CS230

Predicting Company Zombification

Grzegorz Parosa
Graduate School of Business
Stanford University
gparosa@stanford.edu

Omar El Sadany
Graduate School of Business
Stanford University
omare@stanford.edu

Abstract

Zombie companies, i.e. permanently unproductive companies constitute one of the biggest challenges for developed countries. Building on existing literature, in this paper we try to predict company zombification using several neural network architectures. Our features include all variables commonly used in subject literature. Both, company-specific financial data as well as macroeconomic variables are used. We start with logistic regression implementation as the most commonly used in this type of problem in macroeconomics and then additional layers looking to improve the prediction accuracy of our network. Two versions of networks are trained and analyzed – ones with lagged dependent variable included and those without this feature. We find that more complicated networks that work on the normalized dataset, with regularization and varying layer sizes perform better than the logistic regressions generally accepted in economics, reaching an accuracy of 92.6% vs 89.1% for logistic regression when lagged Y is included and 74.7% vs 84.1% when this variable is not included.

1 Introduction

Zombies are companies that would go bankrupt in normal circumstances but instead continue to exist in inefficient forms for years. Zombification is one of the biggest problems tormenting developed economies in the last two to three decades. The problem is complex and not addressed well by traditional econometric methods as there are a lot of nonlinearities and interdependencies involved in the zombification process that are difficult to grasp using simple regreshttps://www.overleaf.com/project/60b93902369d6f5764f50985sions popular in economics. We believe that there is a space for deep learning in economic research that is methodologically stuck in the '80s. The goal of this project is to determine whether a company is a zombie or not based on a set of macroeconomic variables and a few of its financial multiples. The project is seen as an extension of Grzegorz Parosa's PhD dissertation focused on zombie companies in Europe after the year 2000. Because of general expectations for the economics community the problem in the dissertation was explained using multivariate regressions incorporating autoregressive components even though deep analysis of the data revealed significant non-linearities and interactions. Even though these problems were addressed by feature engineering, the predictive abilities of the models were limited. Moreover, because vast amounts of data regressions on the aggregate level were conducted, so the share of zombies in economies was estimated instead of zombification of single companies. We hope to extend that study on company-level data. The output of our algorithm is a prediction of whether a certain company each year is a zombie or not. Because of the lack of implementations of deep learning algorithms in economic research, we try several designs to find the one best suited for our research problem. In the next step, we plan to conduct counterfactual exercises to show how each macroeconomic variable influenced zombification and estimate response functions to changes in each feature.

2 Related work

In recent years, company zombification is an area of economic research. We build on top of the work done in the PhD dissertation mentioned before and research articles cited there. We use those economic papers to 1) understand what features to use and 2) build our architecture in a way that is similar to solutions used in the existing literature and then extend it using more advanced, unused economics techniques. Research used in this project comes mostly from two research centers: OECD and BIS. We want to start close to existing literature so that our results do not end up in a vacuum but can be compared with previously done research and extend our understanding of zombification. As mentioned earlier, economists have not yet adopted deep learning techniques so there are no "gold standards" applications of deep learning in this field. Instead, we approach this problem as classification one and used sources devoted to classification problems.

3 Dataset and Features

The data comes from the Orbis database and consists of 75 million standardized financial statements of companies from 24 European countries in the years 2000-2018. Data quality in the Orbis database is low and the database itself was not created with economics research in mind. Because of that extensive cleaning had to be done before undertaking the next steps. We removed all unrealistic data points (negative assets) or highly unlikely (unheard of profitability, returns on equity and assets, extremely high capital expenditures etc.) Additionally, we merged this dataset with macroeconomic indicators (GDP growth, interest rates, condition of banking sector, indices representing quality and severity of regulations concerning product markets, labor markets, and insolvency regimes for all countries and years). The macroeconomic indicators come from sources such as OECD, IMF, World Bank, ECB, etc. 75 million company-years with about 30 explanatory variables are impossible to process in a reasonable time. Because of that, we decided to select about 1 million observations. As the quality of the Orbis dataset is low, there are many missing data points. We started by removing company-years with missing any financial or macro data points. After this step, about 3 million observations were still left in the dataset. We are aware that this approach skews the dataset towards large companies but accept that as those companies are more important in creating the negative impact zombies have on the whole economies. In the second step, we randomly selected observations so that there are no more than 5.000 coming from one country-year. In this way we make this dataset a bit more balanced (though still far from that) and manage to bring the observation count down to 479,279, a number we feel comfortable with. We kept companies with missing macroeconomic variables as removing those would result in removing whole country-years and thus could distort our results. Instead, we want the neural network to deal with this problem. Originally, the dataset was structured as a panel. We changed that by introducing variables with lagged by one-year observations (more is unnecessary in this case, we are not aware of studies that would include autoregressive models other than AR(0) in this field). Thanks to that approach we are not limited to libraries and network designs capable of interpreting time series. To avoid problems related to omitted variables, we employ a solution often used in econometric regressions and include lagged dependent variables in our dataset. We expect this to greatly increase the accuracy of our model and decrease the importance of other variables as a consequence. We will discuss the results of that move in greater detail later. Results for a version with and without that variable are presented separately. Altogether we have 43 explanatory variables that we use to predict categorical variables representing zombification. There are two groups of variables – financial ratios and macroeconomic indicators, both for the current and the previous year. Moreover, we have 24 dummy variables representing countries and three representing company sizes. The structure of the dataset is presented in Appendix B and variable descriptions in Appendix A. The dataset was later divided into training and test subsamples. We decided to use 95% of observations to train and 5% to test as we had a significant number of observations at our disposition.

