Erasmus University Rotterdam
Erasmus School of Economics

ZANDERS
PERFORMANCE WHEN IT COUNTS

# Dynamic Stress Testing: a Two Stage Clustering Approach Using Interpretable Machine Learning

Virgjil Karaja (566701), Jakub Láža (680687), Kayra Ozyar (652313), Peter Pedersen (481001)

**Abstract**

This paper performs a stress testing exercise on a portfolio of European SME loans. First, we compare the performance of Random Forest against the current standard of Logistic Regression, and show an outperformance in AUC of 0.25. Second, we introduce regional-sector clusters which create additional interpretability on the most important variables that drive defaults. Our results suggests that there is a great heterogeneity in the variables explaining defaults across different regions and sectors. Finally, we model characteristics as a function of macroeconomic variables, and show that not taking into account these dynamics leads to an underestimation of the stressed probability of default by 33%.

ERASMUS UNIVERSITEIT ROTTERDAM

| | |
|---|---|
| Supervisor: | dr. Maria Grith |
| Date final version: | 17th March 2024 |

# Contents

# 1 Introduction

The number of corporate defaults has approached the highest level since the 2008 crisis (S&P Global, 2024). Whilst this sounds alarming, one comforting fact is that banks are nowadays more regulated than they were back in 2008. For example, the emergence of Basel II and Basel III made stress testing an industry requirement. The goal of stress testing is to gain insights on the resilience of banks under adverse economic scenarios. The EBA develops adverse macroeconomic scenarios which are subsequently applied in a probability of default (PD) model. The idea is to investigate how much defaults a bank expects to get when a new crisis would emerge. The predicted PDs can then be used to calculate the Risk Weighted Assets (RWA) which is used to calculate how much capital a bank needs to hold.

The current standard for PD modeling is to use a Logistic Regression (Logit) (Westgaard & van der Wijst, 2001; Tserng et al., 2014; Orlando & Pelosi, 2020). These models are simple to apply, transparent in their predictions and suitable for stress testing scenarios, as shown by Chan-Lau (2006) and Simons & Rolwes (2018). A shortcoming of Logistic Regression models however is that they tend to have lower prediction performance than Machine Learning methods (Brown & Mues, 2012; Hamori et al., 2018; Leo et al., 2019). The first point of investigation of this paper is therefore how tree based Machine Learning (ML) methods perform in PD predictions compared to the commonly used approach of Logistic Regression, specifically we focus on the Random Forest (RF) of Breiman (2001). We overcome the challenge of interpretability for Machine Learning methods (Institute of International Finance, 2019), by using Shapley values (Lundberg & Lee, 2017). In this manner, we utilize a set of inherently interpretable ML approaches in the spectrum of PD modelling and assess their improvement upon the industry-standard approach of Logit.

The set of factors used in the corporate PD literature is dense, consisting of macroeconomic variables and firm-specific variables (Miu & Ozdemir, 2008; Castrén et al., 2010; Alonso & Carbo, 2020, 2022; Barboza et al., 2017). In particular, the work of Miu & Ozdemir (2008) and Castrén et al. (2010) highlights the importance of country and sector specific variables on PD modelling. We argue that assuming that the factors impacting the PD across different countries and or sectors are the same, is a strong assumption, which might not hold in practice. That is why we propose using statistical clustering methods to discover cluster specific factors influencing PD. We use an iterative clustering procedure where we first cluster based on country using a wide range of macroeconomic variables, and subsequently on sectors, allowing for across cluster comparison. Our proposed method is flexible, since one can choose to cluster using different methods at each stage, to account for different data structures.

A further challenge we address is the linkage of macroeconomic indicators and firm-specific variables within the context of stress testing. PD models are based on macroeconomic indicators in order to be able to perform the mandated European Banking Authority (EBA) stress tests. A disadvantage of only using macroeconomic indicators is that it does not consider default heterogeneity at the firm-level. Incorporating firm-specific variables in stress test scenarios can thus be useful to increase the credibility and realism of the stress testing exercise (European Banking Authority, 2023). An important remaining question is however how one should deal with the firm characteristics under the stress test scenarios. For simplicity, it is often assumed that characteristics follow a random walk (Zanders, 2023). We argue this gives overly optimistic stress test results, since it disregards the relations between the macroeconomic variables and the firm characteristics., e.g. a lower GDP is leads to lower sales of a company. We use LASSO Regression and Random Forest to model characteristics as a function of their one year lag and a broad range of macroeconomic variables. This allows us to perform ”dynamic” rather than ”static” stress testing.

Our results show that RF performs substantially better than LASSO Logit in predicting PDs, with an AUC of 0.85 vs 0.60 respectively. These findings are in line with the broader Machine Learning literature which shows the predictive superiority of such methods (Brown & Mues, 2012; Hamori et al., 2018; Leo et al., 2019). We further notice that our regional-sector clusters aid interpretability, as we find different important variables in different clusters. Applying the PD models to the total data would mean that these cluster specific effects would be missed. For example, our Northern Europe Commodities cluster highlights the importance of oil & gas prices, whilst these variables do not play an important role when we apply the model on the total data. With this, we find further support for the importance of sector and region as indicated by Miu & Ozdemir (2008) and Castrén et al. (2010), but add that clustering helps PD models by further highlighting different important variables across regions and sectors. Our clustering results further suggest that region level groupings in our data are elliptical shaped whilst sector level groupings are spherical shaped, which acts as support for the created value of the introduced flexibility of our method and adds on previous work on multi-level clustering (Bijmolt et al., 2004; Mo et al., 2010). Finally, we show that modeling firm characteristics as a function of their lags and macroeconomic variables, leads to an improvement in RMSE over a random walk. For the most important six characteristics, Dynamic LASSO offers an average RMSE improvement of 26%. We apply the predicted characteristics in our stress testing and show that this dynamic approach leads to higher stressed PDs than the static approach (1.72% vs 1.29%). This finding suggests that the static approach underestimates the stress, since it does not take into account any relationships between characteristics and macroeconomic variables. Our Dynamic LASSO approach acts a further extension on work done by Chen (2010); Issah & Antwi (2017); Vieira et al. (2019); Egbunike & Okerekeoti (2018).

The remainder of the paper is organized in the following manner. In Section 2, we provide the relevant literature connected to the topic of credit risk modelling and stress testing. We introduce the data in Section 3. Later, in Section 4 we go over the methodology, starting with the PD modelling, continuing with the iterative clustering approach, ending with the stress testing and performance evaluation criteria. We provide our main findings in Section 5 before concluding in Section 6.

## 2 Literature Review

### 2.1 Clustering

To create regional-sector clusters, we draw on previous work in the field of marketing. The literature on creating multilevel clusters stems from the objective of marketeers to create multi-region customer segments. The motivation for such a clustering approach comes from previous findings which show that the customer segmentation of one region cannot be directly adopted by another region (Mo et al., 2010). A commonly used approach to create these multi-region customer segments is to first group regions by geographic proximity and subsequently perform customer segmentation on the resulting regions (Bijmolt et al., 2004; Mo et al., 2010). Bijmolt et al. (2004) apply a mixture model in both stages. The novelty of this approach is that the relative sizes of latent classes (consumer segments) depend on country segment membership. More recent work by Mo et al. (2010) uses different models in each of the stages. Mo et al. (2010) apply a neural network in the first stage and K-means in the second stage to create regional-customer segments based on Chinese credit card data.

Inspired by this set of marketing research, we argue that the same two-stage relationship plays a role in a PD modeling case. That is, in different regions we expect different sector segmentations, with consequences for PD predictions. Our proposed approach extends on Bijmolt et al. (2004) and Mo et al. (2010) by adding flexibility to the clustering, where at each stage we choose the best model from a set of methods consisting of a Gaussian Mixture Model and K-means. This allows us to combine two of the most popular clustering techniques used in economic and business applications (Yang et al., 2012). It is known that K-Means works best for spherical clusters whilst a Gaussian Mixture Model works with elliptical clusters, which suggests that it is desirable to use different methods at each of the two clustering stages, geographically and sector-wide.

## 2.2 Default Modelling

Logistic Regression (Logit) is a commonly used method in the field of PD modeling (Westgaard & van der Wijst, 2001; Tserng et al., 2014; Orlando & Pelosi, 2020). The method is simple to implement, easily interpretable (Graeve et al., 2008; Simons & Rolwes, 2018) and suitable for stress testing scenarios (Chan-Lau, 2006; Simons & Rolwes, 2018). Gruszczyński (2019) report on the unbalanced nature of PD data and Salas-Eljatib et al. (2018) conclude that Logit struggles in the presence of unbalanced data. The use of over-sampling techniques like the Synthetic Minority Over-sampling Technique (SMOTE) of Chawla et al. (2002) help to solve the issue of a under-represented class and it is what we employ to improve the fit of our Logit model. Furthermore, Habshah Midi & Rana (2010) speak on the arising issues that Logit faces when used over a large set of correlated regressors. The set of factors used in the PD literature is dense, consisting of macro factors (interest rates (Sommar & Shahnazarian, 2018; Jacobson et al., 2005); GDP growth (Pollák & Popper, 2021; Castren et al., 2009); unemployment rates (Pollák & Popper, 2021; Kick & Koetter, 2007); exchange rates and commodities (Castren et al., 2009)) and broader firm characteristics (Alonso & Carbo, 2020, 2022; Barboza et al., 2017). Penalized regressions, like the LASSO of Tibshirani (1996) can help with variable selection by penalizing the coefficient estimate of each regressor. We apply the $L^1$ norm penalization of the LASSO to Logit, allowing it to perform variable selection and overcome multicolinearity problems. In this manner, we make use of a penalized Logit model (LASSO Logit) as our benchmark for PD modelling.

The benefits of using Logit in PD modelling are clear making it the benchmark to beat, but with the more recent boom in ML applications (Castellucio, 2020) there has been an increasing number of scientific papers applying ML methods in the context of PD modelling (Alonso & Carbo, 2020, 2022; Barboza et al., 2017; Dastile et al., 2020). Brown & Mues (2012); Hamori et al. (2018); Leo et al. (2019) show that ML methods have higher predictive power than the Logit benchmark. The Machine Learning approaches show an increase in accuracy of roughly 10% compared to the standard Logit model. Following the arguments of Alonso & Carbo (2020, 2022); Hamori et al. (2018), tree-based methods offer the best accuracy and implementation efficiency trade-off from a broad set of ML methods. A famous tree-based model is the RF of Breiman (2001). RF is good at handling large amounts of data as it is able to capture more complex and non-linear relationships in the data, which a linear model misses. Jung (2018) concludes that reporting errors in financial data in South Korea's markets exist and fundamentally change the underlying properties of a company's returns, which makes a model robust to outliers more attractive for handling financial data. Gruszczyński (2019) report on the unbalanced nature of PD data and RF handles such a dataset better than Logit (Breiman, 2001; Brown & Mues, 2012). More complicated models such as Neural Networks (McCulloch & Pitts, 1943) are also shown to provide better forecasts than standard models, but these models are too complex (even when compared to other ML methods (Lei & Ling, 2023x)) and therefore not preferred from a supervisory perspective (Alonso & Carbo, 2020, 2022).

Most of the literature has focused on the benefits of Machine Learning methods, and the literature on potential costs is thin. Dupont et al. (2020) and Institute of International Finance (2020) offer a qualitative discussion on the risks of Machine Learning methods from a supervisory perspective. The main drawback of ML models in this context comes in the form of a loss in interpretability (Institute of International Finance, 2019). Lundberg & Lee (2017) propose a novel unified approach for computing Shapley values (Shapley, 1951) and interpreting model predictions. We apply their approach to overcome the interpretation issue of the RF. A quantifiable result is provided by (Alonso & Carbo, 2020, 2022), where they calculate a cost metric for ML models based on different risk factors (e.g. stability and interpretability) against the standard Logit regression. They find XGB [1] and RF to offer the best risk-reward trade-off. Despite the valuable contributions, the work of Alonso & Carbo (2022) is prone to some limitations. They only consider 11 explanatory variables, none of them being a macroeconomic variable, and have a relatively small sample size for their analysis. Both issues we overcome by having a very large dataset of defaults and utilizing a large number of regressors picked from relevant literature.

### 2.3 Stress Testing

Having identified a PD model, one applies stress testing using macroeconomic scenarios provided by the EBA to see whether the model is robust to adverse economical conditions. As of 2023, banks have to provide a breakdown of their exposures towards firms and the related impairment by sector of economic activity (European Banking Authority, 2023). The main goal is to increase the credibility and realism of the stress testing exercise. An important remaining question is how one should deal with the firm characteristics under the macroeconomic stress scenarios.

[1] We conduct our analysis for the XGB of T. Chen & Guestrin (2016), but due to its very similar performance to the RF, we opt for putting the XGB results in Appendix B.5

For simplicity, it might be assumed that the firm characteristics follow a random walk (Zanders, 2023). This assumption is limited from two perspectives. First, the data consists of SME companies which are known for experiencing significant changes every year, which are not taken into account with a random walk. Second, a random walk is likely to give too optimistic of a view of the stress scenario. A lower GDP is for example expected to lower the sales of a company, which might mean that the probability of default is underestimated in the case of holding firm characteristics constant. It is thus desirable to predict firm characteristics using macroeconomic variables as well.

The literature on the predictability of firm characteristics is thin. Chen (2010) finds that the sales of large development and construction firms can be forecasted using a linear regression with a set of financial ratios and macroeconomic factors as explanatory variables. Issah & Antwi (2017) takes a broader perspective by not restricting themselves to only one sector. They use a simple linear regression model to regress the Return on Asset on its lag and a set of macroeconomic variables. Using a sample of 116 listed companies in the UK, they find $R^2$s between 0.79 and 0.95. Similar results are found for Portugal (Vieira et al., 2019) and Nigeria (Egbunike & Okerekeoti, 2018), suggesting that modelling firm characteristics using macroeconomic variables is achievable and necessary in the context of stress testing.

The main extensions that we make on this set of the literature are threefold. First, we use Machine Learning methods to improve predictive performance. Second, we study a dataset of SME companies. This makes it more challenging to investigate the relationship between macroeconomic factors and firm characteristics, since SMEs change substantially over time. Machine learning methods might therefore especially be valuable in this setting to capture these complex dynamics. Third, we investigate the role of clustering to improve the predictions, since the role of macroeconomic variables on characteristics might be both sector and region dependent.

# 3 Data

The starting point for our analysis is the Zanders dataset which encompasses 1 075 926 observations from January 2000 to December 2022 of European corporate loans. These observations come from data from 22 countries and 19 sectors. The dataset covers variables such as the default indicator, the country and sector of every loan, and a wide range of firm specific accounting variables with their lags up to three years.

As a first step we remove all variables that are the sum of other variables (according to standard accounting equations), to avoid perfect multicolinearity problems. We retain the individual components rather than the sums, since these are more informative – e.g., we can always construct back the sums from the individual elements. For example, we remove financial p&l since it is the difference between financial revenue and financial expenses. From the 51 unique accounting variables, we eliminate 15 and end up with 36 unique accounting variables with their lags up to 3 years, making it a total of 144 variables.

Second, we remove all variables with more than 33% of missing values. After this step, we are left with 24 unique accounting variables for which we have the current value and the 1 year lag, making up a total of 48 accounting variables. We supplement these accounting variables with the Weight of Evidence of Industry and the Weight of Evidence of Country[2] to end up with a total of 50 firm specific variables.

Other than the firm specific variables, we collect annual macroeconomic data on the national level. The macroeconomic variables can be categorized as follows:

1. Traditional indicators, namely annual gdp growth (%), annual inflation growth (%) and annual unemployment rate (%). These have been sourced from the World Bank.
2. EUR foreign exchange rates, which can be further sub-divided into global, namely EURUSD, EURJPY, EURCNY and EURINR, and local, namely EURGBP, EURNOK, EURCHF and EURTRY. These have been sourced from Bloomberg.
3. Debt, namely household debt (% of gpd), non-financial corporate debt (% of gdp) and general government debt (% of gdp). These have been sourced from the IMF.
4. Government bond yields, namely the 3-month and and 10-year yields (%). These have been sourced from the Federal Reserve Bank of St. Louis.
5. Commodity prices, namely oil, gas, gold and copper, priced in EUR per standard unit. These have been sourced from Bloomberg.

Traditional indicators have been adapted without adjustment. Foreign exchange rates and commodity prices have been sourced as daily time-series, then, daily mid-points have been averaged to obtain annual estimates of the annual rates and prices. Debts have been adapted without adjustments, given availability. Government bond yields had been sourced as daily time-series and daily yields have been averaged to obtain average annual yields, given availability. We remark that government bond yields and debts suffered from missing values for four countries in our sample: Bulgaria, Romania, Croatia and Hungary. The missing values either consisted of single entries missing within the respective annualised time series, or of entire time-series missing for certain countries. Single missing values have been imputed with 5-year-moving averages. Entire time-series missing have been imputed with regional averaging where the region is defined as: Poland, Czech Republic and Slovakia.

The macroeconomic variables are merged with the original firm characteristics data on a row by row approach, where we merge on Country and Year. For example: the GDP entry of an Italian company in 2020 is the Italian GDP in 2020, whilst that of a Dutch company in 2015 is the Dutch GDP in 2015.

