TI 2019-066/VI
Tinbergen Institute Discussion Paper

Data Science in Strategy: Machine learning and text analysis in the study of firm growth

Daan Kolkman[1]
Arjen van Witteloostuijn[2]

[1] Technical University Eindhoven
[2] Vrije Universiteit Amsterdam

Data Science in Strategy

Machine learning and text analysis in the study of firm growth

Daan Kolkman

(Jheronimus Academy of Data Science, the Netherlands)

&

Arjen van Witteloostuijn

(VU Amsterdam, the Netherlands & University of Antwerp / Antwerp Management School, Belgium)

September 2019

Acknowledgements

The data used in this study were provided by Graydon the Netherlands and Graydon Belgium. Graydon is not responsible for any of our conclusions. We are grateful to Bas Bosma, Joeri De Caigny, Wim Coreynen, Pourya Darnihamedani, Marcus Dejardin, Barry Delhez, Julie Hermans, Maurits Kaptein, Simon Noyons, Wouter Stam, Eric Vandenbroele, Diemo Urbig, Johanna Vanderstraeten, Jack van Wijk, and Mark Zwart for their comments and suggestions.

Data Science in Strategy

Machine learning and text analysis in the study of firm growth

Abstract

This study examines the applicability of modern Data Science techniques in the domain of Strategy. We apply novel techniques from the field of machine learning and text analysis. We proceed in two steps. First, we compare different machine learning techniques to traditional regression methods in terms of their goodness-of-fit, using a dataset with 168,055 firms, only including basic demographic and financial information. The novel methods fare to three to four times better, with the random forest technique achieving the best goodness-of-fit. Second, based on 8,163 informative websites of Dutch SMEs, we construct four additional proxies for personality and strategy variables. Including our four text-analyzed variables adds about 2.5 per cent to the R^2. Together, our pair of contributions provide evidence for the large potential of applying modern Data Science techniques in Strategy research. We reflect on the potential contribution of modern Data Science techniques from the perspective of the common critique that machine learning offers increased predictive accuracy at the expense of explanatory insight. Particularly, we will argue and illustrate why and how machine learning can be a productive element in the abductive theory-building cycle.

INTRODUCTION

The central argument of this paper is that Strategy research, illustrated for the case of firm growth models, can be improved upon by employing the novel Data Science techniques that are commonplace in the field of machine learning (or artificial intelligence/AI) and text analysis (cf. Wenzel & Van Quaquebeke, 2018; Kobayashi et al., 2018). Specifically, we apply novel techniques from the field of machine learning and text analysis to explore their potential to improve the performance of SME growth models. The low goodness-of-fit of existing firm growth models stands in sharp contrast with that of state-of-the-art machine learning models that can be used to, for instance, classify images or transfer speech to text. We build on the so-called "Data Science revolution" (Chen et al., 2012; McAfee & Brynjolfsson, 2012) to improve firm growth models' goodness-of-fit.

Specifically, we first compare four machine learning techniques – support vector machines, stochastic gradient-boosted trees, random forest analysis, and multi-layer neural networks – to six traditional parametric regression models (from OLS to elastic net regression) in terms of their goodness-of-fit on a dataset of 168,055 SMEs from Belgium and the Netherlands (cf. van Witteloostuijn & Kolkman, 2018, which only does so for two classic regression methods and random forest analysis). For each of these firms, we have one to six years of historical data. The available data consist primarily of basic demographic and financial information. To avoid overfitting – and to evaluate the performance of the models – we split the data in a training and a validation set. Second, we add text-analyzed scraped data. Based on text scraped from 8,163 informative websites of Dutch SMEs, we construct four additional proxies for personality and strategy variables, illustrating how measures based on text-analyzed scraped data can further improve the explanatory and predictive power of extant firm growth models.

Subsequently, we reflect on a few methodological features commonly associated with machine learning that seem to contrast sharply with standard practices and widespread beliefs in the Strategy domain, suggesting how machine learning can still be a valuable addition to Strategy's empirical toolkit. Key is that machine learning offers increased predictive accuracy at the cost of explanatory insight. That is, econometric efficiency is improved (oftentimes, impressively so) by running a machine learning algorithm that is a black box producing output from input without any insight into throughput. Moreover, machine learning is essentially a non-parametric (or semi-parametric) method, without producing the β-coefficients and p-values the scholarly Strategy community is so used to, and hence is argued to be silent about economic and statistical significance. We offer a threefold response to this widespread critique. First, we argue that, like in many other disciplines such as the Life Sciences and climate studies, prediction deserves a respectful place next to explanation in the Strategy field. Second, within the computer science of artificial intelligence, work is done to produce machine learning output that does offer explanatory insight.

Third, machine learning offers a powerful data-mining tool, being a "quantitative" method of induction, that can be perfectly combined with standard deductive techniques. In so doing, we introduce powerful theory-building and testing practices that fit within the tradition of abduction (cf. Fiss, 2011; Mysangyi & Acharya, 2014). We will extensively argue and illustrate why and how machine learning can be a productive additional element in the abductive theory-building and testing cycle. Specifically, working with our unbalanced panel of 8,163 Dutch SMEs with informative firm websites, we will develop hypotheses suggested by machine learning outcomes (the inductive leg of abduction) that are subsequently tested with traditional multivariate regression analysis (the deductive leg of abduction). In the process of doing so, we add to the Behavioral

Strategy literature (Powell et al., 2011) by focusing on the role of the personality of egocentrism in explaining SME growth.

Next, we very briefly discuss the current literature on machine learning. After that, we provide descriptive statistics of our dataset, and outline our methods. Subsequently, we compare the performance of six traditionally construed linear regression models with that of a support vector machine, stochastic gradient-boosted tree, a random forest analysis (RFA), and a multi-layer neural network. The traditional linear regression models (TLR) serve a baseline against which to evaluate the performance of the other four models regarding the explanation or prediction of firm growth rates. We conclude that RFA outperforms TLR in terms of goodness-of-fit by a ratio of three or four to one. Subsequently, we explore the potential of website scraping and text analysis techniques, revealing that adding four simple text-analyzed proxies for personality and strategy variables does add substantive extra explanatory and predictive power to our basic firm growth model. In a three-step intermezzo, we reflect upon a few critical methodological issues commonly associated with machine learning more generally, which we illustrate for the case of RFA in particular. Finally, we present the implications of our findings for firm growth research and the broader context of Strategy studies.

MACHINE LEARNING

As the volume, velocity, veracity, and variety of the data that society collects are increasing rapidly, the analysis of such so-called Big Data transcends the cognitive capability of people (Kitchin, 2014; Mayer- Schonberger & Cukier, 2013). Consequentially, there is a considerable and growing reliance on algorithms to structure, analyze, and model data. Although there is no consensus on the correct terminology for data-centric technology, the “machine learning” label broadly refers to algorithms that optimize model performance criteria – such as the traditional and

well-established R^2 – by evaluating generated (or "predicted") output against observed (or "true") data. Machine learning, often used interchangeably with "artificial intelligence", is particularly helpful in cases where we do not have the knowledge required to formulate the rules of a target system or where such knowledge is tacit and cannot be readily transferred (Smola & Vishwanathan, 2014).

Alpaydin (2014) offers the example of spoken speech to illustrate that people do some tasks, such as converting acoustic speech signals into written words, but cannot explain how they do this. Machine learning approaches this problem by collecting a large dataset of audio recordings and texts, and feeding this into an algorithm. The algorithm may not necessarily transform the spoken speech to text in the way similar to what people do, but it can learn to do this transformation to produce output that makes sense to people. Machine learning algorithms can, in this fashion, construct a useful approximation that accounts for a surprisingly large part of the input data. A potential downside of machine learning techniques is that the process of transforming input into output, and hence the target system's underlying causal mechanics, can be rather incomprehensible. This is a key area where, from a scholarly perspective, the research community is working hard to produce progress (we extensively return to this issue in our three-step intermezzo).

The origins of machine learning are contested, but most accounts suggest that it developed from the field of Computer Science. While new machine learning algorithms are launched at increasing speed, the origins of the wider classes of algorithms that are used date back several decades. Methods such as regression trees, neural networks, and support vectors have been around for quite for some time now, and will be familiar to the reader (Bishop, 2016). We prefer the term machine learning over data-mining, because the latter also includes descriptive statistics (Kotu & Deshapnde, 2014). Machine learning moves beyond mere description. Moreover, data-mining is sometimes used to refer to algorithms that identify "important relationships and correlations"

amongst large piles of data (Nilsson, 2005) without including problems where a target (or, in traditional terminology, dependent) variable is available. Machine learning algorithms can handle Big Data, and outperform more traditional forms of modeling in a number of domains. Typically, machine learning algorithms have the capacity to uncover non-obvious patterns in data, and facilitate reliable and accurate explanations and / or predictions.

TEN TECHNIQUES

We selected ten models on the basis of their predominance in the Strategy literature or their track record in the field of machine learning: ordinary least squares (OLS), forward stepwise regression (FSR), least absolute shrinkage and selection operator, ridge regression, least-angle regression, support vector machines, elastic net regression, random forest analysis, stochastic gradient-boosted trees, and multi-layer neural networks. The OLS and FSR models are included as our benchmarks, being standard approaches in Strategy research. The other methods are included because of their common application in the Data Science literature. In the current paper, we lack space for a detailed and extensive introduction of all these techniques (for that, see the references). However, we briefly provide the key intuition below. In subsequent sections, by way of illustration, we provide greater detail about the random forest analysis.

1. *Ordinary least squares* (OLS). The fundamental idea of the standard parametric linear (multivariate) regression model is that a target (dependent) variable can be explained or predicted from a collection of (control and independent) variables that are multiplied by parameters, producing β-coefficients (and hence effect sizes) and standard deviations (and hence *p*-values). Optionally, a constant can be added to the model. Linear models can be estimated by using OLS to find the optimal configuration of parameters (including product

terms, known as moderators, and squared variables, to estimate non-linear relationships) and a constant. This type of model is very well known in academia, being its major empirical workhorse, and is added here as the traditional yardstick for cross-method comparison. OLS has no built-in procedure to deal with multicollinearity, the latter increasing the confidence intervals for the parameters, making identification of the unique contribution of variables difficult.

