(1.8985)		
Trade accounts receivable	−1.0190***		
	(0.2789)		
Capital surplus	−0.0367		
	(0.0631)		
Shareholders equity	0.0000	−0.0002***	0.0001***
	(0.0000)	(0.0002)	(0.0000)
Liabilities total	−0.5499***	−0.0476	0.1051
	(0.1984)	(0.1555)	(0.1023)
Retained earnings	−0.1311***	−0.0103	0.0017
	(0.0436)	(0.0078)	(0.0025)
Debt in current liabilities	3.5717***	0.6553**	1.7110***
	(0.5125)	(0.3171)	(0.3528)
Income before extraordinary items	0.0322	0.0198	−0.0025
	(0.0349)	(0.0177)	(0.0263)
Income taxes total	−0.0088	−0.4274	−1.1862
	(0.0106)	(0.6542)	(1.9574)
Pretax income	−0.0967	0.0161	−3.1014
	(0.8175)	(0.0943)	(1.9363)
Long-term debt total	1.8989***	2.6038***	0.4500*
	(0.3425)	(0.2820)	(0.2695)
Inventories total		0.9505***	
		(0.2777)	
Current assets total		−0.0703	
		(0.1039)	
Notes payable		1.4655***	
		(0.4118)	
Deferred charges			1.2092***
			(0.3800)
Revenue total			−0.1757***
			(−0.1757)
Computer Software	0.8918***		
	(0.1159)		
Retail	−0.9981***	−0.8055***	−0.9859***
	(0.2395)	(0.2266)	(0.2101)
Healthcare	1.1634***		
	(0.3563)		
Printing & Publishing		0.9748***	
		(0.2724)	
Wholesale		−0.6518***	−0.7900***
		(0.2179)	(0.2426)
Beer & Liquor		1.6962***	
		(0.2848)	
Real Estate			−0.7830***
			(0.1674)
Long-term debt total Split	−0.8804***	0.4711***	−0.7967***
	(0.1073)	(0.1827)	(0.0962)
Pretax income Split	1.4310***	1.1124***	0.9674***
	(0.1116)	(0.0944)	(0.0823)
Debt in current liabilities Split	−0.2114*	−0.7256***	−0.5102***
	(0.1186)	(0.0810)	(0.1091)
Observations	5,446	5,476	4,966
Pseudo R^2	0.171	0.130	0.128

Table B1: Logistic Regressions Zombie Firms vs. Remaining, Europe. The table shows the logistic regression results for zombies versus the remaining firms for the three cross-sections. The dependent variable is a dummy variable which is equal to 1 whenever a firm is a zombie. The zombie measure is constructed following Acharya et al. (2020). All explanatory variables are lagged with a two year lag. Standard errors are clustered at the firm level. The variables ending with Split are those returned by the trees. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

	2007 (1)	2016 (2)	2020 (3)
Capital surplus	0.1388**		
	(0.0611)		
Shareholders equity	−1.2468	0.3075	0.4355
	(1.0585)	(1.1695)	(1.4413)
Liabilities total	−1.1845	1.0099	−0.4882
	(1.0809)	(1.2685)	(1.5088)
Retained earnings	0.0914	−0.0198	−0.0002
	(0.0630)	(0.0194)	(0.0160)
Debt in current liabilities	−8.3557***	3.0937***	3.8224***
	(1.2209)	(1.1581)	(0.7897)
Income before extraordinary items	−3.4428	−0.0074	2.4274
	(14.2919)	(0.0302)	(1.2639)
Income taxes total	5.7738	0.0813	
	(6.1627)	(0.4186)	
Pretax income	−9.4009	−1.9472**	−0.3419
	(5.9664)	(0.9182)	(2.3258)
Taxes payable	−8.0094*		
	(4.2531)		
Long-term debt total	2.3574***	1.0124	2.3233***
	(0.5314)	(0.6523)	(0.6123)
Work in progress inventories	0.5900**		
	(0.2476)		
Deferred income taxes		1.6890*	
		(1.0294)	
Selling general and administrative expense		−0.6090**	
		(0.2506)	
Current liabilities other		−2.6842***	
		(0.9378)	
Notes payable		2.0600	
		(1.4316)	
Operating activities net cash flow			−0.0530
			(0.2915)
Short-term investments total			1.0422***
			(0.3912)
Apparel	1.5134***		
	(0.3626)		
Measuring & Control Equipment	1.3768***		
	(0.2830)		
Electronic Equipment	0.8734***		
	(0.1726)		
Communication		−0.8018***	−1.1801***
		(0.3068)	(0.4481)
Patroleum and Natural Gas		−1.1387***	−1.2050***
		(0.2556)	(0.3626)
Healthcare			−1.2546**
			(0.4898)
Textiles			2.2255***
			(0.8566)
Long-term debt total Split	−0.7660***	−0.7558***	−0.3791*
	(0.1692)	(0.1840)	(0.1957)
Pretax income Split	1.3593***	1.2116***	0.8068***
	(0.1520)	(0.1811)	(0.1490)
Debt in current liabilities Split	0.6253**		
	(0.3110)		
Deferred income taxes Split		−0.4101**	
		(0.1226)	
Liabilities total Split			−0.2703
			(0.1793)
Observations	2,920	2,323	2,155
Pseudo R^2	0.195	0.138	0.087

Table B2: Logistic Regressions Zombie Firms vs. Remaining, USA. The table shows the logistic regression results for zombies versus the remaining firms for the three cross-sections. The dependent variable is a dummy variable which is equal to 1 whenever a firm is a zombie. The zombie measure is constructed following Acharya et al. (2020). All explanatory variables are lagged with a two year lag. Standard errors are clustered at the firm level. The variables ending with Split are those returned by the trees. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

	2007	2016	2020
	(1)	**(2)**	**(3)**
Long-term debt total	-14.7243^{***}	-11.3904^{***}	9.7673^{***}
	(1.3352)	(1.0927)	(1.2541)
Debt in current liabilities	-14.9976^{***}	-13.2045^{***}	-10.4115^{***}
	(1.6112)	(1.3520)	(1.8426)
Intangibles	-0.5902^{*}		
	(0.3208)		
Pretax income	14.6761^{***}	4.9172^{***}	13.2737
	(1.8878)	(0.6428)	(12.8183)
Shareholders equity	-0.0001	-0.0082^{***}	-0.1688
	(0.0051)	(0.0013)	(1.0194)
Long-term debt due in 1 year	1.0952		
	(1.3599)		
Retained earnings	0.1473^{**}		0.1813^{**}
	(0.0643)		(0.0820)
Liabilities total	0.7768^{**}	-0.6071	-1.1514
	(0.3638)	(0.5470)	(1.1091)
Income taxes total	-0.5326		0.5578
	(0.6086)		(13.5471)
Income before extraordinary items	-0.1787^{*}	-1.5840^{***}	0.1387
	(0.1042)	(0.3747)	(0.2872)
Notes payable		-2.9995^{**}	
		(1.4995)	
Deferred charges		-4.5144^{***}	-4.6399^{***}
		(1.4028)	(1.3926)
Current assets total		0.3546^{*}	
		(0.1838)	
Inventories total		2.0196^{***}	
		(0.5412)	
Income taxes other		10.7655^{*}	
		(5.5112)	
Operating activities net cash flow		1.5242^{***}	
		(0.3308)	
Capital surplus			-0.1278
			(0.1261)
Revenue total			0.5792^{***}
			(0.1111)
Current income taxes			16.0525^{**}
			(6.6639)
Computer Software	-1.4750^{***}		
	(0.2213)		
Non-Metallic & Industrial Metal Mining	-3.7354^{***}		
	(0.7593)		
Textiles	2.8080^{***}		
	(0.7346)		
Steel Works Etc	1.8717^{***}		
	(0.5253)		
Beer & Liquor		-2.5572^{***}	
		(0.5810)	
Pharmaceutical Products		-1.4345^{***}	
		(0.3764)	
Retail		-0.3035	
		(0.3607)	
Long-term debt total Split	0.4388	0.2914	0.4127
	(0.2717)	(0.2831)	(0.2697)
Pretax income Split	-1.3091^{***}	-1.6486^{***}	-0.9352^{***}
	(0.1758)	(0.1698)	(0.1866)
Debt in current liabilities Split	0.3366^{***}		0.5886
	(0.3732)		(0.2817)
Liabilities total Split		0.3583	
		(0.2231)	
Observations	2,565	2,393	2,146
Pseudo R^2	0.601	0.572	0.529

