European Commission | Executive Agency for Small and Medium-sized Enterprises

Concept for an Innovative Data-driven Method to identify Companies in Distress

Deliverable 3.1 (Work Package 3)

Version 1.0

COSME

Call for proposals

Early Warning Europe - 738452

Project acronym:	**EWEurope**
Project duration (months):	**36 months**
Start date:	**01/12/2016**
Coordinating organisation:	**Business Development Centre Central Denmark**
Number of partners:	**15**

European Commission | Executive Agency for Small and Medium-sized Enterprises

TABLE OF CONTENTS

European Commission
Executive Agency for Small and Medium-sized Enterprises

Disclaimer

This model (program) has been developed by the Danish Business Authority for Early Warning Europe. It is at the free disposal of the Early Warning Europe consortium partners for which it has been developed as an optional tool to support the efforts to service companies in distress. It is also at the free disposal of partners in second wave countries.

The program does not advice on a particular course of action or handling in any given company, and the Danish Business Authority is not responsible for the consequences of any decisions or actions taken in reliance upon or as a result of the information provided by the program. The Danish Business Authority shal not be liable to any party for any direct, special, incidental, consequential or other damage regarding the use of the program. Further, the Danish Business Authority is not responsible for any human or mechanical errors or omissions, nor for any further development of the program by partners or third person.

1. Executive Summary

Under Work Package 3 of the Early Warning Europe project, the Danish Business Authority has tested whether it is possible to develop a model that distinguishes companies in distress from financially well-functioning companies through machine learning using publicly available accounting data (annual accounts). The purpose is to identify companies in distress, so that they can be given relevant advice in the early stages of their crisis, reducing the personal, business-related and socio-economic consequences of the distress. The base model is considered to be good at distinguishing the two classes of companies based on its receiver operating characteristics curve (ROC) derived from Danish input data, thus achieving a satisfactory degree of accuracy in its predictions. This result relates to the fact that the public digitisation level is high in Denmark, as training data has been available in sufficient quantity and quality for machine learning purposes. At the same time, it has been realised that there are great differences in the quality of publicly available data across the project's participating countries. This means that derived models for the project's four target countries are less accurate than the Danish base model, but still fairly accurate. Overall, the test is considered to have proven that publicly available annual accounts can be meaningfully modelled for use in automated identification of companies in distress.. Overall, the test is considered to have proven that publicly available annual accounts can be meaningfully modelled for use in automated identification of companies in distress. However, it is important to consider the automated process as a basis, not substitute, for subsequent individual, qualitative processing of the model's results, considering local conditions and the circumstances of the individual company that are not contained in its accounting data. These numbers cannot stand alone.

European Commission | Executive Agency for Small and Medium-sized Enterprises

From damage control to prevention: How machine learning can be used for automated detection of companies in distress

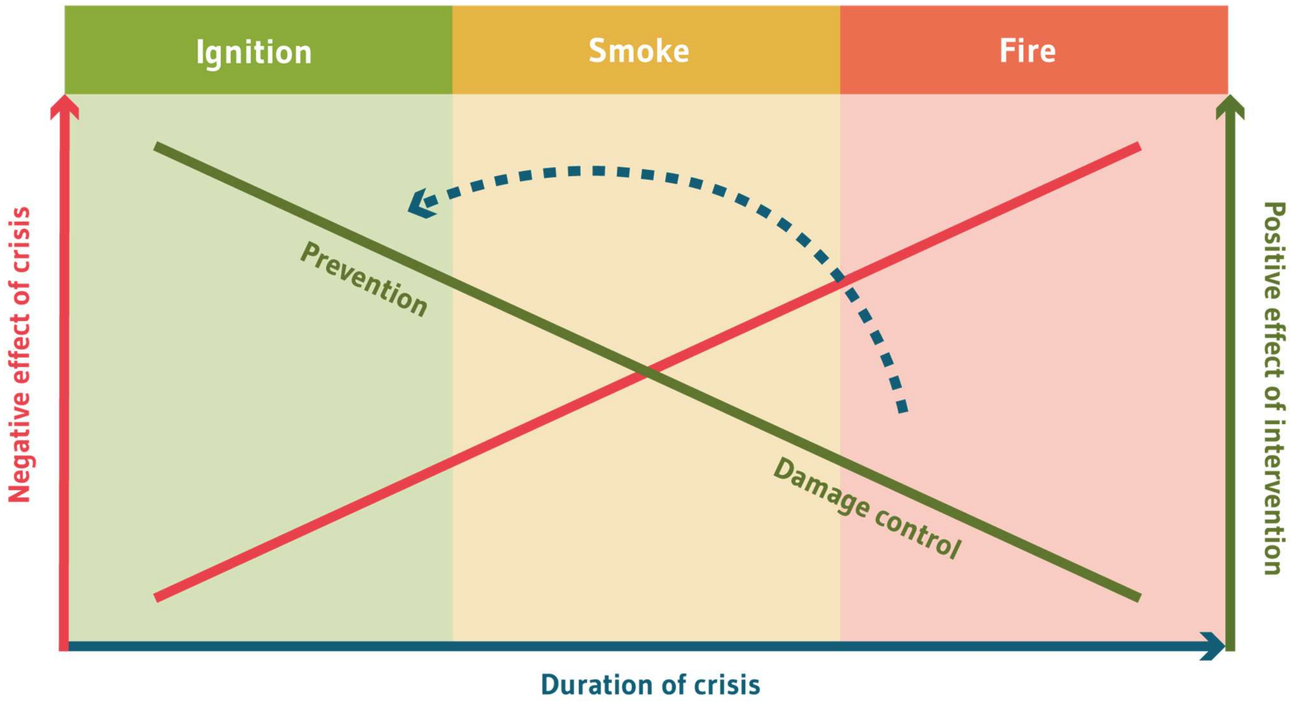
2. Introduction

The digitisation of public accounting data is becoming increasingly widespread in Europe, and the Accounting Directive (2013/34/EU) standardises the requirements for data across Union member states. Combined, these two factors give both public authorities and private actors new opportunities to analyse business economic patterns and trends using advanced calculation methods for a wide range of purposes. Thus, in principle, the methodological and data-related basis needed to automatically identify distressed companies solely on the basis of public data about them is in place. In this regard, machine learning is an obvious approach, as it enables the training of a computer to gradually improve at a specific data-based task using statistical techniques – for example spotting distressed companies in large data sets – without explicitly programming the machine to conduct the identification process in a certain way.

If this is possible, it may potentially create value in Early Warning Europe in two ways. First, it could reduce the costs incurred by organisations in finding relevant companies that can be offered early warning services. This means cost-efficiency and releases resources for actual counselling. Second, it could help establish contact with companies in distress at an earlier stage of their crisis. This buys additional time for advice and guidance to have an impact on the company’s business, thereby increasing the usefulness of the assistance offered.

This time factor is important, as the figure below shows. This is because, in crisis management counselling, the general experience is that advisers' and external consultants' chances of properly helping a company in distress or, alternatively, assisting in its sound, controlled liquidation depend on how early they are involved in the crisis management. The sooner it happens, the greater the positive effect of their work can typically be. Conversely, the negative effect of a company crisis often grows the longer the crisis lasts – for the owners, the employees, the local area and the economy as a whole.

However, it is not uncommon for a crisis to be detected – or recognised – too late. When this happens, external interventions generally take on the nature of firefighting: From the flames, you try to save what can be saved. The burning platform is clear to everyone, so the incentive to act is great. The idea of automatic identification of companies in distress is to shift the possibility of effective advice to an earlier stage of the crisis process, allowing interventions to also be preventive.



Under Work Package 3 of the Early Warning Europe project, the Danish Business Authority has tested whether it is possible to develop a model to conduct automated identification through machine learning. The test and development process, including associated meetings and workshops held along the way, constitute items nos. 3.1 and 3.5 of the consortium's Implementation Plan. This white paper constitutes the plan's item no. 3.6.3 and describes the implementation of the experiment, the data base, the structure and application of the model as well as its predictive key figures.

To increase the relevance of the experiment to the entire consortium behind Early Warning Europe, the Danish Business Authority has first developed a base model based on Danish data and then further developed four derived models for the four target countries in the overall project, i.e. Greece, Spain, Poland and Italy. The four derived models are made available to the relevant country partners, and implementation and training workshops are carried out with them, cf. the implementation plan, item 3.6.2.

3. Constructing a model

3.1. Hypothesis

The Danish Business Authority sought to confirm the hypothesis that by using public accounting figures contained in annual accounts from companies, we can train a machine to distinguish and thereby identify companies in distress and companies that are not in distress using machine learning.