2. *Forward stepwise regression* (FSR). FSR is a procedure for the selection of independent variables in OLS regression. It proceeds by adding independent variables to the model one at a time. At each step, the independent variables not in the model are tested for inclusion in the model. The most significant of these variables is then added to the model, as long as its *p*-value is below some pre-set level (Weisberg, 1980). This threshold is typically set at 0.05. The textbook approach to FSR requires assessment of the level of multicollinearity. Below, for this model, we removed independent variables with an in-time Pearson's correlation higher than 0.8 or lower than -0.8, retaining the first of the two variables with a correlation exceeding this threshold (Field, 2013).
3. *Ridge regression* (RR). RR is variant of OLS that deals with the problem of multicollinearity among independent variables. More specifically, RR applies a form of regularization by adding an additional error term to the model. This error term is typically specified as the L2-norm of the predictor variables, and penalizes the magnitude of the independent variables. RR aims to minimize the impact of irrelevant independent variables on the model (Hoerl & Kennard, 1970).
4. *Least absolute shrinkage and selection operator* (LASSO). Lasso regression is an extension of RR by not just minimizing the impact of irrelevant independent variables, but by setting them to zero. In effect, LASSO automates the selection of relevant independent variables. In

contrast to RR, LASSO employs the L1-norm to penalize independent variables with a large magnitude (Tibshirani, 1996).

5. *Least-angle regression* (LARS). LARS is a stylized version of the FSR regression procedure. It starts with all parameters equal to zero and finds the independent variable that is most correlated with the dependent variable. Below, we take the largest step possible in the direction of this predictor until some other predictor has as much correlation with the current residuals. Instead of continuing along the first independent variable, LARS proceeds in a direction equiangular between the two independent variables until a third variable earns its way into the "most correlated" set (Efron et al., 2004).
6. *Elastic Net Regression* (ENR). Though LARS offers some clear advantages over FSR from a prediction standpoint, a key limitation that LASSO and LARS share with FSR is that they do not handle multicollinearity as well as RR (Tibshirani, 1996). ENR can be viewed as a next-generation extension and optimal combination of the LASSO and RR techniques (Zou & Hastie, 2005). Specifically, ENR provides the benefits afforded by RR in terms of handling multicollinearity of predictors, but also incorporates the slow growth and variable selection features of LASSO and LARS.
7. *Support Vector Machines* (SVM). Of all the modern techniques addressed here, SVM is by far the most abstract and least straightforward to understand—in part because SVM has no clear analogue to simple regression. Nevertheless, SVM, along with stochastic gradient-boosted trees and random forest analysis, is still amongst the most frequently used "out-of-the-box" predictive modeling techniques. SVM comes in many varieties. Our focus here is on SVM for application to regression (Hearst et al., 1998).
8. *Random Forest Analysis* (RFA). RFA is a machine learning algorithm that can be used for regression and classification. An RFA involves a combination of multiple decision trees that

are trained on different sub-sets of the data (Breiman, 2001). This approach of combining different models to improve performance is also referred to as “ensembles”. In addition, using different sub-sets of the data, RFA determines the splits of the constituent decision trees by considering a random sub-set of predictor variables. Final predictions are acquired by aggregating across the constituent decision trees. This prevents the model from overfitting the data. RFA can detect non-linear and high-order interactions between determinants.

9. *Stochastic gradient-boosted trees (SGT).* One potential drawback of RFA is that the trees constructed in each bootstrap sample are built independently of one another. That is, there is no attempt to identify an ensemble of trees that would complement each other well for the purposes of predicting the outcome. This is in contrast with a procedure such as FSR or LARS, where the model is developed incrementally, each step aimed at providing an improvement to the model, given the steps that came before. SGT attempts to address this limitation of RFA (Duchi et al., 2011). The gradient-boosting algorithm will start by using the difference between the outcome value for each individual observation and the mean outcome value across all observations in the development sample as “residuals” to be predicted (i.e., effectively, mean-centered outcome values). The algorithm will then fit a fairly simple regression tree to predict the set of residuals. Unlike bagged trees or random forests, the depth of trees in SGT is typically kept very small (Friedman, 2002).
10. *Multi-layer neural network (MNN).* MNN is loosely based on the way biological nervous systems function. It is composed of many interconnected data-processing elements (neurons) that transform the input data. These neurons often employ non-linear functions (Bishop, 2016). The parameters of the neurons (also referred to as “weights”) can be estimated by using an algorithm such as back propagation (Rummelhart et al., 1986). Many

types of neural networks exist, and the architecture of neural networks constitutes a field of inquiry in itself (Schmidhuber, 2015). One of the most successful neural networks is the multi-layer perceptron, which is also known as a feed-forward neural network (Bishop, 2016). Here, "feed-forward" refers to the absence of cycles or memory neurons in the network architecture (Witten et al., 2011).

DATA AND METHODS

Descriptive analysis

In this study, we start from data from Graydon in Belgium and the Netherlands. Graydon is a provider of business credit information. The data was not collected specifically for this research project, but as part of Graydon's regular business operations. Our initial dataset contains 2,494,784 records that are each associated with one firm in a sample of 533,626 unique SMEs. First, after removing outliers and firms with missing data, we run analyses with 168,055 firms, but without text-analyzed variables (for greater detail, we refer to van Witteloostuijn & Kolkman, 2018). Next, we work with data from 8,163 Dutch SMEs, as only for this subset we could scrape informative information from websites. Given data availability, we operationalized growth rate as an index of total assets growth:

$$Growth_i = \left(\frac{A_{it}}{A_{it-1}}\right) \cdot 100,$$

where A_{it} is the total assets in Euro's of firm i at time t. For records with missing A values, the growth rate was computed based on the number of employees. Figure 1 provides the distribution of firm growth rates in our sample.

[Insert Figure 1 about here]

The Graydon dataset includes financial information (balance sheet total, total equity, working capital current ratio, and solvency ratio), number of employees, legal person, and sector. In our analyses, we add contemporaneous and one-year lagged variables of the financial information and the number of employees, as well as a country dummy. An overview of the correlations between firm growth rates and these variables is displayed in Figure 2. All correlations with growth rates are weak ($r < 0.2$), with most values below 0.01. The number of employees ($r = –0.09$) and age of the firm ($r = -0.06$) have the strongest correlations with firm growth rates.

[Insert Figure 2 about here]

The firms in our sample were labeled according to their industry in line with the EU regulation 'Nomenclature générale des activités économiques dans les Communautés Européennes' or NACE. NACE is a hierarchical classification that consists of four levels: sections, divisions, groups, and classes. In our analysis, we used the second NACE level, or divisions. The largest sectors, in terms of number of companies, in the sample are financial holdings, wholesale, real estate, retail, and construction. The mean growth rate across industries is 8%. The "remediation" sector has a mean growth rate of 30%, which is the maximum in the sample. The lowest growth rate of -25% was found in the "Mining of coal and lignite" sector.

[Insert Tables 1 and 2 about here]

Graydon's base dataset has 31 variables, including two company identification numbers, the company name, start date, postal code, address, and several other "demographic" variables. After sifting through the variables, we ended up with a set of 16 "core" potential predictors in the form of demographic and financial measures, including lagged measures (cf. van Witteloostuijn & Kolkman, 2018). That is, to expand the number of predictor variables, and in line with prior work, we added a one-year time lag for balance sheet total (or total assets), total equity, working capital, current ratio, solvency ratio, and number of employees. The legal person and sector of a firm are

categorical variables, which we transformed into a series of binary dummies. This resulted in a total of 113 independent variables, which make up our base dataset. An overview of the variables can be found in Table 3.

[Insert Table 3 about here]

Estimation procedure

To evaluate and compare the performance of the ten techniques, we use three different datasets. *Dataset (1)* is the base dataset that includes 113 variables. *Dataset (2)* is a dataset that has all base variables, plus quadratic transformations of those variables and first-order interaction product terms. *Dataset (3)* is a dataset with variables selected according to FSR. This results in the following set of eleven base predictors for the regression models: balance sheet total (*t*), balance sheet total (*t*-1), total equity (*t*), total equity (*t*-1), working capital (*t*), working capital (*t*-1), current ratio (*t*), current ratio (*t*-1), solvency ratio (*t*), solvency ratio (*t*-1), and number of employees (*t*). This setup permits us to evaluate whether differences in model performance originate from the variable selection procedure or can be explained by the more flexible specification strategy of machine learning techniques, which can model higher-order interaction effects.

We base our model estimation procedure on that outlined by Putka et al. (2018), with five specific estimation choices. First, we standardize Datasets (1), (2) and (3) by dividing the values by the L2-norm for that variable, using “least absolute deviations”. Although not strictly necessary for linear regression, such normalization is recommended for training neural networks (LeCun et al., 2013). To prevent the normalization from becoming a factor in our comparative analysis, we used the normalized data as input for all the models we estimated (but see below for an exception). Second, we randomly split all three datasets in a training (80% of the data) and a validation (20% of the data) set. Whenever there is a large set of possible relationships, one has to be careful not to use

the resulting freedom to find meaningless patterns in the data. This problem is called overfitting. It is a very general phenomenon, and occurs even when the target function is not at all random. It afflicts every kind of learning algorithm. A typical approach to identify overfitting is to evaluate a model against a set of data that was not used to train the algorithm. This unseen dataset is also known as a hold-out set, validation or test set, and is also used to get a sense of how well a model will generalize (Chico, 2017).

Third, we determine the hyper-parameters by *k*-fold cross-validation on the training set of Dataset (1): *k*-fold cross validation is a bootstrapping technique that draws random samples from the training set with replacement (Refaeilzadeh et al., 2009). To ensure robustness of our results, we employed 30-fold cross-validation. For models with less than 100 degrees of freedom with respect to the hyper-parameters, the maximum degree of freedom is used. Specifically, we first set an input list of hyper-parameters. Next, we conduct a so-called random search to identify hyper-parameters with the highest R^2. We use random search as opposed to grid search because the former finds "better models in most cases and require[s] less computational time" (Bergstra & Bengio, 2012: p. 302). As such, random search allows to explore a larger phase space of hyper-parameters. We set the number of randomly chosen combinations of hyper-parameters to 100. With this setting, random search finds a solution within 5% of the optimal sollution 99% of the time. Subsequently, we select hyper-parameters following the 1-SE rule to "choose the simplest model whose accuracy is comparable with the best model" (Krstajic et al., 2014: p. 11). The selection of the "simplest" set of hyper-parameters depends on model type, and is somewhat arbitrary. For instance, for an RFA, we selected the model with the least trees and highest number of values per leaf. The resulting hyper-parameters per technique are listed in Table 4.