4 Methods

Because there are no research papers employing deep learning to this type of economics problem we start with the simplest possible setup – logistic regression implemented using neural network. This design choice allows us to 1) root our results in existing research, 2) establish reference point to which later we compare accuracy of more complex networks. This network is composed of input

layer and output layer activated by sigmoid function connected with all inputs. The entering data has dimensions of 43x1x455315. This, and all subsequent attempts were build using Keras API and TensorFlow 2.0 library in Python. We decided to employ Adam optimizer with default values of parameters beta1, beta2, and epsilon. We tried different values of alpha from a very wide range of values and found out that smaller values between 0.001 and 0.01 generally worked better, we decided to use 0.001. It turned out later that this value of alpha performed well in all subsequently checked architectures. As this is classification problem and our output is a binary variable we used binary crossentropy as a loss function. Its equation is given below:

$$\text{Loss} = -\frac{1}{\text{m}} \sum_{i=1}^{\text{m}} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i)$$

We used batched optimization with batch size of 10000. We experimented with different values and this proved to be fast and accurate. We also tuned number of epochs starting with lower numbers and increasing them as long as additional runs improved accuracy on the training set. After the accuracy stagnated for a few epochs or started to go down we stopped adding epochs. In the next step we decided to extend our network by four fully connected layers of 43 nodes each. Each of those layers was activated by relu function. In next step normalization of input parameters was added. We decided to implement this step as the results were slightly better even though our input data was fairly regular with mostly binary variables and multiples and indicators ranging between -1 and 1. We used standard normalization to bring all averages of our samples x to 0 and all their standard deviations to unit value:

$$\text{z} = \frac{\text{x} - \text{u}}{\text{s}}$$

where z is the normalized sample, x its original value, u average of training samples, and s standard deviation of training samples. In the next step every relu layer was regularized using L1 and L2 regularization with parameters l_1 = 1e-5 and l_2 = 1e-4.This step changed our loss function by adding:

$$\text{Loss}' = \text{Loss} + l_1 \sum_{i=1}^{\text{N}} |w_i| + l_2 \sum_{i=1}^{\text{N}} w_i^2$$

Experiments with more than four hidden layers did not improve accuracy on the test sample so were not added. However, we managed to slightly improve the network's performance by changing the number of nodes in each hidden layer. The final shape of our network is shown in Figure 1. We come up with node numbers by checking multiple combinations and comparing training sample accuracy.

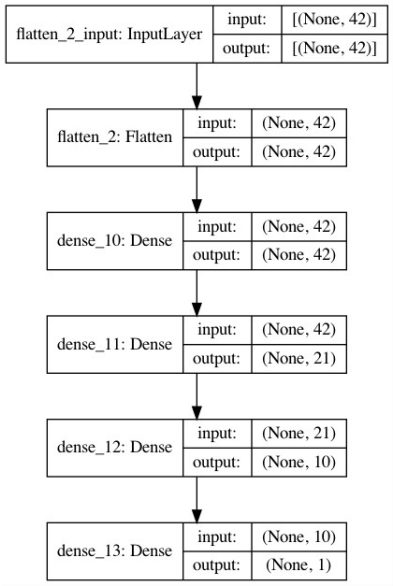

Figure 1: Architecture of final network

5 Results

All variants were tested with lagged zombie variable and without it. We decided to present results this way as the benefits of including lagged variables are not clear to us. On one side it helps to address the problem of omitted variables present in almost on economic studies and significantly increases accuracy. On the other hand, in real-life conditions knowledge about zombie status in the previous year may not be available (data from several previous years is necessary) and the obvious correlation between this and other features makes counterfactuals we intend to do later more difficult. The accuracy obtained on train and test samples for each variant of the network is presented in Table 1 below. Figure 2 displays AUC curves for AR(1) and non-AR versions as well as for logistic regressions and networks presented in Figure 1. Figure 3 presents the confusion matrix for all four variants. The neural network is better at classifying zombies in all cases. The differences are especially visible in the more difficult and useful case without AR(1) component.

