ROBECO
The Investment Engineers

Quantitative Equities

Forecasting stock crash risk with machine learning

Sustainable Investing Expertise by ROBECOSAM

White paper
For professional investors
June 2022

Laurens Swinkels, PhD
Tobias Hoogteijling

Contents

Introduction

Significant developments in big data and machine learning (ML) are pushing back the frontier of quant investing.

This has been facilitated by the increase in computational power, which enables the deployment and use of ML models.[1] In contrast to rule-based models, these adopt a fully data-driven approach and are capable of modelling complex, nonlinear relationships. They can potentially uncover systematic and repeating patterns that simple linear models do not capture.

‘Certain variables might only be able to predict returns when they cross a certain threshold, or when they are combined with other variables, or they might only be able to predict negative performance’

For example, certain variables might only be able to predict returns when they cross a certain threshold, or when they are combined with other variables, or they might only be able to predict negative performance. In this white paper, we delve into how the use of ML techniques can propel quant modelling to the next level. We also look at a concrete example of how they can be used to forecast individual stock price crashes.

The different uses of ML techniques in quantitative investing are increasingly being recognized in the academic literature. These range from relatively simple variable selection models to ones able to identify lead-lag relationships between the returns of different assets. ML techniques are also used in complex deep learning models for statistical arbitrage.[2,3] Overfitting has always been a key concern for quantitative strategies. It is an even bigger issue when the amount of data and the number of possible relationships between variables increases.

The risk here is that one can uncover a coincidental relationship between variables that has been observed in the past and use it as a signal to construct a portfolio, when in fact it will not repeat in the future as there is no real underlying phenomenon.[4] The ML toolbox, however, contains solutions to avoid overfitting, such as regularization (i.e., variable selection), model averaging and cross-validation.

[1] See: Van Vliet, P., and Zhou, W., August 2021, “Moore’s Law is disrupting the world of quant investing”, Robeco article.
[2] For example, LASSO is a frequently used machine learning method that selects explanatory variables when many are available in a statistical regression model. LASSO is the abbreviation for Least Absolute Shrinkage and Selection Operator.
[3] See: Rapach, D. E., Strauss, J. K., and Zhou, G., March 2013, “International stock return predictability: what is the role of the United States?” Journal of Finance; and Fisher, T., and Kraus, C., October 2018, “Deep learning with long short-term memory networks for financial market predictions” European Journal of Operational Research.
[4] See: Martin, I., and Nagel, S., forthcoming, “Market efficiency in the age of big data” Journal of Financial Economics.

Forecasting which firms will face financial distress

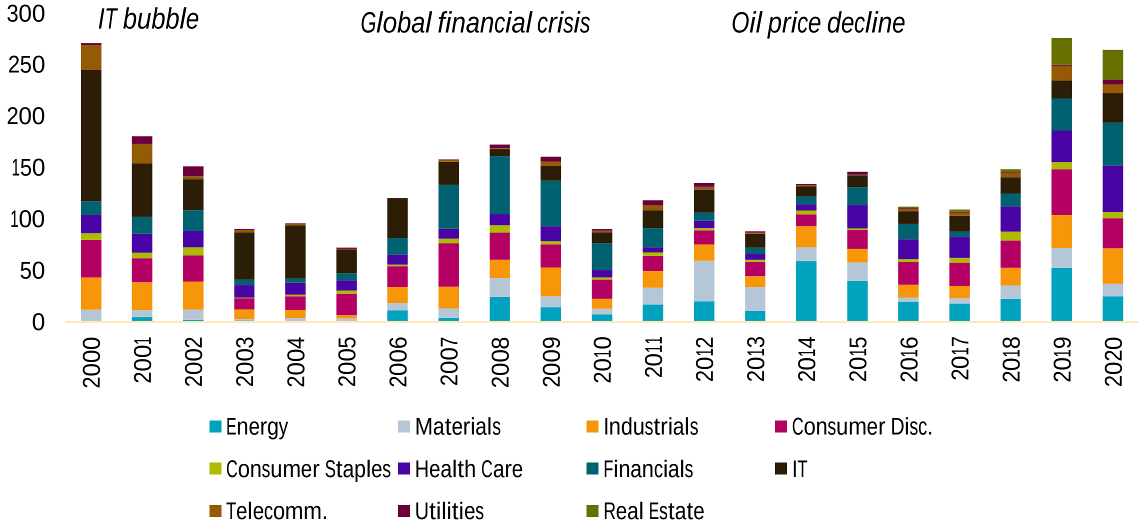
Avoiding investments in companies that will subsequently experience financial distress can help investors outperform the market.

That said, detecting those firms most likely to face distress in the future is far from straightforward. While there are numerous characteristics that can help forecast such outcomes, they perhaps do so most effectively in a nonlinear fashion or in specific combinations. Fortunately, ML techniques are designed to deal with such challenges.

One way to flag distressed firms is by looking at bankruptcy filings or credit downgrades.[5] However, these events can be preceded by falling stock prices. For an equity investor, predicting a stock price decline is more relevant than correctly forecasting a bankruptcy filing or credit downgrade, as these events could already be priced in by the time they take place. Therefore, we define a crash event as a significant drop in a firm's share price relative to those of other stocks. This measure captures idiosyncratic risk rather than market-wide turmoil, and we believe it lends itself well for use in a stock selection model instead of a market or factor-timing model used for asset allocation.