[Insert Table 4 about here]

Fourth, we fit the models with their respective hyper-parameters to the full training set of Dataset (1) , (2) and (3), again using 30-fold cross-validation. Fifth, we produce predicted firm growth rates for each of the models, and evaluate the R^2 (and other fit statistics; see below) of these models on the training set and the validation set for Datasets (1), (2) and (3). All analyses were conducted in Python (3.5.4) using the scikit-learn (0.19.1) and Keras (2.1.2) packages. The MNN employed in this study consists of 3 hidden layers with 48 neurons with a rectifier activation function (see Maas et al., 2013). Below, we will specifically focus on RFA (in combination with text analysis and in the intermezzo) to illustrate what machine learning has to offer to the scholarly Strategy community. We do so for two reasons: first, RFA comes with user-friendly output that adds much in terms of explanatory insight; and second, as we will see below, RFA (by far) outperforms the nine alternative techniques.

TEN TECHNIQUES COMPARED

Goodness-of-fit performance

In van Witteloostuijn and Kolkman (2008), we compared the goodness-of-fit performance of three models (OLS, FSR and RFA). In the current paper, we move beyond that by fitting ten models on firm growth rates to determine which one performs best. Following the recommendations outlined by Legates and McCabe (1999) in their evaluation of goodness-of-fit measures, we consider the performance of our models on four criteria. The first is the coefficient of determination or R^2, which is the classic measure for how well the explanations or predictions approximate the observed firm growth rates. The second is the Mean Absolute Error (MAE), which is the absolute difference between the firm growth rate explanations or predictions and the observed firm growth rates. The third – Mean Squared Error (MSE) – is a similar statistic, measuring the average squared difference between the explained or predicted and actual or observed values. The fourth is the Root

Squared Mean Error (RMSE), being the square root of the average of squared errors. In Table 5, we report the four fit statistics for our ten models.

[Insert Table 5 about here]

Clearly, the RFA performs best across all test statistics on Dataset (1). For instance, the R^2 of 0.23 (training set) or 0.16 (validation set) is impressive vis-à-vis the meagre R^2 of 0.05 or 0.06 for the more traditional regression models. The random forest also outperforms stochastic gradient-boosted trees and the multi-layer neural network in both the training and validation sets. Importantly, the performance of more traditional regression models are almost identical (with an R^2 of 0.05 versus 0.06), implying that the much more flexible specification strategy employed by machine learning techniques cannot explain the underperformance of our parametric linear (multivariate) regression benchmarks. The results on Dataset (2) confirm this. When we fit the models on Dataset (2), the performance of the RFA drops somewhat, but remains vastly superior. In Dataset (3), where we included first-order interaction effects and quadratic terms, the traditional methods start to catch up a little. This suggests that the superior performance of machine learning techniques originates from their flexible capacity to search for fit-enhancing higher-order interaction effects.

With an R^2 of 0.16 / 0.23, the RFA outperforms most top-fitting models in the traditional firm growth literature, even with our limited set of demographic and financial explanatory variables – very impressive indeed. The loss in R^2 from the training to the test or validation set suggests slight overfitting of the data. The scatterplots (of actual versus explained or predicted firm growth) of the validation sets of the FSR vis-à-vis the RFA can be found in Figures 3a and b, respectively. It is evident from the figures that neither model performs perfectly well, which is not surprising, giving our limited set of demographic and financial explanatory variables. The FSR model predicts firm growth rates within the -50 to 50 range, and thus fails to accommodate actual firm growth rates

outside that range. The RFA fares substantially better, with the figure revealing a split at a firm growth rate of 0.

[Insert Figures 3a and b about here]

Note that when fitted on a non-normalized dataset, the R^2 of the random forest on the training set decreased to 0.18, but the R^2 on the validation set increased to 0.17. In the following, we use the RFA fitted on this non-normalized dataset to provide more details on the model's mechanics. We do so partly because the R^2 of this model in the validation set is higher, but also because normalized data can be hard to interpret.

RFA output

The relative importance of the explanatory variables in the random forest model can be measured using the mean decrease in accuracy, or the percentage increase in the MSE. This measure corresponds to the difference between the MSE for including and excluding that variable, averaged over all the trees and divided by the standard deviation of the differences. Machine learning's output gives so-called "feature importance", which indicates the relative weight of each of the listed variables – or "features" in machine learning terminology – in explaining or predicting the "target variable". The five most important variables for the random forest in descending order are: Total assets, working capital, total equity, current ratio, and solvency ratio, all in *t*. The complete list is provided in Figure 4a.

[Insert Figure 4 about here]

The RFA reveals that all the contemporaneous financial variables (as a group) are more important than all the demographic (ranked in-between) or lagged measures, the latter ranking – as a cluster – located at the bottom. However, as the nature of Figure 4's list makes clear, machine

learning is not a parametric method. The standard RFA output does offer insight into the relative importance of all explanatory variables, and it does provide a statistic for the increase in fitness due to adding a specific explanatory variable, but this is different from the familiar β-coefficients and p-values produced in traditional parametric techniques in econometrics. So, the substantially higher overall fitness, an R^2 of ~ 0.05 for the more traditional regression methods vis-à-vis ~ 0.17 for the RFA, is traded off against lack of insight in parametric effect sizes and significance values. Below, we return to this issue in our three-step intermezzo.

TEXT ANALYSIS

Recent advances in machine learning allow for the analysis of massive piles of textual data through natural language processing algorithms. Such algorithms go beyond word counting that have previously been applied in the context of firm growth (see Butscheler et al., 2018), and permit in-depth analysis of the meaning of texts. We investigated the potential of this new technique after scraping the websites of the firms in our dataset. Of the 168,055 firms in the dataset, substantive websites were available for a modest sub-set of 8,163 Dutch SMEs. Note that we removed all French-language SMEs to avoid any translation issues. For many SMEs, we could not find websites at all, or the information on the scraped website was uninformative (e.g., only contact details, and product lists without any explanation). We ran a Latent Dirichlet Allocation (LDA; see Blei et al., 2003) on the textual data for these 8,163 informative websites. LDA is an example of a topic model, which is a class of techniques that identifies groups of words that represent a shared topic across a corpus of texts (Shu et al., 2009). An example of an application in the Business and Management field is Kaplan and Vakili (2014). We first ran an inductive LDA and identified 100 topics – too many to construct a meaningful and workable set of variables. More importantly, none of 100 topics improved goodness-of-fit when added to the models. Hence, we decided to turn to a deductive

LDA, inputting theoretically identified strings of keywords per to-be-measured construct or "custom topic".

For our deductive LDA, by way of illustration, we defined four custom topics: Three based on the competitive strategies of Treacy and Wiersema (1997), and one as a measure of the personality trait of egocentricity (or narcissism; see below). Table 6 provides an overview of the four custom topics, and the word list we used, after careful reading of a sub-sample of the websites, to construct our text-analyzed measures. If the Dutch word is too distant from its English counterpart to reveal its meaning to readers not mastering Dutch, we added a translation in English.

[Insert Table 6 about here]

We tested the added value of these text-based explanatory variables by including this set of four text-analyzed proxies in the RFA model, and re-running the random forest for the sub-sample of 8,163 unique firms. This gives an impressive R^2 of .63. The addition of this set of four variables improves the R^2 of the random forest analysis by .15, which is substantial in the context of Strategy studies (and the Business and Management field broadly, for that matter). The ranking of the independent variables' importance, provided in Figure 4b, reveals that the four text-analyzed measures displaced the time-lagged financial variables as the second most influential group of measures, after the contemporaneous financial variables. This warrants further investigation of text-based variables, using machine learning text analysis algorithms, in future Strategy studies.

MACHINE LEARNING AND THEORY DEVELOPMENT

Prediction and understanding

Machine learning is widespread in many scientific disciplines, being a standard set of methodologies in their research toolkits. For instance, machine learning is applied in Life Sciences to link complex genetic patterns to specific disease symptoms, and in climate studies to fit complex

data to predict climate change outcomes (see, e.g., Boulesteix et al., 2012; Grömping, 2009; Molinaro et al., 2011; Papagiannopoulou et al., 2017). Interestingly, the use of machine learning is much less widespread in the Social Sciences, including Business and Management, with a few exceptions (such as, e.g., Marketing; cf. Burez & van den Poel, 2007). An important reason for the underutilization of machine learning in many sub-disciplines of Business and Management, including Strategy, is probably that machine learning (or artificial intelligence, more broadly) is thought to be associated with prediction without explanation. Many of the Social Sciences are preoccupied with explanation, and not with prediction, the latter allegedly coming with a dislike of predictive black box tools that offer a high R^2 without insight (but see below). In response, we would like to offer a threefold counterargument. This section briefly introduces the first part of the argument. The other two parts are illustrated in the context of random forest analyses in subsequent sections.

The first is that there is no a priori reason not to be interested in high predictive accuracy in the Social Sciences, including Strategy. We argue that, like in many other disciplines such as the Life Sciences and climate studies, prediction deserves a respectful place next to explanation in the Strategy field. In the context of the debate regarding relevance versus rigor, spanning many decades (e.g., Gulati, 2007), investing in predictive accuracy is instrumental in building stronger bridges to Strategy practice. This relates to the recent upsurge of the grassroots movement that aims to promote Responsible Research in Business & Management, or RRBM (https://www.rrbm.network), as active in the Academy of Management (Tsui, 2013). Take the current paper's example of SME growth. Models with an R^2 of 0.15, maximum, are not very helpful for banks and governments that have to decide in which small enterprises to invest. For that, they need tools with high predictive accuracy. If the field is able to produce better and better predictive models, based on machine learning, Strategy studies will engage in solution-oriented Social Science (Watts, 2017). In the

context of Political Science, Muchlinski et al. (2015: p. 3) nicely summarize this essential argument: “Often, a researcher is more interested in the ‘causes of effects’ than in the ‘effects of causes’ (Gelman and Imbens 2013). Public policy considerations may outweigh the value of basic science in a particular domain. Rigorous causal identification may be infeasible for practical or ethical reasons. Large, multidimensional datasets, coupled with theoretical underdevelopment, may undermine the credibility of causal modeling assumptions. For any of these reasons, or more likely a combination of all three, it may be useful to embrace prediction as the explicit goal of research, rather than solely as a criterion of evaluation for causal models.”

Explanation in RFA

The second argument is that within the Computer Science of artificial intelligence, work is done to produce machine learning output that does offer explanatory insight. To substantiate this claim that artificial intelligence has more to offer in terms of explanation than only an input-output black box, we provide illustrative detail for the random forest case. Given space limitations, we cannot but briefly introduce the key intuition behind the output of modern RFA software packages. For insightful introductions, we refer to Boulesteix et al. (2012) and Loh (2011). Here, we focus on decision trees, feature importance lists (including weights and signs), and essential interaction identification, illustrating how and to what extent these can contribute to understanding. In advance, we must emphasize that this type of output mimics what is produced by parametric techniques, but is essentially non-parametric, or what may be referred to as “semi-parametric”. In the next section, we link what machine learning has to offer as an inductive non-parametric input for subsequent deductive parametric multivariate regression. Here, we introduce what machine learning per se can produce by way of explanation, next to prediction.

