ANALYSIS

DANMARKS
NATIONALBANK

21 JANUARY 2019 — NO. 3

Text-based machine learning improves distress modelling

Machine learning improves calculation of probability of distress

Machine learning provides the opportunity to use new modelling methods and to include unstructured data. That increases the accuracy of the calculated probabilities of distress for firms.

Text-based data adds useful information

Inclusion of unstructured data in the form of texts from firms' annual reports expands the information base and improves the possibility of predicting distress.

Auditors' reports have the greatest effect

Especially the auditors' reports contain useful information for calculating the probability of distress.

The ability to predict corporate distress is improved if the assessment is based not only on numerical financial data, but also on texts from the annual reports. That is the conclusion in a working paper[1] published in November 2018, in which machine learning methods are used to calculate probability of distress on the basis of text segments in annual reports.

The paper investigates whether two text segments (auditors' reports and managements' statements) that can be found in the annual reports of most Danish firms can provide useful information about a firm's risk of entering into distress. That is, information which is not already obvious from the numerical financial data.

This analysis reviews the main conclusions of the working paper. Especially the auditors' reports provide useful information in terms of predicting a firm's risk of entering into distress.

Purpose of distress model

The purpose of a distress model is to calculate the probability that a firm enters into distress and ultimately defaults. Such calculations can be used for many different purposes. From a financial stability perspective, they may give an indication of whether some banks are more exposed to firms which are in, or heading for, financial difficulties.

A classical distress model is typically based on a firm's numerical financial data and certain characteristics such as the industry and age of the firm. Besides the numerical financial data, firms' annual reports also include two text segments: auditors' reports and managements' statements, which can be classified as unstructured data.

Box 1 contains an excerpt from an auditors' report. It is seen that the auditors in this case take a negative view of the continued operations of the firm. This could indicate that the firm's probability of distress is heightened. But this information may not be fully reflected in the numerical financial data. So there is a potential for improving the distress model by including information from text segments of this type.

Example of auditors' report

Box 1

... "It is our assessment that there are no realistic options for obtaining funding and we therefore make the caveat that the statement has been submitted on the basis of continued operations. " ...

What is machine learning?

Machine learning is not a new phenomenon. It has existed since the 1940s, when the first theoretical models were developed. But not until recent years, with the rapid development of computing power and the availability of much larger data volumes, has the potential for and interest in applying and developing machine learning really taken off.

Machine learning differs from the more conventional statistical methods, first and foremost by making it possible to model more complex relationships in data that the analyst is not necessarily familiar with beforehand. While conventional methods are about "fitting" data into pre-specified relationships between input and output variables, machine learning methods enable a "freer" approach to the modelling of data as they to a larger extent let the data "speak" rather than trying to force the data to fit into a specific functional form. In addition, machine learning methods provide new opportunities to use "unstructured data" such as pictures and text.

The text-based distress model comprises three types of neural network: a "convolutional neural network", a "recurrent neural network" and a classical neural network. For a description of how a neural network functions, see Box 2.

[1] Hansen, Casper, Christian Hansen, Rastin Matin and Pia Mølgaard, Predicting distresses using deep learning of text segments in annual reports, *Danmarks Nationalbank Working Paper*, No. 130, November 2018 (*link*).

Text analysis

This is a presentation of the machine learning method that estimates probabilities of distress on the basis of text. The modelling comprises the following three steps:

1. word representation,
2. semantics of the word combinations and
3. inclusion of numerical financial data and calculation of probability of distress.

These three steps are outlined below. They all include multiple parameter choices that will be continuously updated as the model is trained. In other words, all parameters have been optimised for precisely our problem, which is calculation of the probability of distress. A diagram of the aggregate model can be found in Chart 1.

