PREDICTING HIGH-GROWTH FIRMS WITH MACHINE LEARNING METHODS

University of Jyväskylä
School of Business and Economics

Master's Thesis

2019

Author: Joosua Virtanen
Subject: Economics
Supervisor: Ari Hyytinen

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

ABSTRACT

Author
Joosua Virtanen

Title
Predicting High-Growth Firms with Machine Learning Methods

Subject Economics	Type of work Master's thesis
Date 02/25/2019	Number of pages 67

Abstract

Motivated by the recently grown political and commercial interest in high-growth firms (HGF) — in this master's thesis — I study whether common machine learning (ML) techniques are useful in predicting which privately owned companies become HGFs in the near future.[1] I employ the Eurostat-OECD definition of HGFs and study this question with a high-dimensional 2005–2016 panel data set of 13,602 unique Finnish firms, of which roughly 5% are defined as HGFs. I also study, which of the 24 predictors included matter the most for prediction. Finally, I examine whether an alternative definition of HGFs, predictors of expert information or studying a sample of young firms only will make a difference in predictive performance. I tackle the questions by developing a predictive scheme similar to a real forecasting scenario, where past values are used to train a set of classifiers, that can be employed to predict unknown future outcomes. Predictive performance is assessed in a separate test sample. My findings indicate that most ML methods offer moderate but statistically significant improvements over benchmarks, depending on the measure of interest. With an out-of-sample area under the ROC curve (AUC) of 0.6422 (equivalent to a 9.4% improvement over benchmark), the best working ML classifier — random forest (RF) — identifies 17.07% of the HGFs with only a 2.19% chance of misclassifying a non-HGF as an HGF. My analysis on variable importance and partial dependence suggests that the current values and past changes in firm size indicators alongside with firm age, contribute the most to predictive performance. Measuring the target variable in turnover rather than in employment improves prediction accuracy, where adding indicators of expert investor information as predictors does not yield any improvements. Finally, the prediction task seems to be considerably more difficult in a sample of young firms. In conclusion, ML methods should be considered for the challenging task of identifying HGFs, when computational costs and model interpretation are of secondary interest to prediction accuracy.

Keywords
High-growth firms, Prediction, Forecasting, Machine learning, Finland

Place of storage
Jyväskylä University Library

[1] I would like to thank Ari Hyytinen, Petri Rouvinen and Mika Pajarinen for interesting and useful conversations regarding the topic of this thesis.

TIIVISTELMÄ

Tekijä Joosua Virtanen	
Työn nimi Nopeakasvuisten yritysten ennustaminen koneoppimismenetelmillä	
Oppiaine Taloustiede	Työn laji Pro-gradu tutkielma
Päivämäärä 25.02.2019	Sivumäärä 67

Tiivistelmä

Kiinnostus nopeakasvuisia yrityksiä kohtaan on viime aikoina kasvanut politiikantekijöiden sekä sijoittajien keskuudessa. Tässä maisterin tutkielmassa tutkin, ovatko koneoppimismenetelmät hyödyllisiä tulevaisuuden nopeakasvuisten yrityksien ennustamisessa.[2] Tutkin tätä kysymystä laajalla 13602:n suomalaisen liikeyrityksen paneeliaineistolla vuosilta 2005–2016 hyödyntäen Eurostat-OECD:n nopeakasvuisen yrityksen määritelmää. Tällä määritelmällä aineistossa noin 5% yrityksistä sijoittuu nopeakasvuisiksi. Tutkin myös, mitkä yhteensä 24:stä ennustavasta muuttujasta myötävaikuttavat ennusteisiin eniten. Viimeiseksi tarkastelen, onko vaihtoehtoisella nopean kasvun määritelmällä, asiantuntijainformaatiota sisältävillä lisämuuttujilla tai vain nuorten yrityksien aineiston käyttämisellä vaikutusta ennustetarkkuuteen. Lähestyn kysymyksiä soveltamalla kehikkoa, joka muistuttaa todellista ennustusskenaariota, missä historiatietoihin perustuvalla aineistolla pyritään ennustamaan tulevaisuuden lopputulemia. Ennustetarkkuutta arvioidaan erillisessä testiaineistossa. Tuloksieni perusteella useimmat koneoppimismenetelmät mahdollistavat lieviä ja tilastollisesti merkitseviä parannuksia ennustetarkkuudessa verrattuna tavanomaisiin menetelmiin. Random forest (RF) -algoritmin opettama luokittelija toimii tässä kontekstissa parhaiten opetusaineiston ulkopuolisella AUC (ROC käyrän rajaaman pinta-alan) -arvolla 0,6422 (mikä vastaa 9,4% parannusta vertailuarvoon) ja tunnistaa 17,07% nopeakasvuisista yrityksistä vain 2,19% riskillä luokitella ei-nopeakasvuinen yritys nopeakasvuiseksi. Yrityksen koon nykyisen hetken ja menneen muutoksen indikaattorit yrityksen iän kanssa myötävaikuttavat eniten ennusteiden muodostamisessa. Kasvun mittaaminen käyttäen liikevaihdon kasvua henkilöstön kasvun sijasta parantaa ennustetarkkuutta. Toisaalta pääomasijoituksien ja yritystukien informaatiota sisältävien muuttujien lisääminen malliin ei paranna tuloksia. Viimeiseksi ennustusongelma osoittautuu vaikeammaksi nuorten yrityksien aineistossa. Yhteenvetona koneoppimismenetelmien soveltamista tulisi harkita nopeakasvuisten yrityksien ennustamisen haastavaan tehtävään, kun ennustetarkkuus on ensisijainen tavoite. Mikäli laskennallisilla kustannuksilla ja mallin tulkittavuudella on painoarvoa, koneoppimismenetelmät eivät välttämättä ole ylivertaisia tässä kontekstissa.