The resulting data covers 7 variables, of which 51 are firm specific (including lags) and 20 are macroeconomic variables. Out of all remaining firm characteristic data points, 23.3% were missing values, most of them stemming from the first lags of the variables. To be able to use all 1 075 926 observations in the sample, we use MICE imputation based on Bayesian Ridge which is extensively discussed in Appendix A.1.

The total number of defaults in the data is 6741 or 0.63%. The average Sales are \$6.7m with a standard deviation of \$2.3m. Furthermore, the average company has total assets worth \$11.7m and a book value of equity of \$3.7m. It should be noted that the standard deviations

[2] More info on Weight of Evidence in Appendix A.2

are quite high for all characteristics (this does also hold in the original data, and does thus not stem from the imputation method). Further descriptive statistics are provided in Appendix B.2.

# 4 Methodology

This chapter discusses the steps we take in our analysis. We start by defining how we impute missing values. Next we discuss our clustering methods, followed upon by our PD models and finally we describe the stress testing approach.

## 4.1 Data Imputation

The data displayed significant sparsity and consequently the use of imputation was required. We considered imputation based on MICE and $k$-Means but ended up using the MICE imputation due to the computational infeasibility of $k$-Means[3].

Multiple Imputation of Chained Equations (MICE)[4] is an imputation method consisting of modelling each feature with missing values as a function of the other features of the data. The imputation is conducted iteratively, with the least sparse features being imputed first and the most sparse features being imputed last; it is important to note that previously imputed features are used for the imputation of the remaining features. This method is shown to have a stable solution (Jadhav et al., 2019).

We used the default version of MICE which is based on Bayesian Ridge, and use the `sklearn.impute` library in Python to perform the imputation. The process was computationally intensive, so we utilized the MetaCentrum computational resources, details on the hardware used can be found in Appendix C.1.

## 4.2 Clustering

After having dealt with the missing values, we focus on creating clusters. Our objective is to define regional-sector clusters. Define *Macros* to be a set of different macroeconomic variables, let *Employees* denote the number of employees of a firm and *WoE_Sector* the weight of evidence of the sectors. *Employees* is included to make the results more stable, since certain sectors typically have a larger workforce than others. Our first round of clustering is done using *Macros* and in the second stage we cluster using *Employees* and *WoE_Sector* on the regional clusters resulting from stage 1. The pseudocode for this procedure is described in detail in Algorithm 1.

[3] We tried K-Means as well but it got a runtime higher than 72 hours, so we did not continue with it

[4] More details about MICE in Appendix A.1

**Algorithm 1:** Two Stage Clustering

```
Step 1: Use $Macros$ to estimate a GMM and K-Means (see Appendix A.5);
Step 2: Choose the optimal number of clusters $m$ for the GMM and construct regional
  clusters Region_GMM = $(Region\_GMM_1, .., Region\_GMM_m)$;
Step 3: Choose the optimal number of clusters $k$ for K-Means and construct regional
  clusters $Region\_KM = (Region\_KM_1, .., Region\_KM_k)$;
Step 4: if interpretability $Region\_GMM \succ Region\_KM$ then
    Step 4a: Construct $C_1, .., C_m$ where $C_i$ contains all observations for $Employees$
      and $WoE\_Industry$ belonging to cluster region $Region\_GMM_i$;
    Step 4b: for $C_i \in 1, .., m$ do
        Estimate on $C_i$ a GMM and K-Means (See Appendix A.5) ;
        Select the optimal number of clusters for both methods and choose the method
          which gives the best interpretability results
    end
end
Step 5: else if interpretability $Region\_GMM \not\succ Region\_KM$ then
    Step 5a: Construct $C_1, .., C_k$ where $C_i$ contains all observations $Employees$ and
      $WoE\_Industry$ belonging to cluster region $Region\_KM_i$;
    Step 5b: for $C_i \in 1, .., k$ do
        Estimate a GMM and K-Means on $C_i$ (See Appendix A.5);
        Select the optimal number of clusters for both methods and choose the method
          which gives the best interpretability results;
    end
end
```

## 4.3 Default Modelling

This section discusses the two PD models that we use in this paper: LASSO Logit and Random Forest. Both models are trained on standardized input variables. The train-test split is 80/20. Finally, for both methods we use SMOTE.

### 4.3.1 LASSO Logit

A common method for modelling the probability of defaults is the Logit. We extend upon that concept by considering a Logit model on a set of LASSO selected variables.

Define *Macros* as a set of macroeconomic variables, *Characteristics* as a set of firm characteristics and *Characteristics_lag* as the lag of *Characteristics*. Our set of explanatory variables can than be coined as $X = (Characteristics, Characteristics\_lag, Macros)^t$.

Logit is a binary classification method based on evaluating a linear model:

$$f(X) = \beta^t X + \varepsilon, \tag{1}$$

with coefficients $\beta \in \mathbb{R}^n$ and data $X \in \mathbb{R}^n$ within the logistic function $p : \mathbb{R} \to (0, 1)$ defined by:

$$p(f(X)) = \frac{1}{1+\exp(-(\beta^t X + \varepsilon))}. \tag{2}$$

Least Absolute Shrinkage and Selection Operator (LASSO) refers to a regularization method based on the $L^1$ norm. Formally, for some linear model $f(X)$ the LASSO regularization consists of solving:

$$\min_{\beta} \left\{ \|Default - f(X)\|_2^2 \right\} \text{ s.t. } \|\beta\|_1 \leqslant C, \tag{3}$$

where Default $\in [0,1]$ denotes the true observations for defaults and $C \in \mathbb{R}$ is a hyperparameter. We combine the two methods above to estimate the LASSO Logit using Maximum Likelihood Estimation (MLE), which consists of solving:

$$\min_{\beta} \left\{ \sum_{i=1}^{n} [Default_i \log(P(Default_i|X_i)) + (1 - Default_i)\log(1 - P(Default_i|X_i))] - \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{4}$$

The LASSO Logit maximization problem is solved by using the `sklearn` package with specification `liblinear`; this method uses a Coordinate Descent (CD) algorithm to iteratively perform approximate minimisations. The hyperparameter $\lambda$ has been set to 0.1, with this value being determined under 5-fold cross-validation (see Appendix C.4).

### 4.3.2 Random Forest

The second PD model we discuss is the Random Forest (RF). The RF algorithm elaborates on the bagging of decision tree methods by proposing to randomly select only a subset of variables to choose from at each splitting. Such approach lessens the correlation of the individual trees, which helps with reducing the variance of predictions leading to higher accuracy and robustness. Again define $X = (Characteristics, Characteristics\_lag, Macros)^t$.

The proposed algorithm of RF based on Breiman (2001) is provided in pseudo-code in Algorithm 2. For estimating Random Forest, we employ the Scikit–Learn API in Python. In order to tune the hyperparameters we use 5-fold stratified cross validation (see details in Appendix C.4).

```
Algorithm 2: Random Forest Prediction Procedure
Step 1: Draw $B$ bootstrap samples from training data $X$.;
Step 2: for $i = 1$ to $B$ do
    Grow a deep Decision Tree (DT) on $b_i$ using the DT algorithm as descripted in Appendix A.6, with the modification that a random subsample $H$ of feature variables is selected for splitting at each node, i.e., $H \subset J$ where $J$ denote the sample of our all features;
    Output: $\widehat{Default}_i^{DT}$ (prediction of the decision tree for boostraped sample $i$)
Step 3a (binary prediction): Apply majority voting on the predictions of B trees to obtain the prediction for RF.;
    $\widehat{Default}^{RF} = \underset{k \in \{1,0\}}{\text{argmax}} \sum_{b=1}^{B} I(\widehat{Default}_i^{DT} = k)$
Step 3b (probability prediction): Compute the probability of default for RF as
    $\widehat{PD}^{RF} = \frac{1}{B} \sum_{b=1}^{B} \widehat{Default}_i^{DT}$
Choose which prediction is relevant for the problem at hand.
```

### 4.3.3 Interpretability

We employ the Shapley value to aid interpretability of our RF model. The Shapley value is a game-theoretic measure used to assess the expected marginal contribution of an individual agent within a cooperative game[5]. The concept of expected marginal contribution has a natural extension within the framework of Machine Learning by replacing cooperative games with models and agents with model features. Formally, let $f : \mathbb{R}^p \to \mathbb{R}^q$ denote a RF model characterised by features $\psi = (\psi_1, \dots, \psi_m)$ for some $m \in \mathbb{N}$. Let $g_i : \mathbb{R}^p \to \mathbb{R}^q$ be a reduced RF model constructed from $f$ by disregarding some feature $\psi_i$ for $i \in \{1, \dots, m\}$. For observations $\{X_t\}$ with $X_t \in \mathbb{R}^p$ and $t \in \{1, \dots, T\}$, the Shapley value can be written as:

$$S_i(f) = \frac{1}{T} \sum_{t=1}^{T} \|f(X_t) - g_i(X_t)\|. \tag{5}$$

### 4.3.4 Synthetic Minority Over-sampling Technique (SMOTE)

Gruszczyński (2019) stress the issue of unbalanced data in credit risk modeling and similarly to Zhou et al. (2013) support the use of oversampling algorithms such as SMOTE to balance the unbalanced samples.

We employ the SMOTE algorithm to synthetically oversample the minority class of defaults. This algorithm, proposed by Zhou et al. (2013), creates new observations by interpolating between clustered minority samples, thus creating artificial data points. This offers greater variability in the minority class and leads to better classifier performance.

[5] More info in Appendix A.3

The `imblearn` package is used to implement SMOTE in our analysis. We set the proportionality, i.e. how much to oversample, at 30%. While it is common to completely balance the classes such that one attains a 1:1 ratio (Abedin et al., 2023; Y. Chen & Zhang, 2021), given the severe lack of balance of our data, the 1:1 ratio would mean we would include too much of a synthetically generated data that would potentially introduce too much noise. Hence, we opt to oversample the default class only to reach 30% of overall proportionality of defaults in the whole dataset.

#### 4.3.5 Evaluation

In order to compare performances among our models we employ the Area Under Curve (AUC) as the evaluation metric. Although, accuracy is usually used for performance evaluation in a classification task, in our case of a highly unbalanced data accuracy its not the best choice as a trivial model that predicts only not defaults would have an accuracy of close to 99 %. The AUC metric measures the area under the receiver operating characteristic (ROC) that plots the True Positive rate against the False Positive rate at different discrimination thresholds. AUC thus provides a metric across all possible thresholds and describes the models ability to discriminate between classes. AUC of 0.5 corresponds to no ability to separate the classes, while AUC = 1 indicates perfect discrimination.

### 4.4 Stress Testing

This section discusses the methods used to perform stress testing. First we define a set of different dynamic stress testing methods which are based on the idea that firm characteristics are predictable using their first lags and a set of macroeconomic variables. Subsequently we discuss the general setup for predicting stressed PDs.

#### 4.4.1 Dynamic Stress Testing Methods

Setup Consider an unbalanced panel data structure with time dimension $t = 1, \dots, T$ and cross sectional dimension (distinct companies) $c = 1, \dots, C$. The goal is to model a set of characteristics as a function of macroeconomic variables to capture a dynamic relationship for stress testing. For example, we might want to know how a company's sales (a characteristic) changes when the macroeconomic circumstances worsen. The challenge is that $N$ is very large, and $T$ relatively short, which makes it difficult to perform a typical panel data analysis. It is thus not feasible to model the relationship between sales and macroeconomic variables separately for every company. Our proposed approach assumes that the most important information rests in the cross section. Assuming that the first lag of a characteristic exists, we can also capture dynamic effects. Define $i = 1, \dots, N$ as the intersection of the time dimension $t$ and $c$:

$$Characteristic_i = Characteristic_{c,t}, \tag{6}$$

$$Characteristic\_lag_i = Characteristic_{c,t-1}. \tag{7}$$

We can model $Characteristic_i$ as a function of $Characteristic\_lag_i$ and a set of macroeconomic variables $Macro_{i,j}$ for $j = 1, \dots, J$ with $J$ being large. Let us illustrate the notation with a fictive example. Define $t = 2010$, $c = \mathit{ItalyCorporation}$ and $j = GDPGrowth$. $\mathit{Characteristic}_i$ corresponds with the 2010 value of the characteristic for Italy Corporation and $\mathit{Characteristic\_lag}_i$ with the 2009 value, whilst $\mathit{Macro}_{i,j}$ in this case will correspond with the 2010 GDP growth of Italy.

Dynamic (Cluster) Lasso We can estimate the relationship between the characteristic, its lag and the macroeconomic variables by using a regularized regression:

$$\hat{\beta}_{char} = \arg\min_{\beta} \sum_{i=1}^{n} (\mathit{Characteristic}_i - \beta_0 - \beta_1 \mathit{Characteristic\_lag}_i - \sum_{j=1}^{J} \mathit{Macro}_{i,j}\beta_{j+1})^2 + \lambda \sum_{j=1}^{J} |\beta_{j+1}|, \tag{8}$$

where *char* denotes the characteristic. Notice that the regularization is only applied to the set of macroeconomic variables and not on the lag, such that we get different characteristic predictions for different companies. Using the estimated parameters, we can perform characteristic predictions for $t = T+1, T+2, \dots, T+M$ for all companies c which have survived until time T. Define $\mathit{Characteristic}_{i_m}$ as:

$$\mathit{Characteristic}_{i_m} = \mathit{Characteristic}_{c,T+m}. \tag{9}$$

Similarly, we define $\mathit{MacroStress}_{i_m,j}$ where the index $i_m$ corresponds again with a time period $T+m$ and a company index $c$, and $j$ to the same set of macroeconomic variables as before. If $T = 2022$, $c = \mathit{ItalyCorporation}$, $j = \mathit{GdpGrowth}$ and $m = 1$, $\mathit{MacroStress}_{i_m,j}$ will correspond to the adverse case GDP prediction for Italy for 2023 provided by the EBA. The prediction at time $T+M$ can be constructed as:

$$\widehat{\mathit{Characteristic}}_{i_M} = \hat{\beta}_{char,0} + \hat{\beta}_{char,1} \widehat{\mathit{Characteristic}}_{i_{M-1}} + \sum_{j=1}^{J} \hat{\beta}_{char,j+1} \mathit{MacroStress}_{i_M,j}, \tag{10}$$

where $\widehat{\mathit{Characteristic}}_{i_{M-1}}$ is estimated using $\mathit{MacroStress}_{i_{M-1},j}$ and $\widehat{\mathit{Characteristic}}_{i_{M-2}}$. This shows that there is a cumulative effect at play where the stressed value of the characteristic at time M also takes into account the stressed macroeconomic values from periods before.

One might expect the dynamic LASSO to perform best when applied to companies that are most similar to each other. For example, an Energy company is likely to be more severely affected by oil prices than a manufacturing company. We can therefore run the described procedure in this section on a cluster level which gives rise to the Dynamic Cluster LASSO model.

Dynamic (Cluster) RF To perform predictions of characteristics using Dynamic RF, we follow a similar procedure as for the Dynamic LASSO. Define the set of training samples $X_{train_i} = (\mathit{Characteristic\_lag}_i,\ \mathit{Macro}_{i,1},\ ..,\ \mathit{Macro}_{i,J})$ for $i = 1, \dots, N$. We estimate a Random Forest as described in subsubsection 4.3.2. We are now however faced with a regression task rather than a classification task, which means that we use averaging to predict rather than applying a majority rule. Using the fitted forest, we can perform characteristic predictions at $t = T+M$ by using the test set $X_{test_{i_M}} = (\mathit{Characteristic}_{i_{M-1}}, \mathit{Macro}_{i_M,1}, \dots, \mathit{Macro}_{i_M,J})$ for $i = 1, \dots, N$.

We again apply the above stated procedure also on a cluster level since we expect there to be differences in reactions to macro changes for different sets of companies. This gives rise to the Dynamic Cluster RF.

### 4.4.2 General Stress Testing Procedure

Having defined methods to predict the characteristics, we can turn to our objective of performing stress testing. Let there be $k = 1, \ldots, K$ characteristics, and denote $a = 1, \ldots, A$ as different ways in which each characteristic is constructed. We can define the set of characteristics at $T + m$ belonging to observation $i$ as:

$$TotalCharacteristics^{a}_{i_m} = (Characteristic^{a}_{i_m,1}, .., Characteristic^{a}_{i_m,K}), \tag{11}$$

furthermore let $TotalCharacteristics\_lag^{a}_{i_m}$ denote the set of first lags of $TotalCharacteristics^{a}_{i_m}$. In a similar fashion as described in 1.3, we define $MacroStress_{i_m,j}$ where the index $i_m$ corresponds again with a time period T+m and where $j$ refers to the same set of macroeconomic variables as before. Thus, we get the set of predictor values at time $T + m$:

$$X^{a}_{i_m} = (TotalCharacteristics^{a}_{i_m}, TotalCharacteristics\_lag^{a}_{i_m}, MacroStress_{i_m,j}), \tag{12}$$

for $a = 1, \ldots, A$, $i = 1, \ldots, N$, $j = 1, \ldots, J$ and $m = 1, \ldots, M$. We use this set of predictor values in our PD models described in subsection 4.3 to predict the stressed PDs.