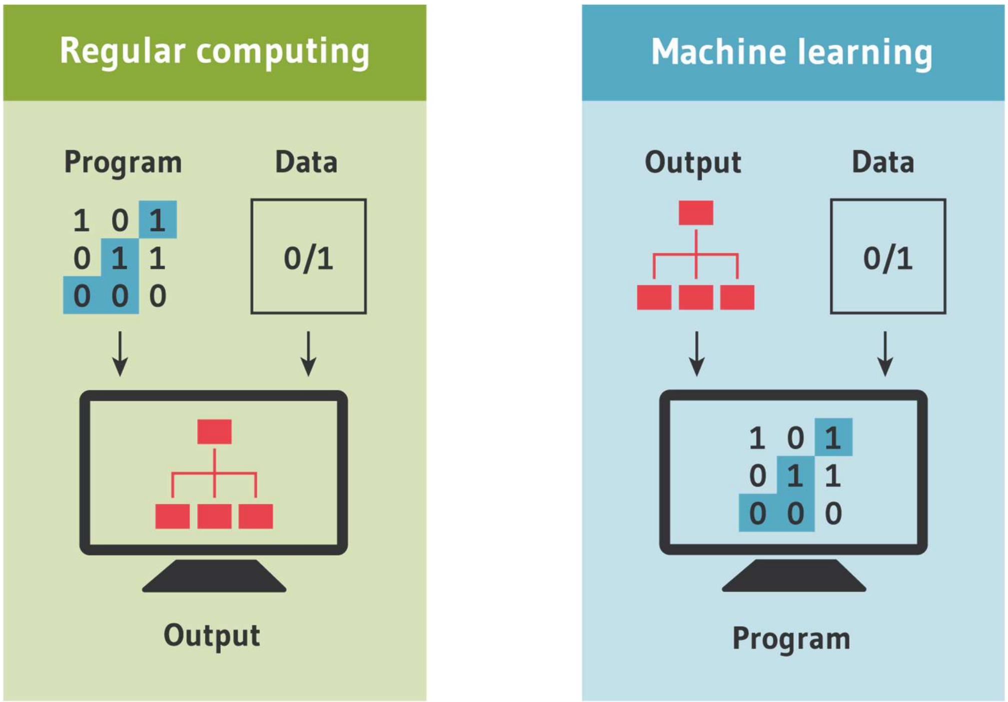
In order for us to test the hypothesis, we need access to both a large number of examples of companies in distress and a large number of examples of companies that are not in distress. In this way, we can develop a model that, through training, learns to recognise the accounting conditions characteristic of the class of companies in distress. You might say that this is an advanced form of pattern recognition. To be technically feasible, accounting figures for both these classes of companies are required. The two data sets must be structured in a uniform manner, either on submission or by subsequent processing, so that the company classes' respective accounting patterns can be compared across, for example, businesses and industries.

3.2. Training of the model

When the machine is trained to only distinguish between companies in distress and companies that are not in distress, the result will be a "model". The model is called "f" because it expresses something as a function or result of something else. It is a simple representation of the reality we are trying to describe.

In machine learning, the computer forms the model itself according to our general instructions by considering the examples we feed it. Put simply, this differs from traditional programming, where you enter data (for example numbers) into a program (Excel) and then explicitly tell the program precisely what to do with each little piece of information in this data (for example perform certain specific calculations in the Excel fields). On this basis, the program performs the calculations on the input data set and produces a result in the form of an output. In contrast, in machine learning we also use data sets, but do not program the computer to perform a specific number of predefined calculations.

This is referred to as the machine not being explicitly programmed. Instead, the computer is presented with a data set with associated output values, i.e. information about whether or not a given company is in distress, and is then instructed to test and adapt – and thereby "learn" from – a variety of computational methods in order to identify companies in distress based on shared accounting characteristics. On this basis, the computer produces a model, i.e. a program. This approach can be illustrated as follows:



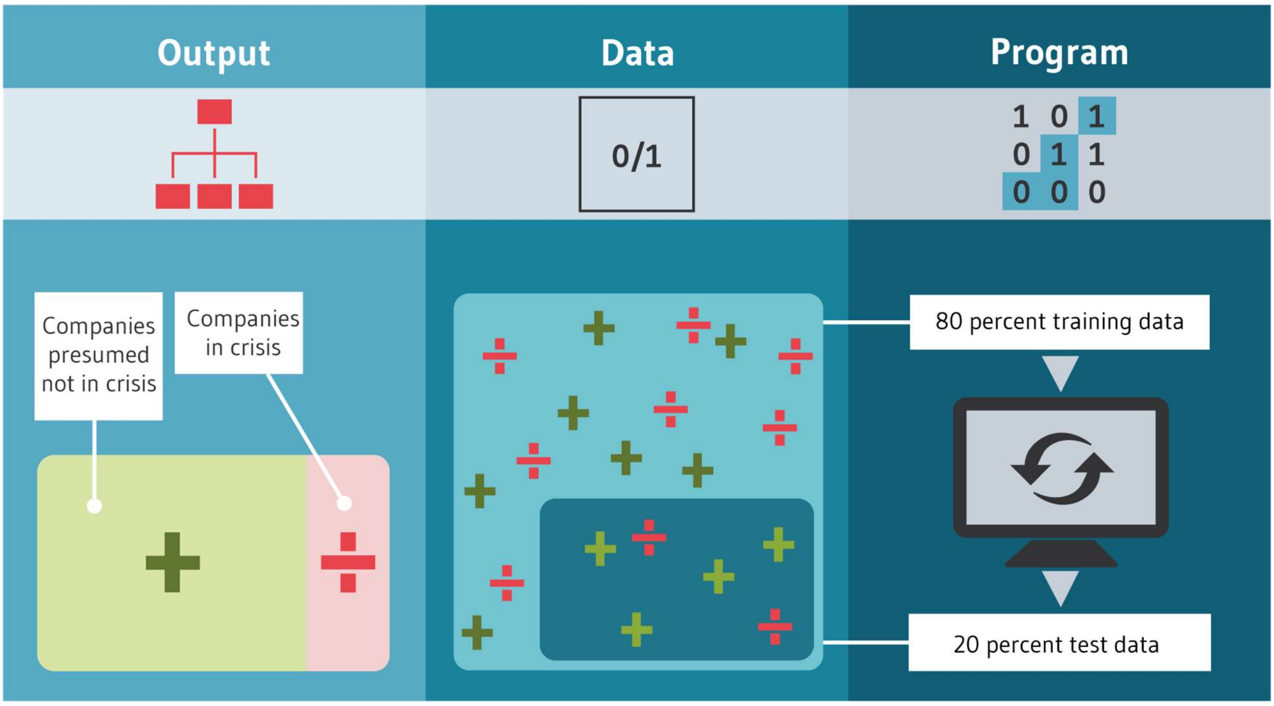
You set up the framework for the model and tell it what it has to learn, but the machine fills out the framework and learns how it should do so on its own. "Framework" means the sum of the technology and the algorithms you choose to employ for machine learning. The quality and transparency of the model depend on the framework we give the machine.

More formally, a machine-learning approach can be expressed as a simple formula, as, based on the input we provide, the model seeks to answer the question of whether it is true or false that a given company is in distress. If we label the answer to the distress question Y, we can express it as the computer forming a model (f) based on the input data X we provide: $y = f(x)$. This can also be expressed as follows: Whether a company is in distress (y) is a function (f) of the company's accounting figures (x).

And this reflects our initial hypothesis. In order for the machine to form the model (f), it requires many examples – the larger the input data sets with examples of both distressed and financially well-functioning companies, the better it is from a machine learning perspective. The machine will try to establish the model that best matches as many examples as possible, thereby maximising its success rate in relation to responding correctly to the question of whether a given company is in distress or not. The model that provides the most correct answers will be preferable.

It is important to note that the machine relates to the examples uncritically. It forms the model on the assumption that each of the examples are true (so-called “ground truth”). That is, the model will reflect any errors, shortcomings and untruths that may be included in the examples.

In order to assess how good a model is, we deliberately choose not to give the machine all our examples. Typically, we will feed the machine just 80 percent of our examples, selected at random, as part of input data set X. Once the machine has formed the model based on the 80 percent, we then test it using the remaining 20 percent of the examples in our total data set. This allows us to assess how well the model is performing on the 20 percent which it has not seen before. We do this by comparing the model's response to what we know in advance about the companies' actual distress status in the test data set. If we zoom out, the process can be illustrated as follows:



4. Data used to train the model

4.1. Training data

The following section describes the training data set used by the Danish Business Authority. These are two data subsets that have been combined so that the aggregate set includes both companies identified as being in distress and financially well-functioning companies.

According to the y = f(x) formula, this means that, for the data subset containing distressed companies, "yes" is answered to the question about distress (y = yes), while, for the subset containing financially well-functioning companies, the answer will be no (y = no). The X values consist of the two company classes' annual accounts. The training data set thus contains:

- Output: Companies in distress – Y = yes/no
- Input: X = annual accounts

4.2. Companies in distress

As examples of companies in distress (y = yes), the Danish Business Authority received an extract of Danish companies which, from 2009 until October 2017, completed an Early Warning process in the Danish Regional Business Development Centres (public business development centres, so-called “Væksthuse”) from the centres. This extract was used as the input data set for distressed companies.

Of the extract's just over 5,000 companies in distress, only 720 could be used for this purpose. In part because many companies were privately owned and thus not obliged to submit accounts, and in part because many of the companies’ distress processes were from before the introduction of digital accounts in Denmark, which means that the required data does not exist in digital form.

The assessment of whether the company is actually in distress is based solely on the identifications by regional business development centres. In practice, this was done through concrete, individual overall assessments of not only the companies' finances but also other relevant factors such as market, products, management, etc. The assessments were carried out by Early Warning consultants in the regional business development centres. In principle, it cannot be ruled out that in some cases there may be erroneous assessments, since the label of "in distress" was not specifically and explicitly defined at the time of the assessments. This had the advantage that it made it possible to accommodate the often unique anatomy of business distress and leave practical definition to the consultants’ discretion. However, it is possible that, in the statistical sense, the individual assessments may also mean that there are a number of false positives in the training data set, i.e. companies labelled as in distress by consultants in the training data, even if the same companies may not have been labelled as in distress under a certain (for example econometric) definition of distress. This factor contributes to unavoidable uncertainty, as the model is trained with data that probably also contains some noise.

4.3. Companies that are not in distress

As examples of financially well-functioning companies (y = no), the Danish Business Authority first used a data extract from the business development centres containing companies identified as not in distress. The extract proved not to be useful as, at the overall level, these companies were not representative of Danish companies. Due to such selection bias, the Danish Business Authority opted not to use this data.

Instead, a random population of 3,000 Danish companies that had submitted accounts for 2015 was extracted from the total population of accounts submitted for the 2015 financial year. Of these, 2,734 continued to have "normal" status in 2017 (two years after submission of the accounts in 2015). These 2,734 companies were used as examples of financially well-functioning companies. The term "normal" status means that, for one reason or another, the company was not liquidated or in the process of being liquidated.