The backbone of the random forest technique is the decision or regression tree. In the words of Grömping (2011: p. 311), the intuition of a decision or regression tree is the following: "A regression tree (…) is built by recursively partitioning the sample (= the 'root node') into more and more homogeneous groups, so-called nodes, down to the 'terminal nodes'. Each split is based on the values of one variable and is selected according to a splitting criterion. Once a tree has been built, the response for any observation can be predicted by following the path from the root node down to the appropriate terminal node of the tree, based on the observed values for the splitting variables, and the predicted response value simply is the average response in that terminal node."

In Figure 5, for illustrative purposes, we present one example of a decision or regression tree, related to our sub-sample dataset.

[Insert Figure 5 about here]

In this part of the tree, the root sample is split on the age of the firm. Rows with a value smaller than or equal to 6.5 descend down the tree, while those with a higher value traverse to another branch. Subsequently, the rows are split on the basis of working capital (*t*), with the left branch splitting again on working capital (*t*) before reaching the terminal nodes and the right branch splitting on number of employees. The predicted firm growth values in the terminal nodes are respectively 1.46 % (with 269 SMEs classified here, with MSE = 1,237), 7.5% (395, MSE = 1,056), -1.13% (437, MSE = 401), and 4.36% (301, MSE = 664).

A random forest is a large number of randomly generated trees, applying a bootstrapping technique (with or without replacement) to the training set, and using specific metrics to identify the optimal fit per tree and across all trees. Training evolves by comparing predicted values from the bootstrapped sub-sample to observed values from an "out-of-bag" (OOB) sub-sample, specific for each tree, often through calculating mean squared distances (or errors, which is intuitively very similar to what OLS does). Note that this OOB sub-sample is from within the training set, and is not

equal to the validation set. Random forest techniques come in different forms and shapes – e.g., classifier and regression, and Classification And Regression Trees (CART) and Conditional Importance (CI) trees, associated with a variety of tuning or hyper-parameters such as the number of trees and bootstrapping sampling size with or without replacement (cf. Table 4).

Petkovic et al. (2017) develop user-centered output, based on a RFA's feature importance outcomes, that adds further explanatory insight. We extend their approach – which was developed for a classification RFA – to our regression analysis. In Table 7, we provide an example of this output, again with reference to our sub-sample data. We selected those decision trees in our RFA with a "good-enough" prediction of firm growth rates per row in the dataset. The threshold for "good-enough" results was put at 1% deviation from the observed values. We then categorized the outputs in five classes ranging from "Strong decline" to "Strong growth".

[Insert Table 7 about here]

This output contains four key pieces of information that, in isolation and jointly, provide explanatory insights. First, independent variables are ranked in order of importance, similar to what is reported in Figure 4. Second, the "Mean threshold" column indicates the mean value for the decision thresholds in the trees with "good-enough" results. Third, the "<" column refers to the percentage of correctly predicted cases in this class that was above the decision threshold. And fourth, MFI (= Mutual Feature Interaction) provides the variables interacting with focal variable *i*. For instance, Table 7 illustrates that total assets growth (*t*-1), country, and solvency ratio (*t*) are the variables that occur in most interactions. For the total assets growth (*t*-1), we see that the mean decision threshold increases with growth rate. In the "Strong decline" class, just 26% of correctly predicted growth rates had a total assets growth above -23.50 in *t*-1. For the "Strong growth" class, 60% of the correctly predicted growth rates had a total assets growth above -5.50 in *t*-1. For all the other variables, the interpretation is much less straightforward. Again, this implies that the RFA

achieves higher performance by modeling higher-order interaction effects that are hard to fathom, let alone interpret.

RFA and multivariate regression

Third, machine learning offers a powerful data-mining tool, being a "quantitative" method of induction, that can be perfectly combined with standard deductive techniques, using the output of the first to identify input for the second. In Business and Management, including Strategy, we tend to associate induction with qualitative methodologies and techniques, such as cases studies and QCA (Quantitative Comparative Analysis). An example in Business and Management is the work of Fiss (e.g., 2011). However, in many other disciplines, data-mining techniques offer a powerful toolkit for quantitative induction. Such techniques are ideal to identify patterns, unknown ex ante, in large and complex data. Subsequently, these inductively derived patterns can inspire theory development, with the associated hypotheses being deductively tested by using well-known parametric techniques. This sequential way of working is akin to abduction. In Business and Management, an example is provided by Dikova et al. (2017).

To illustrate this further, we take the output of our RFA of firm growth as the inductive starting point for subsequent deductive analyses. By inspecting the feature importance list in Figure 4 and the underlying decision trees, by way of example, we inductively formulate the following set of three correlational hypotheses:

> *<u>Hypothesis 1</u> (H1): An SME's strategy of operational excellence (O) is positively associated with firm growth.*
>
> *<u>Hypothesis 2</u> (H2): An entrepreneur's personality of egocentrism (E) is positively associated with firm growth (G).*

*Hypotheses 3 (H3): The positive association of operational excellence and firm growth is positively moderated by egocentrism (E*O).*

The deductive theory is kept simple here, due to lack of space and because this is not the focus of the current paper. Insights from the Psychology of Strategy suggest that the personality of the entrepreneur is a key driver of her or his enterprise's behavior and performance, with the effect of personality being contingent on the nature of the strategy (e.g., Miller & Toulouse, 1986; Wijbenga & van Witteloostuijn, 2007). For instance, Boone et al. (1996) theorize that the entrepreneur's locus-of-control internality has a positive effect on small business performance, particularly so in the case of a product differentiation strategy. They find support for both hypotheses in a sample of 40 Belgian SMEs. We suggest a similar logic, but now related to the personality trait of egocentrism and the strategy of operational excellence. Treacy and Wiersema (1997) argue that not being cost efficient – i.e., not being operationally excellent – is likely to harm performance (H1). Ceteris paribus, egocentrism is associated with strong leadership, which can be expected to be positively associated with firm performance (H2). For example, de Vries (2003) has extensively argued that egocentrism (or narcissism) is a trait that is overrepresented among (top) managers, and that this trait is instrumental in developing an authoritarian and directive leadership style. The latter fits well with an operational excellence strategy, requiring stringent cost control and strict routine application (see, e.g., Gupta & Govindarajan, 1984), pointing to positive moderation (H3).

Next, we specify a regression model with O (H1; expected $\beta > 0$), E (H2; expected $\beta > 0$) and E*O (H3; expected $\beta > 0$), next to the other features as covariates and a constant, that we subsequently step-wise estimate with the data from the sub-sample of 8,163 unique firms: a controls-only Model 1, a Model 2a and b with both main effects added seperately, and a Model 3 with the interaction term included. Note that, ideally, would we have had a sufficiently large Big

Data set, we could have sliced this dataset in three disjoint sub-samples: (1) a training dataset; (2) a validation dataset; and (3) a dataset for deductive testing. In the context of the current paper, running a regression analysis with the same sub-sample suffices. In Table 8, we provide the results.

[Insert Table 8 about here]

The models perform very well, achieving an R^2 of 0.38 – very high in the firm growth literature. This seems like an impressive result, but we should keep in mind that the sub-sample is highly biased toward larger firms for which content-rich websites could be scraped. As such, this R^2 should not be compared to our earlier results. The results in Table 8 illustrate the main effect of operational excellence and egocentrism on a higher growth rate ($p = 0.078$ and $p = 0.008$, respectively), as well as the positive moderation effect ($p = 0.003$), in line with all three hypotheses. Please note that the β-coefficients operate on L1-normalized data, making a straightforward calculation and interpretation of effect sizes difficult.

DISCUSSION

Overview

This study examines whether the performance of firm growth models can be improved by applying modern Data Science techniques. To do this, we constructed a large Big Data set with 168,055 unique firms, each associated with information for one to six years. Our analysis demonstrates that the random forest analysis (RFA) – a machine learning technique – performs best on the training and validation set, and much better so than traditional linear regression models, with an R^2 of 0.16 (the validation set) or 0.23 (the training set) for RFA vis-à-vis an R^2 of ~ 0.05 (both sets) for the traditional linear regression models. The goodness-of-fit of the RFA is on par with or superior to that of top contenders in the firm growth field, despite the very basic set of demographic and financial variables in our dataset. The extant models in the firm growth literature that perform

in the $R^2 = 0.15$ range include a much larger selection of non-demographic and non-financial information (see, e.g., Parker et al., 2010), typically adding data on personality and strategy measures collected through surveys.

Importantly, the RFA outperformed the linear (multivariate OLS) model that was estimated in a traditional parametric fashion, the other types of linear models, and the other machine learning techniques. Moreover, the RFA retains its lead even when a smaller subset of variables is used. This shows that the performance gain of the RFA can be attributed to the machine learning technique, and cannot be solely explained by the larger flexibility of the machine learning algorithm. Only when interaction effects are added to the other models manually, do they start to catch up somewhat in terms of R^2. The superior performance of the RFA can thus be attributed to its flexible capacity to identify higher-order interact effects. Ultimately, this suggests that firm growth cannot be readily explained by a simple set of variables; rather, firm growth exhibits subtle interaction effects, non-linearities, and sensitivity to initial conditions, which are all features of complex systems (cf. McKelvey, 2004).

Our algorithmic text analysis of scraped firm website information provides evidence that text-based variables could further improve machine learning model performance, increasing the R^2 with a not-too-bad 2.5 %, with an illustrative and small set of personality and strategy variables: i.e., egocentrism, product leadership, customer intimacy, and operational excellence. Constructing proxies for personality and strategy features of entrepreneurs and their SMEs by scraping information from the Internet that is subsequently text-analyzed, is yet another pair of techniques from modern Data Science that can be added to the toolbox of Strategy scholarship. In this way, Big Data can be enriched with deep and difficult-to-observe variables that traditionally have to be collected through time-consuming questionnaires or tests, which tend to be associated with low

response rates, high administrative costs, small samples, and a series of biases (e.g., common method, social desirability, self-selection, attrition, and response biases).