Table B3: Logistic Regressions Zombie Firms vs. Recovered, Europe. The table shows the logistic regression results for zombies versus recovered for the three cross-sections. The dependent variable is a dummy variable which is equal to 1 whenever a firm is a zombie. The zombie (recovered) measure is constructed following Acharya et al. (2020). All explanatory variables are lagged with a two year lag. Standard errors are clustered at firm level. Variables ending with Split are those returned by the trees. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

	2007 **(1)**	2016 **(2)**	2020 **(3)**
Long-term debt total	-13.6533^{***} (1.5677)	-14.6848^{***} (1.7718)	-6.7672^{***} (1.4078)
Shareholders equity	3.1534^{***} (2.2449)	1.4094 (2.4068)	-0.8999 (2.0222)
Liabilities total	2.4911 (2.3172)	2.9699 (2.5354)	-0.6160 (2.0993)
Debt in current liabilities	-11.2113^{***} (2.5232)	-13.8730^{***} (3.0920)	-8.3273^{***} (1.6709)
Income before extraordinary items	43.5378 (45.8585)	-0.0349 (0.4541)	0.0744 (0.4263)
Pretax income	20.4700 (16.6163)	3.4581 (2.5613)	0.7404 (0.8346)
Income taxes total	-10.9521 (17.0167)		-0.0726 (0.4249)
Operating activities net cash flow	1.1859 (0.8360)		
Revenue total	0.3194^{***} (0.1115)	0.2379^{*} (0.1254)	0.4044^{**} (0.1663)
Accounts receivable	-3.0506^{**} (1.3549)		
Long-term debt due in 1 year		-5.2206 (4.6116)	
Nonoperating income		-4.4292^{*} (2.6346)	
Retained earnings		0.0915^{**} (0.0410)	0.0432 (0.0547)
Current income taxes			-3.1133 (3.7151)
Short-term investments total			-1.1952 (0.9437)
Accounts receivable total			-0.1475 (1.0394)
Electronic Equipment	-1.0346^{***} (0.3140)		
Communications		3.6269^{***} (0.7899)	
Petroleum & Natural Gas		3.7254^{***} (0.5471)	
Pharmaceutical Products		-0.4866 (0.4102)	-0.5068 (0.4852)
Long-term debt total Split	0.3396 (0.3218)	0.5721 (0.3820)	0.8948^{***} (0.3343)
Pretax income Split	-2.2863^{***} (0.2687)	-1.8268^{***} (0.2398)	
Debt in current liabilities Split	0.7494^{**} (0.3366)	-0.3037 (1.2045)	
Accounts receivable total Split			-0.4772 (0.6276)
Observations	1,456	962	819
Pseudo R^2	0.564	0.543	0.381

Table B4: Logistic Regressions Zombie Firms vs. Recovered, USA. The table shows the logistic regression results for zombies versus recovered for the three cross-sections. The dependent variable is a dummy variable which is equal to 1 whenever a firm is a zombie. The zombie (recovered) measure is constructed following Acharya et al. (2020). All explanatory variables are lagged with a two year lag. Standard errors are clustered at the firm level. The variables ending with Split are those returned by the trees. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B.2 Prediction Results

	Europe						United States					
Years	A	P	R	F1	CM		A	P	R	F1	CM	
2007	78.04	54.88	22.55	31.96	3969	231	74.90	58.84	38.80	46.76	1852	228
					965	281					514	326
2016	77.00	54.26	14.77	23.22	4022	161	78.22	58.94	16.70	26.02	1728	62
					1102	191					444	89
2020	74.49	57.22	30.60	39.87	3679	314	77.50	47.27	5.39	9.67	1644	29
					953	420					456	26

Table B5: Prediction Results Zombies vs. Non-Zombie Firms Logit. This table reports the prediction results for the class of zombie firms for both geographical areas, Europe and United States. The values, in %, show the prediction accuracy (A), precision (P), recall (R), the F1 score (F1), and the confusion matrix (CM) for the logistic regression model for the whole set of observations. For further details, we report a confusion matrix showing the non-zombies in the top line, the zombies in the bottom line, and in the diagonal the correctly estimated versus the incorrectly estimated zombies. Zombie firms are defined following Acharya et al. (2020), (see, Section 3).

	Europe						United States					
Years	A	P	R	F1	CM		A	P	R	F1	CM	
2007	88.88	88.01	90.75	89.36	1083	163	87.57	86.62	83.68	85.12	757	80
					122	1197					101	518
2016	87.76	85.93	87.72	86.76	1135	158	86.49	86.02	83.21	84.59	475	58
					135	965					72	357
2020	87.14	83.00	80.85	81.91	1245	128	82.54	81.09	75.07	77.96	423	59
					148	625					84	253

Table B6: Prediction Results Zombies vs. Recovered Zombie Firms Logit. This table reports the prediction results for the class of zombie firms for both geographical areas, Europe and United States. The values, in %, show the prediction accuracy (A), precision (P), recall (R), F1 score (F1), and confusion matrix (CM) for the logistic regression model for the whole set of observations. For further details, we report a confusion matrix showing the non-zombies in the top line, the zombies in the bottom line, and in the diagonal the correctly estimated versus the incorrectly estimated zombies. Zombie firms are defined following Acharya et al. (2020), (see, Section 3).

References

Acharya, Viral V., Matteo Crosignani, Tim Eisert, and Christian Eufinger. 2020. "Zombie Credit and (Dis-)Inflation: Evidence from Europe." *National Bureau of Economic Research,* no. 27158.

Acharya, Viral V., Tim Eisert, Christian Eufinger, and Christian Hirsch. 2019. "Whatever it Takes: The Real Effects of Unconventional Monetary Policy." *The Reveiw of Financial Studies* 32 (9): 3366–3411.

Allen, Franklin, Michael K.F. Chui, and Angela Maddaloni. 2004. "Financial systems in Europe, the USA, and Asia." *Oxford Review of Economic Policy* 20 (4): 490–508.

Altman, Edward I., Rui Dai, and Wei Wang. 2022. "Global zombies." *Available at SSRN 3970332.*

Banerjee, Ryan, and Boris Hofmann. 2018. "The rise of zombie firms: causes and consequences." *BIS Quarterly Review,* no. September.

———. 2020. "Corporate zombies : Anatomy and life cycle." *BIS Working Papers,* no. 882.

Bargagli Stoffi, Falco, Massimo Riccaboni, and Armando Rungi. 2020. "Machine Learning for Zombie Hunting. Firms' Failures and Financial Constraints."

Breiman, Leo. 2001. "Random forests." *Machine learning* 45 (1): 5–32.

Breiman, Leo, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and regression trees.* CRC press. ISBN: 0412048418.

Brezigar-Masten, Arjana, and Igor Masten. 2012. "CART-based selection of bankruptcy predictors for the logit model." *Expert Systems with Applications* 39 (11): 10153–10159.

Caballero, Ricardo J., Takeo Hoshi, and Anil K. Kashyap. 2008. "Zombie lending and depressed restructuring in Japan." *American Economic Review* 98 (5): 1943–1977.

Chen, Mu Yen. 2011. "Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches." *Computers and Mathematics with Applications* 62 (12): 4514–4524.

Custódio, Cláudia, Miguel A. Ferreira, and Luís Laureano. 2013. "Why are US firms using more short-term debt?" *Journal of Financial Economics* 108 (1): 182–212.

Dichev, Ilia D. 1998. "Is the Risk of Bankruptcy a Systematic Risk?" *Journal of Finance* 53 (3): 1131–1147.

Fama, Eugene F., and Kenneth R. French. 1992. "The Cross-Section of Expected Stock Returns." *The Journal of Finance* 47 (2): 427–465.

———. 1997. "Industry costs of equity." *Journal of Financial Economics* 43 (2): 153–193.

Favara, Giovanni, Camelia Minoiu, and Ander Perez-Orive. 2022. "Zombie Lending to U.S. Firms." *Available at SSRN.*

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2021. "Predictably Unequal? The Effects of Machine Learning on Credit Markets." *Journal of Finance, forthcoming.*

Gepp, Adrian, Kuldeep Kumar, and Sukanto Bhattacharya. 2010. "Business failure prediction using decision trees." *Journal of Forecasting* 29 (6): 536–555.

Hofer, Charles W. 1980. "Turnaround Strategies." *The Journal of Business Strategy* 1 (1): 1–19.

Hoshi, Takeo. 2006. "Economics of the living dead." *Japanese Economic Review* 57 (1): 30–49.

Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. 2010. "Consumer credit-risk models via machine-learning algorithms." *Journal of Banking and Finance* 34 (11): 2767–2787.

Liang, Deron, Chih Fong Tsai, and Hsin Ting Wu. 2015. "The effect of feature selection on financial distress prediction." *Knowledge-Based Systems* 73 (1): 289–297.

McGowan, Müge Adalet, Dan Andrews, and Valentine Millot. 2018. "The walking dead? Zombie firms and productivity performance in OECD countries." *Economic Policy* 33 (96): 685–736.

Platt, Harlan D., and Marjorie B. Platt. 2002. "Predicting Corporate Financial Distress: Reflections on Choice-Based Sample Bias." *Journal of Economics and Finance* 26 (2): 184–199.

Rajan, Raghuram G, and Luigi Zingales. 1995. "What Do We Know about Capital Structure? Some Evidence from International Data." *The Journal of Finance* 50 (5): 1421–1460.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society* 58 (1): 267–288.

Valta, Philip. 2016. "Strategic Default, Debt Structure, and Stock Returns." *Journal of Financial and Quantitative Analysis* 51 (1): 197–229.

Whitaker, Richard B. 1999. "The Early Stages of Financial Distress." *Journal of Economics and Finance* 23 (2): 123–132.

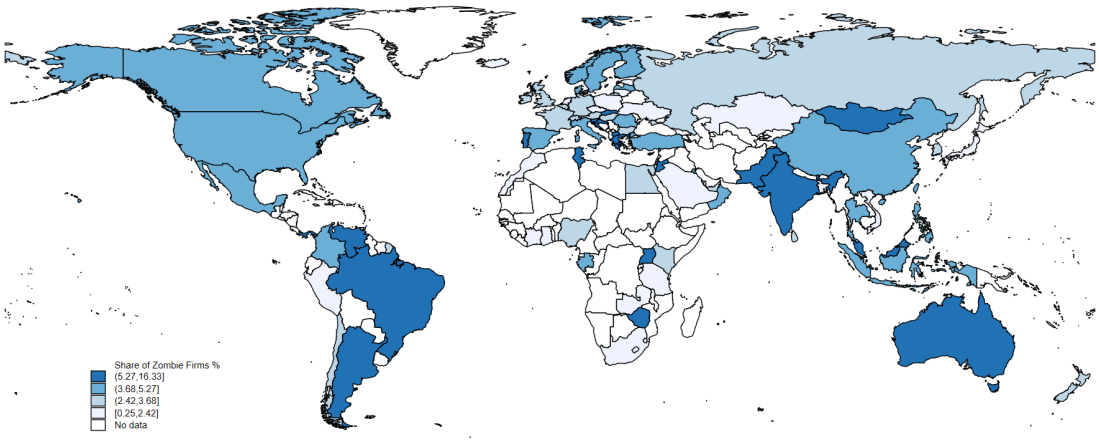

Figure 1: Global Zombie Shares. The map shows the presence of zombie firms by country and visualizes the share of zombie firms in the world. The map is scaled in different shades of blue according to the severity of the phenomenon. The countries for which we have no data are displayed in white color. The countries that register the highest share of zombies are in dark blue. Zombie firms are defined following Acharya et al. (2020), additional information on the identification of zombie firms is provided in Section 3. Source: Authors' projections on Compustat data.

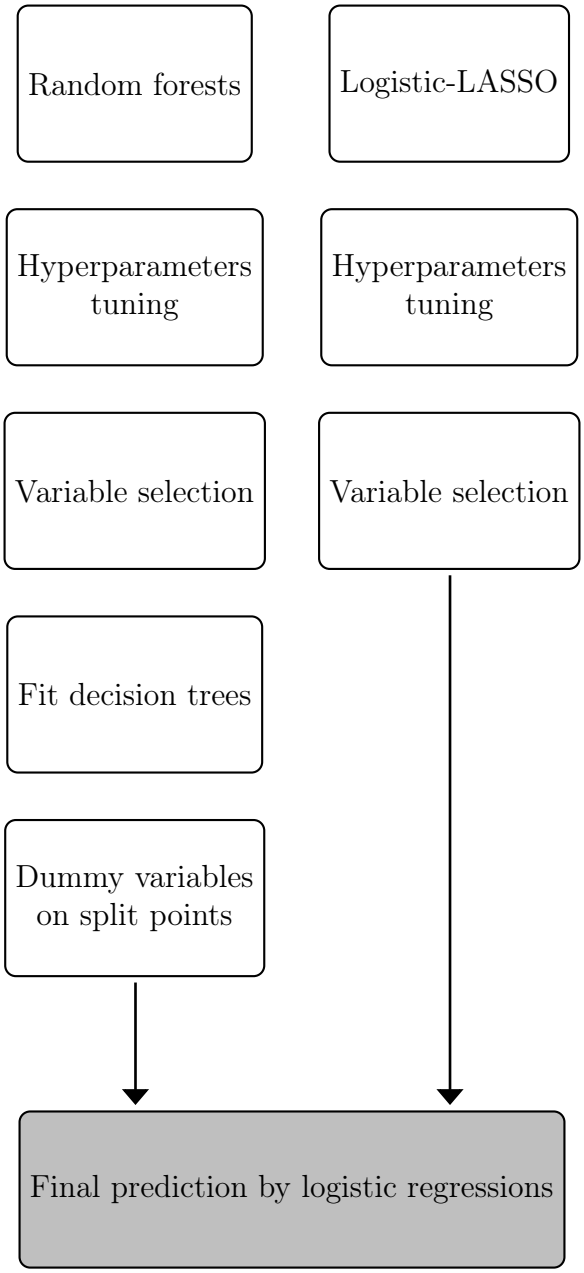

Figure 2: Empirical Strategy. The flow chart outlines our empirical strategy. We rely on variables selected by and derived from machine learning approaches. Our final logistic regression model combines this information and offers a simple but informative tool to predict a firm zombie (recovered) status.

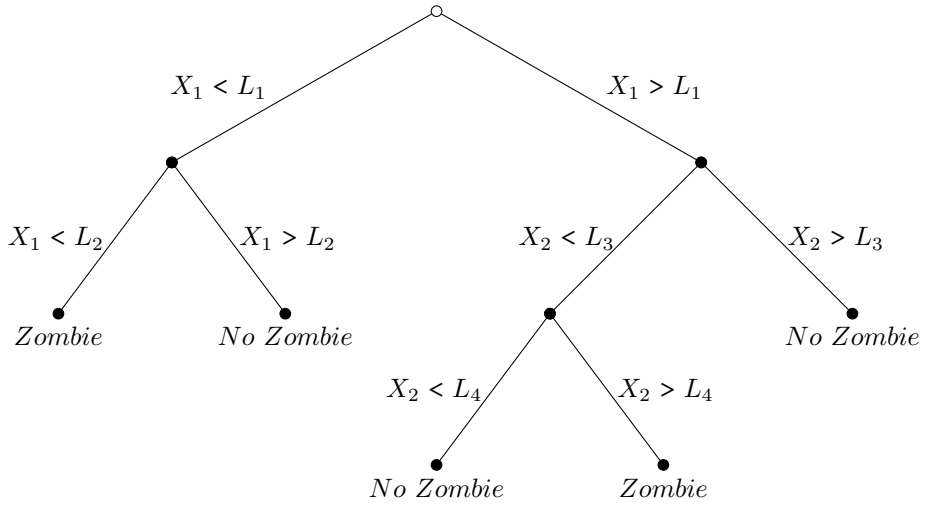

Figure 3: Tree Mechanism Example. The figure provides an example of classification and regression tree (CART) model with two outcomes, zombie firm versus no zombie firm. We use a recursive splitting algorithm that generates decision trees. Tree-based models partition the input space X into segments, where each segment decides for one class, and then fit a simple model into each one of them. The figure shows a tree example corresponding to such partition. As in any tree, the first white node at the top is the most important variable returned by the algorithm as it binary splits the whole dataset to classify a zombie firm from a non-zombie. The algorithm underlying the decision tree finds, at each iteration, the variable that can better classify a zombie firm and separate it from a non-zombie firm.