It should be noted here that not all distress situations can be read in accounts. Therefore, the training data set for financially well-functioning companies may include a number of companies in distress. Statistically expressed, these companies will be false negatives, i.e. they have been incorrectly classified as financially well-functioning and are actually in distress, yet not categorised as such. When extracted based on accounts, such false negatives cannot be avoided. For example, a personal psychological crisis in a company owner cannot be seen directly in the accounts. It will only be visible when – or if – the psychological crisis manifests itself in mismanagement, causing the business to fail. Correspondingly, for example, accidents (loss of a key customer through no fault of one's own) or macro conditions (such as sudden political changes in a key export market affecting demand or volatility of foreign currency receivables) cannot be read directly in the accounts. In other words, such information is not covered by the obligation to submit accounts, and thus will not be directly included in the accounting data. For that reason, the crises that may follow as an effect of such events cannot be predicted on that basis.

Such false negatives, along with the aforementioned false positives, mean that there will always be some uncertainty in the training data that the computer has to work with. This affects the precision of the model that the machine learning produces. For that reason, the precision of the model can, by definition, not be expected to reach 100 percent. It will always have insufficient information to capture all possible signs of distress, which is consistent with a reasonable standard assumption: All business crises have financial *effects*, but only some have financial or economic *causes*.

4.4. Companies' accounting figures (X)

Since the 2012 financial year, companies in Denmark have reported their annual accounts digitally. This requirement means that, in practice, the accounts of over 99 percent of Danish companies, which are subject to the obligation to submit accounts are digitally available. The remaining 1 percent primarily covers certain types of financial companies that are not covered by the digital requirement.

The reporting takes place in the eXtensible Business Reporting Language (XBRL) format, which is standard in the area, allowing reading of the contents of the accounting records at the field level.

The Danish Financial Statements Act contains varying requirements for the annual account, depending on the size of the company – so-called accounting classes. However, all companies must report both their income statement and balance sheet, as well as a range of metadata about the company and the annual account, for example the accounting period, date of adoption, etc. As accounting figures (X), therefore, only the groups of accounting figures that appear for all accounting classes are used in the work with the model. We do this to ensure that the model is based on accounting fields that are available across all companies.

4.5. Input data

An overview of the individual fields used in the model is provided in Annex 1. The fields are selected from the following parameters:

1. **Fields mandatory for all companies to report:** In all European accounts, a number of values must always be reported, cf. Articles 13 and 14 of the Directive. This includes "Annual result" and "Gross profit/loss", as well as balance sheet[1] items indicated by letters and Roman numerals.

2. **Fields that make sense from an accounting perspective:** In connection with the development of the Danish base model, it was estimated through machine learning-based data analysis that the fields "Operating profit" and "Dividends", as well as the period from the balance sheet date to the general meeting date, are relevant for the assessment of the individual company's distress status.

   Furthermore, comparative figures[2] have generally been included, as it is analytically estimated that the possibility of looking at information that extends over two years instead of one is generally more useful.

3. **Fields that add value to the model in connection with training:** Not all required fields are included in the model. This is true of records used so rarely, the model does not take into account their value (e.g. "Assets held for sale" and "Profit (loss) after minority interests' proportionate share").

4.6. Features calculated

If the machine is only presented with the specified accounting fields (see Annex 1), the result will not be satisfactory. This is because the accounts' items themselves have no value in the sense that signs of distress cannot be deduced from them. Conversely, the value lies in the *ratio* between the items. We know these ratios as the company's key ratios: What is the return on equity? How solvent is the company? What is the ratio between debt and balance?

Thus, in the model, all input data is converted into key ratios before being used by the machine. The conversions appear from Annex 2. We calculate a whole host of conditions – "features" – for our model. The ratios show the most important properties of the individual company for accounting purposes. From the very development of these characteristics, our machine must be trained to identify distress. It is implicit in our initial hypothesis that the accounts distressed companies' accounts contain a pattern that the model must learn to recognise.

These calculated properties can be divided into two main groups:

[1] A very small part of Danish accounts presents the balance as "reporting form", cf. Annex IV of the Directive. Similarly, very few presents the income statement as "functionally divided", cf. Annex VI of the Directive. Therefore, the model may be uncertain in situations where the latter rare layouts occur.

[2] Comparative figures mean that companies not only report figures for the accounting period but also the corresponding figures from last year. Last year is thus the comparative period, and the figures are the comparative figures.

1. **Changes (delta):** For all balance sheet items, the change is calculated relative to the preceding year. That is, the difference between this year's figures and last year's figures:

   $$\text{Delta} = \frac{\text{This year's figures} - \text{la} \quad \text{year's figures}}{\text{last year's figures}}$$

   If "last year's figures" are not included in the accounts (e.g. by a new company), delta is set to 0. Delta values are named with the appropriate field name in XBRL followed by "_delta", e.g. "Assets_delta".

2. **Ratios**: For the income statement, the distance from "Gross profit/loss" to "Operating profit" is calculated relative to "Gross profit/loss", and the distance from "Operating profit" to “Annual result” relative to "Gross profit/Gross loss". By "distance" is meant the ratio between "Gross profit/loss" and the other two result items.

   For assets on the balance sheet, the ratio between the asset and the balance sum (assets) is calculated. Since all assets add up to the balance sum, the figures can be interpreted as the distribution of assets.

   For liabilities on the balance sheet, the ratio between the liability and the balance sheet amount minus equity is calculated.

Finally, the solvency ratio of the companies and their return on equity are calculated, as well as the ratio between dividends and annual result. These last three ratios are calculated both for the financial year and any comparative figures.

Together, delta and ratio values provide a total of 35 calculated properties used by the machine to train the model. This list is shown in Annex 2. To the extent that properties cannot be calculated (e.g. because they are missing in the accounts), uniform ways of handling the missing data have been incorporated into the model for each type of data field, so that all properties are always represented with a value that is always determined in the same way, if it is missing from the original data set. Without such uniform filling of data holes, the machine would not be able to handle incomplete data sets.

4.7.Technical toolbox

Based on the 3,454 examples (i.e. 720 companies in distress plus 2,734 financially well-functioning companies) and the calculated properties of said companies' annual accounts, it is now possible to train the machine. The Danish Business Authority has chosen to use open source tools only so that Early Warning Consortium participants can take over the model afterwards free of charge. Python 3.5.x is used as the development language, and Scikit-learn 0.19.x. is used as the machine learning tool.

The model requires our input data for training (i.e. the Xs in the formula) to be structured as a so-called "dictionary" in Python. A "dictionary" can be understood as a collection of data that collects a number of words (called "keys") and associates definitions of them (called "values"). This is a basic feature of Python which can be handled by all Python programmers. From Scikit-learn, the "Gradient boosting" technique is used to build the model. Gradient boosting is a relatively complex machine learning technique in which new models are gradually added to an existing base model automatically until it is no longer possible to improve the predicative properties, i.e. the ability to identify distress, of the overall model by further successions. One advantage of the technique is that, in principle, it is possible to look over the machine's shoulder, i.e. discern how the model has calculated the given result in each case. During the experiment, the Danish Business Authority learned that gradient boosting yielded the best results compared to other proven techniques (including, for example, decision tree and random forest techniques), where it is also partly possible to identify the reasons for each of the model's statements. However, in the final project, the Danish Business Authority chose a model delivery architecture that supports its user-friendliness in Early Warning Europe – the architecture is presented below. As a consequence of the chosen architecture, it is not immediately possible to recreate these reasons.

5. The model in practice

5.1. Use of the model

The model can only be used by a machine that can execute Python code. If you wish to use the model outside Python, this requires additional IT development. In such cases, the approach could be:

- In Python, the model is applied to one or more companies, after which results are exported, for example to Excel, allowing the use of the result on other computers.
- Functionality is developed to display the model to users or other IT systems as a service (for example a website), so the model can be used as a regular service. This is a more demanding process that also requires an IT environment for operation.

A description of possible applications outside Python is not covered by this white paper, as making the model available to this extent is not part of the Danish Business Authority's task in Early Warning Europe.

5.2. Use of the Python model

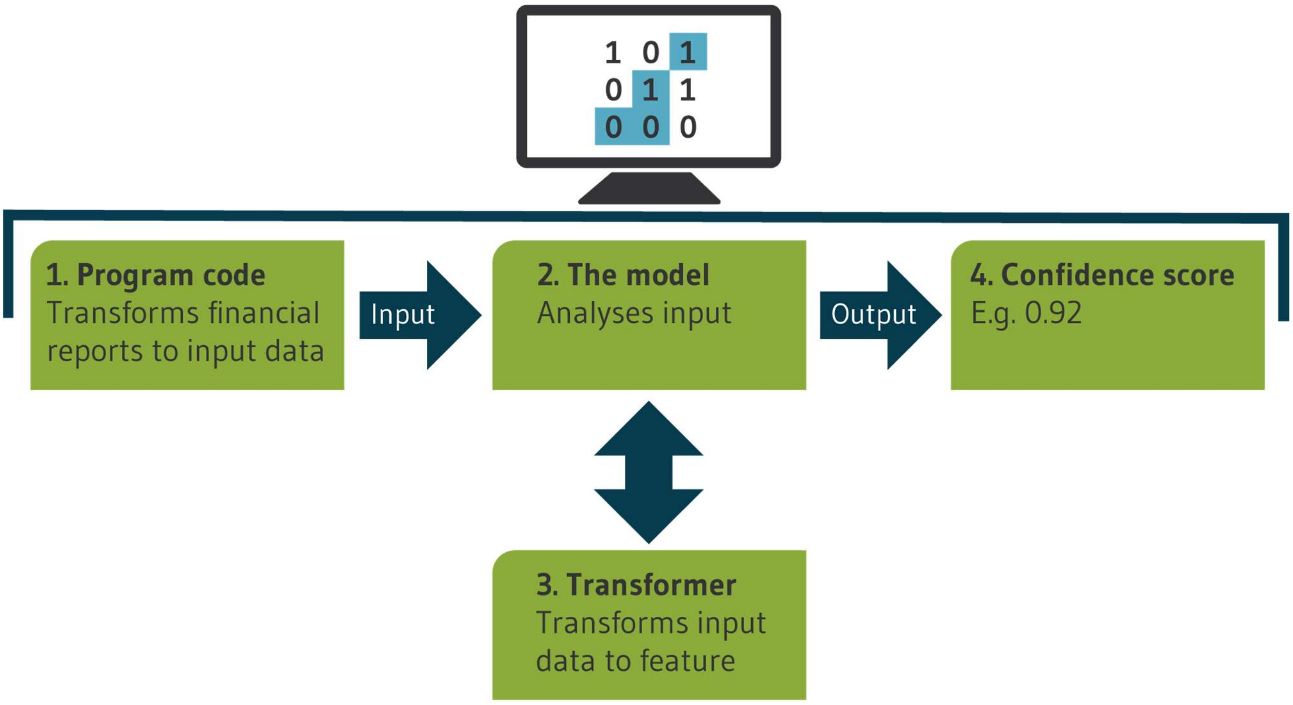
When the model is used in Python, it is done in two steps:

1. Data is converted into useful technical input in the form of a Python dictionary.
2. The model is fed with input data and returns a confidence score (see below).