Theoretical Insights

In addition, we provide a threefold response to the common critique that artificial intelligence or machine learning sacrifices explanatory insight to increase predictive accuracy. First, predictive accuracy is a useful scholarly aim per se, certainly so in context of the rigor versus relevance debate, as well as the convincing pleas for responsible research in Business and Management, and solution-oriented research. Second, modern machine learning software produces output providing more explanatory insight, which may be referred to as "semi-parametric", giving a weight to the importance of features, indicating classification accuracy, and identifying key moderating variables. Third, machine learning and parametric methods can be applied as complementary techniques, the former producing inductively the input for the latter's deductive (hypotheses-testing) regression. We illustrate how all this works out with reference to our firm growth sub-sample with text-analyzed variables.

This third response makes the case for adding machine learning to the theory-building and testing toolbox. In so doing, we provide a powerful mix of techniques that fit within the tradition of abduction, which is a methodology midway deduction and induction (see, e.g., Dikova et al., 2017; Fiss, 2011; Misangyi & Acharya, 2014). Machine learning's very flexible specification strategy quantitatively produces a list of variables that are key in generating high predictive accuracy, oftentimes through very subtle higher-order interaction effects. This is the inductive output that then is the input in the deductive next step. In this inductive step, the identified variables are first taken as the elements with which novel theory can be developed. Subsequently, in the second deductive step, traditional regression is applied to test this new set of hypotheses. Taking egocentrism and

operational excellence as our pair of examples, we illustrate how this abductive methodology can be applied to develop three new hypotheses, all confirmed in a traditional regression analysis with an impressive R^2 of 0.38.

Limitations and Future Decisions

As any other, the current study is prone to a number of limitations that need to be discussed and which point to important avenues for future research. First, our aim was to illustrate the potential of adding modern Data Science techniques to the extant methodological toolbox in Strategy, without any pretention of completeness. There is much more on offer in modern Data Science, of course, than we can discuss in a compact paper format, and we lack the space to explain the nitty-gritty of the techniques that we brought forward. Future research can explore the details of the techniques suggested above, as well as examine other ones from the quickly expanding and progressing Data Science field. Second, our Big Data set was limited to basic demographic and financial information about firms in two countries (Belgium and, to a lesser extent, the Netherlands), and our small list of four text-analyzed proxies could be collected only for a small sub-sample of our SMEs. Future studies could explore other Big Data sets from other countries, encompassing richer information about the entrepreneurs and their ventures, perhaps by investing further in data-scraping opportunities.

Third, while the random forest achieves much better results that traditional linear regression methods and is robust to overfitting, the mechanics of the model are still not easy to understand. So, while machine learning may provide better explanations and predictions of firm growth rates (or other target variables), such techniques are not equally informative about any possible theoretical (let alone causal) relationships between the predictor and target variables. Although outputs related to feature importance lists offer explanatory insight, machine learning is not a parametric technique

producing β-coefficients (economic significance) and *p*-values (statistical significance). This is an issue that is seen as very prominent in the scholarly Data Science community, too. Having techniques that are associated with much higher explanatory power and predictive accuracy implies progress, but this comes, to date, at the expense of less insight into the underlying causal mechanisms.

REFERENCES

Alpaydin, E. (2014). *Introduction to Machine Learning*. Cambridge, MA: MIT Press.

Ardichvili, A., Harmon, B., Cardozo, R. N., Reynolds, P. D., & Williams, M. L. (1998). The New Venture Growth: Functional differentiation and the need for human resource development interventions. *Human Resource Development Quarterly*, 9(1): 55-70.

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13(Feb): 281-305.

Bishop, C. (2016). *Pattern Recognition and Machine Learning*. New York: Springer Press.

Blei, D. M., A. Y. Ng, A., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(1): 993-1022.

Boone, C., Brabander, B., & Witteloostuijn, A. van (1996). CEO Locus of Control and Small Firm Performance: An integrative framework and empirical test. *Journal of Management Studies*, 33(5): 667-700.

Boulesteix, A.-L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics. *WIREs Data Mining Knowledge Discovery*, 2: 493-507.

Breiman, L. (2001), Random Forests. *Machine Learning*, 45(1): 5-32.

Burez, J., & van den Poel, D. (2007). CRM at a Pay-TV Company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications*, *32*(2), 277-288.

Chicco, D. (2017). Ten Quick Tips for Machine Learning in Computational Biology. *BioData Mining*, *10*(1): 35.

Dikova, D., Parker, S. C., & van Witteloostuijn, A. (2017). Capability, Environment and Internationalization Fit, and Financial and Marketing Performance of MNEs' foreign subsidiaries: An abductive contingency approach. *Cross-Cultural and Strategic Management* (formerly known as *Cross Cultural Management*) 24: 405-435.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12(11): 2121-2159.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, 32(2): 407-499.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. Thousand Oaks: Sage.

Fiss, P. C. (2011). Building Better Causal Theories: A fuzzy set approach to typologies in organization research. *Academy of Management Journal*, 54(2): 393-420.

Friedman, J. H. (2002). Stochastic Gradient Boosting. *Computational Statistics & Data Analysis*, 38(4): 367-378.

Garnsey, E., E. Stain, & Hefferman, P. 2006. New Firm Growth: Exploring processes and paths. *Industry and Innovation*, 13(1): 1-20.

Gelman, A., & Imbens, G. (2013). Why Ask Why? Forward causal inference and reverse causal questions. *NBER Working Paper*. Number 19614.

Grömping, U. (2009). Variable Importance Assessment in Regression: Linear regression versus random forest. *The American Statistician*, 63(4): 308-319.

Gulati, R. (2007). Tent Poles, Tribalism, and Boundary Spanning: The rigor-relevance debate in management research. *Academy of Management Journal*, 50(4): 775-782.

Gupta, A. K., & Govindarajan, V. (1984). Business Unit Strategy, Managerial Characteristics, and Business Unit Effectiveness at Strategy Implementation. *Academy of Management Journal*, 27(1): 25-41.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support Vector Machines. *IEEE Intelligent Systems and Their Applications*, 13(4): 18-28.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased estimation for nonorthogonal problems. *Technometrics,* 12(1): 55-67.

Kaplan, S., & Vakili, K. (2014). The Double-Edged Sword of Recombination in Breakthrough Innovation. *Strategic Management Journal*, 36(1): 1435-1457.

Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-Validation Pitfalls When Selecting and Assessing Regression and Classification Models. *Journal of Cheminformatics,* 6(1): 1-10.

Kitchin, R. (2014). Big Data: New epistemologies and paradigm shifts. *Big Data & Society*, 1(1): 1-12

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text Classification for Organizational Researchers: A tutorial. *Organizational Research Methods,* 21(3): 766–7999.

Kotu, V., & Deshpande, B. (2014). *Predictive Analytics and Data Mining: Concepts and practice with rapidminer*. Waltham: Morgan Kaufmann.

LeCun, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier Nonlinearities Improve Neural Network Acoustic Models. *Proceedings ICML*, 30(1): 1-3.

Legates, D. R., & McCabe, G. J. (1999). Evaluating the Use of "Goodness-of-Fit" Measures in Hydrologic and Hydroclimatic Model Validation. *Water Resources Research*, 35(1): 233-241.

Loh, W.-Y. (2011). Classification and Regression Trees. *WIREs Data Mining Knowledge Discovery*, 1: 14-23.

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier Nonlinearities Improve Neural Network Acoustic Models. *Proceedings ICML,* 30(1).

McAfee, A. & Brynjolfsson, E. (2012). Big Data: The management revolution. *Harvard Business Review*, October: 60-68.

McKelvey, B. (2004). Toward a Complexity Science of Entrepreneurship. *Journal of Business Venturing*, 19(3): 313-341.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data–A Revolution That Will Transform How We Live, Think and Work*. New York: Houghton Mifflin Harcourt.

Miller, D., & Toulouse, J. M. (1986). Chief Executive Personality and Corporate Strategy and Structure in Small Firms. *Management Science*, 32(11): 1389-1409.

Molinaro, A. M., Carriero, N. J., Bjornson, R., Hartge, P., Rothman, N., & Chatterjee (2011). Power of Data Mining Methods to Detect Genetic Associations and Interactions. *Human Heredity*, 72: 85-97.

Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2015). Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Political Analysis*, 24(1): 87-103.

Misangyi, V. F., & Acharya, A. G. (2014). Substitutes or Complements? A configurational examination of corporate governance mechanisms. *Academy of Management Journal*, 57(6): 1681-1705.

Nilsson, N. (2005). Introduction to Machine Learning. *Unpublished draft*. Available online at https://ai.stanford.edu/~nilsson/mlbook.html [accessed on 14-06-2018].

Nuscheler, D., Engelen, A., & Zahra, S. A. (2018). The Role of Top Management Teams in Transforming Technology-Based New Ventures' Product Introductions into Growth. *Journal of Business Venturing* (forthcoming).

O'Gorman, C. (2001). The Sustainability of Growth in Small and Medium-Sized Enterprises. *International Journal of Entrepreneurial Behavior & Research*, 7(2): 60-75.

Parker, S.C., Storey, D., & van Witteloostuijn, A. (2010). What Happens with Gazelles?: The role of dynamic management strategies. *Small Business Economics*, (35): 203-226.

Petkovic, D., Altman, R., Wong, M., & Vigil, A. (2017). Improving the Explainability of Random Forest Classifier: User centered approach. *Pacific Symposium Biocomputing*, 23: 204-215.

Powell, T. C., Lovallo, D., & Fox, C. R. (2011). Behavioral Strategy. *Strategic Management Journal*, 32(13): 1369-1386.

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern Prediction Methods: New perspectives on a common problem. *Organizational Research Methods*, 21(3): 689–732.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. In *Encyclopedia of Database Systems* (pp. 532-538). Chicago, IL: Springer.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations by Back-Propagating Errors. *Nature*, 323(6088): 533.

Schmidhuber, J. (2015). Deep Learning in Neural Networks: An overview. *Neural Networks*, 61(1): 85-117.

Shu, L., Long, B., & Meng, W. (2009, March). A Latent Topic Model for Complete Entity Resolution. In *Proceedings of the 2009 IEEE International Conference on Data Engineering* (pp. 880-891). IEEE Computer Society.

Smola, A., & Vishwanathan, S. (2014). *Introduction to Machine Learning*. Cambridge, UK: Cambridge University Press.

Strobl, C., Malley J., & Tutz G. (2009). An Introduction to Recursive Partitioning: Rationale, application and characteristics of classification and regression trees, bagging and Random Forests. *Psychological Methods*, 14(4): 323-348.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Treacy M., & Wiersema, F. (1997). Customer Intimacy and Other Value Disciplines. *Harvard Business Review*, January-February: 84-95

Tsui, A. S. (2013). 2012 Presidential Address—On compassion in scholarship: Why should we care? *Academy of Management Review*, 38(2): 167-180.

de Vries, M. F. K. (2003). *Leaders, Fools and Impostors: Essays on the psychology of leadership*. San Fransisco: Jossey-Bass.