In our paper, we take $A = 5$, with $a = 1$ corresponding to Dynamic LASSO; $a = 2$ with Dynamic RF; $a = 3$ with Dynamic Cluster LASSO; $a = 4$ with Dynamic Cluster RF and $a = 5$ as a Random Walk (Benchmark). $M$ is taken to be 3 since we are interested in stress test predictions for the years 2023, 2024 and 2025.

# 5 Results

This chapter discusses our main results. We first show our resulting clusters which are subsequently used throughout the paper. Second, we apply PD models on the resulting clusters and identify what variables are the most important. Third, we share our results on the predictability of firm characteristics which allows us to perform (dynamic) stress testing in the final step.

## 5.1 Clustering

The clustering on region-level results in three clusters with approximate geographical interpretation: Southern Europe, Eastern Europe and Northern Europe. We find that GMM gives more interpretable clusters than K-Means which is an indication that country level groupings are of an elliptical rather than spherical shape.[6]. The reason the geographical interpretation is only approximate lies in the fact that three geographical outliers defy this classification. The three outliers are Iceland, Belgium and Slovenia, with Iceland and Belgium being categorised as Eastern Europe and Slovenia being categorised as Northern Europe. While the classification of Iceland and Slovenia can be justified both statistically and macroeconomically, Belgium seems

[6] More details about the estimation procedure in the Appendix C.2

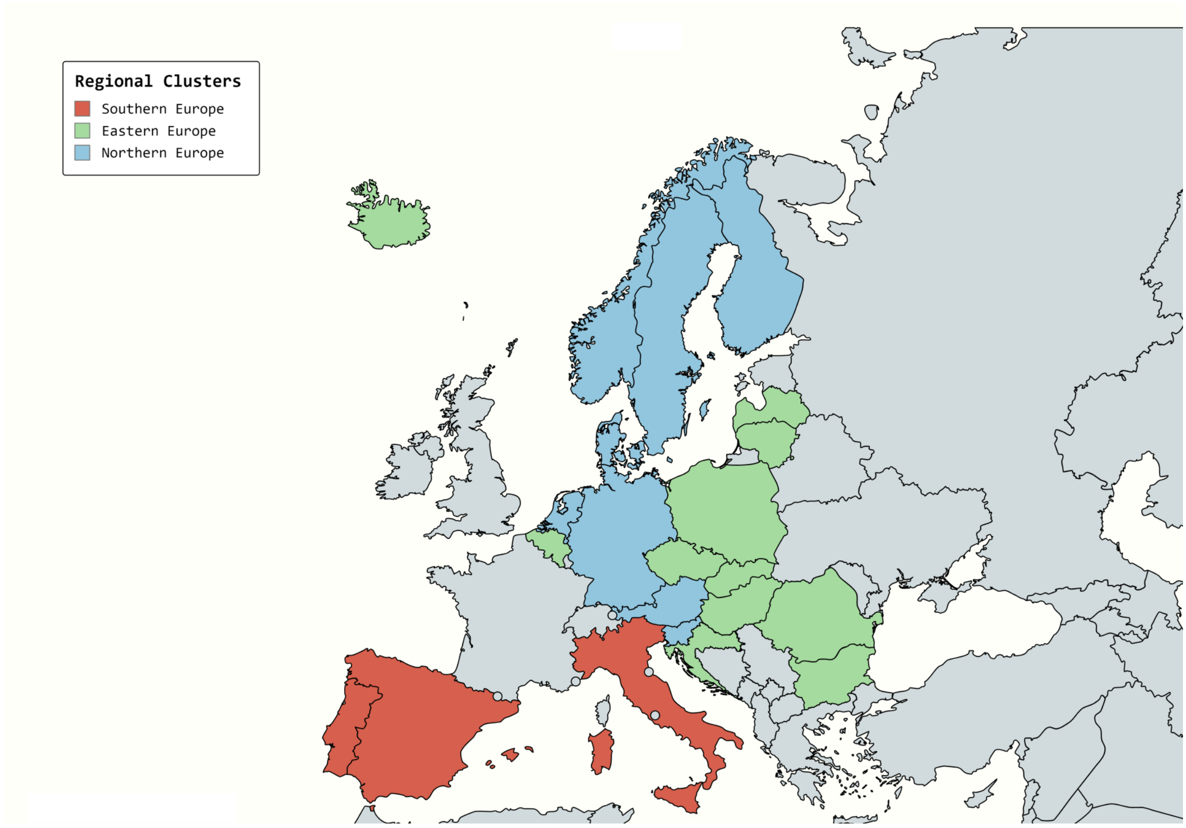


Figure 1: Regional Statistical Clusters. Illustration of the three clusters resulting from clustering procedure with geographical outliers consisting of Iceland, Belgium and Slovenia.

to be an outlier among outliers. First, we note that Slovenia's unorthodox classification stems partly from the fact that it is approximately evenly distributed in multiple clusters, displaying a 61/39 split in Northern/Eastern. An analogous argument can be used to justify Iceland, which displays a 86/14 split in Eastern/Northern. Belgium however displays a 100/0 split in Eastern. Secondly we note that Slovenia and Iceland display macroeconomic similarities with their corresponding clusters, but Belgium does not. As illustrated in the Figure 2, we can see that Slovenia's historical 3-month yields align with the average Northern European historical 3-month yield and Iceland's historical 3-month yields align with the average Eastern European historical 3-month yield. In the case of Belgium, this does not hold however; in fact for many macroeconomic variables Belgium does not display alignment with any cluster (as for example indicated by Household and Corporate debt in the Figure 2), adding to the peculiarity of the outlier. We decide to keep all outliers in as to not intervene with the statistical methods and to compare their performance in a fair manner.

Next we wish to highlight macroeconomic differences between the clusters. As illustrated in figure 3, we see that indeed significant macroeconomic differences among the clusters exist. We note that the Southern European cluster differs most significantly from the other two clusters, as exemplified by the elevated unemployment rate and 10 year yield levels post-2008.

The second clustering step consists of clustering on sector-level. We find across all regions that K-Means rather than GMM offers the more interpretable clusters, which shows that sector level groupings are spherical shaped whilst they appear to be elliptical shaped for the region

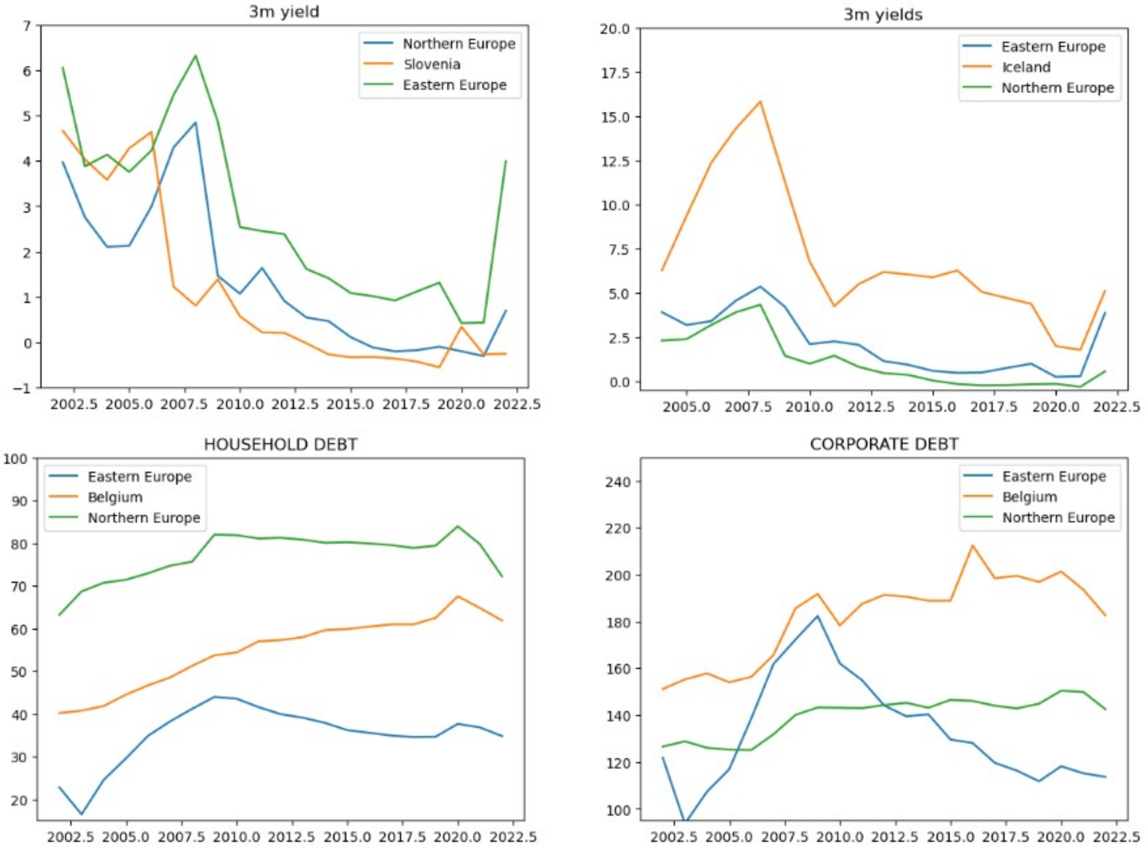


Figure 2: OUTLIERS. This figure displays the 3m yields for Slovenia and Iceland compared to the Northern and Eastern European clusters; and the household and corporate debt for Belgium compared to the Northern and Eastern European clusters. Northern Europe covers Denmark, Finland, Germany, Netherlands, Austria, Norway, Sweden (note no Slovenia); Eastern Europe includes Bulgaria, Croatia, Czech Republic, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia (Note no Belgium and Iceland

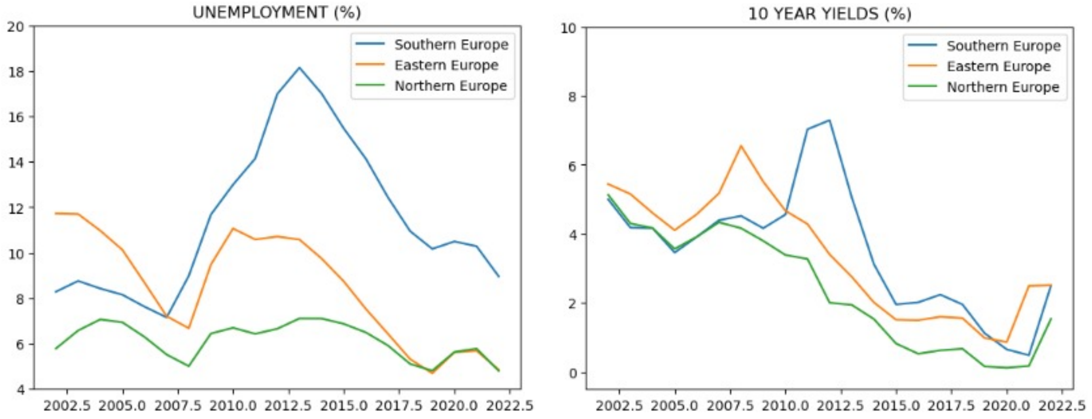


Figure 3: MACROECONOMICS This figure displays the unemployment rate and 10 year yields over time for three cluster regions. Southern Europe covers Italy, Portugal and Spain; Northern Europe covers Denmark, Finland, Germany, Netherlands, Austria, Norway, Slovenia, Sweden (note no Slovenia); Eastern Europe includes Bulgaria, Croatia, Czech Republic, Hungary, Latvia, Lithuania, Poland, Iceland, Romania, Slovakia, Belgium

clusters. This difference acts as support for our proposition that different clustering methods might be useful at each of the clustering stages, due to the different nature of the data. For all regions we get an Infrastructure cluster and a Commodities cluster. The third cluster for Southern and Northern Europe is Services, whilst clusters three and four of Eastern Europe are made up of Cyclical Services and Stable Demand Services. We thus find that more sector clusters are needed for Eastern Europe to account for the heterogeneity among firms, which offers some (weak) support for the need of different industry segmentations for different regions (Bijmolt et al., 2004; Mo et al., 2010).

Consequently, we end up with 10 final clusters[7] for which we show the PDs and number of observations in table Table 2. It is visible that two clusters have structurally higher PDs than the rest of the data. Southern European Services and Southern European Infrastructure have PDs of 0.72 and 1.09 respectively against 0.63% in the total data. Furthermore, we notice some disparity in the cluster size with Southern Europe Services having the most observations (364k) and Northern Europe Commodities the lowest (26k). We will come back to this disparity when we discuss the predictive power of our PD models in subsubsection 5.2.1.

| | Services | Commodities | Infrastructure | All Sectors |
|---|---|---|---|---|
| Southern Europe | **0.72%** | 0.45% | **1.09%** | 0.78% |
| Eastern Europe | 0.33% | 0.39% | 0.62% | 0.43% |
| Northern Europe | 0.56% | 0.33% | 0.54% | 0.53% |
| All Regions | 0.64% | 0.39% | 0.80% | 0.63% |

| | Services | Commodities | Infrastructure | All Sectors |
|---|---|---|---|---|
| Southern Europe | **364k** | 46k | 118k | 528k |
| Eastern Europe | 55k | 161k | 67k | 284k |
| Northern Europe | 156k | **26k** | 82k | 264k |
| All Regions | 575k | 234k | 267k | 1076k |

Table 2: REGION-SECTOR-CLUSTERS. The upper table presents the PDs for the resulting Region and Sector Clusters, whilst the lower table shows the number of observations. The Region clusters are formed on *Macros* using GMM, whilst the Sector Clusters are formed using *WoE_Sector* and *Employees* on the resulting region clusters using K-Means. Services for Eastern Europe is the sum of Cyclical and Stable Services. A detailed overview of the countries and sectors that fall in each cluster is provided in Appendix B. The numbers in **bold** highlight the most remarkable results.

## 5.2 Default Modelling

Having identified the clusters in the previous section, we can now apply our PD models to identify the predictive performance and the most important variables.

[7] A full list with all countries and sector in each cluster is provided in Appendix B

### 5.2.1 PD Models Predictive Power

In Table 3, we display the predictive powers of LASSO Logit and Random Forest across clusters in terms of the AUC evaluation metric. The Random Forest model outperforms LASSO Logit in every cluster marking the non-linear Machine Learning approach superior to the linear approach of Logistic Regression. We argue that based on our complex dataset with a lot of variables across a plethora of observations, simple linear relations are not enough to capture the underlying mapping of firm characteristics and macro variable to probability of default.

Another key takeaway is in relation to the performance across clusters. We see that the Machine Learning models do better for the Southern region and then see a drop in performance, especially for Northern Europe. Our explanation links this variance in performance to the number of observations and proportion of defaults in each cluster as is visible in Table 2. We use SMOTE to over-sample the minority default class; having more defaults equals more information and thus SMOTE interpolations result in better performance. Therefore, each cluster is estimated on the same proportion of defaults (30%). However, it is crucial to note that SMOTE depends on the data it samples from. Hence, in the case of Southern Europe, which contain the highest number of defaults and percentage of defaults, it is possible to sample from a wide range of information on defaults. On the other hand, for example, cluster of commodities sector in Northern Europe that ranks last in number of observations and proportion of defaults does not allow SMOTE to gain substantial benefit as the underlying distribution of defaults to sample from is limited and at the same time does not match well to the test dataset.

| Region Cluster | Sector Cluster | LASSO | Random Forest |
|---|---|---|---|
| Southern Europe | Services | 0.77 | **0.90** |
| | Commodities | 0.76 | **0.88** |
| | Infrastructure | 0.70 | **0.88** |
| Eastern Europe | Stable Services | 0.61 | 0.83 |
| | Cyclical Services | 0.61 | 0.78 |
| | Commodities | 0.69 | 0.79 |
| | Infrastructure | 0.66 | 0.77 |
| Northern Europe | Services | 0.60 | **0.71** |
| | Commodities | 0.55 | **0.63** |
| | Infrastructure | 0.55 | **0.76** |
| All Data | | 0.60 | 0.85 |

Table 3: AUC VALUES FOR DIFFERENT MODELS. This table presents AUC values of LASSO and Random Forest for different clusters. Both models are run on the cluster level, on standardized input variables and with SMOTE. The Region clusters are formed on the total data using GMM, whilst the Sector Clusters are formed on observations from the resulting region clusters using K-Means. A detailed overview of the countries and sectors that fall in each cluster is provided in Appendix B. The numbers in **bold** highlight the most remarkable results.

Despite this phenomenon, we retain an AUC score for the total data (by merging results from all models run at the cluster level) of 0.85. This is slightly higher than the AUC score of 0.84 which we get when one model is trained on the total data. The main improvement which we attain with clustering is however that we gain additional interpretability, which is discussed in the next section.

### 5.2.2 Important Variables

In this subsection we discuss the most important variables selected by LASSO Logit and RF in each cluster. For LASSO Logit, the important variables are selected based on their standardized coefficients whilst Shapley values are used for RF. We select the top three most important variables using their mean absolute SHAP value. For these variables, we retrieve the beeswarm plots to get insights into the actual relationships between those variables and the predicted outcome. Figure 4 provides an example using Southern Europe Services.