Based on our definition of financial distress, Figure 1 shows that the identified distress events are not concentrated in a handful of sectors, but across multiple. That said, the more distressed sectors usually have a higher tally of events. Overall, the number of stock price crashes varies over time as cross-sectional volatility is typically higher during crisis periods.

Figure 1 | Average number of stock price crashes according to our definition of financial distress, for the period January 2000 to December 2020



Source: Robeco Quantitative Research. The bars contain the average number of distress events per calendar year split by sector. The developed markets universe consists of the largest 3,000 stocks from the S&P Broad Market Index and MSCI World Index. The emerging markets universe consists of the largest 25% constituents of the S&P Emerging Broad Market Index and MSCI Emerging Markets Index.

[5] See: Campbell, J Y., Hilscher, J., and Szilagyi, J., November 2008, "In search of distress risk" Journal of Finance.

If we want to identify firms that will face financial distress through the use of a ML model, we need to feed it a set of variables that may contain predictive information on stock price crashes. In ML terminology, these predictive variables are called features, while the variable that is being predicted is called the target.

Selecting the set of features that serve as inputs for the algorithm that aims to find a predictive relationship with the target is an important modelling step for any predictive model. For our purposes, the variables we chose include classical financial distress indicators such as distance-to-default, volatility, and return on equity.[6] But since ML models are designed to pick the most important features from a larger set of variables, we also decided to include more granular items from financial statements such as cash flow from operations as well as short and long-term debts.

In our model, we only incorporate features that we believe are potentially relevant to identify firms with stock price crash risk. For instance, features such as the first letter of a company name or the color of its logo are not included, given that we are not able to link them to an economic rationale that could explain why they may be able to forecast crash risk. Since many company characteristics are only a snapshot at a point in time, for example profitability, we have also included several lagged versions for some of the features in our model. In our view, this allows the algorithm to identify how series of events, such as declining profitability, influence distress risk.

Although most ML algorithms are designed to filter out irrelevant predictors from a large set of features, this process is also useful in avoiding overfitting as the number of features is reduced. We believe fewer redundant features lowers the chance of making investment decisions based on noisy data. Furthermore, decreasing the number of features speeds up the training process of the algorithm. For our application, we train three statistical models: a regularized logistic regression, a random forest, and an extreme gradient boosted tree.

A regularized logistic regression is a method based on a classical linear regression model, but through a logistic transformation (where measurements on a linear scale are converted into probabilities) that is adapted to predict the probability of a binary outcome: in our case, whether a company is in financial distress or not. Regularization is ML terminology for model selection, i.e., this technique only selects those variables that help predict this binary outcome.

Meanwhile, the random forest classification is a nonlinear model based on multiple decision trees – hence the term forest – with randomly chosen features as the nodes, where the majority vote decides the classification (determination of which group an observation belongs to), which in our application is a binary decision. Finally, the extreme gradient boosted tree algorithm is also based on decision trees, but instead of relying on majority voting, it changes the weighting of the features in a manner that reduces the number of false predictions.[7]

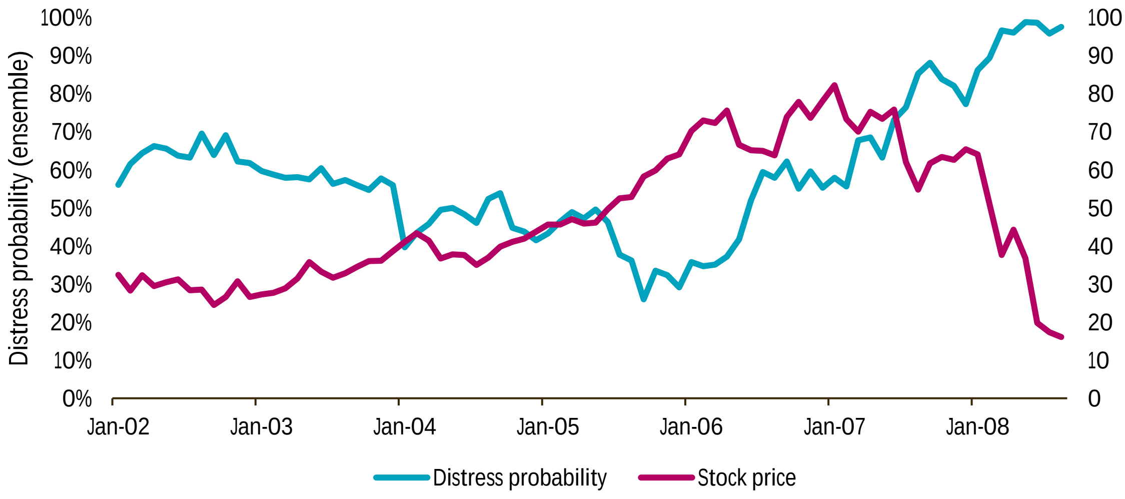
> ‘Data leakages lead to better backtested results that cannot be achieved in practice as information that will be accessible in the future will obviously not be available’

These three prediction models are parameterized and retrained each year based on their cross-validation metrics. After retraining the models, we construct the signal for the upcoming year based on previously unseen data. When setting up this training-testing configuration, it is extremely important to ensure there are no data leakages. This refers to the occurrence of overfitting that happens when previously unseen data is leaked in the model training process. Data leakages lead to better backtested results that cannot be achieved in practice as information that will be accessible in the future will obviously not be available.