Watts, D. J. (2017). Should Social Science Be More Solution-Oriented? *Nature Human Behaviour*, 1(1): 0015.

Weisberg, S. (1980). *Applied Linear Regression*. New York: Wiley.

Wenzel, R., & Van Quaquebeke, N. (2018). The Double-Edged Sword of Big Data in Organizational and Management Research: A review of opportunities and risks. *Organizational Research Methods*, 21(3): 548-591.

Wijbenga, F. H., & van Witteloostuijn, A. (2007). Entrepreneurial Locus of Control and Competitive Strategies: The moderating effect of environmental dynamism. *Journal of Economic Psychology*, 28(5): 566-589.

van Witteloostuijn, A., & Kolkman, D. (2018). Is Firm Growth Random?: A machine learning perspective. *Journal of Business Venturing Insight*s, 10: e00107.

Witten, I., Frank, E., & Hall, M. (2011). *Data Mining: Practical machine learning tools and techniques*. San Francisco: Elsevier.

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301-320.

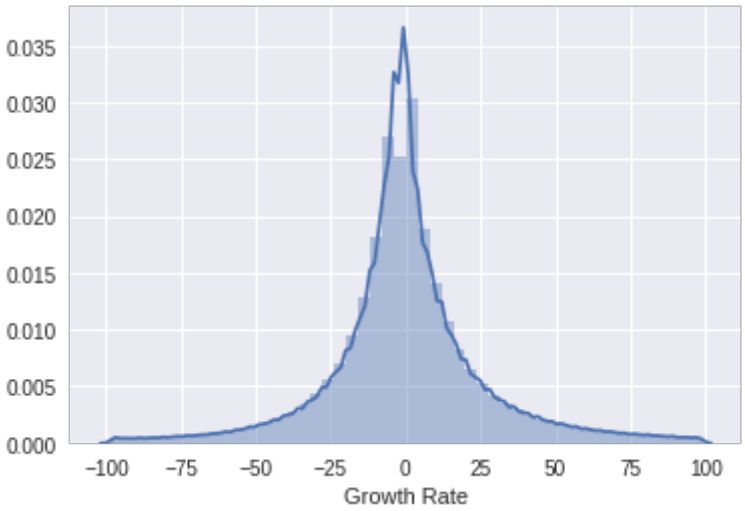
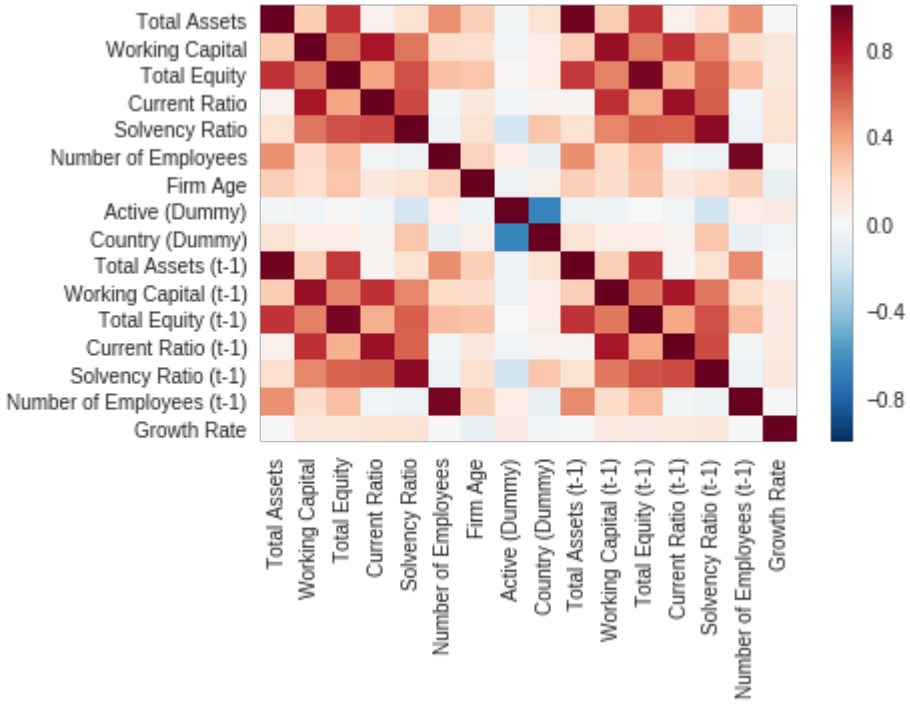


Figure 1: Distribution of growth rates in the dataset



All variables measured at *t*, except if indicated otherwise.

Figure 1: Correlations (Spearman *r*) between the variables and firm growth rates

NACE2	Count	Description
70	17682	Activities of head offices; management consultancy activities
46	17269	Wholesale trade, except of motor vehicles and motorcycles
68	15064	Real estate activities
47	15050	Retail trade, except of motor vehicles and motorcycles
43	15016	Specialised construction activities
86	10205	Human health activities
64	8502	Financial service activities, except insurance and pension funding
41	7011	Construction of buildings
69	6702	Legal and accounting activities
56	6277	Food and beverage service activities
62	6118	Computer programming, consultancy and related activities
71	5020	Architectural and engineering activities; technical testing and analysis

Table 1: The top ten largest industries in the sample

NACE2	Growth rate (%)	Description
39	30	Remediation activities and other waste management services
98	24	Undifferentiated goods- and services-producing activities of private households for own use
60	18	Programming and broadcasting activities
80	15	Security and investigation activities
78	15	Employment activities
9	0	Mining support service activities
36	-1	Water collection, treatment and supply
8	-2	Other mining and quarrying
99	-5	Activities of extraterritorial organizations and bodies
5	-26	Mining of coal and lignite

Table 2: The industries with the five highest and five lowest growth rates

Category	Name	Description	Type
Financial	Total Assets *(t)*	Balance sheet total	Continuous (ratio)
	Working Capital *(t)*	Working capital	Continuous (ratio)
	Total Equity *(t)*	Total equity	Continuous (ratio)
	Curent Ratio *(t)*	Current ratio	Continuous (ratio)
	Solvency Ratio *(t)*	Solvability ratio	Continuous (ratio)
Financial (lagged)	Total Assets *(t-1)*	Balance sheet total in the previous year	Continuous (ratio)
	Working Capital *(t-1)*	Working capital in the previous year	Continuous (ratio)
	Total Equity *(t-1)*	Total equity in the previous year	Continuous (ratio)
	Current Ratio *(t-1)*	Current ratio in the previous year	Continuous (ratio)
	Solvency Ratio *(t-1)*	Solvability ration in the previous year	Continuous (ratio)
Other	Number of Employees *(t)*	Number of employees	Continuous (interval)
	Number of Employees *(t-1)*	Number of employees in the previous year	Continuous (interval)
	NACE2	The firm's industry according to the NACE2 classification	Categorical (nominal, 90 options)
	Legal Person	The firm's legal person	Categorical (nominal, 7 options)
	Active (Dummy)	An indicator of the firm's current activity	Categorical (binary)
	Country (Dummy)	Country	Categorical (binary)

Table 3: Base list of variables

Model	Hyper-parameter search space	Optimized hyper-parameter setting
OLS	Intercept (True, False)	Intercept (False)
FSR	Intercept (True, False)	Intercept (False)
RR	Intercept (True, False); Alpha (5-e5 – 5+e5); Selection method (random, cyclic); Tolerance (1-e10 – 1)	Intercept (False); Alpha (0.000125); Selection method (cyclic); Tolerance (0.0001)
LASSO	Intercept (True, False); Max number of Alphas (1 – 1e+10); Selection method (random, tyclic); Tolerance (1-e10 – 1)	Intercept (False); Number of Alphas (1000); Selection method (random); Tolerance (0.0001)
LARS	Intercept (True, False); Max number of Alphas (1 – 1e+10); Eps (1-e20 – 1)	Intercept (True, False); Eps (1e-10)
ENR	Intercept (True, False); L1 Ratio(0 - 1); Eps (1-e20 – 1); Selection method (random, cyclic); Tolerance (1-e10 – 1)	Intercept (False); L1 Ratio(0 .55); Eps (0.06); Selection method (cyclic); Tolerance (0.001)
SVM	Intercept (True, False); C (1-e10 – 1); Eps (1-e20 – 1); Selection method (random, cyclic); Tolerance (1-e10 – 1)	Intercept (False) ;C (0.05); Eps (0.03); Selection method (cyclic); Tolerance 0.001)
RFA	Bootstrap (True, False); Max depth (1 – 10000), Number of trees (1, 10000); Minimum Samples per Leaf (1, 10000); Minimum Samples for Split (2 – 10000), Max features (auto, sqrt, log2, None)	Bootstrap (True); Max depth (10), Number of trees (1000); Minimum Samples per Leaf (30); Minimum Samples for Split (100); Max features (auto)
SGT	Bootstrap (True, False); Learning Rate (1-e10- 1); alpha (1-e10- 1), Max depth (1 – 10000), Number of trees (1, 10000); Minimum Samples per Leaf (1, 10000); Minimum Samples for Split (2 – 10000); Max features (auto, sqrt, log2, None)	Bootstrap (True); Learning Rate (0.001); Max depth (100), Number of trees (550); Minimum Samples per Leaf (20); Minimum Samples for Split(100); Max features (auto)
MNN	Optimizer (Adam, SGD, Adagrad, Adadelta, Adam, Adamax, Nadam)	Optimizer (Nadam)