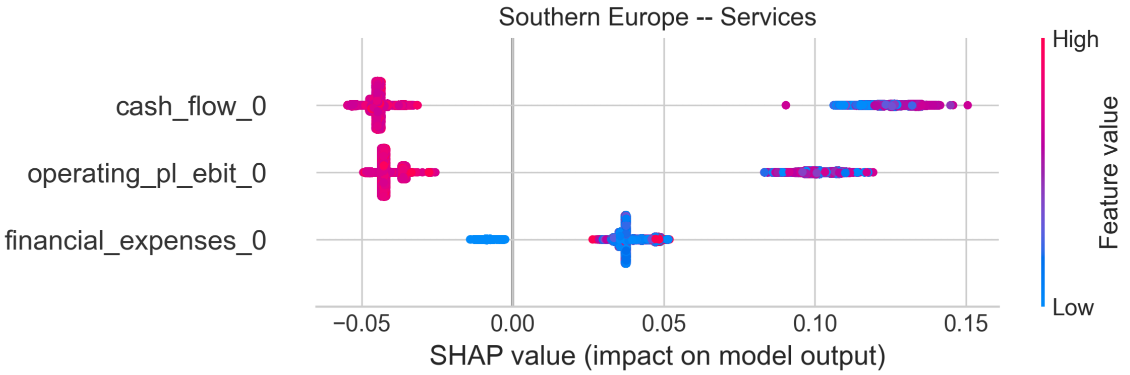


Figure 4: Shapley Values. This figure illustrates the Shapley values for Southern Europe Services. A high absolute value indicates a strong relative impact of the variable. Methodology behind Shapley Values is described in subsubsection 4.3.3.

In Southern Europe Services, we notice that higher values of Cash Flow and Operating PL EBIT, have a negative effect on the PD (the points extending towards the left are increasingly red). For Financial Expenses we find an effect of the opposite sign: medium to high values of financial expenses are associated with higher SHAP values (the points extending towards the right are mostly dark blue/red). We perform a similar procedure for all clusters and report the results in Table 4. Table 4 shows the top three selected variables by LASSO Logit and Random Forest for the different clusters. A more detailed overview of the important variables is provided in Appendix B.4 for LASSO Logit and Appendix B.3 for RF.

To aid interpretability we classify the characteristics into: operating performance variables, capital structure variables and other variables. Operating performance relates to items which show how much money a company has made over the last period. Example variables are Sales, Cash Flow and Operating PL EBIT. Capital structure variables relate to the financing decisions of companies - e.g., how they financed themselves. Broadly speaking a company can finance itself with Equity (as indicated Shareholder Funds and Capital) or Debt (as indicated by Creditors) (Hovakimian et al., 2001).

It can be observed that LASSO Logit generally selects more macroeconomic variables than Random Forest. This finding might be explained by the fact that LASSO Logit is a linear model whilst RF is a non-linear model, which means that RF is able to capture additional dynamics as indicated by the firm characteristics. A further difference between the models is highlighted by the fact that there are only 4/10 clusters with an exact match in the top three variables and in none of the clusters the top three important variables share at least 2 variables. In the rest of this section, we focus on further interpreting the important variables from RF since this model was shown to perform substantially better in PD predictions.

The signs of the coefficients for the characteristics are in most cases of the expected direction. We see that the better the operating performance (the higher Cash Flow, Sales or Operating PL EBIT), the lower the PD. The higher the leverage (the higher Creditors), the higher the PD and the higher the equity (the higher Shareholder Funds or Capital) the lower the PD since equity is a substitute for debt (Cathcart et al., 2020; Hovakimian et al., 2001). For the macro variables, we get as expected that a higher FX rate corresponds with higher PDs since exporting becomes more expensive (Kandil & Mirzaie, 2005). One unexpected finding is the positive coefficient of Sales lag in Eastern Europe Services, where we would have expected a negative sign. This might potentially be caused by the fact that many values for Sales Lag had to be imputed.

Next we turn to comparing the most important variables across each cluster. Starting with Southern Europe, we notice that for Services the most important variables relate to a company's operating performance (Cash Flow and Operating PL EBIT). For Commodities and Infrastructure, we notice both capital structure variables (Shareholder Funds and Creditors Lag) and operating performance variables (Cash Flow). This difference might be explained by the relatively high earnings volatility of firms in Commodity and Infrastructure in comparison to those in Services, which makes leverage more costly for these firms than for those in Services (Strebulaev & Yang, 2013).

For Eastern Europe, both Cyclical and Stable Services pick up operating performance variables which is in line with Services in Southern Europe. For Commodities, we notice again the importance of capital structure variables (Creditors and Creditors lag). For Infrastructure in Eastern Europe, we observe the importance of Household Debt. A larger household debt indicates higher customer mortgages which is accompanied with more Real Estate investment and a lower PD. It is interesting to note that Household debt does not play an as important role in the Infrastructure clusters in Southern and Northern Europe. One potential explanation for this is the relatively low maturity of the Real Estate sector in Eastern Europe, in comparison to the West, which means that there is more opportunity to grow (Chmelar, 2013). This could also be seen in Figure 3 where Eastern Europe has substantially lower current houshold debt levels than Northern Europe.

Third, we look at Northern Europe. In the Commodities cluster we find as expected Gas and Oil as important variables. The signs are positive since it is likely that the SMEs in our sample ”source” commodities rather than produce them due to their small size. This means that higher commodity prices result in higher costs and thus also higher PDs. One thing to note is that, these variables were not as important in Commodities for Southern and Eastern Europe. This difference might be explained by the presence of Norway with a large oil sector and the Netherlands being a large gas producer in Europe (IEA, 2021a,b). In general, we notice that macroeconomic variables play a more important role in Northern Europe than in the other regions. One other example of this is the importance of FX rates such as EURUSD and EURTRY. Both the US and Turkey are large trade partners for Northern European countries, and with an appreciation of the EURO the competitive position of Northern European exports worsen (Kandil & Mirzaie, 2005). The reason why FX rates are more important in Northern Europe than that they are in other regions, can be explained by by the fact that the countries in our Northern Europe cluster are countries such as the Netherlands and Germany which have high reliance on exports (Statista, 2022).

| **Region** | **Sector** | **LASSO Logit** | **Random Forest** |
|---|---|---|---|
| **Southern Europe** | Services | 3m yield (-)<br>**Sales (-)**<br>Government debt (+) | **Cash flow (-)**<br>*Financial expenses lag (+)*<br>**Operating pl ebit (-)** |
| | Commodities | *Shareholder funds (-)*<br>10y yield (+)<br>Government debt (+) | **Cash flow (-)**<br>*Creditors lag (+)*<br>*Shareholder funds (-)* |
| | Infrastructure | 10y yield (+)<br>**Cash flow (-)**<br>EURJPY (+) | *Shareholder funds (-)*<br>Goverment debt (+)<br>**Cash flow (-)** |
| **Eastern Europe** | Cyclical Services | Depreciation & Amortization (-)<br>*Creditors lag (+)*<br>**Sales (-)** | **Material cost lag (+)**<br>**Operating pl ebit (-)**<br>*Creditors lag (+)* |
| | Stable Services | 3m yields (-)<br>**Sales (-)**<br>**Added value (-)** | **Sales lag (+)**<br>**Added value (-)**<br>**Operating pl ebit (-)** |
| | Commodities | Cash equivalent (-)<br>EURxJPY (-)<br>3m yield (-) | *Creditors lag (+)*<br>**Operating pl ebit (-)**<br>*Creditors (+)* |
| | Infrastructure | 3m yield (-)<br>EURJPY (-)<br>Household debt (-) | *Shareholder funds (-)*<br>Household debt (-)<br>Gold (-) |
| **Northern Europe** | Services | *Shareholder funds lag (-)*<br>*Other current liabilities (-)*<br>*Capital (-)* | EURUSD (+)<br>EURTRY (+)<br>*Creditors lag (+)* |
| | Commodities | **Cash flow (-)**<br>**Sales (-)**<br>EURNOK (+) | Gas (+)<br>EURTRY (+)<br>Oil (+) |
| | Infrastructure | Debtors (-)<br>Tangible fixed assets (-)<br>Corporate debt (+) | *Shareholder funds (-)*<br>Gas (+)<br>**Operating pl ebit lag (-)** |

Table 4: Most important variables per cluster. The table illustrates the variables with the most impact on PD for each cluster. The Region clusters are formed on the total data using GMM, whilst the Sector Clusters are formed on observations from the resulting region clusters using K-Means. A detailed overview of the countries and sectors that fall in each cluster is provided in Appendix B. The symbol (+) indicates a positive effect on PD, the symbol (-) indicates a negative effect on PD. Blue variables denote macroeconomic variables, **red** variables denote operating performance variables, *red* show the capital structure variables and red other characteristics. Note that **red** + *red* + red are equal to the total set of firm characteristics.

To sum up, we find that defaults in Services (incl Cyclical and Stable) in Southern Europe and Eastern Europe are primarily caused by operating performance variables, whilst in Northern Europe exchange rates play a more important role. This difference is likely to be caused by differences in reliance on exports of the regions. For Commodities, we find Gas and Oil prices to be important predictors in Northern Europe but less so in other regions which is in line with the large oil industry in Norway and gas industry in Netherlands. In the other regions, capital structure variables play an important role for PD predictions in Commodities - which is not surprising since high leverage is more costly for these companies due to large fluctuations in commodity prices. For Infrastructure, we find importance of Household Debt in Eastern Europe but not in other regions, potentially stemming from differences in maturity of the Real Estate sector across regions.

A final point can be made about the created interpretability by the clusters. We notice that the important variables per cluster differ and are tailored to each different cluster. When solely looking at the total data, RF provides Cash Flow, Operating PL EBIT and Added Value as the top three important variables. Our clusters show that the importance of these variables is primarily driven by Southern and Eastern Europe and less so by Northern Europe. Furthermore, only focusing on the total data would mean that we ignore sector specific information, which is shown to vary across regions. This shows that our two stage clustering procedure creates additional insights.

## 5.3 Stress Testing

We can use the best identified PD model from the previous section to perform stress testing. Before doing that, we investigate whether firm characteristics are predictable using macro variables, to find out whether we can introduce some dynamics in the stress testing.

### 5.3.1 Firm Characteristic Predictability

This subsection compares the RMSE values for firm characteristic predictions using a Random Walk (the Benchmark), Dynamic LASSO, Dynamic RF, Dynamic Cluster LASSO and Dynamic Cluster RF. Due to the fact that these models have been estimated for all 24 firm characteristics, we use the default values in Python and do not apply an extensive grid search procedure. This should not have a big effect on the results, since if the performance without grid search is better than a random walk, the performance with grid search is likely to be even better.

The results are displayed in Table 5 for the most important characteristics picked up by our PD methods. A complete table with all variables is available in Appendix B.6. It can be observed that Dynamic LASSO performs slightly better than Dynamic Random Forest. 83% of the key variables have a lower RMSE than a Random Walk (vs 67% for Dynamic Random Forest) and the average RMSE improvement is 26% (vs 24 % for Dynamic Random Forest). The clustering methods Dynamic Cluster LASSO and Dynamic Cluster Random Forest perform worse than the models applied on the total data. Dynamic Cluster LASSO has for 30% of the key variables a lower RMSE than the Random Walk, whilst this is 50% for Dynamic Cluster Random Forest.

We conclude that Dynamic LASSO is the best performing model for firm characteristic predictions in our sample. Another advantage of this model in comparison with Dynamic Random Forest is that it is relatively easy to interpret. For this reason, we will use the Dynamic LASSO method to perform dynamic stress testing. The results for this approach will be compared to "static" stress testing where we perform stress testing assuming a Random Walk for the characteristics.

| | RW | DL | DRF | DCL | DCRF |
|---|---|---|---|---|---|
| Shareholder funds | 44.01 | 17.54 | 15.87 | 19.67 | 22.48 |
| Creditors | 1.81 | 1.46 | 1.31 | 2.73 | 2.42 |
| Sales | 6.93 | 1.78 | 1.64 | 2.90 | 1.87 |
| Operating PL EBIT | 2.99 | 2.96 | 2.43 | 3.37 | 1.97 |
| Cash flow | 4.07 | 3.90 | 4.22 | 4.66 | 4.57 |
| Financial Expenses | 2.14 | 2.17 | 2.95 | 5.81 | 4.87 |
| % of key variables with lower RMSE | | 83% | 67% | 33% | 50% |
| Average RMSE improvement | | 26% | 24% | -23% | -3% |

Table 5: RMSE. The table illustrates the RMSE for different firm characteristics, which are selected based on their frequency of appearing in Table 4. The predicted value is the firm characteristic of interest. The predictors consist of the lag of the firm characteristic of interest and a range of macroeconomic variables described in section 3. The benchmark is set by the Random Walk (RW). Dynamic Cluster LASSO (DCL) and Dynamic Cluster RF(DCRF) consists of the Dynamic LASSO (DL) and Dynamic RF (DRF) applied on the regional-sector clusters, which are described in detail in Appendix B.

### 5.4 Stressed PDs

Table 6 shows the PD predictions for 2023-2025. We show the base case, the adverse case under RW and the adverse case under DL. It can be noticed that DL gives higher stressed PDs than RW which is in line with our proposed idea that a RW underestimates defaults in times of crisis. The additional effect of DL over RW is largest for clusters where characteristics rather than macro variables are the most important factors, such as all Southern European clusters. In Southern Europe, we demonstrated the importance of characteristics such as Sales, Cash Flow, Operating PL EBIT, Creditors and Shareholder Funds. These variables were further shown in Table 5 to have relatively good levels of predictability. The DL model incorporates the adverse macro scenarios in these characteristics, which gives a higher PD estimate. On the other hand, we saw in Table 4 that the most important variables in Northern Europe are macro variables, which explains the little added value (arguably adds noise instead) of DL over RW in Northern European clusters, especially for the commodity sector, which under static adverse case attain PD of 0.56, while in the dynamic one we do not see any defaults for this cluster.

We notice that most of the stressed PDs are expected to fall in Southern Europe, which is not surprising since our model has better predictive performance in these regions (as denoted by AUC) than in the others. Another explanation is related to the relatively low creditworthiness of Southern European Countries, which results in higher credit risk for companies in this region (TradingEconomics, 2024; Breckenfelder, 2018). Furthermore, we observe across all regions that Commodities face higher PD risks than Infrastructure, which can be explained by the fact that companies in the Commodity cluster typically face the highest levels of earnings volatility due to frequent fluctuations in commodity prices (Pindyck, 2004; Strebulaev & Yang, 2013). For Services, we notice a further distinction for Cyclical and Stable in Eastern Europe. Cyclical Services has a PD of 0.34% against 0 for Stable Services, which shows that a further distinction of Services in Cyclical and Stable, offers additional insights.

| | | Static Case | | | Dynamic Case | |
|---|---|---|---|---|---|---|
| Region | Industry | Historic PD | Base case | Adverse Case | Adverse Case | RWA |
| **Southern** | Services | 0.72% | **2.02%** | **2.92%** | **3.79%** | **0.35** |
| | Commodities | 0.45% | **0.31%** | **1.63%** | **1.83%** | **0.21** |
| | Infrastructure | 1.08% | **0.37%** | **1.24%** | **2.02%** | **0.22** |
| Eastern | Infrastructure | 0.61% | 0.00% | 0.00% | 0.00% | 0.00 |
| | Stable Services | 0.38% | 0.00% | 0.00% | 0.00% | 0.00 |
| | Cyclical Services | 0.38% | **0.14%** | **0.22%** | **0.34%** | 0.05 |
| | Commodities | 0.24% | **0.06%** | **0.55%** | **0.82%** | 0.11 |
| Northern | Services | 0.55% | 0.00% | 0.01% | 0.00% | 0.00 |
| | Commodities | 0.33% | 0.00% | **0.56%** | 0.00% | 0.00 |
| | Infrastructure | 0.53% | 0.02% | 0.00% | 0.14% | 0.02 |
| All Data | | 0.62% | 0.75% | 1.29% | 1.72% | 0.17 |

Table 6: Probability of Defaults for Stress Test over 2023 – 2025. This table presents the results for the Stress Test based on EBA scenarios for 2023-2025. The stress test is performed with the same Random Forest model for which we showed the AUC in Table 3. The given percentage corresponds to the proportion of defaulted companies over the 3 years in a given cluster. Historic PD is the unconditional probability of default in our dataset spanning from 2000 to 2020. The RWA for dynamic case is computed based on Basel Framework (Appendix A.4), assuming LGD = 25% and EAD = 1. The text and numbers in **bold** highlight the most remarkable results.

All in all, we get a base case PD for the total data (weighted average of the clusters) of 0.75%, which is slightly higher than the historic PD of 0.62%. The Adverse RW PD for the total data is 1.29% and for DL this is 1.72%. We thus conclude that a RW approach for the characteristics leads to a substantial underestimation of the stressed PD by 33% (or 0.43 percentage points).

The DL Adverse Case PD's can be used to calculate Risk Weighted Assets (RWA). Eight Percent of RWA indicates how much capital a bank should hold for every 1 euro. We get an RWA for the total data of 0.17. In line with our PD predictions, the bank has to hold most capital for Southern Europe Services (RWA of 0.35) and Southern Europe Infrastructure sector (RWA of 0.22).