It is possible to either apply the model to one company at a time or to a list of companies. This two-step approach, which is independent of the number of companies processed, has been chosen as it is expected to be the easiest to implement in Early Warning contexts as described below. Use of the model therefore requires that the following code is present:

1. Program code that converts accounting data into input data.
2. The model as trained by the Danish Business Authority on Danish data.
3. A transformer that converts input data into calculated properties in Python.

The following figure shows how the model works in use:



1. **Program code that converts accounting data into input data**. The Danish Business Authority has only had access to data from Denmark and small data samples, which were specifically purchased for the purpose for the four target countries (as described below in the section on models and data for other countries). As mentioned, data is structured in XBRL in Denmark. For the four target countries, however, data has only been available on the private data market in various Excel formats.

   For Denmark, input data can be generated using a code library located online and freely accessible as open source. For each of the four target countries, Python program code which converts from Excel to input data is required. As part of the project, the Danish Business Authority therefore wrote that code, cf. the following section on access to models for other countries. The advantage of the chosen structure – i.e. the model's structure in building blocks 1, 2 and 3 – is that it is subsequently easy to replace program code (block no. 1), should the data structure change, for example if you change the structure in Excel. The replacement can thus be implemented without affecting the model and its use, as long as the required input is still present.

2. **The model as trained by the Danish Business Authority on Danish data**. As described above, the Danish Business Authority trained a model based on Danish data – and thus Danish input – collected from Early Warning Denmark.

   Similarly, the Danish Business Authority trained one "derived" model for each of the four target countries based on the Danish "base model". The derived models are all different, as the available data varies in target countries. Each derived model depends on the respective data subsets in both the base model and the target country concerned. The individual derived models and obtained data for these are described in separate sections below. It is important to emphasise that the derived models should be retrained if new accounting items in a country's data set are made available, or if certain items are deleted. Similarly, the supplied transformer should be updated. The model only works if it has access to the supplied transformer for the country/model in question.

3. **A transformer converting input data into calculated properties in Python**: The conversion of input data into calculated properties takes place for each derived model in the transformer. The transformer is made available as a piece of Python code which must be available to the model.

4. **Confidence score**: Output from the model is a so-called "confidence score". This expresses how accurate the model is in its prediction of whether a given company is in distress based on the available data, thus indicating the model's accuracy. It consists of two numbers: A number for "no distress" and another for "in distress". The sum of the two is always 1 (for example 0.00133183 + 0.99866817), because we know logically that the company is always either in a state of financial distress or economically well-functioning.

   The advantage of this kind of output is that you can set a limit as a threshold for when, according to local or national priorities and circumstances, companies are deemed likely to be in distress, and therefore appropriate recipients of Early Warning services. The principle of the threshold is reviewed separately in a later section.

   The model also has the option of simply returning Yes/No, by the model merely determining whether a company scores above or below 0.5. However, the Danish Business Authority recommends that you use confidence score rather than the simpler yes-no form, thus achieving a more nuanced representation of reality. When using confidence scores, priority can be given to the companies for which the model is more certain rather than those for which the model has greater doubts. The next section describes a proposal for how to prioritise in practice.

5.3. Local use of model output

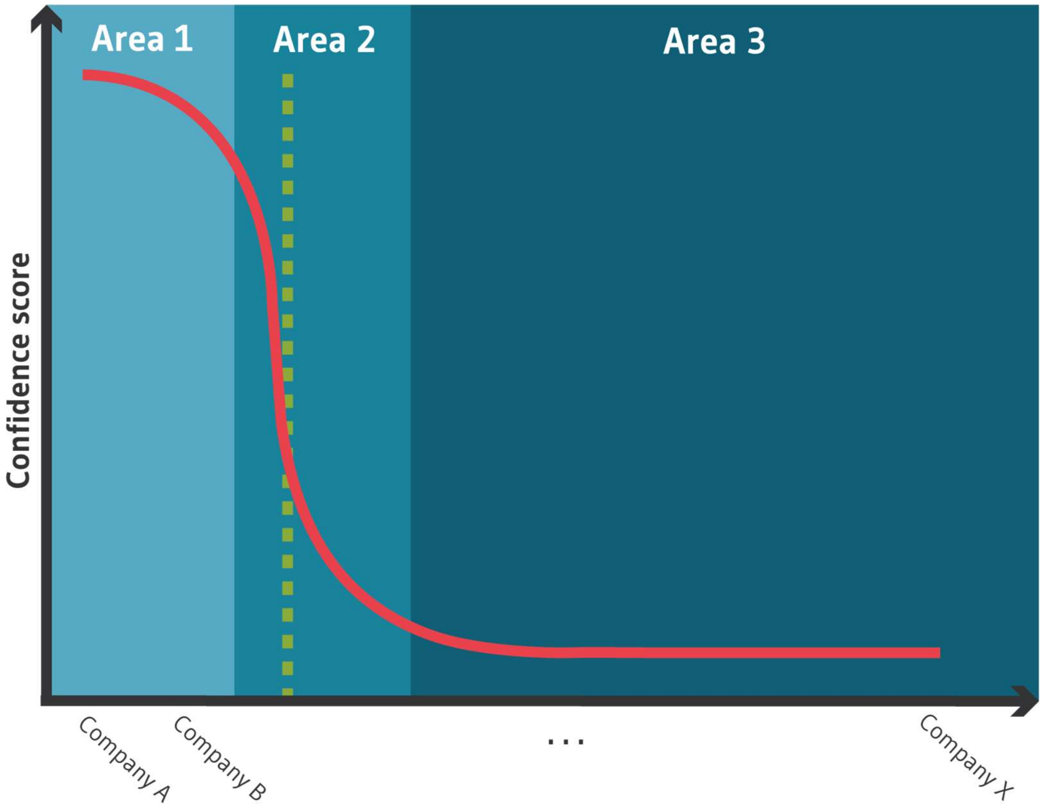
It is beyond the scope of the Danish Business Authority’s Work Package 3 to develop detailed implementation manuals for the model, adapted to local circumstances in the home countries of organisations receiving the model. However, the following sections can provide a general starting point for the further local work. If the model is applied to a data set that includes X number of companies, an Excel sheet is returned with X number of rows. Each row shows the company's confidence score, which expresses in simple terms the likelihood that it is in a state of distress. As mentioned, the number is between 0 and 1: The higher the number, the greater the likelihood of distress. The output for an imaginary data set can look like this, but it might be quite long if, for example, the accounts for a whole municipality or region were processed:

Company name	Probability
A	0.98
B	0.15
C	0.56
D	0.24
...	...

In order to increase the practical usability of this output, it is recommended to sort the output set by probability based on the model test from high to low. This returns:

Company name	Probability
A	0.98
C	0.56
D	0.24
B	0.15
...	...

Assuming that, for example, 10,000 companies have been assessed, we can illustrate the descending confidence score down through the list. It is the expectation of the Danish Business Authority that the derived models will produce an output of which the visualisation will follow the logic of the following basic outline:



In area (1), we have a relatively limited number of companies for which the model finds a significant probability of distress. In area (2), we also have a relatively limited number of companies for which the probability of distress ranges from medium to significant, but for which the curve also has a high slope, i.e. the gap between individual companies is large. In area (3), we have the clear majority of companies in the data set, for which the model finds a very limited probability of distress.