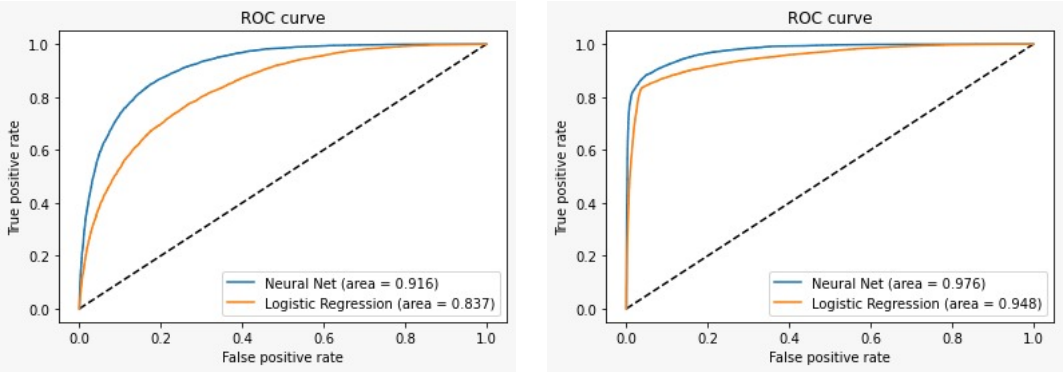

Figure 2: Left: AUC graph for logistic regression and best neural network with no AR component Right: AUC graph for logistic regression and best neural network with AR(1) component

Architecture	AR(1)	Regularization	Normalization	Train Accuracy	Test Accuracy
Logistic regression	No	No	No	76.8%	74.7%
Logistic regression	Yes	No	No	89.1%	89.1%
Relu 42x42x42	No	Yes	No	83.8%	83.7%
Relu 43x43x43	Yes	Yes	No	91.6%	91.4%
Relu 200x1000x200	No	Yes	Yes	83.7%	83.8%
Relu 200x1000x200	Yes	Yes	Yes	92.6%	92.4%
Relu 42x21x10	No	Yes	Yes	84.1%	84.1%
Relu 43x21x10	Yes	Yes	Yes	92.8%	92.6%

Table 1: Prediction accuracy on test and train samples using different network architectures

		No AR(1)		AR(1)	
		non-zombie	zombie	non-zombie	zombie
Neural Net	non-zombie	56%	7%	61%	2%
	zombie	9%	28%	6%	31%
Logit	non-zombie	57%	6%	60%	3%
	zombie	18%	19%	6%	31%

Figure 3: Confusion matrix. Results for logistic regression and best architecture are shown

6 Conclusion/Future Work

In this project, we managed to prove that deep learning algorithms can outperform techniques commonly used in economic research and solve problems hard to address using those methods due to data high volumes. Our still simple networks significantly outperformed logistic regressions both in AR(1) variant and without autoregressive component by, respectively 2% and 10% as measured by prediction accuracy on the test set. Predicting zombification itself has limited applications though it may help policymakers make the right decisions faster. In the next step, we plan to use this model to conduct counterfactual exercises and estimate response functions to changes in features we employed. This should deepen our understanding of zombification and allow us to prepare policy recommendations. We are especially interested in the impact of ultra-low interest rate policy as that was the main focus of Grzegorz's dissertation.

7 Contributions

Omar El Sadany - additional data cleaning, coding, architecture design, managing github repository and AWS

Grzegorz Parosa - definition of the research problem, dataset preparation and cleaning, tweaking network architectures, fine tuning hyperparameters

8 References

McGowan, Muge Adalet, and Dan Andrews. 2018. 'Design of Insolvency Regimes across Countries', September.

Banerjee, Ryan Niladri, and Boris Hofmann. 2018. 'The Rise of Zombie Firms: Causes and Consequences'. BIS Quarterly Review.

Brownlee, Jason. 2021. 'Gentle Introduction to the Adam Optimization Algorithm for Deep Learning'. https://machinelearningmastery.com/

Caballero, Ricardo J., Takeo Hoshi, and Anil K. Kashyap. 2006. 'Zombie Lending and Depressed Restructuring in Japan'. Working Paper 12129. National Bureau of Economic Research. Chollet, F. others, 2015. Keras. Available at: https://github.com/fchollet/keras.