### 5.4.1 Reverse Stress Testing

In the Reverse Stress Testing (RST) we focus on stressing the variables that are most important for modelling probability of default, whilst keeping all other macroeconomic variables at their base case prediction. We decid to employ *government debt, GDP growth, inflation growth, gas and oil.* This scenario could happen if the Ukraine-Russia war would escalate and demand the more active involvement of other European countries. As a result, governments would accumulate debt to fund the war, and as a consequence of government spending, inflation would grow. Overall, the GDP would plummet, and short-term yields would increase since investors would demand higher risk premia. Lastly, oil and gas prices would rise due to the increased geopolitical uncertainty (Zhang et al., 2023). The aforementioned variables are stressed over the EBA adverse case as displayed in Table 7. For example, in 2023 we assume that the GDP

Growth will be 5% worse than is currently stated in the EBA adverse case.

| Stressed variables | 2023 | 2024 | 2025 |
|---|---|---|---|
| Government Debt | 20% | 25% | 30% |
| GDP Growth | 5% | 10% | 15% |
| Inflation Growth | 10% | 20% | 25% |
| 3m yields | 10% | 20% | 30% |
| Oil & Gas | 20% | 30% | 40% |

Table 7: War Scenario. The table illustrates the war scenario increase of the variables we selected for RST over the EBA adverse case.

By stressing only this subset of variables we are able to surpass the probability of default of the static adverse case, as is shown in Figure 5. We thus show that stressing only a subset of important variables is enough to substantially increase the probability of default of a portfolio of loans. It can further be noticed that the the PD increases the most in Eastern Europe. Overall, the RWA for the Reverse Stress Test is 0.148, and the RWA for the Base Case is 0.085. This corresponds to a more than 20% increase in the RWA, which, based on Zanders (2023), in most cases suggests a default of a bank holding such loan portfolio.

Note that the analysis for RVT is done on the static case (firm characteristics follow a RW) as was advised by Zanders (2023), because the dynamic approach would require a lot of iterations to obtain the dynamic characteristics when we look for the ideal scenario.

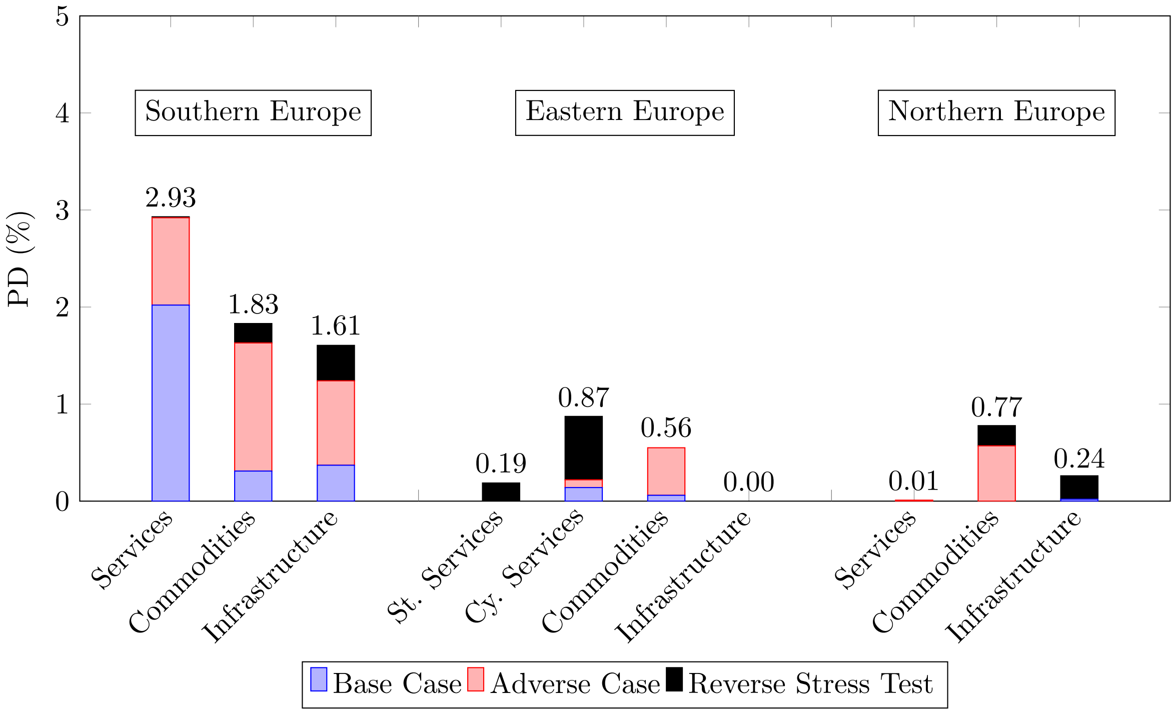


Figure 5: Results for War Scenario Reverse Stress Test. This figure illustrates the increase in probability of defaults across clusters over the EBA Base Case stress test scenario. The stressed variables and their magnitude for the Reverse Stress Test are visible in Table 7.

# 6 Conclusion

This paper performs a stress testing exercise on a portfolio of European SME loans. We introduce three novelties to this exercise. First, we compare the performance of a Machine Learning method against the current standard of Logistic Regression. Second, we deploy clusters to create additional interpretability on the most important variables that drive defaults. Finally, we model characteristics as a function of macroeconomic variables and compare the results of dynamic stress testing against a static approach.

Our results show that Random Forest performs substantially better than LASSO Logit in predicting PDs, with an AUC of 0.85 against 0.60. We further notice that our regional-sector clusters aid interpretability, as we find different important variables in different clusters. Applying the PD models to the total data would mean that these custom effects would be missed. For example, our Northern Europe Commodities cluster highlights the need of oil & gas prices, whilst these variables do not play an important role when we apply the model on the total data. Finally, we show that modeling firm characteristics as a function of their lags and macroeconomic variables, leads to an improvement in RMSE over a random walk. For the most important six characteristics, Dynamic LASSO offers an average RMSE improvement of 26%. We apply the predicted characteristics in our stress testing and show that this dynamic approach leads to higher stressed PDs than the static approach (1.72% vs 1.29%). This finding suggests that the static approach underestimates the stress, since it does not take into account any relationships between characteristics and macroeconomic variables.

The analysis performed in this paper is prone to some limitations which can lead to future research. First, a significant chunk of the lagged values we use consists of imputed values. We think these lags add value because they allow us to study dynamic effects, but nonetheless it should be noted that this may make these variables less reliable. An example of this was seen in Table 4 where Sales Lag has a positive effect on the PD. Future research can implement alternative imputation methods to investigate the robustness of our findings. Second, due to the large number of observations and an unbalanced panel, we "ignored" the panel data structure and treated each observation as it's own instance. This means for example that we ignore firm fixed effects in our analysis. Third, our clustering results consisted of some outliers which we did not remove to not intervene with the methods. Future research can try alternative clustering methods and investigate the robustness of our findings when outliers are removed. One related point is the further investigation of region only vs region-sector clusters. It might also be interesting to investigate clustering based on other characteristics such as sales or operating profit. Finally, we compare our predictions using metrics such as RMSE and AUC but do not test for statistical differences. Future research can investigate whether the substantial differences we find in predictions are also statistically significant.

# References

Abedin, M. Z., Guotai, C., Hajek, P. & Zhang, T. (2023). Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. , *9*(4), 3559–3579.

Alonso, A. & Carbo, J. M. (2020). Machine learning in credit risk: Measuring the dilemma between prediction and supervisory cost. *Banco de Espana*, 134.

Alonso, A. & Carbo, J. M. (2022). Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. *Financial Innovation*, *8*(1), 135.

Barboza, F., Kimura, H. & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405-417.

Bijmolt, T. H., Paas, L. J. & Vermunt, J. K. (2004). Country and consumer segmentation: multi-level latent class analysis of financial product ownership. *International Journal of Research in Marketing*, *4*(21), 324-340.

Breckenfelder, J. (2018). How is a firm's credit risk affected by sovereign risk? *European Central Bank*.

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

Brown, I. & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert systems with applications*, *39*(3), 3446-3453.

Castellucio, M. (2020). Machine learning to boom in 2020. *Strategic Finance*, *101*(8), 53-54.

Castren, O., Dées, S. & Zaher, F. (2009). Stress-testing euro area corporate default probabilities using a global macroeconomic model. *Journal of Financial Stability*, *6*(2), 64-78.

Castrén, O., Dées, S. & Zaher, F. (2010). Stress-testing euro area corporate default probabilities using a global macroeconomic model. *Journal of Financial Stability*, *6*(2), 64-78.

Cathcart, L., Dufour, A., Rossi, L. & Varotto, S. (2020). The differential impact of leverage on the default risk of small and large firms. *Journal of Corporate Finance*, *60*(1).

Chan-Lau, J. A. (2006). *Fundamentals-based estimation of default probabilities: a survey.*

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321357.

Chen. (2010). Using financial and macroeconomic indicators to forecast sales of large development and construction firms. *The Journal of Real Estate Finance and Economics*, *40*(1), 310-331.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system.

Chen, Y. & Zhang, R. (2021). Research on credit card default prediction based on k-means SMOTE and BP neural network. , *2021*, 1–13. Retrieved 2024-03-16, from https://www.hindawi.com/journals/complexity/2021/6618841/ doi: 10.1155/2021/6618841

Chmelar, A. (2013). Household debt and the european crisis. *ECRI Research Report No. 13*.

Costa, V. G. & Pedreira, C. E. (2023). Recent advances in decision trees: an updated survey. , *56*(5), 4765–4800. Retrieved 2024-02-15, from https://link.springer.com/10.1007/s10462-022-10275-5 doi: 10.1007/s10462-022-10275-5

Dastile, X., Celik, T. & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, *91*, 121.

Dupont, L., Fliche, O. & Yang, S. (2020). *Governance of artificial intelligence in finance.* Banque De France.

Egbunike, C. & Okerekeoti, C. (2018). Macroeconomic factors, firm characteristics and financial performance: A study of selected quoted manufacturing firms in nigeria. *Asian Journal of Accounting Research*, *3*, 142–168.

European Banking Authority. (2023). 2023 EU-wide stress test: Frequently Asked Questions.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. , *29*(5), 1189–1232. Retrieved 2024-02-15, from https://www.jstor.org/stable/2699986 (Publisher: Institute of Mathematical Statistics)

Graeve, F., Kick, T. & Koetter, M. (2008). Monetary policy and financial (in) stability: An integrated micro–macro approach. *Journal of Financial Stability*, *4*(3), 205-231.

Gruszczyński, M. (2019). On unbalanced sampling in bankruptcy prediction. *International Journal of Financial Studies*, *7*(2), 113.

Habshah Midi, S. S. & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, *13*(3), 253-267.

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

Hovakimian, A., Opler, T. & Titman, S. (2001). The debt-equity choice. *Journal of Financial and Quantitative analysis*, *36*(1), 1-24.

IEA. (2021a). *Crude oil production, regional ranking.* (https://www.iea.org/regions/europe/oil [Accessed: (16 March 2024)])

IEA. (2021b). *Natural gas supply, regional ranking.* (https://www.iea.org/countries/the-netherlands/natural-gas [Accessed: (16 March 2024)])

Institute of International Finance. (2019). Machine Learning in Credit Risk Report.

Institute of International Finance. (2020). Machine Learning: recommendations for policymakers.

Issah, M. & Antwi, S. (2017). Role of macroeconomic variables on firms' performance: Evidence from the uk. *Cogent Economics & Finance*, *5*.

Jacobson, T., Lindé, J. & Roszbach, K. (2005). Exploring interactions between real activity and the financial stance. *Journal of Financial Stability*, *1*(3), 308-341.

Jadhav, A., Pramod, D. & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, *33*(10), 913-933.

Jung, N. H., H. (2018). The analysis of data errors in financial information databases: New evidence from the korean financial markets. *South African Journal of Business Managemen*, *49*(1).

Kandil, M. & Mirzaie, I. (2005). The effects of exchange rate fluctuations on output and prices: evidence from developing countries. *The Journal of Developing Areas*, 182-219.

Kick, T. & Koetter, M. (2007). Slippery slopes of stress: ordered failure events in German banking. *Journal of Financial Stability*, *3*(2), 132-148.

Lei, G. & Ling, G. (2023x). Interpretability of machine learning: Recent advances and future prospects.

Leo, M., Sharma, S. & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. *Risks*, *7*(1), 129.

Lundberg, S. & Lee, S.-I. (2017). A unified approach to interpreting model predictions.

McCulloch, W. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 115-133.

Miu, P. & Ozdemir, B. (2008). Stress-testing probability of default and migration rate with respect to basel ii requirements.

Mo, J., Kiang, M. & P Zou, Y. L. (2010). A two-stage clustering approach for multi-region segmentation. *Expert Systems with Applications*, *10*(37), 7120-7131.

Orlando, G. & Pelosi, R. (2020). Non-performing loans for italian companies: When time matters. an empirical research on estimating probability to default and loss given default. *International Journal of Financial Studies*, *8*(4).

Pindyck, R. S. (2004). Volatility and commodity price dynamics. *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, *24*(11), 1029-1047.

Pollák, Z. & Popper, D. (2021). Stress tests in hungarian banking after 2008. *Acta Oeconomica*, *71*(3), 451-463.

Salas-Eljatib, C., Fuentes-Ramirez, A., Gregoire, T. G., Altamirano, A. & Yaitul, V. (2018). A study on the effects of unbalanced data when fitting logistic regression models in ecology. *Ecological Indicators*, *85*, 502-508.

Shapley, L. (1951). Notes on the n-person game ii: The value of an n-person game. *RAND RM*, 670.

Siddiqi, N. (2006). *Credit risk scorecards: developing and implementing intelligent credit scoring.*

Simons, D. & Rolwes, F. (2018). Macroeconomic default modeling and stress testing. *International Journal of Central Banking*, *18*, 129.

Sommar, P. A. & Shahnazarian, H. (2018). Interdependencies between expected default frequency and the macro economy. *International Journal of Central Banking*, *18*, 128.

S&P Global. (2024). *Global defaults: A growing threat.* (https://www.spglobal.com/marketintelligence/en/news-insights/research/global-defaults-a-growing-threat [Accessed: (16 March 2024)])

Statista. (2022). *Leading export countries worldwide in 2022.* (https://www.statista.com/statistics/264623/leading-export-countries-worldwide/ [Accessed: (16 March 2024)])

Strebulaev, I. A. & Yang, B. (2013). The mystery of zero-leverage firms. *Journal of financial Economics*, *109*(1), 1-23.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267-288.

TradingEconomics. (2024). *Government debt ratings.* (https://tradingeconomics.com/country-list/rating?continent=europe [Accessed: (16 March 2024)])

Tserng, H. P., Chen, P. C., Huang, W. H., M. C., Lei & Tran, Q. H. (2014). Prediction of default probability for construction firms using the logit model. *Journal of Civil Engineering and Management*, *20*(2), 247-255.

Vieira, E., Neves, M. & Dias, A. (2019). Determinants of portuguese firms' financial performance: panel data evidence. *International Journal of Productivity and Performance Managemen*, *68*, 1323–1343.

Westgaard, S. & van der Wijst, N. (2001). Default probabilities in a corporate bank portfolio: A logistic model approach. *European Journal of Operational Research*, *135*(2), 338-349.

Yang, Miin-Shen, Lai, C.-Y. & Lin., C.-Y. (2012). A robust em clustering algorithm for gaussian mixture models. *Pattern Recognition*, *45*(11), 3950-3961.

Zanders. (2023). *Seminar case studies: kick-off presentation.*

Zhang, Q., Hu, Y., Jiao, J. & Wang, S. (2023). The impact of russia–ukraine war on crude oil prices: an EMC framework. , *11*(1), 8. Retrieved from https://doi.org/10.1057/s41599-023-02526-9 doi: 10.1057/s41599-023-02526-9

Zhou, B., Yang, C., Guo, H. & Hu, J. (2013). A quasi-linear svm combined with assembled smote for imbalanced data classification. *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 1–7.

# A Background Information

## A.1 Mice

The firm characteristic data provided by Zanders suffered from significant sparsity, with 23.3% of the accounting data missing. In order to proceed with our analysis, a complete dataset was necessary and consequently the need for imputation arose. We opted to use the procedure known as Multivariate Imputation by Chained Equations (MICE), summarized in Algorithm 3.