Panel A: Europe								
	Zombie Firms				Recovered Firms			
	Full	2007	2016	2020	Full	2007	2016	2020
Log(tot. assets)	5.643	5.392	5.372	5.502	5.253	4.997	5.478	5.671
Log(employment)	-0.205	-0.534	-0.790	-1.191	0.009	-0.274	-0.084	-0.734
Profit rate	0.004	0.018	0.005	-0.007	0.062	0.075	0.059	0.052
Size	5.103	4.758	4.551	4.650	5.308	4.994	5.362	5.378
Tangibility	0.199	0.174	0.137	0.175	0.166	0.123	0.105	0.103
Profitability	0.054	0.055	0.043	0.032	0.127	0.128	0.113	0.111
Leverage	0.352	0.341	0.341	0.367	0.046	0.037	0.044	0.056
Investments	0.167	0.252	0.157	0.100	0.219	0.299	0.212	0.124
Ebit ICR	1.301	1.701	1.424	0.434	14.235	18.781	22.874	18.621
Cash holdings	0.070	0.084	0.077	0.110	0.137	0.167	0.148	0.188

Panel B: United States								
	Zombie Firms				Recovered Firms			
	Full	2007	2016	2020	Full	2007	2016	2020
Log(total assets)	4.276	5.923	7.016	6.983	4.190	5.240	5.521	6.776
Log(employment)	-0.062	-0.040	0.897	0.660	-0.183	-0.454	-0.262	0.828
Profit rate	0.016	0.003	0.004	-0.035	0.075	0.067	0.041	0.043
Size	4.345	5.647	6.782	6.700	4.463	5.328	5.743	6.799
Tangibility	0.287	0.180	0.189	0.205	0.203	0.102	0.091	0.115
Profitability	0.091	0.073	0.078	0.043	0.162	0.129	0.101	0.096
Leverage	0.363	0.322	0.364	0.372	0.023	0.000	0.000	0.110
Investments	0.186	0.271	0.170	0.103	0.256	0.293	0.229	0.130
Ebit ICR	1.515	1.350	1.796	0.312	14.182	25.655	18.073	10.797
Cash holdings	0.048	0.070	0.066	0.123	0.162	0.265	0.222	0.168

Table 1: Zombies and Recovered Zombie Firms. The table reports descriptive statistics on a set of performance measures. Panel A presents the statistics for the European sample, while Panel B those for the U.S. sample. In both panels we report median values for the full observation period (1996-2020), the year 2007, 2016, and 2020. Profit rate is Net income/total assets, size is log(sale), tangibility is Total property, plant and equipment (net)/total assets, profitability is Operating income before depreciation/total assets, leverage is (Total liabilities/total assets), investments is Capital expenditures/lagged property, plan and equipment, Ebit ICR is Operating income after depreciation/interest expenses, cash holdings is cash and short-term investments/lagged total assets. Zombie firms are defined following Acharya et al. (2020), additional information on the identification of zombie, and recovered, firms is provided in Section 3. Source: Authors' calculations on Compustat data.

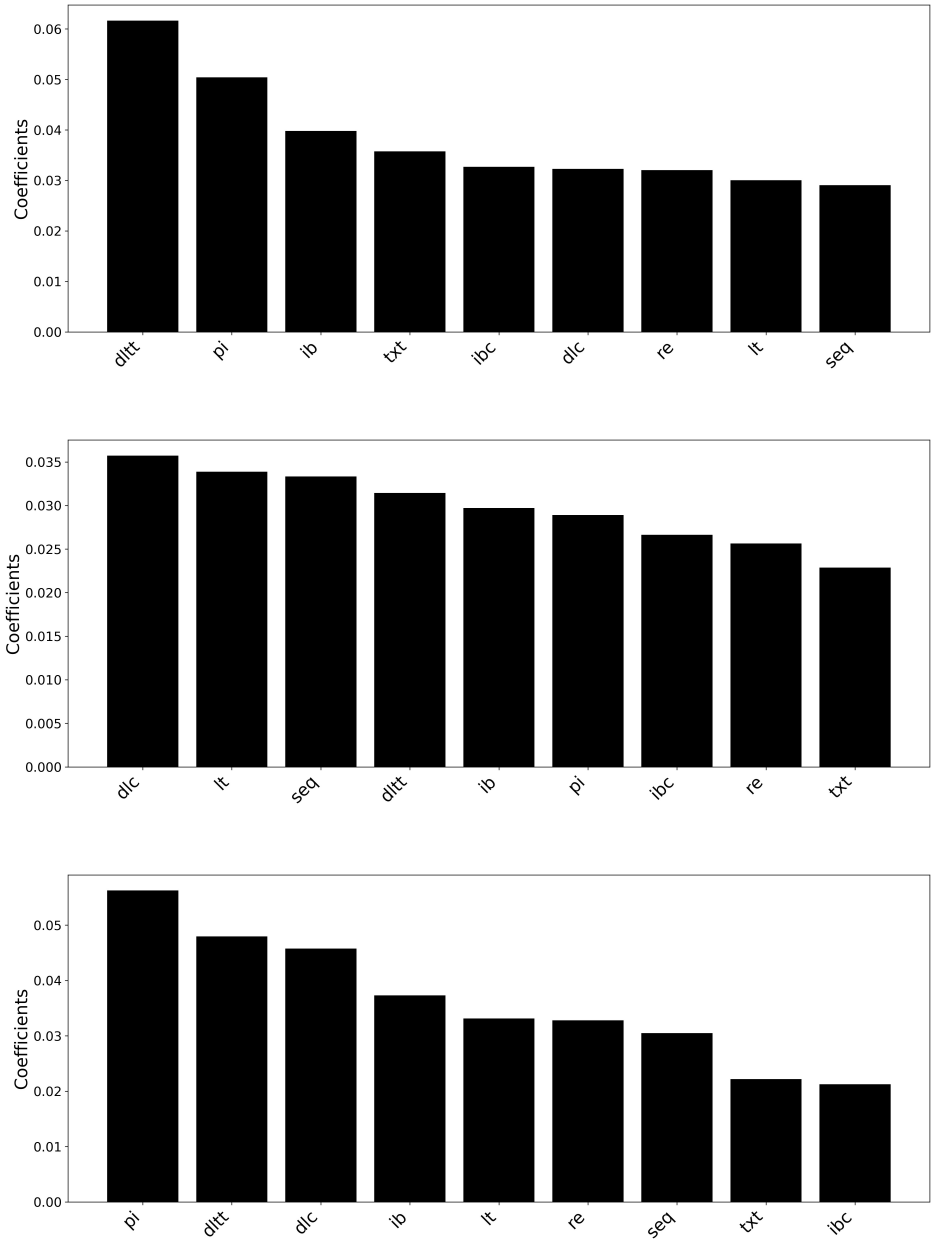

Figure 4: Random Forests Zombie Firms' Selected Features Importance - Europe 2007, 2016, and 2020. The figures report the results of the most important features returned by the random forests categorizing zombie firms in Europe in 2007, upper graph, 2016 middle, and 2020 bottom. The reported coefficients relate to the entropy, the higher the coefficient the higher the importance of the variables and their informativeness. Zombie firms are defined following Acharya et al. (2020), (see, Section 3).

Legend: *pi* Pretax Income, *ib* Income Before Extraordinary Items, *dltt* Total Long-Term Debt, *dlc* Debt in Current Liabilities, *lt* Total Liabilities, *seq* Total Shareholders Equity, *txt* Total Income Taxes, *ibc* Income before Extraordinary Items, *re* Retained Earnings.