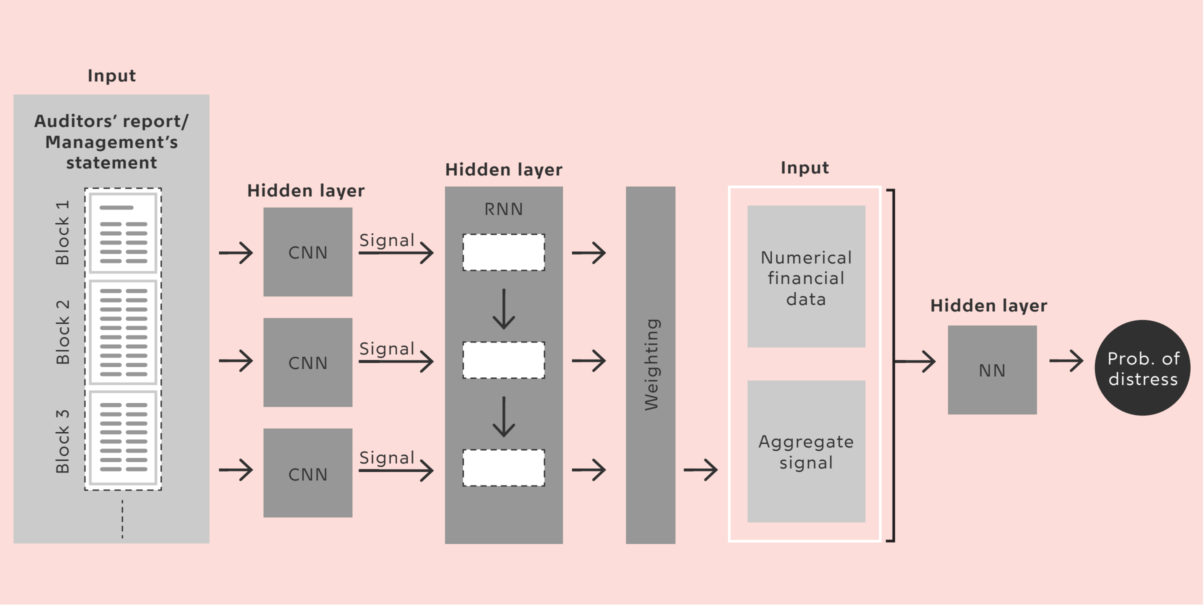
Step 1: Word representation

First of all we need to create a word representation, which is done by means of "word2vec". This method is used to convert all words from text to numbers, which are easier for the model to work with. All (unique) words are converted into series of numbers that uniquely identify each word.

As part of the word representation, words that "resemble" each other semantically will also resemble each other in the numerical representation, so that calculations can be made on the words in relation to their meanings. For example, the numerical representation of the following word combination *king* – *man* + *woman* will be very close to the number representation of the word *queen*.

Diagram of text analysis model

Chart 1



Note: In the first input layer, the text is broken down into blocks that are run through separate CNNs which produce signals. These signals are fed into the RNN, which takes into account the relative positions of the text blocks. The signals are then weighted relative to each other, and the aggregate signal is combined with numerical financial data and run through a classical neural network, which ultimately produces a probability of distress.

Step 2: Semantics of the word combinations
While the first step involves understanding and representing the individual words in the text from the annual report, this step is about understanding their *context*. So we are moving away from simple text analysis, which merely considers the individual words, towards understanding the meaning of full sentences or paragraphs in the text segments.

More specifically, this is done by first splitting the auditors' reports and managements' statements into smaller, overlapping text blocks, illustrated by the input layer on the far left in Chart 1. Combined, all the blocks in the layer make up one auditor's report or one management's statement. Using a convolutional neural network, CNN, a low-dimensional representation of the semantics of the whole text block is created.

The signals from the individual blocks are then combined into an aggregate semantic signal representing the whole text segment. This is done in a recurrent neural network, RNN, which sequentially reads the individual signals from the blocks and combines them into a final signal. In this way, the network "reads" the whole text and creates an aggregate signal in which the sequence of the individual text segments may be of importance. If, say, there is a negation before a positive word, the neural network understands that the meaning of the sentence is negative.

Step 3: Inclusion of numerical financial data and calculation of probability of distress
In this step, the above aggregate signal from the text is combined with the "conventional" numerical financial data, and training of the model is completed on both numerical and text-based data in a neural network, NN.

Training of the model consists in feeding it with numerous examples of firms, with a combination of their numerical financial data and texts from their annual reports and an indicator of whether or not they have entered into distress. Based on all these "known" examples, the model learns which combinations of numerical financial data and text signals typically characterise a distressed firm compared with a non-distressed firm. The learning method is, roughly speaking, trial and error. Based on the input fed into the model, it makes a "guess" at whether or not a firm is distressed. Since we already know the answer, the model can reassess its guess in relation to the correct answer by changing the network parameters and checking whether this reassessment brings it closer to the correct answer. This is done again and again, until the model gets as close as possible to the correct answer.

Training of the model involves the whole system. In other words, when the model "tries out" a new solution, it will also change the signals created in step 2.