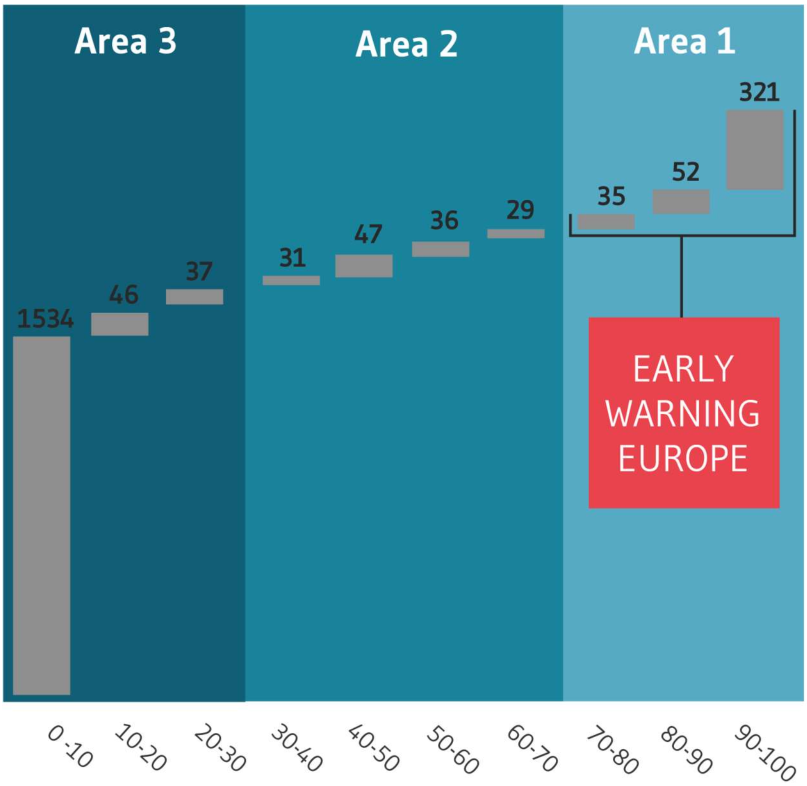
For practical reasons, it may typically be appropriate to define a threshold for distress probability (green dotted line in the figure) as an expression of how to reduce the number of companies on which to focus local Early Warning resources – i.e. area (1). A threshold confidence score of approx. 0,80 seems relatively sound, but will have to be determined based on an overall qualitative assessment from country to country, depending on factors such as local business conditions, industrial structures and data quality.

In the subset of companies in area (1), local political priorities and economic circumstances, which are compatible with Early Warning Europe's objectives, may also be taken into account. For example, the analytical focus can be on certain types of companies, industries, number of employees or combinations thereof, as it makes it possible to target the outreach work of the consortium and to better adapt the communication to the recipients. This can help increase companies' perceived relevance of information from Early Warning Europe and ultimately help ensure that the right companies get the right offers at the right time. Therefore, it is recommended that the exact focus in area (1) is determined locally in the individual departments of Early Warning Europe, taking into account the consortium's common objectives.

The approach can be summarised in the following steps:

1. Output is sorted by confidence score (gross list)
2. Sorted output list is sorted by descending values
3. Threshold is determined as the transition between area (1) and (2) to create a net list
4. Within the framework of Early Warning Europe, any local focal points for the net list above the threshold value are determined.

With this approach, an overview of the model's output may be established. For example, this can be illustrated as in the following figure, in which, as part of an internal partial test of the model, 2,168 company accounts were processed in the model, and the output (confidence score) is sorted in 10 percent intervals as independent variables. For example, in the figure, the threshold value is set at 70 percent, returning a total of 408 companies that should then be qualitatively reviewed, and are identified as *potentially* relevant recipients of Early Warning Europe services.



6. Results

6.1. The model's strength for Danish data

Before using the above approach, we are of course interested in knowing how good the model is, i.e. to what extent its assessments are correct. To establish its accuracy, the model, which, as mentioned above, relies on a training data set, is tested on the test data set which it has not previously been presented with. Next, the model's assessment of the companies' status in the test data set is compared to their actual status, which we kept isolated.[3]

Basically, the range of possibilities is divided into four categories:

- True negative (TN): The company is not in distress, and that is also the model's assessment.
- False positive (FP): The company is not in distress, but the model erroneously finds that it is.
- False negative (FN): The company is in distress, but the model erroneously estimates that it is not.

[3] Previously, it was mentioned that 80 percent of the total data set is typically used to train the model, and 20 percent is then used to test the accuracy of the model. This was also the case in connection with the work done to create the Danish base model. In this section, however, we instead apply "Leave one out cross-validation" (LOOCV) to assess the model, thus avoiding the 20% loss of training data.

- True positive (TP): The company is in distress, and that is also the model's assessment.

Within the range of possibilities, the accuracy of the model can be set as follows:

		Model assessment		
		Not in distress (0)	In distress (1)	
Actual data	Not in distress (0)	2,606 (TN)	128 (FN)	**95 percent**
	In distress (1)	201 (FP)	519 (TP)	**72 percent**
		93 percent	**80 percent**	

By considering the above, the accuracy of the model can now be calculated:

- **The model identifies 95 percent of companies not in distress:** TN/(TN+FN). Thus, the model erroneously finds 5 percent of healthy companies to be in distress. As such cases will exist, it is important not to equate the model's output with the actual circumstances of the company, for example in communication with potential recipients of Early Warning services. The model can substantiate a presumption that a company is in distress, but not state it as fact. Qualitative assessments are required – so are reservations (recall).

- **The model identifies 72 percent of the companies in distress:** TP/(FP+TP). Thus, the model erroneously finds 28 percent of the distressed companies to be healthy. In other words, it overlooks 28 percent of distress cases which, as mentioned, can be interpreted as an indication that not all forms of distress can be identified solely based on (limited) accounting data (recall).

- **In 80 percent of cases where the model identifies distress, it is correct:** TP/(TP+FN). Conversely, 20 percent of them are actually healthy (precision).

- **In 93 percent of cases where the model identifies no distress, it is correct:** TN/(TN+FP). Conversely, 7 percent of them actually are in distress (precision).

Overall, the accuracy of the model can be calculated as it being correct in 90 percent of all cases:

$$\text{Accuracy} = \frac{\text{TN+TP}}{\text{TN+FN+FP+T}}$$

It is the Danish Business Authority's assessment that the accuracy of the model can justify a conclusion that, in line with the initial hypothesis, it is possible to train a model through machine learning based on publicly available, digitised accounting data and have it distinguish companies in distress from companies not in distress to a satisfactory degree of accuracy. The conclusion is based on the Danish base model's receiver operating characteristic curve (ROC), a method used for assessing the performance of discriminative machine learning models classifying groups in datasets, according to the standards of which it is considered to be a good test for distinguishing between the two classes of companies. Hence, the model is deemed to be a sufficiently valid basis for the experiment to be considered successful, i.e. the prototype works and can be exported to other organisations, possibly for local adaptation and further development. In addition, the work indicates that a higher confidence score can be advantageously sought by supplementing public accounting figures with other data, including economic, and that despite the potential of digitisation, it is both useful and appropriate to supplement the automatic identification process with qualitative processing of the model output.

6.2. Protection of companies from training data

As previously described, it is necessary to have access to the Python model in order to use it. Although the model is a binary file, data from the training data set can theoretically be derived from it, including accounting data from companies that have been in dialogue with the business development centres, and thus may be assumed to be in distress. This poses a data security issue, if the Danish Business Authority's internal original base model was to be disclosed to third parties.

In order to avoid the disclosure of confidential business information, the Danish Business Authority has applied the model to 24,924 other companies' 2016 accounts – a kind of pseudo data. Based on this data, we have built a new base model, and this is the version made available to Early Warning Europe. Consequently, there is no traceability between data provided from the business development centres and the model shared with other organisations. When the new base model based on pseudo data is tested with training data, the following results are obtained, which are quite close to the original internal base model:

		Model assessment		
		Not in distress (0)	In distress (1)	
Actual data	Not in distress (0)	2,592 (TN)	142 (FN)	**94 percent**
	In distress (1)	174 (FP)	546 (TP)	**78 percent**
		94 percent	**79 percent**	

This model is generally of the same accuracy as the original base model. This resolves the potential data security problem with a minimal loss of accuracy. Hence, the pseudo data base model is used for the subsequent follow-on development of derived models for target countries.

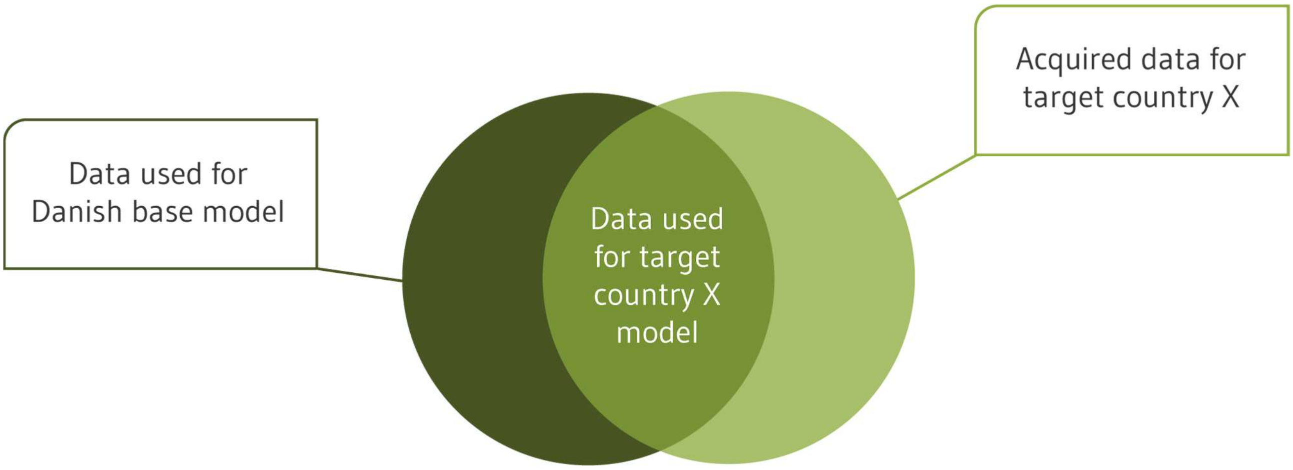
6.3. Model and data for other countries

In order to create derived models for the four target countries, the Danish Business Authority trained a unique model for each country based on Danish data. It was necessary to create four different models because different data sets were available in the respective countries. In particular, the challenge was that, for each target country, the data structure for input data deviated from the Danish structure. The deviations include:

1. Data was obtained in Excel worksheets instead of XBRL format
2. Data fields have different names
3. The gross list of fields differs from the Danish fields

For items 1 and 2, respectively, this means that input data must be transformed from target country format to a format that can be used by the Python model.