[6] Distance-to-default is the number of asset return standard deviations that a firm can handle with its equity before it will default. If the number is high, this means that the firm’s equity buffer is large compared to its business risk. See: Duffie, D., Saita, L., and Wang, K., March 2007, “Multi-period corporate default prediction with stochastic covariates“, Journal of Financial Economics.
[7] For more information on machine learning models see: Snow, D., January 2020, “Machine learning in asset management – part 1: portfolio construction – trading strategies“, Journal of Financial Data Science; and Snow, D., April 2020, “Machine learning in asset management – part 2: portfolio construction – weight optimization“, Journal of Financial Data Science.

Each of these models make independent predictions, which are then combined as an ensemble prediction. In our view, ensembles are the only free lunch in ML, similar to what diversification is in terms of investing. Figure 2 illustrates the share price of a US financials stock as well as its ensemble forecasted distress risk over time. In this example, we see that its probability of distress increases beyond 75% post mid-2007, and eclipses 90% in the months that follow. What we also observe is that its stock price peaks in January 2007, but only experiences a sharp drop from March 2008 onwards before the company files for bankruptcy in September 2008.

Figure 2 | Illustrative example of a stock price and its forecasted distress risk, for the period January 2002 to August 2008



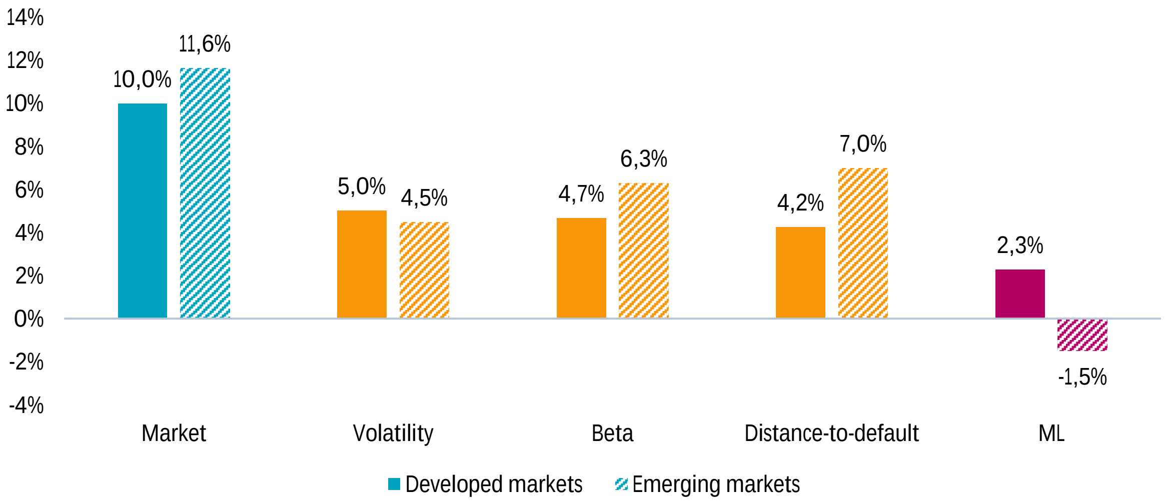
Source: Robeco Quantitative Research. This is an illustration of the forecast distress probability of a US financials stock (left-hand-side y-axis) and its realized stock price (right-hand-side y-axis).

We applied this methodology to generate financial distress probabilities for separate samples of developed and emerging markets stocks for the period 2000 to 2020. For comparison purposes, we also considered other distress risk predictors such as stock return volatility, equity market beta and distance-to-default. For each sub-sample, we formed 20 portfolios based on the ranking of these stock price crash probabilities and calculated their respective returns in the following period.

Figure 3 depicts the performance of the market as well as the portfolios with the highest financial distress probabilities for each of the four distress risk measures. In terms of developed markets, the average market return over this period was 10.0%. Meanwhile, the bottom portfolio for the strategy based on the ML method only delivered 2.3%. This was considerably lower than the gains achieved by the portfolios constructed on the highest distress predictions based on the other conventional variables, which ranged from 4.2% to 5.0%.

We saw an even larger improvement in terms of emerging markets. The average market return was 11.6%, while the portfolios built on the other conventional distress predictors delivered gains between 4.5% to 7.0%. On the other hand, the portfolio based on the ML method generated a loss of 1.5%, thus producing a substantially better result than the other three alternatives.