Avoid an "over-fitting" model
Because machine learning models contain large data volumes and many different parameters, there is a large risk that they will "over-fit". This means that the model becomes very good at predicting the probability of distress of the firms in the training set, but does not necessarily perform equally well on a new set of firms. To avoid over-fitting, the model's predictive powers are continuously evaluated on a validation set. Training of the model only continues so long as it improves with regard to both the training set and the validation set. This method is known as cross-validation and is commonly used in machine learning.

Results and conclusion

Model evaluation and basis for comparison
All in all, three different neural networks containing text are implemented: 1. a model containing both auditors' reports and managements' statements, 2. a model containing auditors' reports only, and 3. a model containing managements' statements only. Furthermore, two models based only on numerical financial data are implemented: 1. a neural network (without text) exploiting the good properties of machine learning models – but on conventional numerical financial data only, and 2. logistic regression, which is a conventional statistical method for predicting distress. These two models are to be used as a basis for comparison to investigate whether the models that include text become better at predicting a firm's probability of distress.

To compare the predictive powers of the models, AUC (Area Under the receiver operating characteristics Curve) is applied. AUC measures the probability that a model attaches a higher risk to a random firm that enters into distress in a given year than to a random firm that does not. So AUC primarily gives an indication of the *ranking* of firms by their probability of distress. The AUC score will always be between 0.5

and 1. The closer it is to 1, the better the model is at ranking firms by their risk of entering into distress.

Results

Chart 2 shows AUC for the five different models. There are several interesting results. If AUC for the logistic regression is compared with the neural network without text, it is seen that the predictive powers are improved substantially by implementing a machine learning model rather than a conventional statistical model, even though the two models apply the same data.

Furthermore, we see that the models including text perform even better than the neural network which includes numerical financial data only. The differences in AUC between the text models and the neural network without text are all statistically significant at a 1 per cent level. So we can conclude that the texts from the annual reports contain useful information that can be exploited in a distress model to improve its predictive powers.

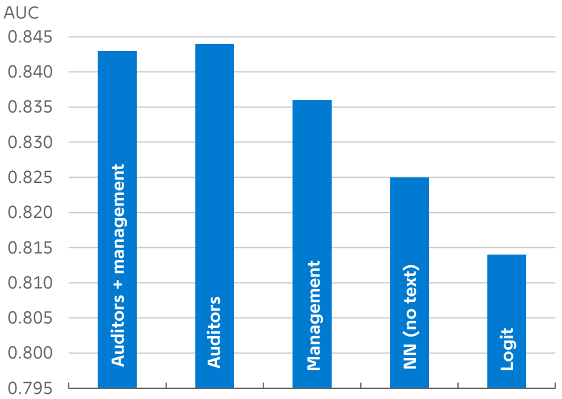
It is also interesting to note that the model with auditors' reports performs better than the two other text-based models, but not statistically significantly better than the model containing both managements' statements and auditors' reports. On the basis of this model it can therefore be concluded that there is no useful information in the managements' statements beyond that already found in the auditors' reports. The reason may be that an auditors' report provides a more objective assessment of a firm's status, while the management may tend to focus only on areas where the firm is performing well.

Conclusion

It is seen that machine learning methods are useful when it comes to predicting distress. In particular, the predictive powers are improved by including text-based analysis of auditors' reports.

Relative predictive powers of the models

Chart 2



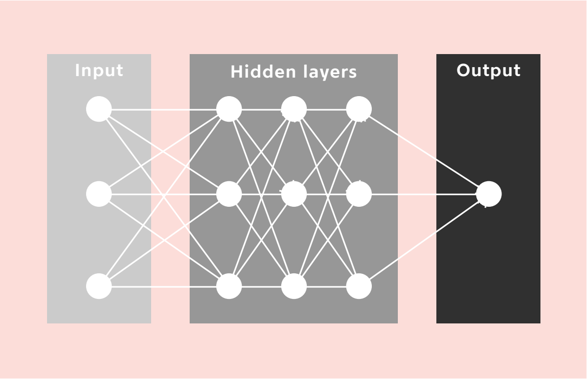
Note: AUC for the five different models. AUC is a measure of how well the model ranks firms by their probability of distress. A higher AUC score means that the model has better predictive powers.

Source: Bisnode and own calculations.