For item 3, this means that the model does not have access to all the fields for which the Danish model is trained. The consequence of this is that for each target country, the Danish model is retrained on Danish fields, but based only on the overlapping fields between Denmark and the country in question. The common data volume, of course, provides less input data and thus a less accurate model, although there was a reasonable degree of overlap. The limitation can be illustrated as follows:


Acquired data for target country X
Data used for Danish base model
Data used for target country X model

An overview of the data fields available for each target country is shown in Annex 3.

A further challenge was the quality of available training data for the four target countries. As mentioned, a number of requirements for such data are made regarding their structure, naming, file packaging, etc. Freely available public Danish data complies with these requirements, and can therefore be used for machine learning, but this is not the case in the respective countries, where public data is not immediately available in the required volume, structure or degree of digitisation. Therefore, it was necessary to acquire the relevant data from private providers which have compiled and structured relevant public data in the respective countries into merchantable goods.[4] These four training data sets consist of balance sheets and/or operating accounts for two consecutive years collected in each country from 100 randomly selected companies (totalling 400) which are not publicly listed, and which were active in 2015.

A shared feature of the four derived models is that only after practical application of the individual models in each target country can it be concluded whether – and to what extent – the respective models' results also apply to accounts from companies from the countries concerned.

6.4. Access to the model for other countries

In order to provide access to the derivative models of the four target countries, the Danish Business Authority chose to use the GitHub file sharing tool. The code is available here: https://github.com/Niels-Peter/EW4E. The code for using the models is thus publicly available, while the models themselves are delivered directly from the Danish Business Authority to the relevant partners in the four target countries. A procedure for further dissemination to relevant partners will be determined by the Consortium Steering Committee.

For each target country, Github contains:

- Code for using the model, including code for converting from worksheet to input data (for example, for Poland "poland_pred_from_xlsx.py")
- Example worksheet for input data (for example "Poland_Database.xlsx")
- Dummy model that can be used for testing the code (for example "EW_DK_POKAND_dummy.pkl")
- Python code used by the model to transform input data into calculated features (for example "poland_transformation.py")

Thus, it is possible, for each target country:

- To retrieve and test the program code, the transformer and the dummy model before transfer of the derived model from the Danish Business Authority.
- To view the data structure for input data, the code for transformation from worksheet to input, and the code for transformation of input data into calculated features.
- To incorporate the correct derived model after transfer from the Danish Business Authority, without having to make further corrections.

[4] Training data sets were acquired from Findustria Srl (Italy), Infobank Hellastat SA (Greece), Axesor Conoser para Decisor SA (Spain) and Iwona Surdykowska-Huk/Infocredit (Poland).

6.5. Model for Poland

Based on 100 companies' accounting data, the Danish Business Authority prepared an Early Warning model for Poland. The code for using the model is available online.[5] The list of data included in the Polish model is shown in Annex 3. On this basis, the Danish Business Authority trained a model on Danish data, where only these fields are included in the data base. The accuracy of the model for Polish companies is as follows:

		Model assessment		
		Not in distress (0)	In distress (1)	
Actual data	Not in distress (0)	(TN)	(FN)	**92 percent**
	In distress (1)	(FP)	(TP)	**62 percent**
		90 percent	**68 percent**	

The accuracy of the model is as follows:

- It identifies 92 percent of companies that are not in distress: TN/(TN+FN). Thus, the model mistakenly assumes that 8 percent of the healthy companies are in distress.

- It identifies 62 percent of companies in crisis: TP/(FP+TP). The model thus erroneously identifies 38 percent of distressed companies as healthy.

- It is right in 68 percent of the cases where it finds a company to be in distress: TP/(TP+FN). Conversely, 32 percent of them are actually healthy.

- It is correct in 90 percent of the cases in which it considers a business to be healthy: TN/(TN+FP). Conversely, 10 percent of them are actually in distress.

Overall, the accuracy of the Polish model can be calculated as being correct in 86 percent of all cases, based on the following formula:

$$\text{Accuracy} = \frac{\mathrm{TN+TP}}{\mathrm{TN+FN+FP+TP}}$$

6.6. Model for Spain

Based on 100 companies' accounting data, the Danish Business Authority prepared an Early Warning model for Spain. The code for using the model is available online.[6] The list of data included in the Spanish model is

[5] https://github.com/Niels-Peter/EW4E/tree/master/Poland

[6] https://github.com/Niels-Peter/EW4E/tree/master/Spain

shown in Annex 3. On this basis, the Danish Business Authority trained a model on Danish data, where only these fields are included in the data base. The accuracy of the model for Spanish companies is as follows:

		Model assessment		
		Not in distress (0)	In distress (1)	
Actual data	Not in distress (0)	(TN)	(FN)	**92 percent**
	In distress (1)	(FP)	(TP)	**61 percent**
		90 percent	**68 percent**	

The accuracy of the model is as follows:

- It identifies 92 percent of companies that are not in distress: TN/(TN+FN). Thus, the model mistakenly assumes that 8 percent of the healthy companies are in distress.

- It identifies 61 percent of companies in distress: TP/(FP+TP). The model thus erroneously identifies 39 percent of distressed companies to be healthy.

- It is right in 68 percent of the cases where it finds a company to be in distress: TP/(TP+FN). Conversely, 32 percent of them are actually healthy.

- It is correct in 90 percent of the cases in which it considers a business to be healthy: TN/(TN+FP). Conversely, 10 percent of them are actually in distress.

Overall, the accuracy of the Spanish model can be calculated as being correct in 86 percent of all cases, based on the following formula:

$$\text{Accuracy} = \frac{\mathrm{TN+TP}}{\mathrm{TN+FN+FP+T}}$$

6.7. Model for Greece

Based on 100 companies' accounting data, the Danish Business Authority prepared an Early Warning model for Greece. The code for using the model is available online.[7] The list of data included in the Greek model is shown in Annex 3. On this basis, the Danish Business Authority trained a model on Danish data, where only these fields are included in the data base. The accuracy of the model for Greek companies is as follows:

[7] https://github.com/Niels-Peter/EW4E/tree/master/Greece

		Model assessment		
		Not in distress (0)	In distress (1)	
Actual data	Not in distress (0)	(TN)	(FN)	**92 percent**
	In distress (1)	(FP)	(TP)	**63 percent**
		90 percent	**68 percent**	

The accuracy of the model is as follows:

- It identifies 92 percent of companies that are not in distress: TN/(TN+FN). Thus, the model mistakenly assumes that 8 percent of the healthy companies are in distress.

- It identifies 63 percent of companies in distress: TP/(FP+TP). The model thus erroneously identifies 37 percent of distressed companies to be healthy.

- It is right in 68 percent of the cases where it finds a company to be in distress: TP/(TP+FN). Conversely, 32 percent of them are actually healthy.

- It is correct in 90 percent of the cases in which it considers a business to be healthy: TN/(TN+FP). Conversely, 10 percent of them are actually in distress.

Overall, the accuracy of the Greek model can be calculated as being correct in 86 percent of all cases, based on the following formula:

$$\text{Accuracy} = \frac{\mathrm{TN+TP}}{\mathrm{TN+FN+FP+TP}}$$

6.8. Model for Italy

Based on 100 companies' accounting data, the Danish Business Authority prepared an Early Warning model for Italy. The code for using the model is available online.[8] The list of data included in the Italian model is shown in Annex 3. On this basis, the Danish Business Authority trained a model on Danish data, where only these fields are included in the data base. The accuracy of the model for Italian companies is as follows:

[8] https://github.com/Niels-Peter/EW4E/tree/master/Italy

		Model assessment		
		Not in distress (0)	In distress (1)	
Actual data	Not in distress (0)	(TN)	(FN)	**91 percent**
	In distress (1)	(FP)	(TP)	**66 percent**
		90 percent	**69 percent**	

The accuracy of the model is as follows:

- It identifies 91 percent of companies not in distress: TN/(TN+FN). The model thus erroneously identifies 9 percent of healthy companies to be in distress.

- It identifies 66 percent of companies in distress: TP/(FP+TP). The model thus erroneously identifies 34 percent of distressed companies to be healthy.

- It is right in 69 percent of the cases in which it finds a company to be in distress: TP/(TP+FN). Conversely, 31 percent of them are actually healthy.

- It is correct in 90 percent of the cases in which it considers a business to be healthy: TN/(TN+FP). Conversely, 10 percent of them are actually in distress.

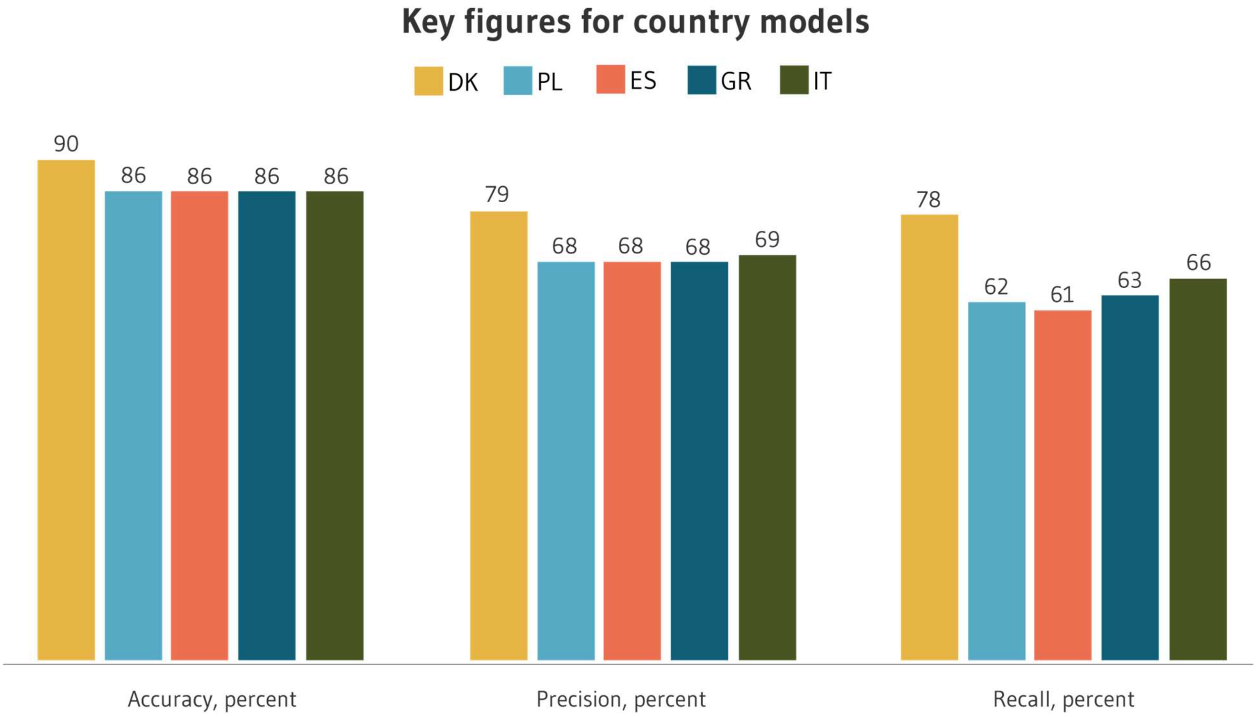
Overall, the accuracy of the Italian model can be calculated as being correct in 86 percent of all cases, based on the following formula:

$$\text{Accuracy} = \frac{\mathrm{TN+T}}{\mathrm{TN+FN+FP+TP}}$$

6.9. Overall model across countries

The predicative key figures of the models are summarised in the following charts:

	DK	PL	ES	GR	IT
Accuracy, percent	90	86	86	86	86
Precision, percent	79	68	68	68	69
Recall, percent	78	62	61	63	66


Key figures for country models
DK
PL
ES
GR
IT
90
86
86
86
86
79
68
68
68
69
78
62
61
63
66
Accuracy, percent
Precision, percent
Recall, percent

Annex 1: List of input data

The Danish base model uses 19 fields, for which it also uses 17 comparative numbers (Xs), i.e. a total of 36 fields.

No.	Label	XBRL element name	Comparative figures
Income statement			
1	EN: Gross profit or loss DK: Bruttofortjenester/Bruttotab	GrossProfit (fsa:GrossResult or fsa:GrossProfitLoss)	X
2	EN: Profit (loss) from ordinary operating activities DK: Resultat af ordinær drift	fsa:ProfitLossFromOrdinaryOperatingActivities	X
3	EN: Profit (loss) DK: Årets resultat	fsa:ProfitLoss	X
Distribution of net profit			
4	EN: Distributions DK: Uddelinger	fsa: ProfitLoss_DistributionsMember	X
The balance sheet			
5	EN: Assets DK: Aktiver	fsa:Assets	X
6	EN: Non-current assets DK: Langfristede aktiver	fsa:NoncurrentAssets	X
7	EN: Intangible assets DK: Immaterielle anlægsaktiver	fsa:IntangibleAssets	X
8	EN: Property, plant and equipment DK: Materielle anlægsaktiver	fsa:PropertyPlantAndEquipment	X
9	EN: Investments DK: Financielle anlægsaktiver	fsa:LongtermInvestmentsAndReceivables	X
10	EN: Current assets DK: Kortfristede aktiver	fsa:CurrentAssets	X
11	EN: Inventories DK: Varebeholdninger	fsa:Inventories	X
12	EN: Receivables DK: Tilgodehavender	fsa:ShorttermReceivables	X
13	EN: Short-term investments DK: Værdipapirer og kapitalandele	fsa:ShorttermInvestments	X
14	EN: Cash and cash equivalents DK: Likvide beholdninger	fsa:CashAndCashEquivalents	X
15	EN: Equity DK: Egenkapital	fsa:Equity	X
16	EN: Liabilities other than provisions DK: Gældsforpligtelser	fsa:LiabilitiesOtherThanProvisions	X
17	EN: Provisions DK: Hensatte forpligtelser	fsa:Provisions	X
Metadata			
18	EN: Date of general meeting DK: Generalforsamlingsdato	gsd:DateOfGeneralMeeting	
19	EN: Reporting period end date DK: Regnskabsperiodens slutdato	gsd:ReportingPeriodEndDate	

European Commission | Executive Agency for Small and Medium-sized Enterprises

No.	Label	XBRL element name	Comparative figures
No.	Label	XBRL element name	Comparative figures
Income statement			
1	EN: Gross profit or loss DK: Gross profit/loss	GrossProfit (fsa:GrossResult or fsa:GrossProfitLoss)	X
2	EN: Profit (loss) from ordinary operating activities DK: Profit from ordinary operations	fsa:ProfitLossFromOrdinaryOperatingActivities	X
3	EN: Profit (loss) DK: Profit for the year	fsa:ProfitLoss	X
Distribution of net profit			
4	EN: Distributions DK: Dividends	fsa: ProfitLoss_DistributionsMember	X
The balance sheet			
5	EN: Assets DK: Assets	fsa:Assets	X
6	EN: Non-current assets DK: Non-current assets	fsa:NoncurrentAssets	X
7	EN: Intangible assets DK: Intangible assets	fsa:IntangibleAssets	X
8	EN: Property, plant and equipment DK: Property, plant and equipment	fsa:PropertyPlantAndEquipment	X
9	EN: Investments DK: Investments	fsa:LongtermInvestmentsAndReceivables	X
10	EN: Current assets DK: Current assets	fsa:CurrentAssets	X
11	EN: Inventories DK: Inventories	fsa:Inventories	X
12	EN: Receivables DK: Receivables	fsa:ShorttermReceivables	X
13	EN: Short-term investments DK: Investments	fsa:ShorttermInvestments	X
14	EN: Cash and cash equivalents DK: Cash and cash equivalents	fsa:CashAndCashEquivalents	X
15	EN: Equity DK: Equity	fsa:Equity	X
16	EN: Liabilities other than provisions DK: Liabilities	fsa:LiabilitiesOtherThanProvisions	X
17	EN: Provisions DK: Provisions	fsa:Provisions	X
Metadata			
18	EN: Date of general meeting DK: Date of general meeting	gsd:DateOfGeneralMeeting	
19	EN: Reporting period end date DK: Reporting period end date	gsd:ReportingPeriodEndDate	

Annex 2: List of calculated features

No.	Feature for the model	Calculation of feature	Comparative figures
1+2	Gross profit to ordinary operating activities ratio	(Gross profit – Profit loss from ordinary operating activities) Gross profit	X
3+4	Ordinary operating activities to profit loss ratio	(Profit loss from ordinary operating activities – Profit loss) Gross profit	X
5+6	Return on equity	Profit loss Equity	X
7+8	Profit loss distributions ratio	Profit loss distributions member Profit loss	X
9	Assets delta	Assets – Assets previous year Assets previous year	
10	Non-current assets delta	Non-current assets – Non-current assets previous year Non-current assets previous year	
11	Non-current assets ratio	Non-current assets Assets	
12	Intangible assets delta	Intangible assets – Intangible assets previous year Intangible assets previous year	
13	Intangible assets ratio	Intangible assets Assets	
14	Property, plant and equipment delta	Property, plant and equipment – Property (…) previous year Property, plant and equipment previous year	
15	Property, plant and equipment ratio	Property, plant and equipment Assets	
16	Long-term investments and receivables delta	Long-term investments and receivables – Long-term (…) previous year Long-term investments and receivables previous year	
17	Long-term investments and receivables ratio	Long-term investments and receivables Assets	
18	Current assets delta	Current assets – Current assets previous year Current assets previous year	
19	Current assets ratio	Current assets Assets	
20	Inventories delta	Inventories – Inventories previous year Inventories previous year	
21	Inventories ratio	Inventories Assets	
22	Short-term receivables delta	Assets – Assets previous year Assets previous year	
23	Short-term receivables ratio	Short-term receivables Assets	
24	Short-term investments delta	Short-term receivables – Short-term receivables previous year Short-term receivables previous year	
25	Short-term investments ratio	Short-term investments Assets	
26	Cash and cash equivalents delta	Cash and cash equivalents – Cash and cash equivalents previous year Cash and cash equivalents previous year	
27	Cash and cash equivalents ratio	Cash and cash equivalents Assets	
28	Equity delta	Equity – Equity previous year Equity previous year	
29	Solvency ratio	Equity Assets	X
30	Liabilities other than provisions delta	Liabilities other than provisions – Liabilities other than provisions previous year Liabilities other than provisions previous year	

No.	Feature for the model	Calculation of feature	Comparative figures
31	Liabilities other than provisions ratio	$\frac{\text{Liabilities other than provisions}}{\text{Assets – Equity}}$	
32	Provisions delta	$\frac{\text{Assets – Assets previous year}}{\text{Assets previous year}}$	
33	Provisions ratio	$\frac{\text{Provisions}}{\text{Assets – Equity}}$	
34	Comparative figures?	Do the financial statements contain figures from both this year and last year? Yes/No	N/a
35	Days end to meeting	No. days between date of general meeting and reporting period end date	N/a