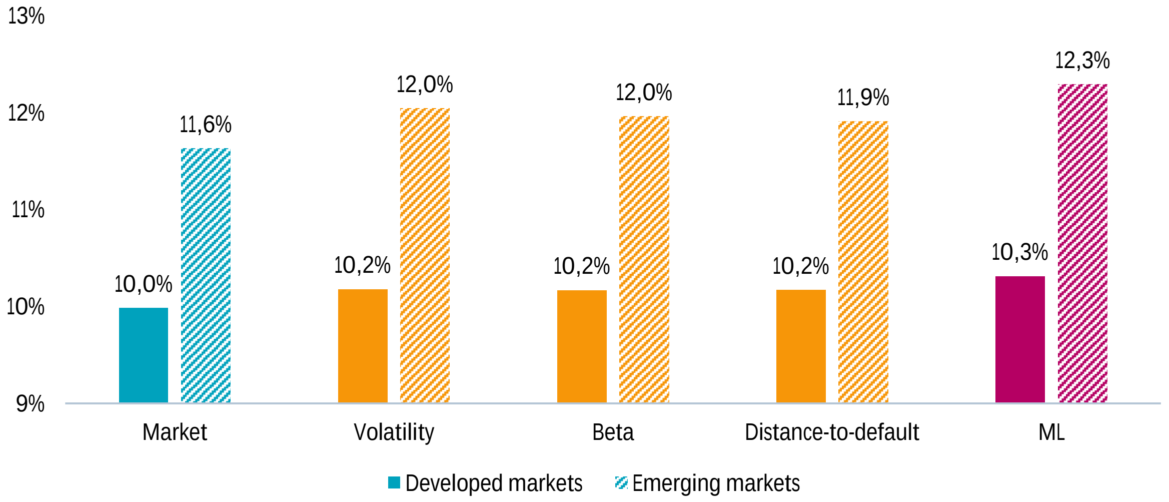
Figure 3 | Performance of developed and emerging markets portfolios with highest distress probability, for the period January 2000 to December 2020



Source: MSCI, S&P and Robeco Quantitative Research. The developed markets universe consists of the largest 3,000 stocks from the S&P Broad Market Index and MSCI World Index. The emerging markets universe consists of the largest 25% constituents of the S&P Emerging Broad Market Index and MSCI Emerging Markets Index. The solid colors represent developed markets and the diagonally striped colors represent emerging markets. The portfolios are sorted based on historical volatility, market beta, distance-to-default and machine learning (plum) distress probabilities. The holding period is 1 month and the returns are annualized in USD.

Figure 4 incorporates the findings from Figure 3 and illustrates the potential gains for an investor using these predictive methods for financial distress. Just by excluding 5% of the stocks with the highest ML-based distress probabilities, while keeping the remaining 95% of the universe, we saw an improvement of 33 and 66 basis points for developed markets and emerging markets respectively over our sample period.

Figure 4 | Performance of the market portfolio excluding 5% of stocks with highest distress probability, for the period January 2000 to December 2020



Source: MSCI, S&P and Robeco Quantitative Research. The developed markets universe consists of the largest 3,000 stocks from the S&P Broad Market Index and MSCI World Index. The emerging markets universe consists of the largest 25% constituents of the S&P Emerging Broad Market Index and MSCI Emerging Markets Index. The solid colors represent developed markets and the diagonally striped colors represent emerging markets. The market returns exclude 5% of the stocks with the highest distress risk probabilities. The portfolios are sorted based on historical volatility, market beta, distance-to-default and machine learning (plum) distress probabilities. The holding period is 1 month and the returns are annualized in USD.

This indicates that advanced ML techniques can potentially help us to identify stocks with elevated distress risk. Put differently, we can potentially enhance the returns of our quantitative equity portfolios if we avoid investing in these stocks.

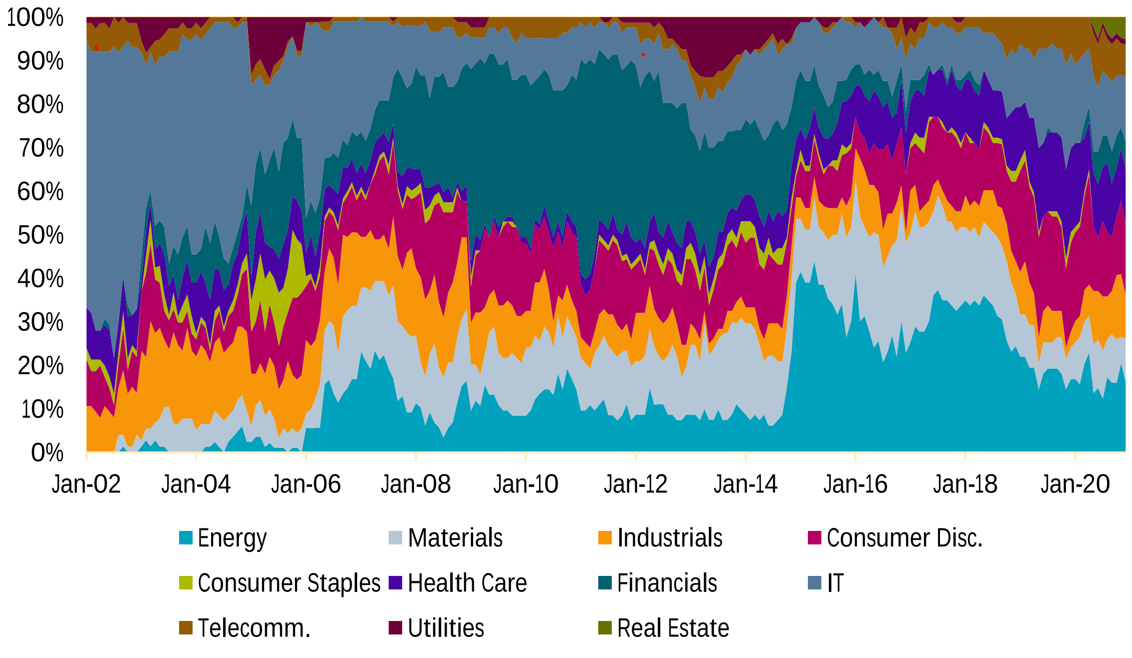
Uncovering how ML helps to forecast financial distress

To get a better understanding of how ML helps to forecast financial distress, we started by examining the sector breakdown of the 5% of stocks with the highest distress probability.