Cook, Thomas. 2019. 'Macroeconomic Indicator Forecasting with Deep Neural Networks'. 2019 Meeting Papers 402. Society for Economic Dynamics.

Coulomb, Philippe Goulet, Maxime Leroux, Dalibor Stevanovic, Stéphane Surprenant.2019. 'How is Machine Learning Useful for Macroeconomic Forecasting?'. University of Pennsylvania, Université du Québec à Montréal.

Diederik, Kingma, Jimmy Ba. 2015. 'Adam: A Method for Stochastic Optimization'. Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego.

Liu, Ernest, Atif R. Mian, and Amir Sufi. 2020. 'Low Interest Rates, Market Power, and Productivity Growth'. SSRN Scholarly Paper ID 3320551. Rochester, NY: Social Science Research Network.

Maehashi, Kohei Shintani, Mototsugu. 2020. 'Macroeconomic forecasting using factor models and machine learning: an application to Japan', Journal of the Japanese and International Economies, Elsevier, vol. 58(C).

McGowan, Muge Adalet, Dan Andrews, and Valentine Millot. 2017. 'The Walking Dead?: Zombie Firms and Productivity Performance in OECD Countries'. OECD Economics Department Working Paper 1372. OECD Publishing.

Parosa, Grzegorz. 2021. 'Zombie companies and ultra low interest rate policy in Europe'. Warsaw School of Economics. Unpublished.

Tölö, Eero. 2020. 'Predicting systemic financial crises with recurrent neural networks,' Journal of Financial Stability, Elsevier, vol. 49(C).

Yadav, Saurabh. 2018. 'All you need to know about Regularization'.towardsdatascience.com.

9 Appendix A: Data description

Variable	Description
ctryiso	24 dummies representing countries for each company-year
year	year
size	Three dummies representing small, medium and large companies for a company-year
zombie_moj_fixed_3y	1 if zombie and 0 otherwise for each company-year
ROA_EBIT	Return on assets for each company-year
t	Capital expenditure to total assets for each company-year
Eq	Shareholder funds to total assets for each company-year
LP.diff	Year on year change in index representing labour protection regulations for each country-year
IR	Number of years it takes to liquidate a company for each country-year
PBV	Average price to book value multiple for listed banks for each country-year
GDP.GR	GDP growth for each country-year
STIR.positive	Short term interest rate for each country-year increased by a constant so that all observations are positive
min.SD	Optimal risk level of new investment projects measured by its standard deviation calculated in other study
.l	All above variables lagged by one year

10 Appendix B: Data example

	closdate_year	zombie_moj_fixed_3y	ROA_EBIT	t	Eq	min.SD.l	LP.diff	IR	PBV	GDP.GR	GDP.GR.l	STIR.positive	PMR.SC.diff	PMR.BE.diff	PMR.BT.diff
0	2002	0	0.01	0.17	0.15	0.33	0	0	0	0.02	0.02	0.04	0	0	0
1	2002	0	0.07	0.11	0.17	0.33	0	0	0	0.02	0.02	0.04	0	0	0
2	2002	1	0.02	0.19	0.13	0.33	0	0	0	0.02	0.02	0.04	0	0	0
3	2002	0	0.07	0.10	0.40	0.33	0	0	0	0.02	0.02	0.04	0	0	0
4	2002	0	0.06	0.15	0.30	0.33	0	0	0	0.02	0.02	0.04	0	0	0

	zombie_moj_fixed_3y.l	Eq.l	t.l	ROA.EBIT.l	STIR.positive.l	PBV.l	ctryiso_BE	ctryiso_CH	ctryiso_CZ	ctryiso_DE	ctryiso_DK	ctryiso_ES	ctryiso_FI	ctryiso_FR	ctryiso_GB
0	0	0.13	0.06	-0.02	0.05	0	0	0	0	0	0	0	0	0	0
1	0	0.20	0.07	0.10	0.05	0	0	0	0	0	0	0	0	0	0
2	1	0.14	0.04	0.02	0.05	0	0	0	0	0	0	0	0	0	0
3	0	0.40	0.04	0.08	0.05	0	0	0	0	0	0	0	0	0	0
4	0	0.28	0.10	0.06	0.05	0	0	0	0	0	0	0	0	0	0

	ctryiso_GR	ctryiso_HU	ctryiso_IE	ctryiso_IS	ctryiso_IT	ctryiso_LT	ctryiso_LU	ctryiso_LV	ctryiso_NL	ctryiso_PL	ctryiso_PT	ctryiso_SE	ctryiso_SI	ctryiso_SK	size_large
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

	size_medium	size_micro	size_small
0	0	0	0
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0