**Algorithm 3:** Bayesian Ridge Imputation Algorithm

```
Specify a Bayesian Ridge imputation model
  $P(Characteristic_j^{\text{mis}} \mid Characteristic_j^{\text{obs}}, Characteristic_{-j}, R)$ for variable $Characteristic_j$
  with $j = 1, \dots, p$;
foreach $j = 1$ to $p$ do
    Fill in starting imputations $\widehat{Characteristic}_j^{0}$ by random draws from $Characteristic_j^{\text{obs}}$;
end
for $t = 1$ to $M$ do
    for $j = 1$ to $p$ do
        Define $\widehat{Characteristic}_{-j}^{t} =$
          $(\widehat{Characteristic}_1^{t}, \dots, \widehat{Characteristic}_{j-1}^{t}, \widehat{Characteristic}_{j+1}^{t-1}, \dots, \widehat{Characteristic}_p^{t-1})$
          as the currently complete data except $Characteristic_j$;
        Draw $\hat{\phi}_j^t \sim P(\phi_j^t \mid Characteristic_j^{\text{obs}}, \widehat{Characteristic}_{-j}^{t}, R)$;
        Draw imputations $\widehat{Characteristic}_j^{t} \sim P(Characteristic_j^{\text{mis}} \mid$
          $Characteristic_j^{\text{obs}}, \widehat{Characteristic}_{-j}^{t}, R, \hat{\phi}_j^t)$;
    end
end
```

Formally, let $Characteristic \in \mathbb{R}^{n \times p}$ denote the data and let $R \in \{0,1\}^{n \times p}$ denote the indicator matrix such that $r_{ij} = 0$ if $Characteristic_{ij}$ is missing and $r_{ij} = 1$ of $Characteristic_{ij}$ is observed. Let $Characteristic_j$ denote the $j^{\text{th}}$ column of $Characteristic$ and let $Characteristic_{-j}$ denote the data less $Characteristic_j$. Finally, for $j \in \{1, \dots, p\}$ let $Characteristic^{\text{mis}}$ denote the missing data characterised by $Characteristic_{ij}$ such that $r_{ij} = 0$ and $Characteristic^{\text{obs}}$ the observed data characterised by $Characteristic_{ij}$ such that $r_{ij} = 1$. Finally, $\{\phi_j\}$ are parameters of the model. We use the default setting of MICE which corresponds with a Bayesian Ridge regression.

## A.2 Weight of Evidence (WoE)

For the categorical variables, country and sector, we construct the Weight of Evidence (WoE) to use in our models. This helps us in reducing computational time since we have only 1+1 rather than 22+19 variables to use. Another advantage of the WoE is that it increases interpretability by taking into account the number of good and bad cases of the outcome variable (Siddiqi, 2006). This is especially useful for our objective of clustering, since we can include default information in our cluster creation. We do not apply the WoE on the continuous variables, since this requires a binning procedure that may lead to a loss of information (Siddiqi, 2006). For every country and sector category, we calculate the WoE as:

$$WoE = \ln\left(\frac{\text{Number of non-defaults per category}}{\text{Number of defaults per category}}\right) \tag{13}$$

## A.3 Shapley Values

The Shapley value is a game-theoretic measure used to asses the contribution of an individual agent withing a cooperative game. Formally, let $N$ be a set of agents with cardinality $n$ and let $v : \mathcal{P}(N) \rightarrow \mathbb{R}$ be a score function such that $v(\emptyset) = 0$. Then, for an agent $i \in N$ the Shapley value is defined as:

$$S_i(N, v) = \sum_{A \subseteq N \setminus \{i\}} \frac{(n - 1 - |A|)!|A|!}{n!}[v(A \cup \{i\}) - v(A)]. \tag{14}$$

Intuitively, $v(A \cup \{i\}) - v(A)$ denotes the marginal contribution provided by agent $i \in N$ to a predetermined group $A$, while $(n-1-|A|)!|A|!/n!$ acts as a weight accounting for all potential permutations arising from the construction of the coalition $A \cup \{i\}$. Hence, by construction, the Shapley value can be interpreted as the expected marginal contribution any agent $i \in N$ provides within the cooperative game.

## A.4 Basel Framework

The Basel Framework gives specific formula for calculating risk weighted assets:

$$R = 0.12 \cdot \frac{1 - e^{-50}}{1 - e^{-50 \cdot PD}} + 0.24 \cdot \left(1 - \frac{1 - e^{-50 \cdot PD}}{1 - e^{-50}}\right), \tag{15}$$

$$K = LGD \cdot N\left(\frac{G(PD}{\sqrt{1 - R}} + \sqrt{\frac{R}{1 - R}} \cdot G(0.999)\right) - PD \cdot LGD. \tag{16}$$

The RWA is then computed as:

$$RWA = K \cdot 12.5 \cdot EAD. \tag{17}$$

## A.5 Clustering Algorithms

**Algorithm 4:** Gaussian Mixture Model Estimation

```
Initialize $\pi^{(0)}$ and $\theta^{(0)}$ with some values;
Set $k \leftarrow 0$;
repeat
    // E-Step;
    for $i \leftarrow 1$ to $m$ and $j \leftarrow 1$ to $J$ do
        Calculate $p_j^{(k)} = P(z_j = i|x_j; \pi^{(k)}, \theta^{(k)})$;
    // M-Step;
    Construct the likelihood function
     $Q(\pi, \theta|\pi^{(k)}, \theta^{(k)}) = \sum_{j=1}^{n} \sum_{i=1}^{m} p_j^{(k)} [\log(\pi_i) + \log(\Phi(x_j; \theta_i))]$;
    Update the parameters: $(\pi^{(k+1)}, \theta^{(k+1)}) = \arg\max_{\pi,\theta} Q(\pi, \theta|\pi^{(k)}, \theta^{(k)})$;
    $k \leftarrow k + 1$;
until convergence;
```

**Algorithm 5:** K-Means Estimation

```
Choose initial means $\mu_1^{(0)}, \dots, \mu_k^{(0)}$;
Set $t \leftarrow 0$;
repeat
    // Assignment step: Assign each observation $x_1, \dots, x_n$ uniquely to the cluster with
      the nearest mean;
    $S_i^{(t)} \leftarrow \left\{ x_\alpha : \left\| x_\alpha - \mu_i^{(t)} \right\|^2 \leq \left\| x_\alpha - \mu_j^{(t)} \right\|^2, \forall j \in \{1, \dots, k\} \right\}$;
    // Update step:;
    $\mu_i^{(t+1)} \leftarrow \frac{1}{\left|S_i^{(t)}\right|} \sum_{x \in S_i^{(t)}} x$;
    $t \leftarrow t + 1$;
until convergence;
```

## A.6 Desicion Trees

The idea of DT is to partition the feature space into smaller regions until a final region is reached; the prediction of default or non-default then equals to the most prevailing class in a given region. If most companies in that region did not default then this region predicts no-default. The goal in training the Decision Tree is to find such splitting rules of the feature space that provide the most consistent leafs for the predictions. This is done by minimizing the respective loss function (criterion). A criterion is used to evaluate the quality of splits during the tree-building process. For classification purposes, the following two criteria are most often used (Friedman, 2001):

- Gini Index

$$c_{\text{Gini}}(\mathcal{T}) = |I_\mathcal{T}| \sum_k p_\mathcal{T}(k)\big(1 - p_\mathcal{T}(k)\big)$$

  where $\mathcal{T}$ denotes the leaf (region), $|I_\mathcal{T}|$ is the set of features in leaf $\mathcal{T}$ and $p_\mathcal{T}$ represent the average probability for class $k$ in a region $\mathcal{T}$. This measure thus describes how often a randomly chosen training point would be misclassified if it was randomly labeled based on $p_{\mathcal{T}}$.

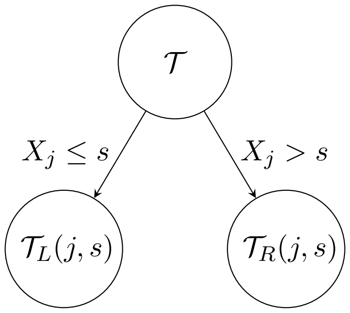
- Entropy Criterion defined as

$$c_{\text{entropy}}(\mathcal{T}) = -|I_{\mathcal{T}}| \sum_{\substack{k \\ p_{\mathcal{T}}(k)\neq 0}} p_{\mathcal{T}}(k) \log p_{\mathcal{T}}(k)$$

measures the uncertainty of training points in leaf $\mathcal{T}$ with respect to their class labels.

The training algorithm of DT as introduced in Breiman et al. (1984), in their famous book Classification and Regression Tree (CART), runs in a following way:

1. At the beginning, all training data $X = [Characteristics, Characteristics', Macros]^{\intercal}$ are in a single node, called the root node. Here starts the node splitting, it is done in a so called greedy way, i.e., the algorithm for loop over all features $(J)$ and their unique values $(S)$ to find such a splitting rule that minimizes the criterion value. The Figure 6 shows splitting of a node $\mathcal{T}$ into $\mathcal{T}_L(j,s) = \{X|X_j \leq s\}$ and $\mathcal{T}_R(j,s) = \{X|X_j > s\}$. The goal is to obtain new nodes such that after the split the overall criterion is the lowest possible, this corresponds to minimizing the difference of after split and before split criterion value over all possible splits, i.e., $\min c_{\mathcal{T}_L} + c_{\mathcal{T}_R} - c_{\mathcal{T}}$, where $c_{\mathcal{T}_L} + c_{\mathcal{T}_R}$ describes the overall criterion value after split and $c_{\mathcal{T}}$ is the before split criterion.



2. *Note:* This figure depicts the splitting of node $\mathcal{T}$ into $\mathcal{T}_L$ and $\mathcal{T}_R$ based on variable $j$ and its value $s$.

Figure 6: Example of Node Splitting

3. Once the optimal splitting point is established based on the criterion value, two new child nodes emerge. The same procedure as in step 1. is repeated on the newly created nodes.

4. The splitting of new nodes is done until a pre-defined constraint is reached: the tree reaches maximum depth, minimum number of training points on a node to split on is matched or a maximum number of leaf nodes is attained.

5. In order to prevent overfiting, prunning based on cost complexity is usually preformed

$$C_\alpha(\mathcal{T}) = \sum_{m=1}^{|\mathcal{T}|} \sum_{x_j \in \mathcal{T}_m} c_m + \alpha|\mathcal{T}|$$

where $m$ denotes the respective nodes and $c_m$ is the criterion used for splitting the tree. This approach prevent overfiting by penalizing for a large trees, the constant $\alpha$ determines the strength of the penalty on tree size. Note that $\alpha$ needs to be tuned to obtain its value – k-fold cross validation is used in most cases.

6. Predictions are obtained by employing the majority voting of company labels in each final leaf. That is, for companies falling into the terminal leaf $\mathcal{T}_m$, obtain the proportion of class $k \in \{\text{default, non-default}\}$ in the given leaf as

$$\hat{PD}_{mk} = \frac{1}{|\mathcal{T}_m|} \sum_{x_i \in \mathcal{T}_m} I(PD_i = k)$$

where $I(PD_i = k)$ is the indicator function that return 1 if label of $X_i$ belongs to class $k$. Next, we look for the class that gives the highest proportion in the leaf $\mathcal{T}_m$:

$$\hat{PD}_m = \underset{k \in \{\text{default, non-default}\}}{\text{argmax}} \hat{PD}_{mk}$$

where $\hat{PD}_m$ defines the class prediction for leaf $\mathcal{T}_m$.

In this way, the tree building process is finished – splitting points are determined and final leaves have assigned labels based on majority voting that determine the predictions. Hence, upon inference, when new data are inputted into the trained model the prediction is obtained based on in which final leaf the new test point will end.

As Costa & Pedreira (2023) summarize, Decision Trees are a versatile method for modelling non-linear relationships and as a consequence are used in a wide range of application. What is more, DTs tend to be popular due to its good relative interpretability compared to other Machine Learning methods. The interpretability stems from the setting of the algorithm that splits data based on a specific splitting rules that are easy to comprehend and visualize; the interpretability however decreases with higher dimensionality of feature data and complexity of a tree. One of the main disadvantage that is well-recognized, is its instability, i.e., deep trees tend to have high variance, while shallow trees produce high bias. In order to overcome this drawback, ensembles of many trees constructed. Bagging aggregation averages over predictions of many trees, where each tree is trained on a bootstrapped subset of the dataset. The individual trees are grown to reach a low bias, i.e., deep trees are fitted. The averaging over fitted deep trees reduces the variance. However, this method assumes that trees are independent, which may not hold in the reality, hence to relax this assumption Breiman (2001) proposes to use Random Forests that reduces correlation among trees by randomly selecting a pool of variables which will by considered for splitting.

# B Results Extra

## B.1 Full Depiction of Clusters

1. **Southern Europe (SE):** Italy, Portugal, Spain

   (a) **Services (SERV):** Wholesale and Retail Trade, Manufacturing, Professional Services, Information and Communication, Art & Entertainment, Finance

   (b) **Commodities, Food and Agriculture (COMD):** Agriculture, Energy, Utilities, Defense activities, Mining, Accommodation and Food Services, Human Health and Social Work

   (c) **Infrastructure Development and Construction (INFR):** Construction, Transportation, Technical Services, Real Estate

2. **Eastern Europe (EE):** Bulgaria, Croatia, Czech Republic, Hungary, Latvia, Lithuania, Poland, Iceland, Romania, Slovakia, Belgium,

   (a) **Infrastructure Development and Construction (INFR):** Construction, Transportation, Technical Services, Real Estate

   (b) **Stable Demand Services (StSERV):** Professional services, Information and Communication, Utilities, Energy, Entertainment, Finance

   (c) **Commodities, Food and Agriculture (COMD):** Agriculture, Social Work Activities, Accommodation and Food, Mining, Defense

   (d) **Cyclical Services (CySERV:** Wholesale and Retail Trade, Manufacturing, Arts & Entertainment

3. **Northern Europe (NE):** Denmark, Finland, Germany, Netherlands, Austria, Norway, Slovenia, Sweden.

   (a) **Services (SERV):** Wholesale and Retail Trade, Manufacturing, Professional Services, Information and Communication, Art & Entertainment, Finance

   (b) **Commodities, Food and Agriculture (COMD):** Agriculture, Energy, Utilities, Defense activities, Mining, Accommodation and Food Services, Human Health and Social Work

   (c) **Infrastructure Development and Construction (INFR):** Construction, Transportation, Technical Services, Real Estate

## B.2 Descriptive Statistics for the clusters

| | SE_SERV | SE_COMD | SE_INFR | EE_INFR | EE_StSERV | EE_CySERV |
|---|---|---|---|---|---|---|
| Intangible Fix. Assets | 0.22 (1.48) | 0.59 (3.26) | 0.41 (6.55) | 0.12 (1.81) | 0.36 (3.74) | 0.11 (1.67) |
| Tangible Fix. Assets | 1.32 (4.09) | 6.65 (14.93) | 3.23 (15.37) | 5.02 (16.19) | 4.74 (45.89) | 5.73 (8.67) |
| Oth. Fix. Assets | 0.7 (15.05) | 1.99 (15.75) | 3.19 (23.73) | 2.91 (70.33) | 3.82 (123.94) | 0.94 (14.44) |
| Stock | 1.08 (4.45) | 0.51 (2.27) | 2.97 (19.78) | 0.88 (17.11) | 0.38 (3.93) | 0.9 (1.34) |
| Debtors | 1.59 (1.81) | 1.52 (2.07) | 1.91 (3.62) | 1.39 (5.61) | 1.75 (9.83) | 1 (1.67) |
| Oth. Curr. Assets | 1.49 (6.38) | 2.49 (6.31) | 2.82 (54.97) | 15.89 (2324.4) | 20.77 (1691.04) | 1.49 (6.55) |
| Cash Equivalents | 0.69 (2.66) | 1 (3.39) | 1.09 (17.8) | 0.94 (2.64) | 1.42 (9.71) | 0.89 (4.49) |
| Capital | 0.48 (4.98) | 1.44 (6.11) | 1.53 (17.75) | 1.93 (10.89) | 2.6 (29.15) | 2.24 (6.48) |
| Shareholder Funds | 1.93 (13.73) | 3.64 (12.57) | 3.32 (21.26) | 2.89 (69.8) | 4.01 (120.16) | 3.21 (16.03) |
| Long Term Debt | 0.75 (9.34) | 3.65 (12.41) | 3.32 (19.09) | 3.26 (143.49) | 2.8 (125.13) | 1.54 (5.39) |
| Oth. Noncurr. Liabilities | 0.44 (4.1) | 1.24 (4.92) | 0.99 (5.79) | 1.77 (28.73) | 1.44 (44.44) | 0.84 (3.88) |
| Loans | 0.62 (2.09) | 0.86 (3.13) | 1.25 (11.47) | 0.7 (4.3) | 0.47 (4.9) | 0.46 (1.78) |
| Creditors | 1.11 (1.27) | 1.01 (1.75) | 1.22 (2.74) | 0.87 (1.64) | 0.85 (4.09) | 0.71 (1.12) |
| Oth. Curr. Liabilities | 1.09 (5.11) | 1.94 (6.91) | 2.94 (46.17) | 1.65 (18.42) | 2.73 (36.78) | 1.21 (2.96) |
| Sales | 6.78 (1.49) | 6.48 (1.71) | 6.46 (2.27) | 6.54 (2.08) | 6.62 (1.91) | 6.18 (1.74) |
| EBIT | 0.27 (1.69) | 0.55 (1.93) | 0.35 (5.91) | 0.57 (1.78) | 0.53 (2.57) | 0.41 (1.98) |
| Financial Revenue | 0.05 (1.4) | 0.06 (2.41) | 0.11 (2.62) | 0.17 (1.78) | 0.32 (8.25) | 0.1 (0.73) |
| Financial Expenses | 0.08 (1.57) | 0.26 (1.6) | 0.26 (2.63) | 0.31 (2.34) | 0.16 (21.05) | 0.15 (2.06) |
| Taxation | 0.08 (0.26) | 0.12 (0.54) | 0.11 (1.09) | 0.07 (0.66) | 0.1 (0.6) | 0.06 (0.22) |
| Material Costs | 4.15 (2.2) | 2.09 (2.33) | 2.62 (5.49) | 2.17 (2.09) | 1.81 (7.35) | 2.5 (1.78) |
| Costs of Employees | 1.05 (1.05) | 1.68 (1.5) | 1.41 (1.49) | 1.02 (1.26) | 1.57 (1.88) | 1.55 (1.32) |
| Dep.&Amort. | 0.18 (0.45) | 0.51 (0.86) | 0.28 (1.67) | 0.32 (0.85) | 0.35 (4.04) | 0.44 (0.66) |
| Cash Flow | 0.33 (2.13) | 0.78 (2.67) | 0.38 (6.74) | 0.66 (2.48) | 0.88 (10.33) | 0.74 (2.48) |
| Number of employees | 29 | 62 | 41 | 74 | 76 | 113 |
| Number of defaults | 2387 | 633 | 1283 | 330 | 178 | 152 |
| Number of observations | 364383 | 45983 | 117536 | 67271 | 32338 | 22796 |