Neural network

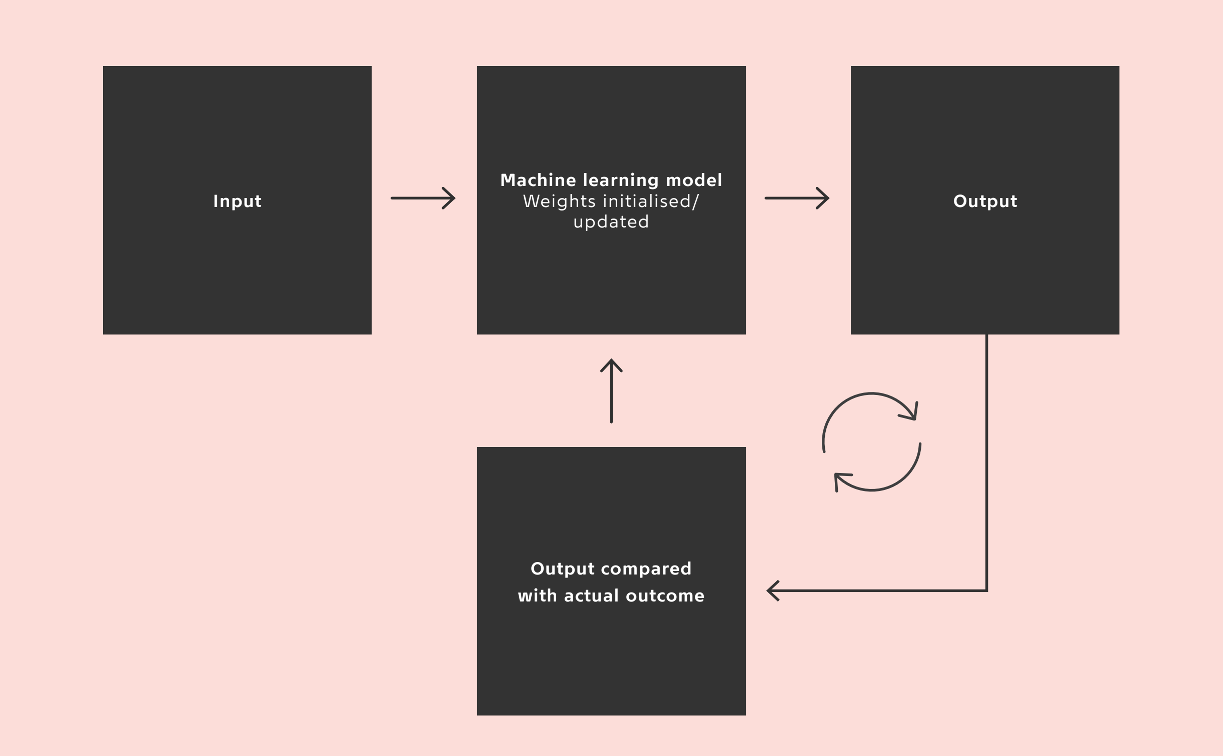
Box 2

A neural network is somewhat similar to the structure of the human brain, and particularly the brain's way of processing information and learning. A neural network consists of an input layer, one or more "hidden layers" and an output layer, as shown in the top chart. Each layer consists of a number of "neurons" containing unique information. In image analysis, for example, each neuron in the input layer can represent a pixel in a picture. Information from this layer is weighted and combined in the next layer of neurons. In this way, information propagates through the network and ultimately reaches the output layer, where the final outcome is stated, e.g. whether or not it is a picture of a dog.



By feeding a neural network with numerous examples of known outcomes, the network can be trained to recognise patterns and contexts that can be used to predict new, unknown outcomes. So if you want to teach the network to recognise a dog in a picture, it is fed a lot of different pictures of dogs, a lot of different pictures of other animals and an indicator of whether or not the individual pictures are of dogs.

The model then starts by "guessing" an outcome (e.g. dog/not dog) based on the given input. An evaluation is then performed of how close the guess is to the actual outcome, and the weights of the neural network are updated accordingly, and then the model once again tries to guess an outcome. The process is repeated until the network's guess, which gradually becomes more and more educated, is as close as possible to the actual outcome. The training process is illustrated in the lower chart.



ABOUT ANALYSIS

As a consequence of Danmarks Nationalbank's role in society we conduct analyses of economic and financial conditions.

Analyses are published continuously and include e.g. assessments of the current cyclical position and the financial stability.

The analysis consists of a Danish and an English version. In case of doubt regarding the correctness of the translation the Danish version is considered to be binding.

DANMARKS NATIONALBANK
HAVNEGADE 5
DK-1093 COPENHAGEN K
WWW.NATIONALBANKEN.DK

This edition closed for contributions on 18 January 2019

Anna Kirstine Hvid
Quantitative Risk Analyst

Pia Mølgaard
Quantitative Risk Analyst

FINANCIAL STABILITY

DANMARKS NATIONALBANK