While we would like the ML methodology to automatically recognize the sectors that are likely to experience distress, we do not want it to be dominated by sector selection. In terms of our ML model, we have not placed such restrictions on it as we have not included sector information in its set of features.

Figure 5 shows the sector distribution of this subset of stocks over time. We observed that most of the financially distressed stocks were found in the information technology sector in the early 2000s, then the financials sector during the 2008 to 2010 period, followed by the energy sector during the 2015 to 2019 interval. Although there have been clear tilts towards these distressed sectors over time, we did note that the identified distress events are not concentrated only in a few sectors. Therefore, we believe this makes our ML technique a good candidate for use within stock selection processes.

Figure 5 | Sector distribution of the portfolio with the highest distress risk, January 2002 to December 2020



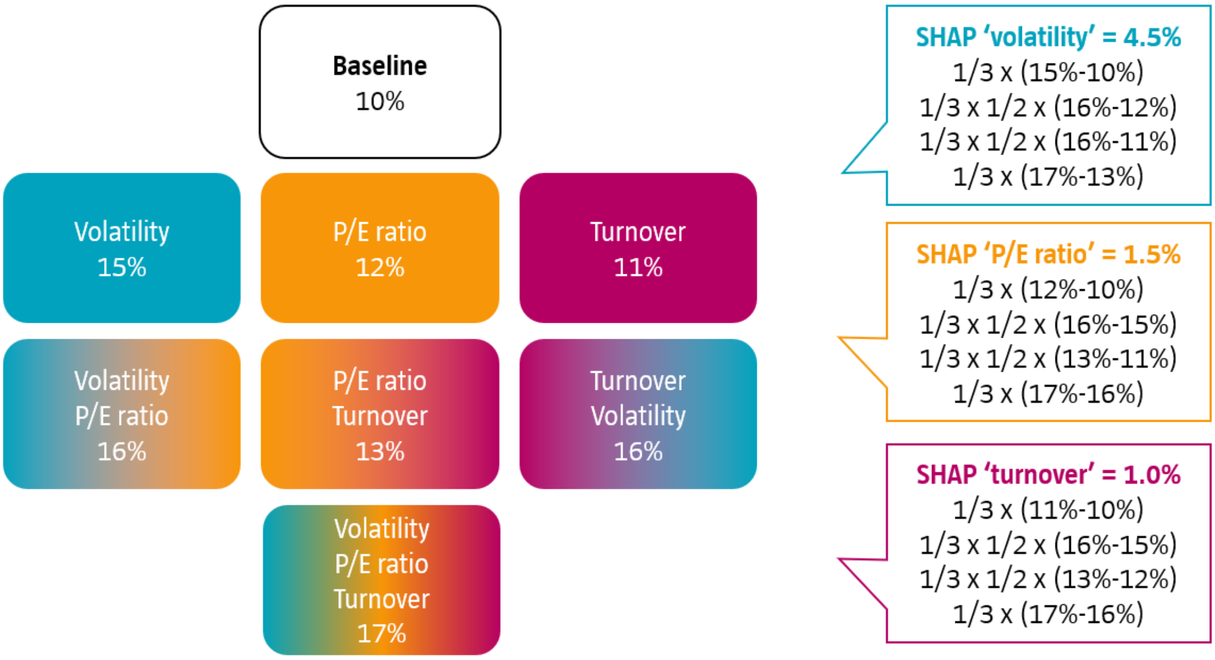
Source: MSCI, S&P and Robeco Quantitative Research. The developed markets universe consists of the largest 3,000 stocks from the S&P Broad Market Index and MSCI World Index. The emerging markets universe consists of the largest 25% constituents of the S&P Emerging Broad Market Index and MSCI Emerging Markets Index. The portfolio is sorted on machine learning distress probabilities and consists of 5% of the stocks with the highest distress risk probabilities.

In our view, it is also crucial to evaluate the importance of each of the selected features, and we did so for a particular year in our sample period. This is indicated by the SHapley Additive exPlanations (SHAP) value.[8] The SHAP value of each individual prediction is the average of the marginal contribution of the feature when predicting the target. This methodology allows us to understand the economic rationale behind our predictions.

In Figure 6, we illustrate this with a stylized example, in which our predictive model consists of only three features: volatility, price/earnings (PE) ratio, and share turnover. Without taking any feature into account, let us say the model would predict the average distress probability of the training sample to be 10%. Then, with the inclusion of only volatility, PE ratio or share turnover information, we would see the distress probability of a stock increase to 15%, 12% or 11% respectively. Furthermore, we show the changes in the predicted probability when two variables are simultaneously included in the second row, as well as when all three features are taken into account in the bottom one.

The average marginal contribution of each feature, or SHAP value, is calculated on the right-hand side. In this example, the stock's volatility has the highest contribution at 4.5%. This consists of the average of the 5% increase when it is added to the baseline, the 4.5% average increase when it is added as a second feature on the next row,[9] and the 4% increase when it is added as a third variable on the bottom one. Similarly, the SHAP value for the PE ratio is 1.5%, while the marginal contribution of turnover is the least important with a SHAP value of 1.0%. Altogether, the SHAP values add up to 7% and constitute the difference between the model's prediction of 17% versus the baseline of 10%. Through this method, we can deconstruct the importance of each feature for each observation.