Table 4: Hyper-parameter search space and results of random search

	DATASET (1)				DATASET (2)				DATASET (3)			
TRAINING SET												
Model	R^2	**MAE**	**MSE**	**RSME**	R^2	**MAE**	**MSE**	**RSME**	R^2	**MAE**	**MSE**	**RSME**
OLS	0.06	17.25	781.63	28.22	0.05	17.63	829.10	28.80	0.08	17.28	79.54	28.22
FSR	0.05	17.56	821.69	28.27	0.05	17.56	821.69	28.27	0.06	17.48	821.05	28.66
RR	0.05	17.63	828.85	28.79	0.05	17.63	828.85	28.79	0.06	17.52	820.59	28.65
LASSO	0.05	17.64	829.00	28.79	0.05	17.64	829.31	28.80	0.06	17.52	821.01	28.65
LARS	0.05	17.59	823.61	28.71	0.05	17.63	829.22	27.80	0.05	17.60	824.96	28.72
ENR	0.05	17.64	830.10	28.81	0.05	17.65	830.05	28.81	0.05	17.54	823.43	28.70
SVM	0.03	17.19	842.99	29.00	0.03	17.19	843.02	29.03	0.04	17.13	837.85	28.95
RFA	0.23	15.76	665.61	25.80	0.14	16.78	749.38	27.37	0.17	16.45	722.47	26.88
SGT	0.12	16.97	770.10	27.75	0.10	17.04	780.09	27.93	0.11	17.00	771.75	27.78
MNN	0.12	16.97	720.33	27.71	0.08	17.02	781.08	28.01	0.10	17.24	786.83	28.05
VALIDATION SET												
Model	R^2	**MAE**	**MSE**	**RSME**	R^2	**MAE**	**MSE**	**RSME**	R^2	**MAE**	**MSE**	**RSME**
OLS	0.06	17.25	781.63	28.22	0.05	17.71	828/67	28.79	0.03	17.48	825.86	29.10
FSR	0.05	17.64	821.90	28.67	0.05	17.64	821.90	28.67	0.04	17.37	824.11	28.98
RR	0.05	17.71	828.90	28.79	0.05	17.71	828.90	28.79	0.06	17.62	822.53	28.68
LASSO	0.05	17.71	828.96	28.79	0.05	17.71	829.00	28.79	0.06	17.61	822.91	28.69
LARS	0.06	17.66	823.61	28.70	0.05	17.71	828.79	28.79	0.05	17.68	825.70	28.73
ENR	0.05	17.72	829.76	28.81	0.05	17.71	829.74	28.81	0.05	17.63	825.27	28.73
SVM	0.03	17.26	842.71	29.00	0.03	17.26	842.72	29.03	0.04	17.21	776.44	28.91
RFA	0.16	16.69	733.36	27.08	0.13	16.96	760.00	27.59	0.15	16.81	746.35	27.31
SGT	0.12	17.10	772.72	27.80	0.10	17.13	783.66	28.00	0.11	17.11	771.75	27.86
MNN	0.10	17.03	721.43	27.79	0.08	17.09	782.03	28.12	0.09	17.42	799.47	28.27

Table 5: Model fit statistics

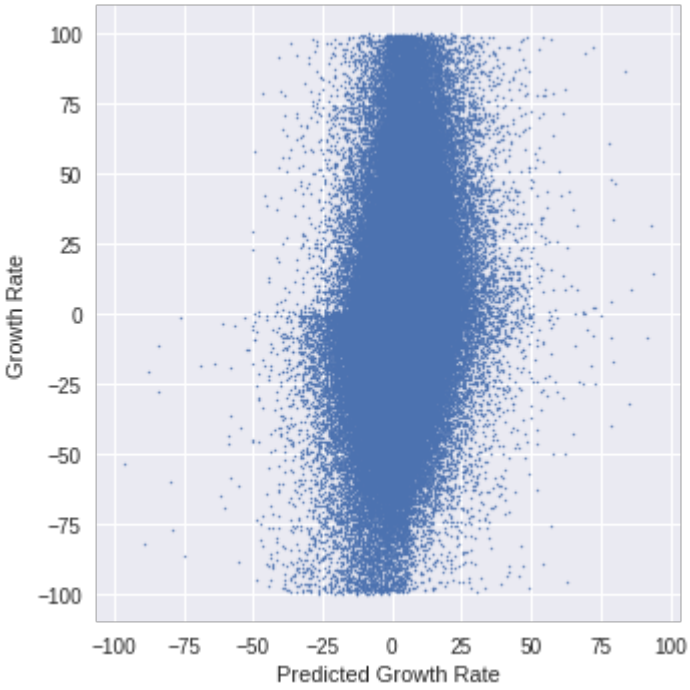


Figure 3a: Scatterplot of the FSR

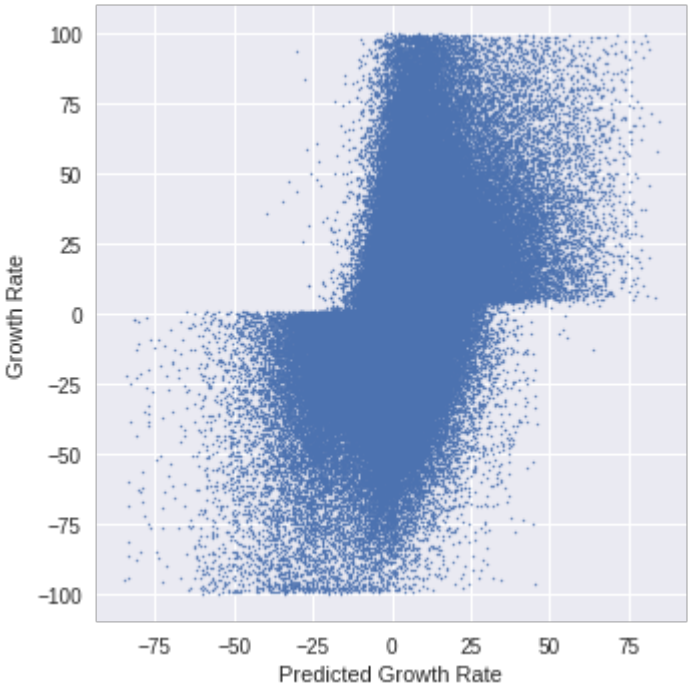
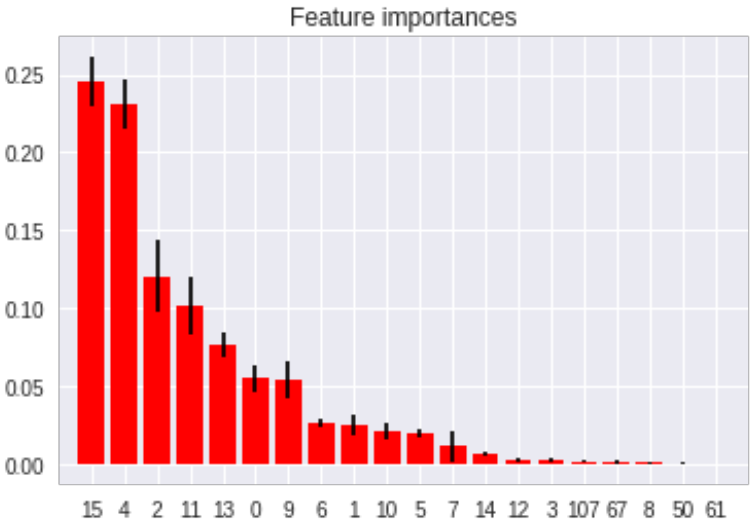


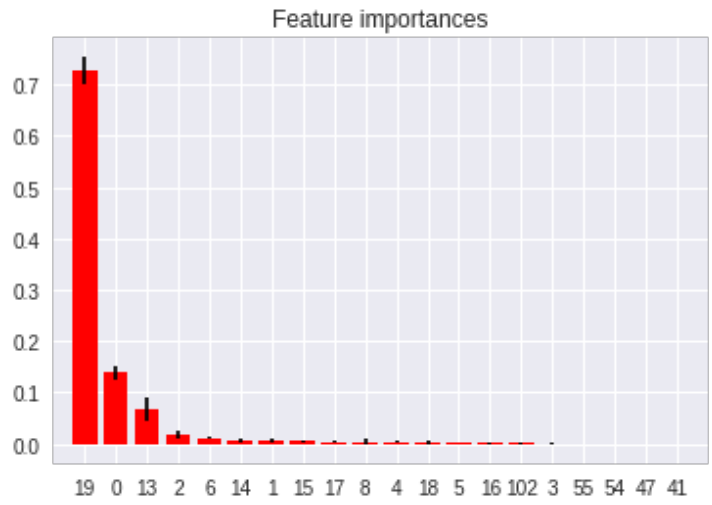
Figure 3b: Scatterplot of the RFA



Ranking	Description	Feature #
1	Total Assets	15
2	Working Capital	4
3	Total Equity	2
4	Current Ratio	11
5	Solvency Ratio	13
6	Number of Employees	0
7	Firm Age	9
8	Active (Dummy)	6
9	Country (Dummy)	1
10	Total Assets (t -1)	10
11	Working Capital (t-1)	5
12	Total Equity (t-1)	7
13	Current Ratio (t-1)	14
14	Solvency Ratio (t-1)	12
15	Number of Employees (t-1)	3
16	Total Asset Growth (t-1)	107
17	NACE2 Code = 1 (Dummy)	67
18	NACE2 Code = 10 (Dummy)	8
19	NACE2 Code = 11 (Dummy)	50
20	NACE2 Code = 12 (Dummy)	61

All variables measured at *t*, except if indicated otherwise.

Figure 4a: Plot of relative importance of firm growth's explanatory variables



Ranking	Description	Feature #
1	Total Assets	19
2	Working Capital	0
3	Total Equity	13
4	Current Ratio	2
5	Solvency Ratio	6
6	Number of Employees	14
7	Firm Age	1
8	Active (Dummy)	15
9	Country (Dummy)	17
10	Egocentrism	8
11	Product Leadership	4
12	Customer Intimacy	18
13	Operational Excellence	5
14	Total Assets (t -1)	16
15	Working Capital (t-1)	102
16	Total Equity (t-1)	3
17	Current Ratio (t-1)	55
18	Solvency Ratio (t-1)	54
19	Number of Employees (t-1)	47
20	Total Asset Growth (t-1)	41

All variables measured at *t*, except if indicated otherwise.