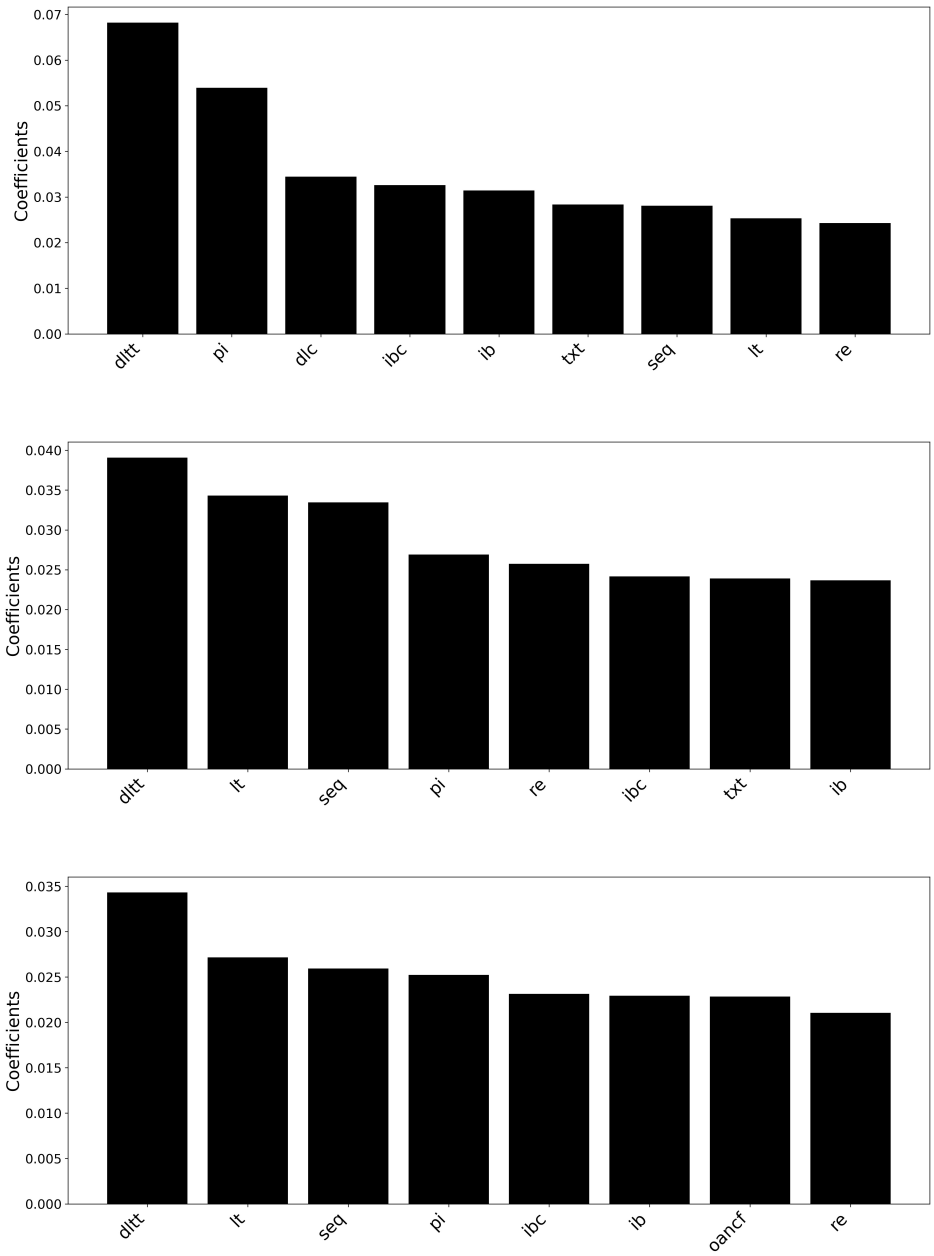

Figure 5: Random Forests Zombie Firms' Selected Features Importance - USA 2007, 2016, and 2020. The figures report the results of the most important features returned by the random forests categorizing zombie firms in the U.S. in 2007, upper graph, 2016 middle, and 2020 bottom. The reported coefficients relate to the entropy, the higher the coefficient the higher the importance of the variables and their informativeness. Zombie firms are defined following Acharya et al. (2020), (see, Section 3).

Legend: *pi* Pretax Income, *ib* Income Before Extraordinary Items, *dltt* Total Long-Term Debt, *dlc* Debt in Current Liabilities, *lt* Total Liabilities, *seq* Total Shareholders Equity, *txt* Total Income Taxes, *ibc* Income before Extraordinary Items, *re* Retained Earnings, *oancf* Operating Activities Net Cash Flow.

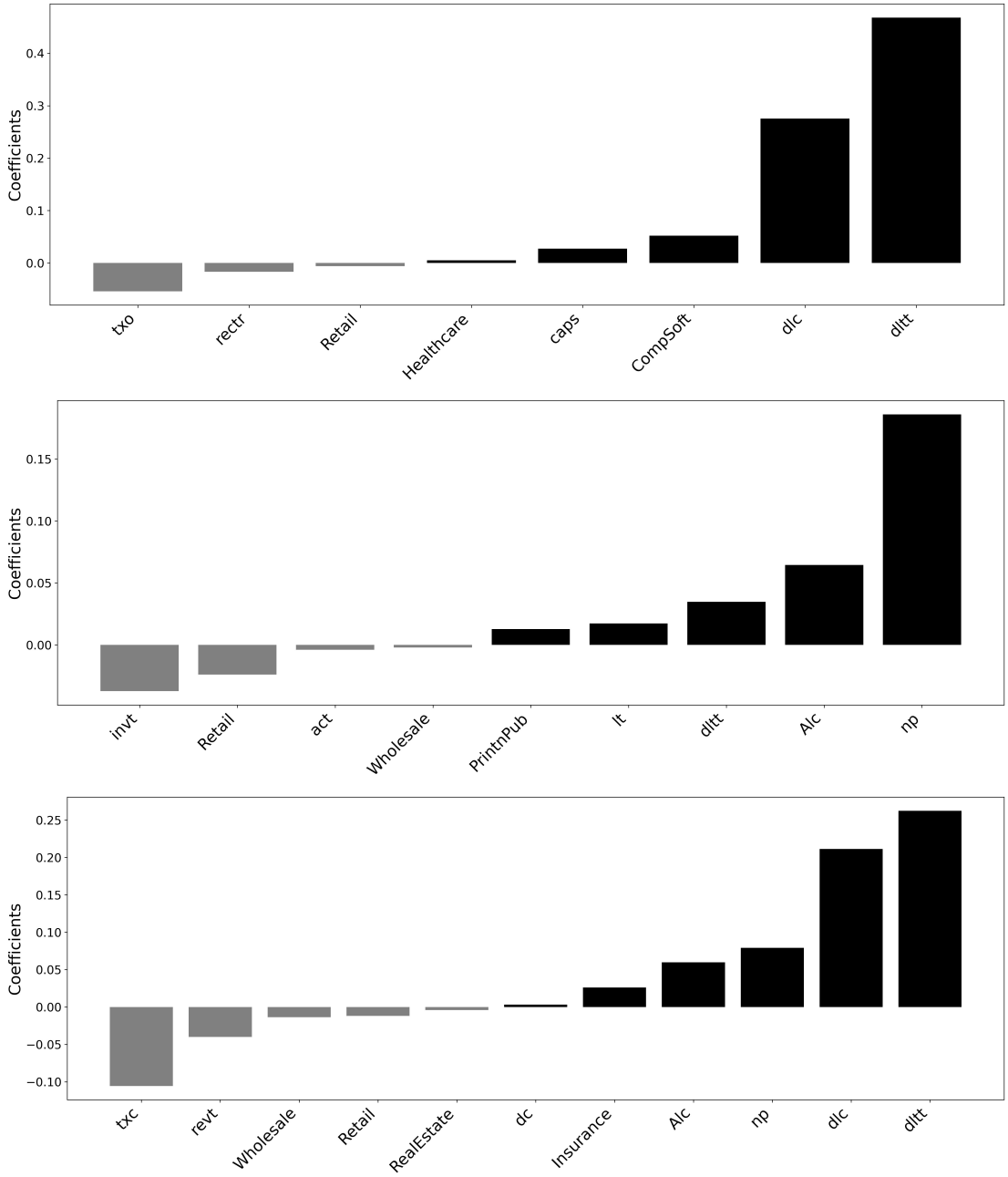

Figure 6: Lasso Zombie Firms' Selected Features Importance - Europe 2007, 2016, and 2020. The figures report the results of the most important features returned by Lasso categorizing zombies in Europe in 2007, upper graph, 2016 middle, and 2020 bottom. The higher the coefficient the higher the importance of the variables and their informativeness. Zombie firms are defined following Acharya et al. (2020), (see, Section 3).

Legend: *dltt* Total Long-Term Debt, *dlc* Debt in Current Liabilities, *lt* Total Liabilities, *np* Notes Payable, *caps* Capital Surplus, *dc* Deferred Charges, *rectr* Trade Accounts Receivable, *revt* Total Revenue, *txo* Other Income Taxes, *invt* Total Inventories, *txc* Current Income Taxes, $CompSoft$ Computer Software, $PrintnPub$ Printing and Publishing, *Alc* Beer and Liquor.