Figure 6 | Stylized example to calculate SHAP values



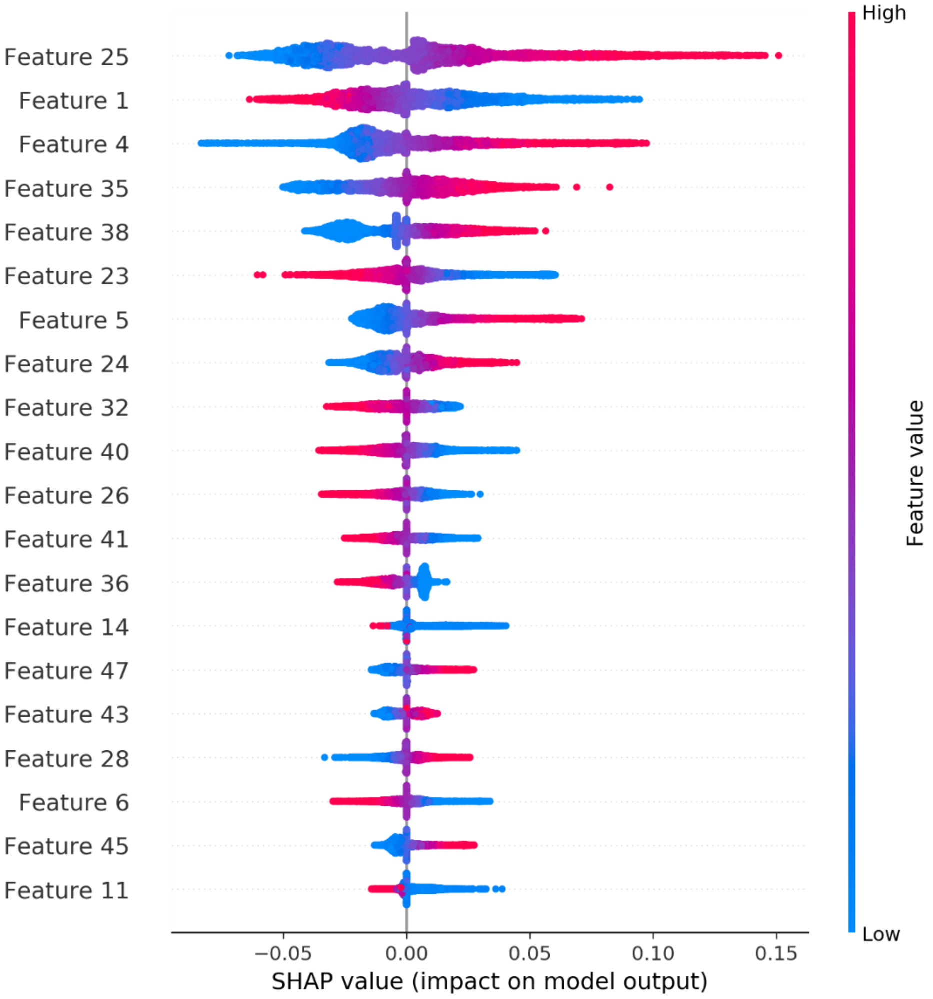
Source: Robeco Quantitative Research.

The importance of the anonymized features in our ML model is shown in Figure 7. The color of each dot indicates the sign and magnitude of a feature, where red signals a high feature value and blue denotes a low feature value. The features are ranked from top to bottom in terms of importance. For example, feature 25 is the most important and this is indicated by it having the largest spread in SHAP values.

[8] This measure was first introduced by Lundberg, S, M., and Lee, S., December 2017, "A unified approach to interpreting model predictions", NIPS 17: Proceedings of the 31st Conference on Neural Information Processing Systems 4768–4777.
[9] Adding it to the PE ratio results in an increase of 4%, while adding it to the turnover results in an increase of 5%, leading to an average of 4.5%.

A high value for feature 25, as indicated by the red dots, leads to a large increase in financial distress probability, as these dots are further to the right than for any other feature. On the other hand, a low value, as highlighted by the blue dots, amounts to a decrease in the probability of financial distress, as the blue dots are found on the negative side of the scale. By contrast, the lowest values of feature 4 are furthest to the left, so these can potentially have a larger negative impact on the probability of distress than the low values of feature 25. The effect is nonlinear for feature 5 as high values have a large positive effect on the probability, while low values only have a small negative effect.

Figure 7 | SHAP values of the 20 most important features for prediction of financial distress



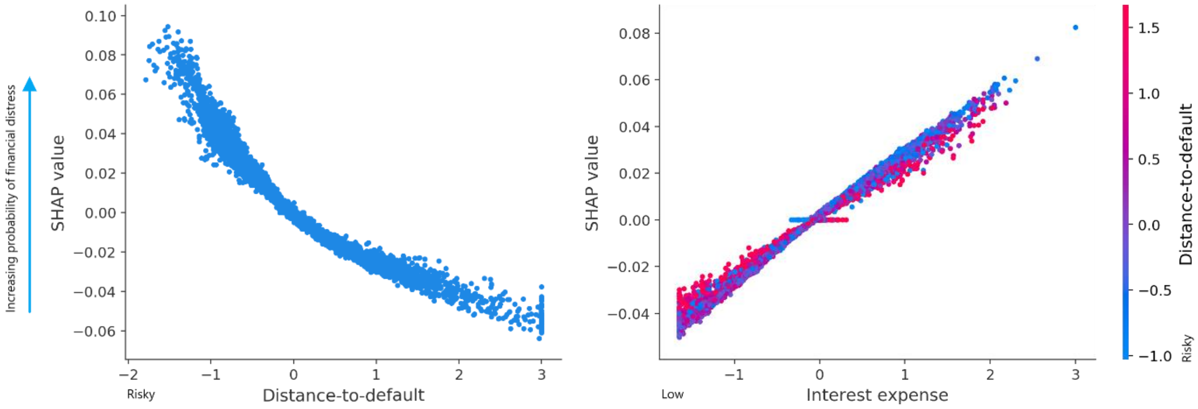
Source: Robeco Quantitative Research. These are SHAP values for a specific year in our sample period. Values to the right indicate increases in the probability of financial distress, while values to the left indicate decreases in the probability. The red dots are high values of the feature, while the blue dots are low values of the feature. The features are ranked in order of importance, from top to bottom. Only the 20 most important features are displayed.