Appendix 3: Lists of fields in the derived models

No.	Label	XBRL element name	DK	PL	ES	GR	IT
Income statement							
1	EN: Gross profit or loss DK: Bruttofortjeneste/Bruttotab	GrossProfit (fsa:GrossResult or fsa:GrossProfitLoss)	X			X	
2	EN: Profit (loss) from ordinary operating activities DK: Resultat af ordinær drift	fsa:ProfitLossFromOrdinaryOperatingActivities	X		X	X	X
3	EN: Profit (loss) DK: Årets resultat	fsa:ProfitLoss	X	X	X	X	X
Distribution of net profit							
4	EN: Distributions DK: Uddelinger	fsa: ProfitLoss_DistributionsMember	X				
The balance sheet							
5	EN: Assets DK: Aktiver	fsa:Assets	X	X	X	X	X
6	EN: Non-current assets DK: Langfristede aktiver	fsa:NoncurrentAssets	X	X	X	X	X
7	EN: Intangible assets DK: Immaterielle anlægsaktiver	fsa:IntangibleAssets	X	X	X	X	X
8	EN: Property, plant and equipment DK: Materielle anlægsaktiver	fsa:PropertyPlantAndEquipment	X	X	X	X	X
9	EN: Investments DK: Finansielle anlægsaktiver	fsa:LongtermInvestmentsAndReceivables	X	X	X	X	X
10	EN: Current assets DK: Kortfristede aktiver	fsa:CurrentAssets	X	X	X	X	X
11	EN: Inventories DK: Varebeholdninger	fsa:Inventories	X	X	X	X	X
12	EN: Receivables DK: Tilgodehavender	fsa:ShorttermReceivables	X	X	X	X	X
13	EN: Short-term investments DK: Værdipapirer og kapitalandele	fsa:ShorttermInvestments	X	X	X	X	X
14	EN: Cash and cash equivalents DK: Likvide beholdninger	fsa:CashAndCashEquivalents	X	X	X	X	X
15	EN: Equity DK: Egenkapital	fsa:Equity	X	X	X	X	X
16	EN: Liabilities other than provisions DK: Gældsforpligtelse	fsa:LiabilitiesOtherThanProvisions	X	X	X	X	
17	EN: Provisions DK: Hensatte forpligtelser	fsa:Provisions	X	X	X	X	
Metadata							
18	EN: Date of general meeting DK: Generalforsamlingsdato	gsd:DateOfGeneralMeeting	X				
19	EN: Reporting period end date DK: Regnskabsperiodens slutdato	gsd:ReportingPeriodEndDate	X				
No.	**Label**	**XBRL element name**	**DK**	**PL**	**ES**	**GR**	**IT**
Income statement							
1	EN: Gross profit or loss DK: Gross profit/loss	GrossProfit (fsa:GrossResult or fsa:GrossProfitLoss)	X			X	
2	EN: Profit (loss) from ordinary operating activities DK: Profit from ordinary operations	fsa:ProfitLossFromOrdinaryOperatingActivities	X		X	X	X
3	EN: Profit (loss) DK: Profit for the year	fsa:ProfitLoss	X	X	X	X	X
Distribution of net profit							

No.	Label	XBRL element name	DK	PL	ES	GR	IT
4	EN: Distributions DK: Dividends	fsa: ProfitLoss_DistributionsMember	X				
The balance sheet							
5	EN: Assets DK: Assets	fsa:Assets	X	X	X	X	X
6	EN: Non-current assets DK: Non-current assets	fsa:NoncurrentAssets	X	X	X	X	X
7	EN: Intangible assets DK: Intangible assets	fsa:IntangibleAssets	X	X	X	X	X
8	EN: Property, plant and equipment DK: Property, plant and equipment	fsa:PropertyPlantAndEquipment	X	X	X	X	X
9	EN: Investments DK: Investments	fsa:LongtermInvestmentsAndReceivables	X	X	X	X	X
10	EN: Current assets DK: Current assets	fsa:CurrentAssets	X	X	X	X	X
11	EN: Inventories DK: Inventories	fsa:Inventories	X	X	X	X	X
12	EN: Receivables DK: Receivables	fsa:ShorttermReceivables	X	X	X	X	X
13	EN: Short-term investments DK: Investments	fsa:ShorttermInvestments	X	X	X	X	X
14	EN: Cash and cash equivalents DK: Cash and cash equivalents	fsa:CashAndCashEquivalents	X	X	X	X	X
15	EN: Equity DK: Equity	fsa:Equity	X	X	X	X	X
16	EN: Liabilities other than provisions DK: Liabilities	fsa:LiabilitiesOtherThanProvisions	X	X	X	X	
17	EN: Provisions DK: Provisions	fsa:Provisions	X	X	X	X	
Metadata							
18	EN: Date of general meeting DK: Date of general meeting	gsd:DateOfGeneralMeeting	X				
19	EN: Reporting period end date DK: Reporting period end date	gsd:ReportingPeriodEndDate	X				

Annex 4: Experience notes from pilot testing in Skive Municipality

- *By Vaeksthus Midtjylland/Business Development Centre of Central Denmark*

The pilot testing was conducted in Skive Municipality, selected for testing on the basis of its ranking as a close-to-average municipality in terms of inhabitants and number of companies.

The testing for this experience paper was conducted by Early Warning Denmark consultants Michael Lynge Hansen and Bent Aage Madsen by using an Excel file received from DBA containing an unfiltered gross list of 11,230 company registration numbers, extracted from the machine learning tool.

At first, the list may seem overwhelming as no filtering or sorting opportunities are (yet) available in the Excel format. These will be available as the programme further advances. The filtering was thus done manually by the Early Warning Denmark Consultants on the basis of the default parameters listed by the machine learning tool in the Excel sheet. The choices on how to filter the companies and how to separately classify these was a difficult task, as this had much importance to the following step. This following step was to sort the list on behalf of a number of predetermined data filters.

In an effort to maintain and secure the overriding usefulness of the testing the Consultants decided to maintain a focus on organisations which had submitted annual accounts for the fiscal years 2016 and 2017. Additionally, the Excel file generated by the machine learning tool includes a number of organisations for whom no potential distress probability was calculated. This category includes financial holding companies, companies which have gone bankrupt in the meantime, merged companies, companies which have been judicially wound up etc.

On behalf of these requirements (rendering of annual accounts and model calculation) the number of organisations with a potential of further examination was reduced to 307.

The output of the model is explained as follows:

> *With results/outputs below 0.50 the model exhibits a certain degree of doubt. With results/outputs approaching 1.00 the model exhibits an increasing amount of certainty.*

With this in mind, the researchers started looking at companies with a model output above (>)0.50. This led to a further decrease in the number of companies for further examination to 79. All 79 companies were examined manually on the basis of annual accounts, company webpage and other insights into the organisation – the human factor. This manual examination looked at factors such as the existence of other companies in the same corporation or factors which have a negative impact on the solidity of the company, but which do not always signal a crisis, including large purchases for stock building, the payment of dividends to company owners or the failure of one debtor. In short, this examination was about understanding the numbers before contacting the companies in question. In addition to this, the Consultants noticed a number of companies on the filtered list with stable balance sheets, a strong net capital etc. which in general would signal a healthy organisation. These “false positives” with no apparent signs of crisis were removed from the list.

On the basis of this examination of these more specified factors we made a further decrease of the list of companies potentially in distress, now down to 37.

In terms of business sectors, the list showed the following split:

- 11 organisations within the construction industry,
- 9 organisations within the manufacturing industry,
- and the remaining 17 in different industries.

The companies on the list were distributed evenly to the two Consultants. This was done in a manner allowing opportunity for gathering information on any potential special circumstances within a specific business sector.

The next step was to contact these companies by telephone. Ahead of this telephonic contact we considered several opportunities in terms of approaching them effectively and with the highest possible hit rate. We gave high priority to this first approach to the companies on the list as it may have great impact on the impression of our contact by the company owners and thereby on our hit rate. It is worth noting that both consultants have many years of knowledge and experience in the field of canvassing and commercial outreach. This meant that both consultants could approach company owners with confidence and knowhow. Nonetheless, canvassing is a special discipline that calls for a sense of the occasion for meeting the company owners with empathy and intuition, so a slow, well considered approach is necessary. We have chosen to distinguish between two approaches; *the direct and the indirect.* Examples below:

The direct approach: "Hello. I am calling from Early Warning. An analysis has been conducted on your organisation, showing potential pitfalls and signs of crisis."

The indirect approach: "Hello. I am calling from Væksthus Midtjylland. We help organisations with a broad variety of problems and challenges."

It is noteworthy that in both cases much attention and importance are put to the fact that Early Warning offers a free and non-committal dialogue with the company owners.

Based on the manual examination of the individual companies described above, our consultants chose the direct approach in the majority of cases, but both approaches proved to be efficient.

We chose to contact the company owners by phone. We could have written to them instead, but our experience from Early Warning Europe shows that the direct, interactive contact tends to give the best results while at the same time allowing the company owners to ask questions and express their views.

What are the results?

Cold canvass can be a difficult task. But our round of phone calls on the basis of the filtered list generated by the machine learning tool yielded a hit rate (agreed meeting appointments with company owners) of approximately 50 percent. We are very satisfied with this result of the outreach effort.

Our recommendations

It is our suggestion that the filtered lists generated on the basis of the results from the machine learning tool are worked up in advance, making it less time-consuming for the individual consultant to gain understanding of the data by filtering and sorting. Simultaneously, every consultant must be carefully selected and coached. The consultants must have a willingness and ability to perform cold canvas tasks, to confront company owners with potential problems and to speak openly about difficult topics.

In choosing the method for contacting the company owners, we recommend taking into account the resource spending for this part of the process but also the bias that may lie in the communication approach. For instance, written communication may be perceived very differently across national and sectoral barriers.

Used in the Danish system, the machine learning tool will not include the many proprietorships showing signs of crisis, as these companies do not render annual accounts to be publicly published. Thereby there is no foundation for examining them manually as they simply are not visible in the model.

A golden opportunity deriving from the approach of contacting companies by phone on the basis of the probabilities identified by the machine learning tool is to create focused 'marketing'. This could potentially be done on an industry level based upon geographical differentiation, meetings in diverse associations (craftsmen, farmers etc.), industry associations and so on.

In this sense, the machine learning tool performs the arduous task of prequalifying companies from a very large quantity of data and narrowing the list down to a limited number of potential candidates for contacting. Such prequalification will lead to an optimised resource spending and a more focused approach for getting into contact with companies in distress.