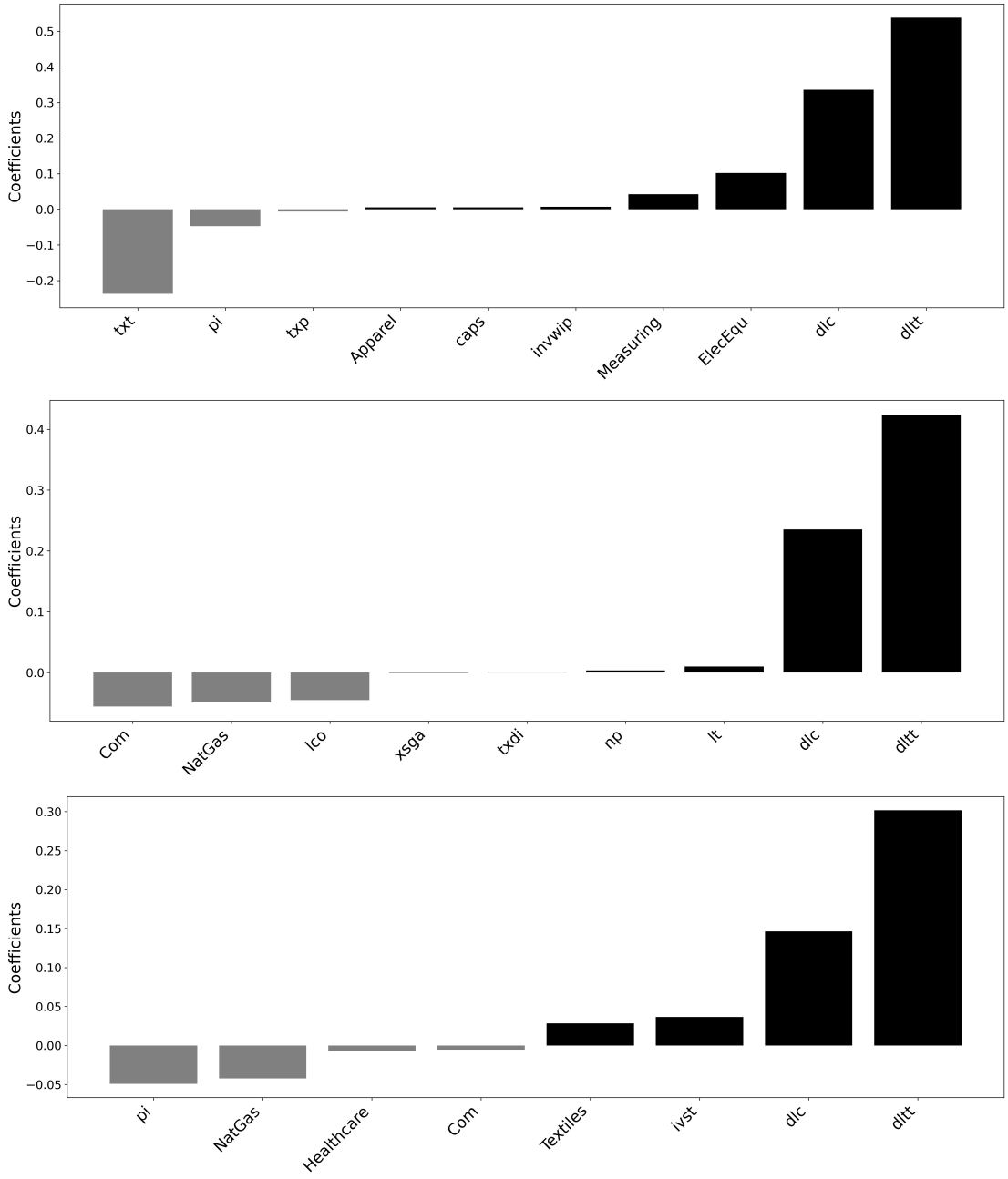

Figure 7: Lasso Zombie Firms' Selected Features Importance - USA 2007, 2016, and 2020. The figures report the results of the most important features returned by Lasso categorizing zombies in the U.S. in 2007, upper graph, 2016 middle, and 2020 bottom. The higher the coefficient the higher the importance of the variables and their informativeness. Zombie firms are defined following Acharya et al. (2020), (see, Section 3).

Legend: *dltt* Total Long-Term Debt, *dlc* Debt in Current Liabilities, *lt* Total Liabilities, *ivst* Total Short-Term Investments, *invwip* Inventories Work in Progress, *np* Notes Payable, *txdi* Deferred Income Taxes, *caps* Capital Surplus, *xsga* Selling General & Administrative Expense, *txp* Taxes Payable, *lco* Other Current Liabilities, *pi* Pretax Income, *txt* Total Income Taxes, *Measuring* Measuring and Control Equipment, *ElecEqu* Electronic Equipment, *Com* Communications, *NatGas* Petroleum and Natural Gas.

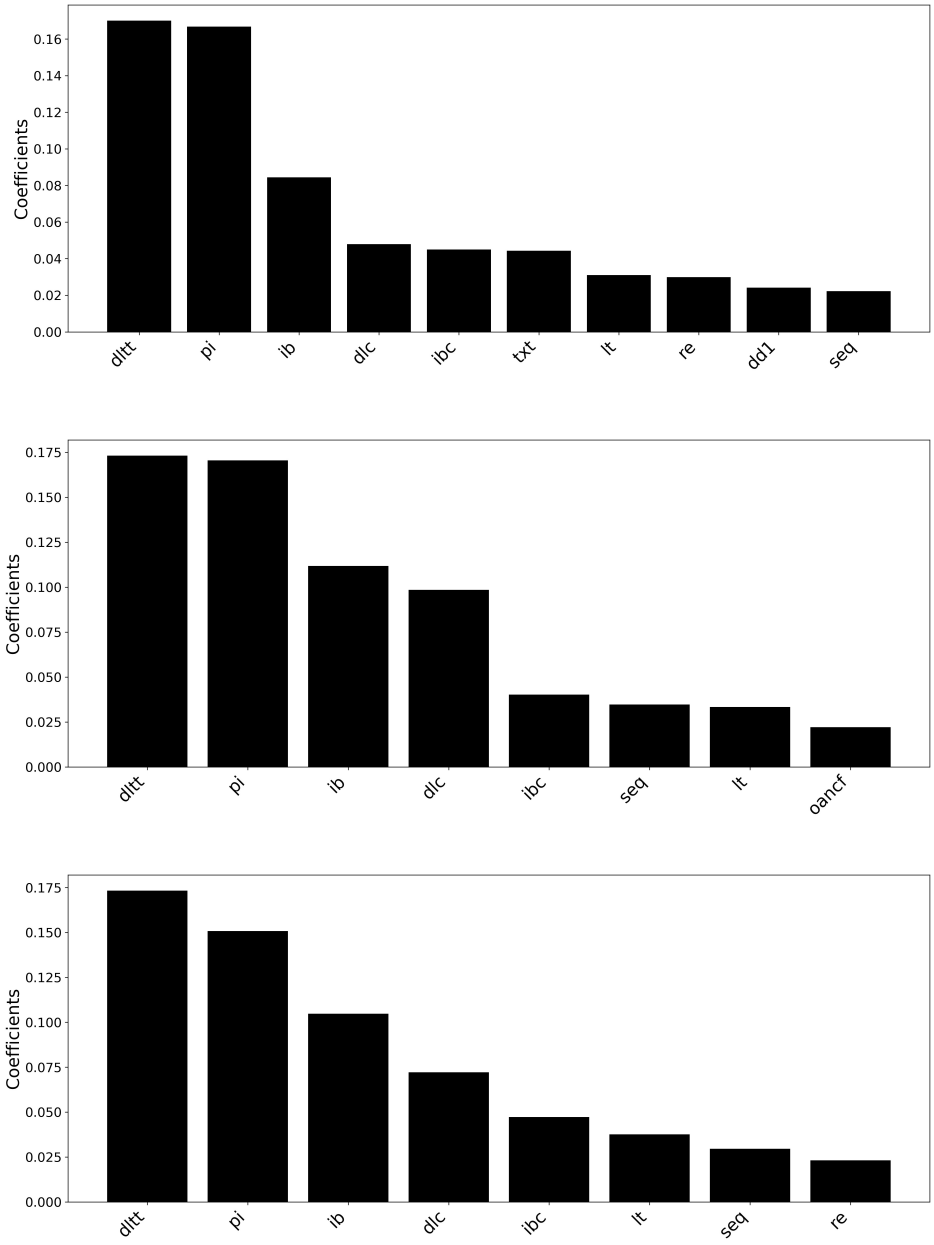

Figure 8: Random Forests Recovered Zombie Firms' Selected Features Importance - Europe 2007, 2016, and 2020. The figures report the results of the most important features returned by the random forests categorizing recovered zombies in Europe in 2007, upper graph, 2016 middle, and 2020 bottom. The reported coefficients relate to the entropy, the higher the coefficient the higher the importance of the variables. Information on the identification of recovered zombie firms is provided in Section 3.