To better understand the nonlinear and interaction effects between features and financial distress probabilities, we looked at partial dependence plots in Figure 8. These plots show the marginal effect of a single feature on predicted distress probabilities.

The left-hand side chart in Figure 8 shows that a lower distance-to-default value (DtD, left on the horizontal axis) increases the distress probability (up on the vertical axis). The impact of DtD on distress probability accelerates for lower DtD , amounting to a nonlinear effect. This is a clear example of why nonlinear models are useful as the entire range of the cross-sectional values is not equally informative.

The distress probability increases faster at lower DtD than for higher DtD. This means that a single unit rise in the DtD barely reduces the distress probability when a stock has a high DtD (i.e., it is safe), but a single unit drop in DtD significantly increases the distress probability if a stock has a low DtD (i.e., it is risky).

Figure 8 | Partial dependence plot of features



Source: Robeco Quantitative Research. The horizontal axis contains the distance-to-default (left on the horizontal axis) and interest expense (right on the horizontal axis) and the SHAP value (vertical axis). The added coloring dimension on the right-hand side refers to the distance-to-default, where high values (safe) are in red and low values (risky) in blue.

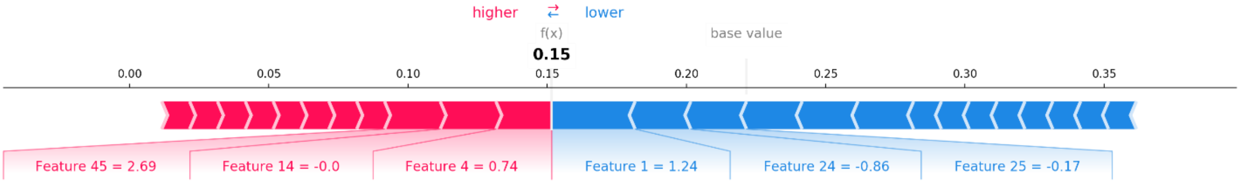
The right-hand side chart in Figure 8 depicts that a high interest expense increases distress probability, but to a lesser extent if the DtD is elevated. The colors of the dots indicate the DtD levels. The distress probability rises by less for high interest paying companies if the DtD is high (safe, in red) in comparison to when the DtD is low (risky, in blue). This interaction effect can be viewed as the difference between safe and risky leverage, which makes sense and provides an economic understanding of the modelling performed by the algorithm.

So far, we have shown the (possibly nonlinear) relationship between features and the probability of financial distress, which helps us to understand the behavior of the predictive model based on its inputs. However, it is also possible to dive deeper into what appeared to be a black box to explain individual distress predictions. Drilling down to the source data enables us to have a clear understanding of why a particular stock is or is not identified as distressed.

For each observation, we can show precisely what features drive up the distress probability and which ones pull it down. This is depicted in Figure 9 for one specific stock in our investment universe. This example takes into account the January 2021 period; the firm's overall probability of default is 0.15.

For each of the relevant features, we illustrate the input value and the effect it has on the estimated probability of distress. The red bars contain the feature values which push up the distress probability, while the blue bars comprise the feature values which pull down the probability. Through the use of these visualizations, we ensure that our ML application is a glass box rather than a black box, and we can drill down to the source data to better understand the distress probability for each stock in our universe.

Figure 9 | Deconstruction of the probability of distress for a specific stock in January 2021 using a force plot.



Source: Robeco Quantitative Research. The red and blue bars indicate the features that push up or pull down the distress probability respectively. For this example, the overall probability of distress is 15%.

As outlined in this section, we believe ML algorithms that are appropriately chosen to forecast stock price crash risk can indeed lead to economically interpretable results. Furthermore, ML tools are available to analyze the core of the predictive model such as feature importance, nonlinearities, and the interaction between features. Additionally, for each individual stock in our investment universe, we can dissect the source data of each feature to better understand the drivers of its distress probability. This can also help us to explain why a specific stock is either held or not held in our quantitative equity portfolios. In our view, the latter gives us more confidence that our ML model is an effective way of identifying stocks that will face financial distress.

The future of ML for stock selection

Several ML applications have already made their way into our stock selection models.

For example, we use natural language processing techniques to interpret large volumes of news to detect sentiment, or to analyze corporate disclosures to determine the contribution companies may make to the United Nations Sustainable Development Goals. This is in addition to the ML technique used to forecast stock price crash risk as described in this article. We believe these tools can help us to better identify companies that will likely underperform in the future. Avoiding investments in such stocks can potentially enhance investment performance.