| | EE_COMD | NE_SERV | NE_COMD | NE_INFR | Total Data |
|---|---|---|---|---|---|
| Intangible Fix. Assets | 0.07 (1.03) | 0.25 (3.94) | 0.84 (21.42) | 0.15 (3.24) | 0.24 (4.59) |
| Tangible Fix. Assets | 1.64 (21.1) | 1.39 (9.03) | 7.53 (29.85) | 10.59 (162.29) | 3.1 (47.37) |
| Oth. Fix. Assets | 0.43 (29.94) | 2.03 (53.24) | 2.83 (130.84) | 3.19 (75.7) | 1.66 (48.24) |
| Stock | 0.95 (1.25) | 0.83 (1.72) | 0.29 (1.45) | 0.88 (14.36) | 1.13 (9.24) |
| Debtors | 1.02 (1.55) | 0.84 (35.07) | 0.76 (9.24) | 0.77 (14.06) | 1.32 (14.26) |
| Oth. Curr. Assets | 1.55 (251.51) | 2.1 (25.44) | 2.6 (13.62) | 3 (23.16) | 3.4 (658.59) |
| Cash Equivalents | 0.48 (1.99) | 0.92 (6.66) | 1.27 (4.65) | 1.29 (8.91) | 0.85 (7.41) |
| Capital | 0.75 (12.26) | 0.98 (29.87) | 1.91 (39.97) | 2.15 (16.17) | 1.1 (17.01) |
| Shareholder Funds | 1.56 (35.35) | 2.43 (37.48) | 4.79 (43.9) | 4.11 (58.81) | 2.56 (39.59) |
| Long Term Debt | 0.49 (22.93) | 1.28 (23.54) | 3.24 (17.85) | 6.09 (147.96) | 1.9 (60.64) |
| Oth. Noncurr. Liabilities | 0.39 (6.21) | 0.66 (12.09) | 2.41 (83.71) | 2.71 (18.28) | 0.9 (18.6) |
| Loans | 0.34 (1.09) | 0.26 (3.42) | 0.4 (4.09) | 0.77 (13.35) | 0.61 (5.85) |
| Creditors | 0.82 (1.51) | 0.5 (2.58) | 0.5 (3.85) | 0.48 (1.21) | 0.89 (2.01) |
| Oth. Curr. Liabilities | 0.7 (3.75) | 1.45 (24.39) | 1.78 (21.56) | 2.32 (18.22) | 1.52 (20.77) |
| Sales | 6.82 (1.55) | 6.87 (2.11) | 6.6 (2.21) | 6.58 (5.3) | 6.7 (2.28) |
| EBIT | 0.34 (0.89) | 0.3 (2.55) | 0.34 (6.96) | 0.7 (2.26) | 0.37 (2.84) |
| Financial Revenue | 0.11 (12.21) | 0.21 (6.58) | 0.29 (13.36) | 0.34 (10.49) | 0.13 (6.73) |
| Financial Expenses | 0.11 (4.06) | 0.17 (9.05) | 0.32 (5.29) | 0.39 (3.21) | 0.17 (5.59) |
| Taxation | 0.07 (1.27) | 0.09 (0.79) | 0.06 (2.57) | 0.11 (0.57) | 0.09 (0.85) |
| Material Costs | 3.63 (2.92) | 3.56 (4.32) | 2 (5.86) | 2.68 (5.58) | 3.34 (3.87) |
| Costs of Employees | 0.79 (0.77) | 1.54 (1.81) | 2.24 (2.16) | 1.53 (1.65) | 1.24 (1.39) |
| Dep.&Amort. | 0.18 (2.19) | 0.2 (2.18) | 0.55 (6.12) | 0.41 (2.82) | 0.25 (1.97) |
| Cash Flow | 0.47 (2.87) | 0.42 (5.44) | 0.66 (4.56) | 0.81 (5.52) | 0.48 (4.35) |
| Number of employees | 56 | 38 | 66 | 44 | 46 |
| Number of defaults | 766 | 600 | 107 | 305 | 6741 |
| Number of observations | 161367 | 155667 | 26329 | 82256 | 1075926 |

Table 8: Descriptive statistics. This table displays the descriptive statistics for each cluster and the total data. The balance sheet items are shown in millions, whilst all other variables are in raw numbers. The clusters are named as ”Region_Sector”, e.g. Southern Europe (SE) Services (SERV) is annotated as SE_SERV.