Legend: *pi* Pretax Income, *ib* Income Before Extraordinary Items, *dltt* Total Long-Term Debt, *dlc* Debt in Current Liabilities, *lt* Total Liabilities, *seq* Total Shareholders Equity, *txt* Total Income Taxes, *ibc* Income before Extraordinary Items, *re* Retained Earnings, *oancf* Operating Activities Net Cash Flow.

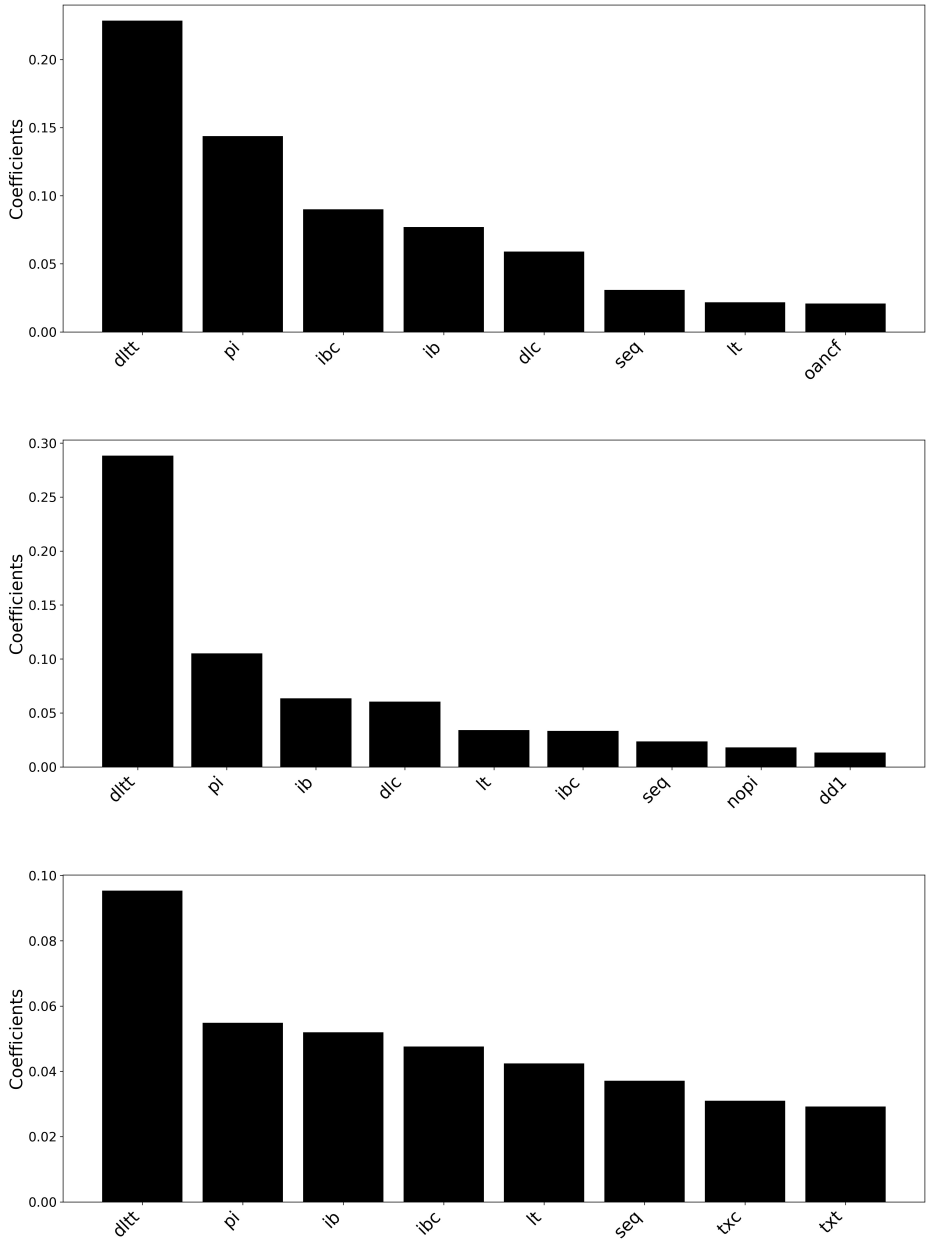

Figure 9: Random Forests Recovered Zombie Firms' Selected Features Importance - USA 2007, 2016, and 2020. The figures report the results of the most important features returned by the random forests categorizing recovered zombies in the U.S. in 2007, upper graph, 2016 middle, and 2020 bottom. The reported coefficients relate to the entropy, the higher the coefficient the higher the importance of the variables. Information on the identification of recovered zombie firms is provided in Section 3.

Legend: *pi* Pretax Income, *ib* Income Before Extraordinary Items, *dltt* Total Long-Term Debt, *dlc* Debt in Current Liabilities, *lt* Total Liabilities, *seq* Total Shareholders Equity, *txt* Total Income Taxes, *ibc* Income before Extraordinary Items, *nopi* Nonoperating Income, *oancf* Operating Activities Net Cash Flow, *txc* Current Income Taxes, *dd*1 Long-Term Debt due in 1 Year.

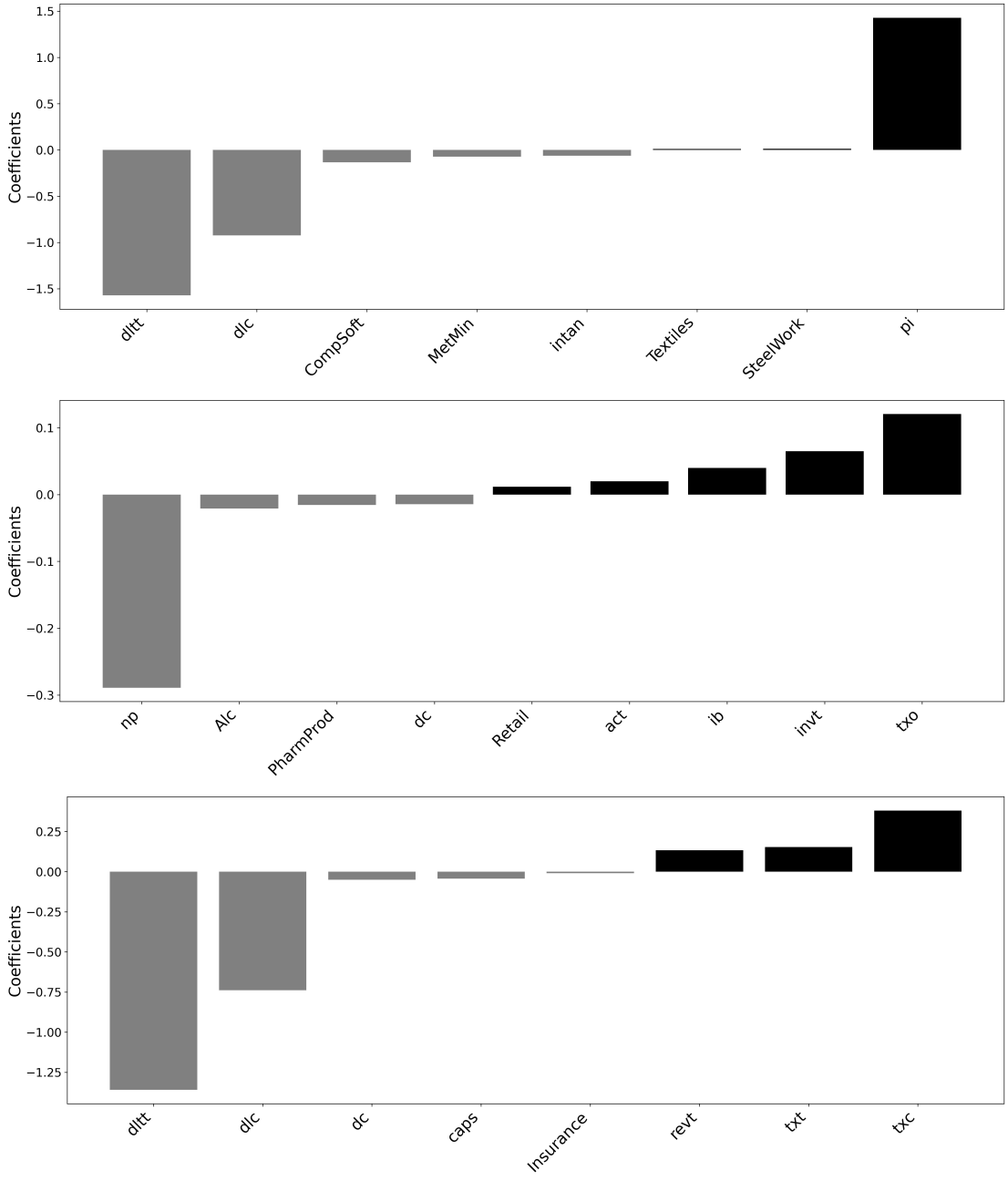

Figure 10: Lasso Zombies vs. Recovered Zombie Firms' Selected Features Importance - Europe 2007, 2016, and 2020. The figures report the results of the most important features returned by Lasso categorizing zombies in Europe in 2007, upper graph, 2016 middle, and 2020 bottom. The higher the coefficient, the higher the importance of the variables. Zombie firms are defined following Acharya et al. (2020), (see, Section 3).