In our view, new quantitative research techniques – such as ML – challenge conventional wisdom in the field of quantitative investing and can potentially improve investment results. In this note, we have illustrated how ML can help an investor to spot distressed firms prior to distress events (for example bankruptcy filings or credit downgrades) in both developed and emerging markets. We have assessed the added value from ML models given that they can take into account complexities such as nonlinearities and interactions, especially when compared to simpler conventional linear models.

One key strength of ML algorithms is that we can feed them with economically sensible and well-behaved input parameters and analyze in detail whether their choices are logical over time. As we remain cognizant of the risk of overfitting, we make use of a cross-validation framework that we believe leads to a prudent data-driven variation in predictability over time. Moreover, we have the ability to scrutinize the source data which allows us to explain why a specific stock in our portfolio has a particular distress risk probability. In short, our ML-based distress risk prediction application is consistent with Robeco's quant investment philosophy.

The likelihood that future innovative research projects will uncover complex predictive relationships or lead to better interpretation of unstructured data is high. The drawback of financial markets data is that stock prices may be influenced by many known and unknown factors, while relationships between firm characteristics are complex and unstable over time. This makes it difficult to predict stock returns.

Thus, standard statistical methods may not be able to uncover complex patterns, or complex patterns in the data can turn out to be random, and therefore may not be repeated in the future. Our ML application for distress prediction illustrates that the ML toolbox potentially contains methods to reveal true signals of future financial distress, while ignoring the noisy data present in asset prices and firm characteristics.

> 'When one moves from rule-based models to ML-based ones, the role of the researcher changes from instructor to orchestrator'

When one moves from rule-based models to ML-based ones, the role of the researcher changes from instructor to orchestrator. With the traditional approach, the researcher instructs the computer to test specific rules on input data to see whether they can help predict the output. In terms of ML, the researcher feeds both the input and output data to the computer for it to assess what the best rule is. This is referred to as supervised learning in ML terminology.

This role change allows researchers to deal with more complexities. But one needs to be cautious and pay attention to model explainability and overfitting. Only then can the quantitative researcher stay in the driver's seat and orchestrate ML processes, rather than instructing an algorithm on how to behave.

Therefore, we believe ML techniques can propel quant modelling to the next level within the realms of our proven investment philosophy. Indeed, evidence-based research, economic rationale and prudent investing are crucial ingredients to make ML work in the context of investments.

Important Information

Robeco Institutional Asset Important Information

Robeco Institutional Asset Management B.V. (Robeco B.V.) has a license as manager of Undertakings for Collective Investment in Transferable Securities (UCITS) and Alternative Investment Funds (AIFs) ("Fund(s)") from The Netherlands Authority for the Financial Markets in Amsterdam. This document is solely intended for professional investors, defined as investors qualifying as professional clients, who have requested to be treated as professional clients or who are authorized to receive such information under any applicable laws. Robeco B.V and/or its related, affiliated and subsidiary companies, ("Robeco"), will not be liable for any damages arising out of the use of this document. The contents of this document are based upon sources of information believed to be reliable and comes without warranties of any kind. Any opinions, estimates or forecasts may be changed at any time without prior notice and readers are expected to take that into consideration when deciding what weight to apply to the document's contents. This document is intended to be provided to professional investors only for the purpose of imparting market information as interpreted by Robeco. It has not been prepared by Robeco as investment advice or investment research nor should it be interpreted as such and it does not constitute an investment recommendation to buy or sell certain securities or investment products and/or to adopt any investment strategy and/or legal, accounting or tax advice. All rights relating to the information in this document are and will remain the property of Robeco. This material may not be copied or used with the public. No part of this document may be reproduced, or published in any form or by any means without Robeco's prior written permission. Investment involves risks. Before investing, please note the initial capital is not guaranteed. This document is not directed to, nor intended for distribution to or use by any person or entity who is a citizen or resident of or located in any locality, state, country or other jurisdiction where such distribution, document, availability or use would be contrary to law or regulation or which would subject Robeco B.V. or its affiliates to any registration or licensing requirement within such jurisdiction.

Additional Information for US investors

This document may be distributed in the US by Robeco Institutional Asset Management US, Inc. ("Robeco US"), an investment adviser registered with the US Securities and Exchange Commission (SEC). Such registration should not be interpreted as an endorsement or approval of Robeco US by the SEC. Robeco B.V. is considered "participating affiliated" and some of their employees are "associated persons" of Robeco US as per relevant SEC no-action guidance. Employees identified as associated persons of Robeco US perform activities directly or indirectly related to the investment advisory services provided by Robeco US. In those situation these individuals are deemed to be acting on behalf of Robeco US. SEC regulations are applicable only to clients, prospects and investors of Robeco US. Robeco US is wholly owned subsidiary of ORIX Corporation Europe N.V. ("ORIX"), a Dutch Investment Management Firm located in Rotterdam, the Netherlands. Robeco US is located at 230 Park Avenue, 33rd floor, New York, NY 10169.

Additional Information for investors with residence or seat in Canada

No securities commission or similar authority in Canada has reviewed or in any way passed upon this document or the merits of the securities described herein, and any representation to the contrary is an offence. Robeco Institutional Asset Management B.V. is relying on the international dealer and international adviser exemption in Quebec and has appointed McCarthy Tétrault LLP as its agent for service in Quebec.