### B.3 RF interpretability: Beeswarm Plots for Clusters

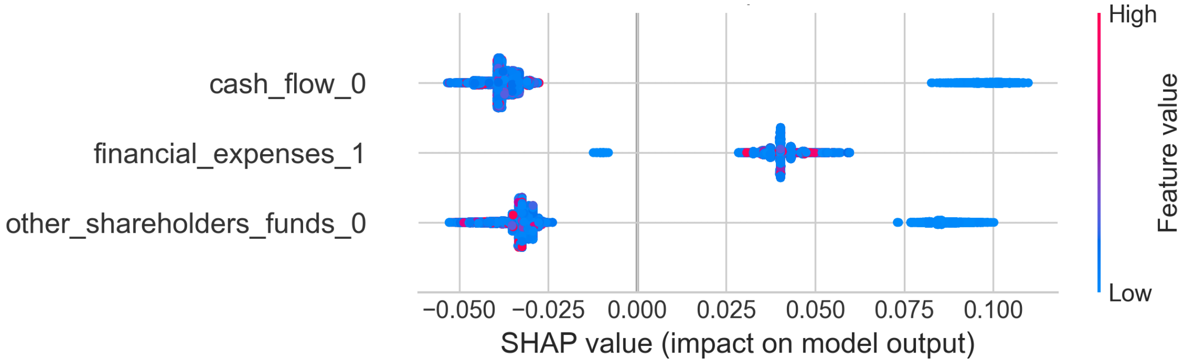


Figure 7: Beeswarm Plot for Southern Europe Services cluster

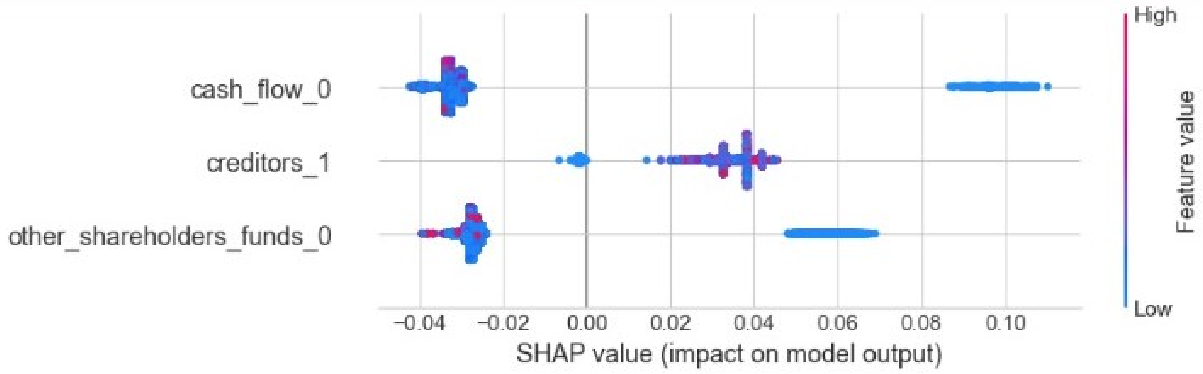


Figure 8: Beeswarm Plot for Southern Europe Commodities cluster

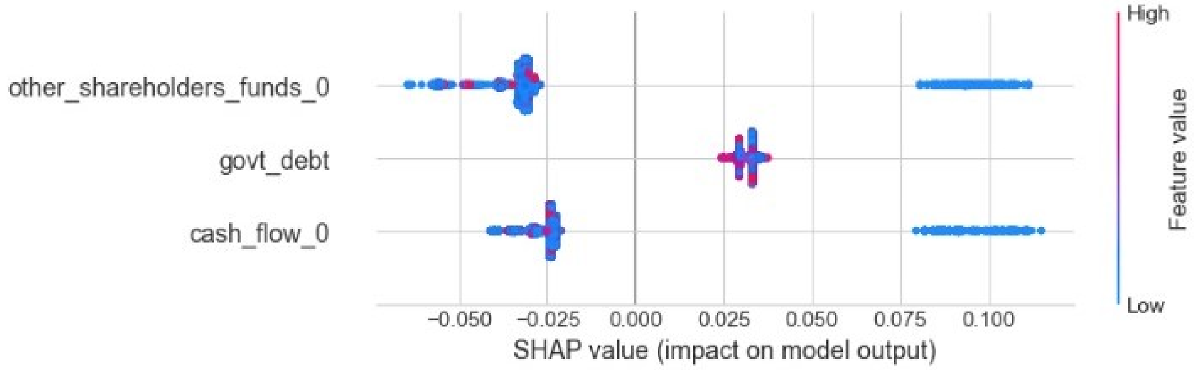


Figure 9: Beeswarm Plot for Southern Europe Infrastructure cluster

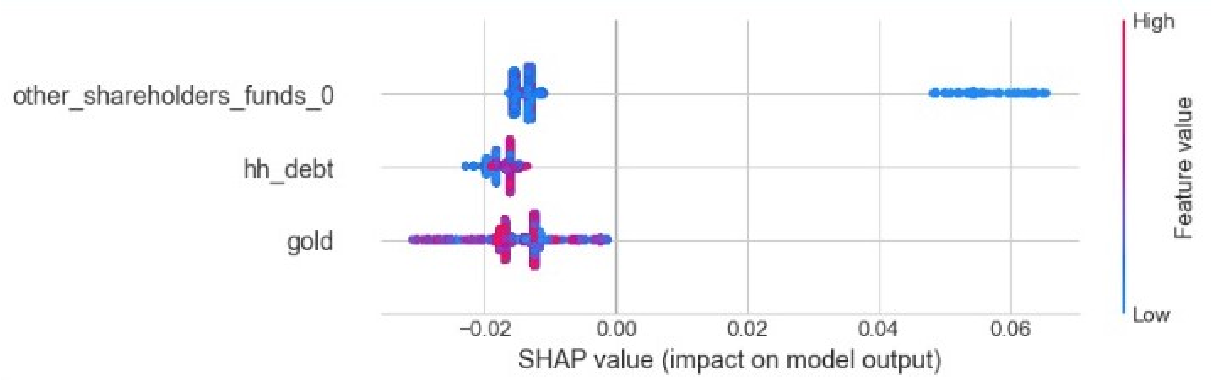


Figure 10: Beeswarm Plot for Eastern Europe Infrastructure cluster

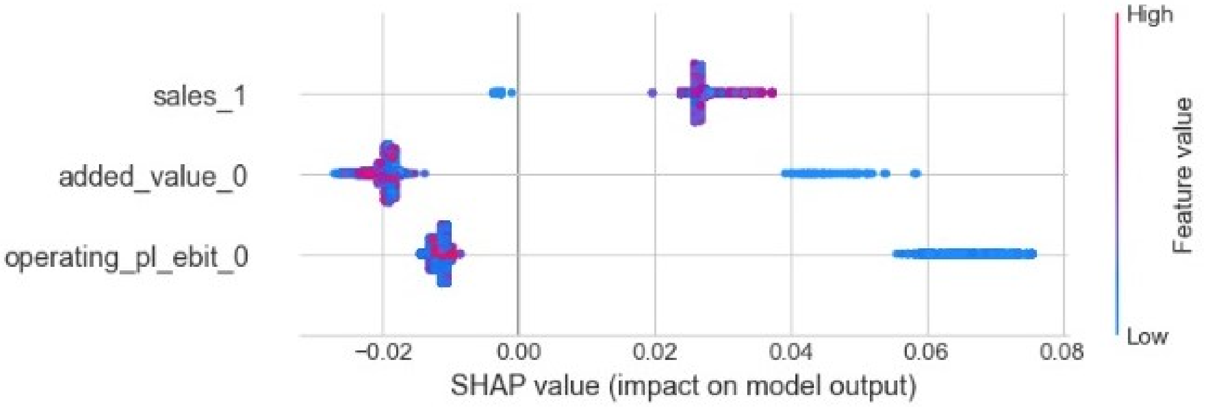


Figure 11: Beeswarm Plot for Eastern Europe Stable Services cluster

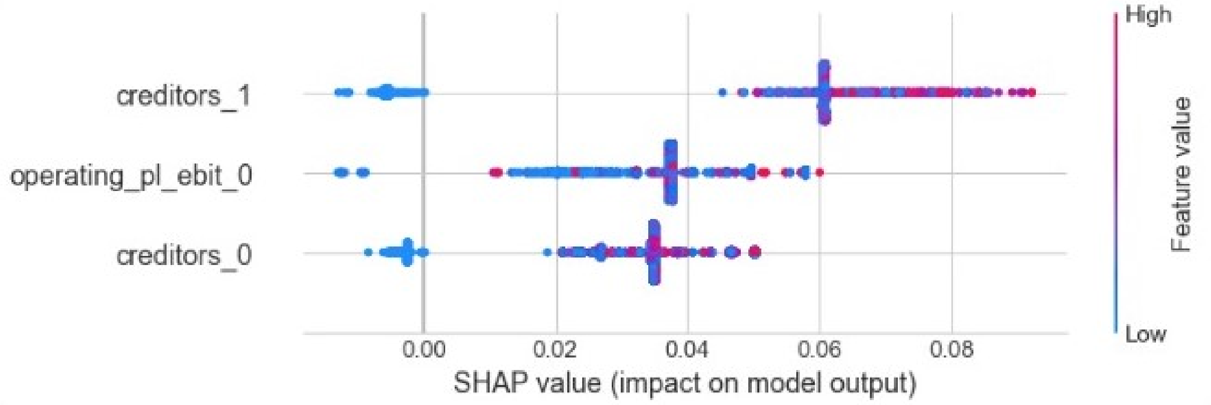


Figure 12: Beeswarm Plot for Eastern Europe Commodities cluster

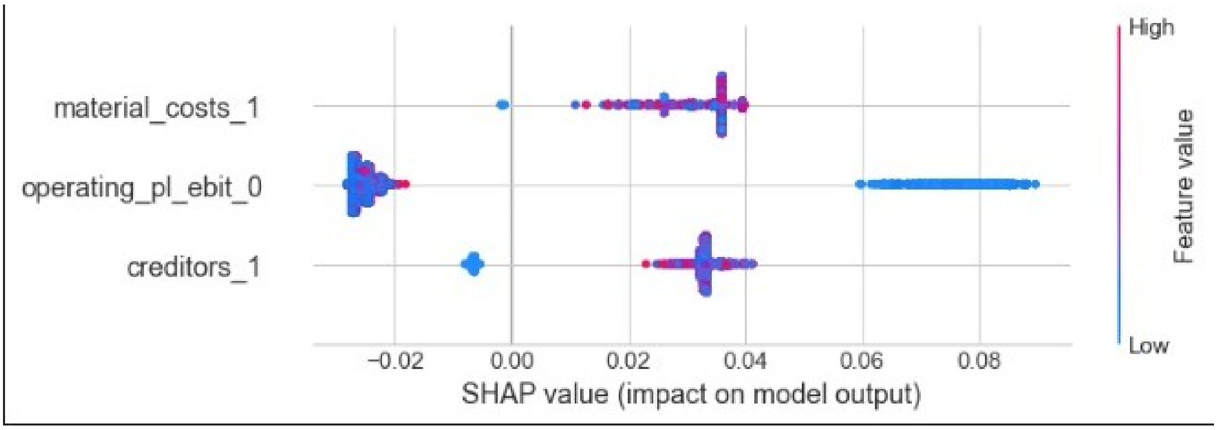


Figure 13: Beeswarm Plot for Eastern Europe Cyclical Services cluster

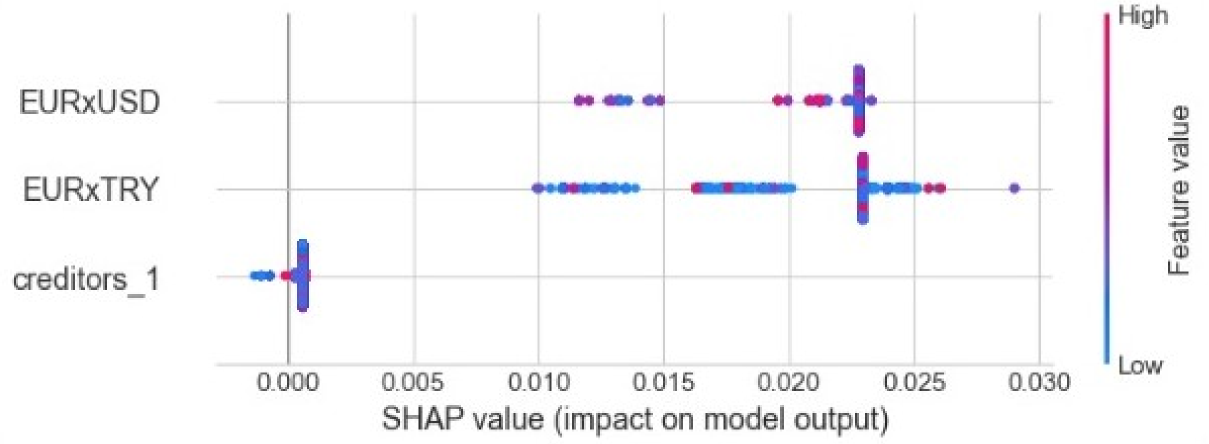


Figure 14: Beeswarm Plot for Northern Europe Services cluster

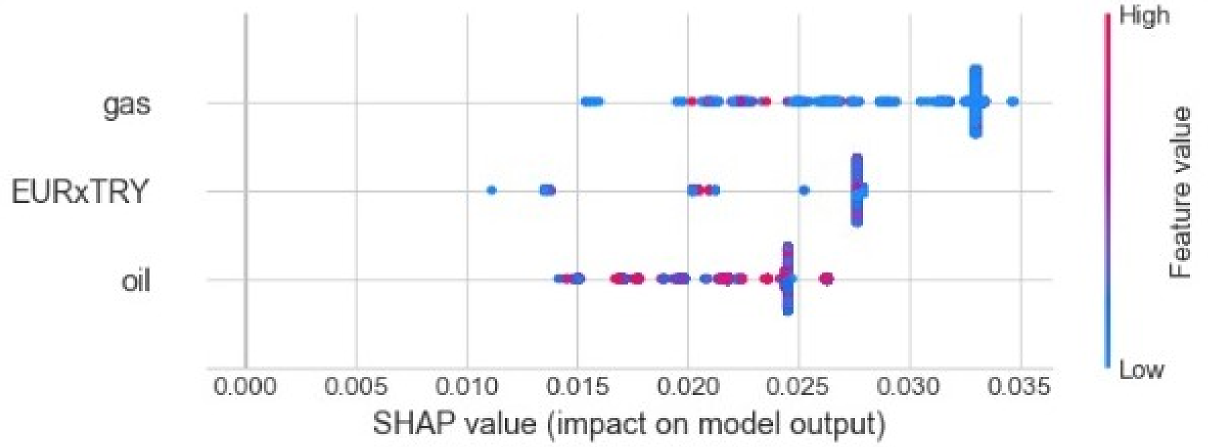


Figure 15: Beeswarm Plot for Northern Europe Commodities cluster

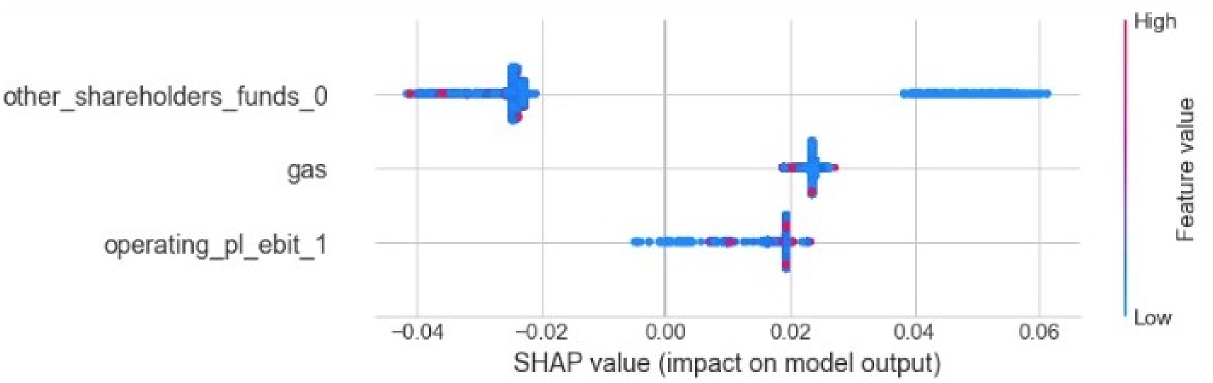


Figure 16: Beeswarm Plot for Northern Europe Infrastructure cluster

## B.4 LASSO Logit interpretability: Important Variables Detailed

| SE_SERV | SE_COMD | SE_INFR | EE_INFR |
|---|---|---|---|
| Govt Debt (0.28) | Govt Debt (0.36) | 10y Yield (0.87) | 3m Yield (-0.56) |
| GDP Growth (-0.14) | 10y Yield (0.42) | Govt Debt (0.37) | Hhold Debt (-0.46) |
| Unemployment (0.13) | Material Costs (0.34) | EURxJPY (-0.7) | Sales (-0.42) |
| Sales (-0.34) | Shareholder Funds (-0.56) | Cash Flow (-0.73) | EURxTRY (-0.21) |
| 3m Yield (-0.47) | EURxJPY (-0.28) | Oth. Fix. Assets lag (-0.45) | EURxJPY (-0.47) |

Table 9: Feature importance for RF (1/3). The table display the five most significant regressors for these clusters: Southern Region Services, Commodities, Infrastructure and Eastern Region Infrastructure, in terms of coefficient absolute value as found by the penalized Logistic Regression. The name of each variable is given with its estimate in brackets.

| EE_StSERV | EE_CySERV | EE_COMD | NE_SERV |
|---|---|---|---|
| Sales (-0.29) | Debtors lag (0.08) | Cash Equivalent (-1.52) | Corporate Debt (0.34) |
| Added Value (-0.28) | Creditors lag (0.14) | 3m Yield(-0.44) | Oth. Curr. Liabilities (-0.47) |
| Cash Flow lag (-0.24) | Sales (-0.08) | EURxJPY (-0.49) | Shareholder Funds lag (-0.7) |
| Hhold Debt (-0.26) | Op.PL EBIT lag (-0.08) | Hhold Debt (-0.33) | Capital (-0.37) |
| 3m Yield (-0.47) | Dep.&Amort. (-0.33) | Sales (-0.33) | Govt Debt (0.15) |

Table 11: Feature importance for RF (2/3). The table display the five most significant regressors for these clusters: Southern Region Services, Commodities, Infrastructure and Eastern Region Infrastructure, in terms of coefficient absolute value as found by the penalized Logistic Regression. The name of each variable is given with its estimate in brackets.

| NE_COMD | NE_INFR | All Data |
|---|---|---|
| Cash Flow (-0.17) | Tangible Fix. Assets (-0.38) | Cash Equivalent lag(-1.18) |
| Sales (-0.11) | Debtors (-0.57) | Shareholders Funds (-0.95) |
| EURxNOK (0.1) | Oth. Shareholder Funds (-0.31) | Capital (-0.65) |
| Copper (-0.11) | EURxCHF: (-0.24) | EURxJPY (-0.62) |
| Op.PL EBIT lag (-0.07) | Corporate Debt (0.38) | Debtors (-0.49) |

Table 13: Feature importance for RF (3/3). The table display the five most significant regressors for these clusters: Southern Region Services, Commodities, Infrastructure and Eastern Region Infrastructure, in terms of coefficient absolute value as found by the penalized Logistic Regression. The name of each variable is given with its estimate in brackets.

### B.5 XGBoost AUC Results

| Region Cluster | Sector Cluster | LASSO | Random Forest | XGBoost |
|---|---|---|---|---|
| Southern Europe | Services | 0.77 | **0.90** | **0.91** |
|  | Commodities | 0.76 | **0.88** | **0.89** |
|  | Infrastructure | 0.70 | **0.88** | **0.89** |
| Eastern Europe | Stable Services | 0.61 | 0.83 | 0.73 |
|  | Cyclical Services | 0.61 | 0.78 | 0.79 |
|  | Commodities | 0.69 | 0.79 | 0.78 |
|  | Infrastructure | 0.66 | 0.77 | 0.79 |
| Northern Europe | Services | 0.60 | 0.71 | 0.75 |
|  | Commodities | 0.55 | 0.63 | 0.59 |
|  | Infrastructure | 0.55 | 0.76 | 0.81 |
| All Data |  | 0.60 | 0.85 | 0.83 |

Table 14: auc values for different models. This table presents AUC values of LASSO and Random Forest for different clusters. Both models are run on standardized input variables and with SMOTE. The Region clusters are formed on the total data, whilst the Sector Clusters are formed on observations from the resulting region clusters. Services for Southern and Northern Europe covers the sectors Wholesale and Retail Trade, Manufacturing, Professional Services, Information and Communication, Art & Entertainment, Finance. Services for Eastern Europe is the sum of Cyclical Services and Stable Services where Cyclical Services covers Wholesale and Retail Trade, Manufacturing, Arts & Entertainment and Stable Services includes Professional services, Information and Communication, Utilities, Energy, Entertainment, Finance. Infrastructure covers the same sectors across all regions: Construction, Transportation, Technical Services, Real Estate. Commodities in Southern and Northern Europe include Agriculture, Energy, Utilities, Defense activities, Mining, Accommodation and Food Services, Human Health and Social Work; in Eastern Europe: Agriculture, Social Work Activities, Accommodation and Food, Mining, Defense.

### B.6 Dynamic Stress Testing Methods For All Characteristics

| | Random Walk | Dynamic LASSO | Dynamic RF | Dynamic Cluster LASSO | Dynamic Cluster RF |
|---|---|---|---|---|---|
| Intangible Fixed Assets | 2.79 | 3.11 | 2.89 | 7.21 | 7.50 |
| Tangible Fixed Assets | 100.69 | 61.21 | 92.48 | 23.01 | 8.93 |
| Other Fixed Assets | 51.49 | 25.54 | 24.61 | 25.16 | 23.86 |
| Stock | 5.84 | 5.95 | 4.70 | 6.96 | 8.93 |
| Debtors | 10.65 | 3.32 | 5.40 | 4.91 | 20.91 |
| Other Current Assets | 85.45 | 316.00 | 728.27 | 232.69 | 166.90 |
| Cash and Cash Equivalents | 7.25 | 4.69 | 5.20 | 3.38 | 3.35 |
| Capital | 12.59 | 8.28 | 7.07 | 7.63 | 8.34 |
| Shareholder Funds | 44.01 | 17.54 | 15.87 | 19.67 | 22.48 |
| Long Term Debt | 91.90 | 66.43 | 66.84 | 28.07 | 23.13 |
| Other Noncurrent Liabilities | 12.78 | 12.64 | 11.65 | 16.67 | 12.13 |
| Loans | 8.13 | 6.88 | 6.88 | 3.43 | 3.13 |
| Creditors | 1.81 | 1.46 | 1.31 | 2.73 | 2.42 |
| Other Current Liabilities | 14.86 | 12.85 | 14.72 | 14.33 | 15.57 |
| Sales | 6.93 | 1.78 | 1.64 | 2.90 | 1.87 |
| EBIT | 2.99 | 2.96 | 2.43 | 3.37 | 1.97 |
| Financial Revenue | 7.70 | 4.55 | 6.53 | 7.28 | 11.16 |
| Financial Expenses | 2.14 | 2.17 | 2.95 | 5.81 | 4.87 |
| Taxation | 0.46 | 0.48 | 0.48 | 1.36 | 1.41 |
| Material Costs | 4.72 | 3.28 | 2.83 | 4.31 | 4.32 |
| Costs of Employees | 1.81 | 1.21 | 0.73 | 1.61 | 0.88 |
| Depreciation and Amorization | 2.65 | 2.75 | 2.67 | 2.67 | 2.50 |
| Cash Flow | 4.07 | 3.90 | 4.22 | 4.66 | 4.57 |
| **% of variables with lower RMSE vs benchmark** | | **71%** | **71%** | **54%** | **54%** |
| **Median RMSE Improvement vs benchmark** | | **17.36%** | **17.05%** | **4.51%** | **5.37%** |

Table 15: RMSE for different models (in millions)

## C Estimation and Validation

| | |
|---|---|
| GRID NAME | turin37.grid.cesnet.cz |
| CPU | 8x AMD EPYC 7543 |
| RAM | 64GB |
| DISC | 2x 3.84TB NVMe |
| INPUT DIMENSION | 1 075 926 x 63 |
| NANs IMPUTED | 14 418 671 |
| RUN TIME | 38 hours |

Table 16: MICE Imputation Job Specification. The specification of MICE imputation job on the MetaCentrum computing grid.

## C.1 MetaCentrum

MetaCentrum Virtual Organization (MetaVO) offers distributed computing infrastructure consisting of computing resources owned by Czech academic institutions such as Charles University or Czech Technical University.[8] The Table 16 provide the specification used for MICE imputation.

## C.2 Elbow Plots Clusters

The first stage of our two-step clustering procedure consists of estimating GMM and K-Means on macroeconomic data. Figure 17 shows the within cluster dissimilarity for K-Means and the AIC/BIC values for GMM, for different numbers of clusters. For K-Means, we notice a sharp decrease in the elbow plot for k=2. For GMM this is after m=3 clusters. Next, we compare the interpretability of the resulting clusters. For K-Means we get one cluster consisting of Italy, Portugal, Spain and Belgium and a second cluster consisting of all other countries. This would result in a Southern European and a rest of Europe cluster. For GMM, we get three clusters which are displayed in Figure 1. We judge the three cluster approach to yield more interpretability and thus choose GMM in the first stage. A similar procedure is followed for the second level clustering.

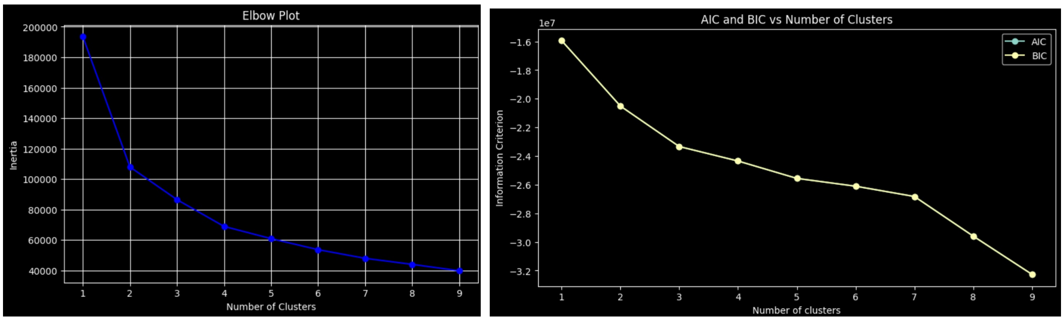


Figure 17: Choosing the number of regional clusters Left figure shows elbow plot for K-Means and right AIC/BIC values for GMM

## C.3 Grid Search for LASSO Logit

| Models | Parameter Values |
|---|---|
| $\lambda$ | $\{0.01, 0.1, 0.5\}$ |
| criterion | {RMSE} |

Table 17: Grid of Parameters. This table displays all the hyperparameters used in the Grid Search for LASSO Logit

We ran a stratified 5-fold grid search because of the highly unbalanced nature of our dataset. We try three distinct values of the parameter $\lambda$ and use RMSE as the measure of prediction

[8] For more information see https://metavo.metacentrum.cz/en/about/index.html.

error across the validation sets. The Table 17 gives the hyperparameter space used during the search.

### C.4 Grid Search for RF

We ran a similar grid search as C.3. The Table 18 gives the hyperparameter space that was in the grid search. Notably, we focus on the *n_estimators* and *max_depth* variables as these parameters tend to have a strong influence on the model performance. Further, we aim to find what criterion function used to evaluate the splits during tree building works best on our data. We choose from gini and entropy criterion functions. Lastly, we look for the best parameter of *class_weight*, which, if set to 'balanced', assign weights to the classes in the calculation of the criterion in such way that the minority class (in our case defaults) gets assigned higher weight. In this way, the model should be more sensitive to the minority class and thus slightly offset the imbalance. The Grid Search selected the following parameters: *n_estimators* = 500, *max_depth* = 6, *criterion* = gini and *class_weight* = None. However, upon running the models with theses parameters we got very similar results based on AUC and confusion matrix to the specification of *n_estimators* = 50, *max_depth* = 3, while the computational time significantly increased. Hence, to save computational resources we estimate the Random Forest models with the following specification: *n_estimators* = 50, *max_depth* = 3, *criterion* = gini and *class_weight* = None.

| Models | Parameter Values |
|---|---:|
| n_estimators | {10, 50, 100, 500} |
| max_depth | {1, 3, 6, 12} |
| criterion | {gini, entropy} |
| class_weight | {None, balanced} |

Table 18: Grid of Parameters. This table displays all the hyperparameters used in the Grid Search for RF