Legend: *txo* Other Income Taxes, *txc* Current Income Taxes, *txt* Total Income Taxes, *invt* Total Inventories, *pi* Pretax Income, *ib* Income Before Extraordinary Items, *revt* Total Revenue, *act* Total Current Assets, *intan* Intangibles, *caps* Capital Surplus, *dc* Deferred Charges, *dltt* Total Long-Term Debt, *dlc* Debt in Current Liabilities, *np* Notes Payable, $CompSoft$ Computer Software, $MetMin$ Non-Netallic and Industrial Metal Mining, *Alc* Beer and Liquor, $PharmProd$ Pharmaceutical Products.

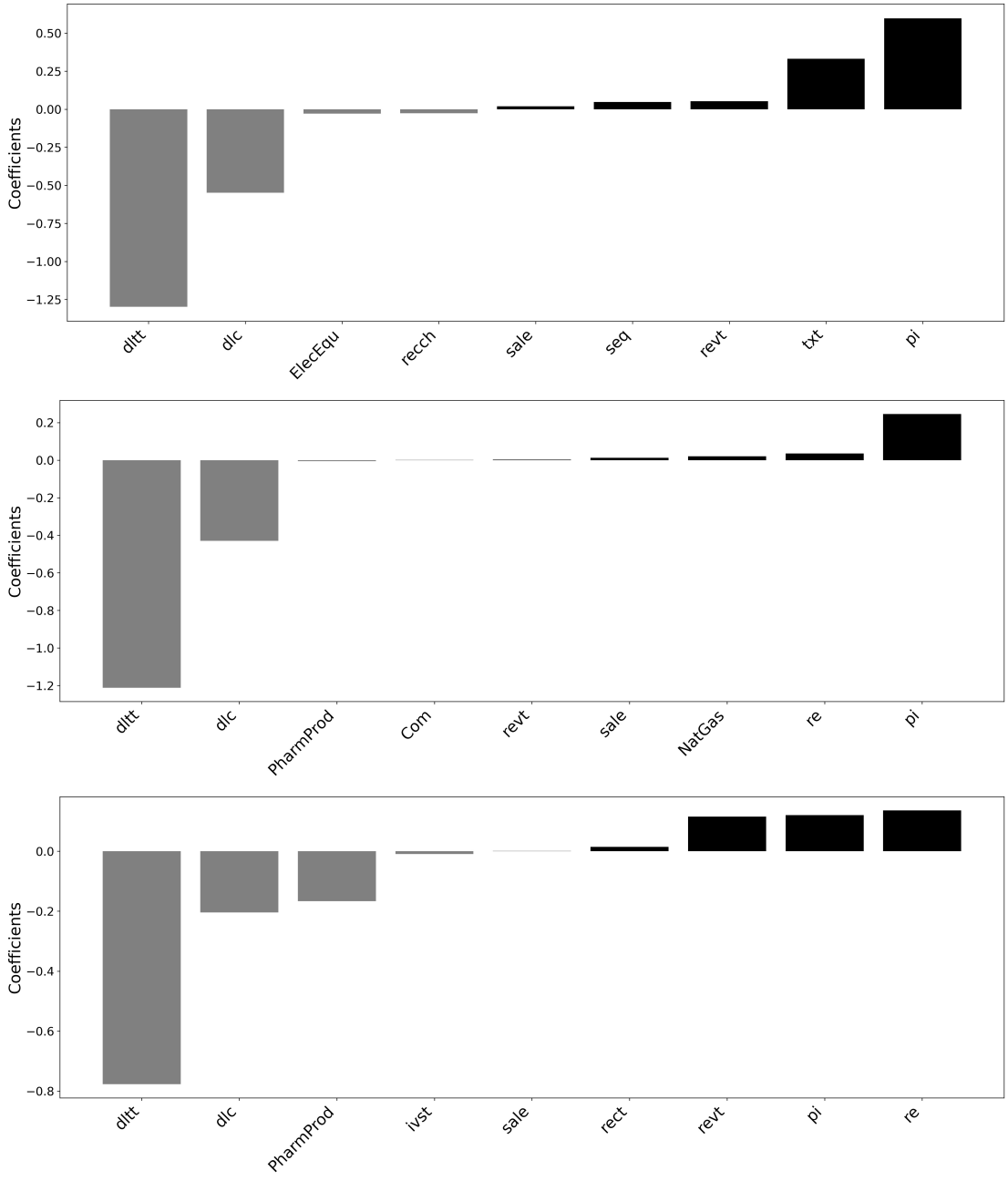

Figure 11: Lasso Zombies vs. Recovered Zombie Firms' Selected Features Importance - USA 2007, 2016, and 2020. The figures report the results of the most important features returned by Lasso categorizing zombies in the U.S. in 2007, upper graph, 2016 middle, and 2020 bottom. The higher the coefficient the higher the importance of the variables. Zombie firms are defined following Acharya et al. (2020), (see, Section 3).

Legend: *pi* Pretax Income, *txt* Total Income Taxes, *revt* Total Revenue, *rect* Total Accounts Receivable, *sale* Sales/Turnover, *recch* Accounts Receivable, *ivst* Total Short-Term Investments, *dltt* Total Long-Term Debt, *dlc* Debt in Current Liabilities, *re* Retained Earnings, *ElecEqu* Electronic Equipment, *PharmProd*, Pharmaceutical Products, *NatGas* Petroleum and Natural Gas.

	Europe						United States					
Years	A	P	R	F1	CM		A	P	R	F1	CM	
2007	77.36	63.93	19.31	29.66	593	22	75.80	61.54	49.23	54.70	268	40
					163	39					66	64
2016	75.18	51.72	14.56	22.72	588	28	76.80	50.00	16.05	24.30	255	13
					176	30					68	13
2020	75.44	60.81	22.61	32.96	517	29	78.70	75.00	4.22	8.00	252	1
					154	45					68	3

Table 2: Prediction Results Zombies vs. Non-Zombie Firms Logit. This table reports the prediction results for the class of zombie firms for both geographical areas, Europe and United States. The values, in %, show the prediction accuracy (A), precision (P), recall (R), the F1 score (F1), and the confusion matrix (CM) for the logistic regression model for a 15% out-of-sample test set. For further details, we report a confusion matrix showing the non-zombies in the top line, the zombies in the bottom line, and in the diagonal the correctly estimated versus the incorrectly estimated zombies. Zombie firms are defined following Acharya et al. (2020), (see, Section 3).

	Europe						United States					
Years	A	P	R	F1	CM		A	P	R	F1	CM	
2007	88.05	91.04	86.73	88.83	156	18	86.76	87.91	81.63	94.28	110	11
					28	183					18	80
2016	88.58	84.66	91.41	87.90	169	27	86.21	78.57	84.61	81.48	81	12
					14	149					8	44
2020	85.71	81.08	77.60	79.30	185	21	88.62	90.00	78.26	83.72	73	4
					26	90					10	36

Table 3: Prediction Results Zombies vs. Recovered Zombie Firms Logit. This table reports the prediction results for the class of zombie firms for both geographical areas, Europe and United States. The values, in %, show the prediction accuracy (A), precision (P), recall (R), F1 score (F1), and confusion matrix (CM) for the logistic regression model for a 15% out-of-sample test set. For further details, we report a confusion matrix showing the non-zombies in the top line, the zombies in the bottom line, and in the diagonal the correctly estimated versus the incorrectly estimated zombies. Zombie firms are defined following Acharya et al. (2020), (see, Section 3).