Figure 4b: Top twenty relative importance after adding text-analyzed measures

Egocentrism word list = ['ik' (‘I’), 'mij' (‘me’), 'mijn' (‘mine’), 'mijzelf' (‘myself’)]
Product leadership word list = ['technologie', 'innovatie', 'vernieuwing’ (‘renewal’)]
Customer intimacy word list = ['service', 'maatwerk' (‘tailor-made’), 'kwaliteit' (‘quality’)]
Operational excellence word list = ['korting' (‘discount), 'uitverkoop' (‘sales’), 'sale']

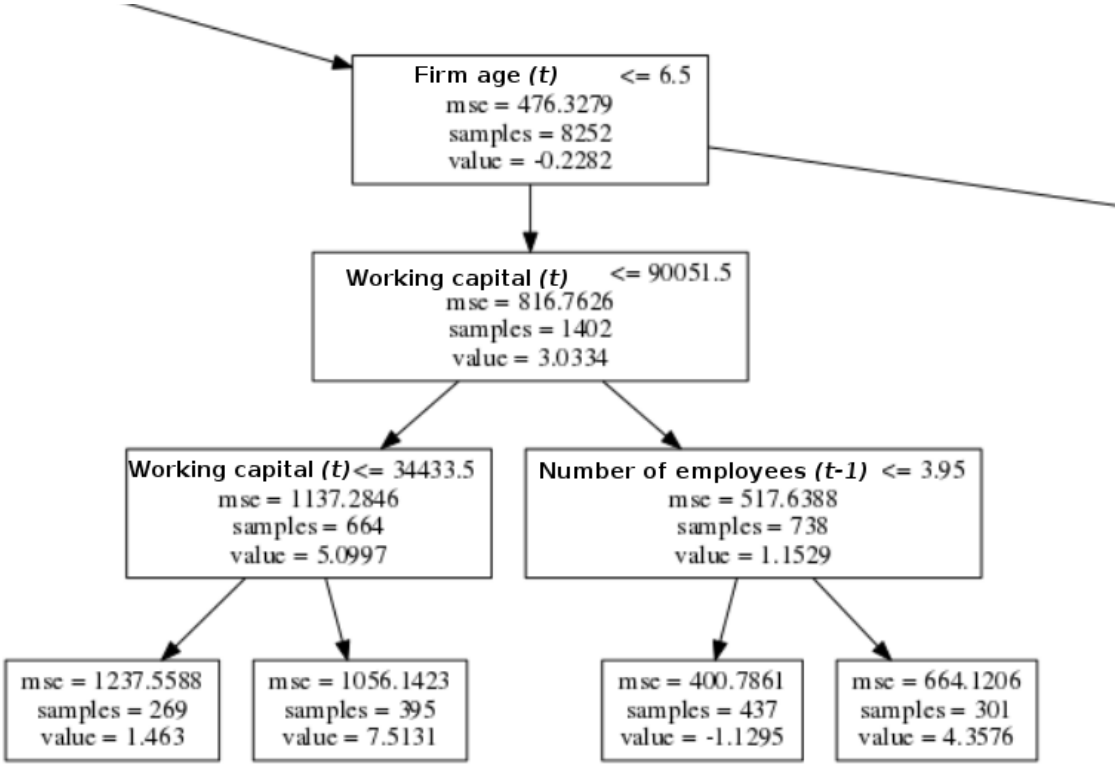
Table 6: Overview of custom topics and word mappings

#	Variable	MFI	Strong decline (< -10%) Mean threshold	<	Decline (-10% > -2.5%) Mean threshold	<	Stale (2.5% > -2.5%) Mean threshold	<	Growth (10% > 2.5%) Mean threshold	<	Strong growth (>10%) Mean threshold	<
1	Total Assets (*t*)	16,9,5	270190.00	64%	193678.00	61%	247718.00	73%	270214.50	73%	201602.50	56%
2	Working Capital (*t*)	16,9,5	10697.75	53%	-8393.00	42%	15127.50	62%	16051.50	73%	13235.00	67%
3	Total Equity (*t*)	16,9,5	50835.50	59%	18533.00	55%	42127.00	71%	42127.00	78%	30166.00	66%
4	Current Ratio (*t*)	16,9,5	1.14	45%	0.89	32%	1.32	55%	1.22	63%	1.26	62%
5	Solvency Ratio (*t*)	16,9,10	0.36	76%	0.10	26%	0.28	64%	0.25	63%	0.32	72%
6	Number of Employees (*t*)	16,9,5	3.35	43%	2.65	26%	2.85	32%	3.05	46%	2.85	35%
7	Firm Age (*t*)	16,9,5	11.50	67%	10.50	57%	10.50	67%	11.50	60%	9.50	44%
8	Active (Dummy)	16,9,5	0.50	4%	0.50	84%	0.50	61%	0.50	99%	0.50	60%
9	Country (Dummy)	13,15,8	1.50	97%	n.a.	n.a.	1.50	33%	n.a.	n.a.	1.50	37%
10	Total Assets (*t* -1)	16,9,5	225426.00	69%	189989.00	64%	244194.00	73%	224679.00	74%	209052.50	58%
11	Working Capital (*t*-1)	16,9,5	5290.00	56%	-21425.00	53%	16104.50	59%	12191.50	73%	12191.50	64%
12	Total Equity (*t*-1)	16,9,5	62350.50	58%	53681.50	47%	71189.50	69%	78983.00	75%	54723.50	58%
13	Current Ratio (*t*-1)	16,9,5	1.31	40%	1.23	36%	1.44	51%	1.39	57%	1.32	51%
14	Solvency Ratio (*t*-1)	8,5,7	0.36	82%	0.13	33%	0.28	64%	0.25	60%	0.28	65%
15	Number of Employees (*t*-1)	9,16,5	2.50	46%	2.50	31%	2.50	38%	2.50	54%	2.55	38%
16	Total Assets Growth (*t*-1)	15,7,9	-23.50	26%	-11.50	44%	-10.50	48%	-8.50	53%	-5.50	60%

Some variables were not used in decision trees with good predictions. As such, the descriptive statistics for NACE2 Dummies 1,10,11,12,13 could not be computed.

Table 7: Explanatory RFA output

Figure 5: Example of branches in a decision tree



Model 1: Baseline model	***β***	**SE**	***t***	***p***	**CI 0.025**	**CI 0.975**
Total Assets *(t)*	1.094.630	1.880	58.230	0	105.778	113.148
Total Assets *(t-1)*	-1.082.186	1.857	-58.275	0	-111.859	-104.579
Total Equity *(t)*	-84.586	2.550	-3.317	0.001	-13.456	-3.461
Total Equity *(t-1)*	98.860	2.591	3.815	0	4.807	14.965
Working Capital *(t)*	239.933	2.699	8.889	0	18.702	29.284
Working Capital *(t-1)*	-181.002	2.852	-6.348	0	-23.690	-12.511
Current Ratio *(t)*	-2.275.397	180.463	-1.261	0.207	-581.276	126.196
Current Ratio *(t-1)*	94.650	194.843	0.049	0.961	-372.457	391.387
Solvency Ratio *(t)*	143.080	138.278	0.103	0.918	-256.739	285.355
Solvency Ratio *(t-1)*	802.705	255.491	0.314	0.753	-420.530	581.071
Number of Employees *(t)*	40.667.894	4.964.480	0.819	0.413	-5.664.348	1.38E+04
Model 2a: Operational Excellence model	***β***	**SE**	***t***	***p***	**CI 0.025**	**CI 0.975**
Total Assets *(t)*	1.094.406	1.880	58.222	0	105.756	113.125
Total Assets *(t-1)*	-1.081.507	1.857	-58.231	0	-111.791	-104.510
Total Equity (*t)*	-84.853	2.550	-3.328	0.001	-13.483	-3.488
Total Equity *(t-1)*	98.718	2.591	3.810	0	4.793	14.951
Working Capital (t)	240.316	2.699	8.903	0	18.741	29.322
Working Capital (t-1)	-180.970	2.851	-6.347	0	-23.686	-12.508
Current Ratio *(t)*	-2.260.303	180.450	-1.253	0.21	-579.741	127.680
Current Ratio *(t-1)*	29.859	194.862	0.015	0.988	-378.972	384.944
Solvency Ratio *(t)*	139.814	138.267	0.101	0.919	-257.043	285.006
Solvency Ratio *(t-1)*	732.686	255.501	0.287	0.774	-427.551	574.089
Number of Employees *(t)*	37.445.834	4.967.441	0.754	0.451	-5.992.358	1.35E+04
Operational Excellence *(t)*	1.75E+07	9.92E+06	1.760	0.078	-1.99E+06	3.69E+07
Model 2b: Egocentrism model	***β***	**SE**	***t***	***p***	**CI 0.025**	**CI 0.975**
Total Assets *(t)*	1.095.303	1.880	58.275	0	105.846	113.214
Total Assets *(t-1)*	-1.080.873	1.857	-58.198	0	-111.728	-104.447
Total Equity *(t)*	-85.119	2.549	-3.339	0.001	-13.509	-3.515
Total Equity *(t-1)*	99.532	2.591	3.842	0	4.875	15.032
Working Capital *(t)*	240.152	2.699	8.899	0	18.725	29.305
Working Capital *(t-1)*	-181.418	2.851	-6.364	0	-23.730	-12.554
Current Ratio *(t)*	-2.682.844	181.071	-1.482	0.138	-623.211	86.642
Current Ratio *(t-1)*	-90.964	194.921	-0.047	0.963	-391.172	372.979
Solvency Ratio *(t)*	116.408	138.249	0.084	0.933	-259.348	282.630
Solvency Ratio *(t-1)*	663.667	255.483	0.26	0.795	-434.419	567.153
Number of Employees *(t)*	33.039.854	4.971.590	0.665	0.506	-6.441.088	1.30E+04
Egocentrism *(t)*	2.71E+10	1.02E+07	2.656	0.008	7.09E+06	4.70E+07
Model 3: Operational Excellence * Egocentrism model	***β***	**SE**	***t***	***p***	**CI 0.025**	**CI 0.975**
Total Assets *(t)*	109.4623	1.934	56.606	0	105.672	113.253
Total Assets *(t-1)*	-107.8437	1.897	-56.85	0	-111.562	-104.125
Total Equity *(t)*	-8.6807	2.609	-3.327	0.001	-13.795	-3.566
Total Equity *(t-1)*	9.4331	2.662	3.544	0	4.216	14.65
Working Capital *(t)*	27.4897	2.768	9.93	0	22.063	32.916
Working Capital *(t-1)*	-21.6613	2.951	-7.34	0	-27.446	-15.877
Current Ratio *(t)*	-295.857	202.031	-1.464	0.143	-691.869	100.155

Current Ratio *(t-1)*	-72.8403	192.992	-0.377	0.706	-451.134	305.453
Solvency Ratio *(t)*	42.4468	154.136	0.275	0.783	-259.684	344.578
Solvency Ratio *(t-1)*	816.7946	410.634	1.989	0.047	11.889	1621.7
Number of Employees *(t)*	2271.6331	5231.152	0.434	0.664	-7982.222	1.25E+04
Operational Excellence *(t)*	6.33E+06	1.11E+07	0.57	0.569	-1.54E+07	2.81E+07
Egocentrism *(t)*	1.12E+07	1.04E+07	1.081	0.28	-9.12E+06	3.15E+07
Operational Excellence *(t)* * Egocentrism *(t)*	1.74E+10	5.93E+09	2.927	0.003	5.73E+09	2.90E+10

Table 8: Results of the multivariate regression analysis