Asiasanat
Nopeakasvuiset yritykset, ennustaminen, koneoppiminen, Suomi

Säilytyspaikka
Jyväskylän yliopiston kirjasto

[2] Haluan kiittää Ari Hyytistä, Petri Rouvista ja Mika Pajarista mielenkiintoisista ja hyödyllisistä keskusteluista tämän maisterin tutkielman aiheeseen liittyen.

CONTENTS

LIST OF TABLES AND FIGURES

1 INTRODUCTION

The interest in high-growth firms (HGF) has increased extensively on behalf of policymakers, academics and private investors (Coad, Daunfeldt, Hölzl, Johansson, & Nightingale, 2014; Henrekson & Johansson, 2010). Any motion to directly target policy measures or investments towards potential HGFs requires being able to reliably identify them.

Policymakers have an interest in societal outcomes (such as employment and wages) and social and economic welfare, which are oftentimes originated from blooming business activities in the economy. Generally, firm growth is desired. Moreover, a small number of HGFs seem to create the most net jobs in economies (Henrekson & Johansson, 2010). Also, HGFs' have a tendency to lead technological innovations (Birch and Medoff, 1994), which are key factors for productivity, competitiveness and nationwide economic growth. Policy measures towards HGFs are likely to create welfare, which has created an urge to support potential future HGFs (European Comission, 2010). Two important questions have arisen (OECD, 2010): what policy measures should be used to foster HGFs, and which firms should be targeted for these measures?

The two questions above are both issues of interest and need further research, although, their nature is vitally different. The issue on specific policy measures is a question of causality, as in how and which measures will affect firm growth. The latter one requires prediction; which firms are the most probable to experience high growth in the future, and therefore the most potential targets for policy? This is what Kleinberg, Ludwig, Mullainathan, and Obermeyer (2015) call a 'prediction policy problem'. However, being able to predict HGFs does not mean that the growth of these firms can necessarily be affected with some measure of policy. Purely predictive methods provide probabilities of outcomes, but they do not take a stance on the more complex question of how to optimally allocate resources, which is central for policy decisions (Athey, 2017).

Identifying potential HGFs is of significant interest from a private investor's point of view as well. Private equity investors make investment decisions to maximize their expected return on investment, usually based on having a superior view on the firm's opportunities for growth. Investments in start-ups and potential HGFs are mostly evaluated using traditional discounted cash-flow analysis. However, the lack of historical data for individual companies complicates the reliability of this kind of assessment (Gompers, Gornall, Kaplan, & Strebulaev, 2016). Therefore, the field of private equity investments lacks robust prediction frameworks, for which the growing academic literature on identifying HGFs can perhaps fill a gap.

The HGF literature—based on regression studies—has not succeeded to extract accurate predictions of potential HGFs. Also, there is a considerable lack of studies implementing a truly predictive scheme. It has been argued that potential HGFs are impossible to predict due to their heterogeneous characteristics and stochastic nature of growth (Coad et al., 2014). Nevertheless, ML methods have proven effective in prediction policy problems in various applications and therefore provide chances on improving predictions in previously challenging tasks (Athey, 2018; Kleinberg et al., 2015; Mullainathan & Spiess, 2017).

Machine learning (ML) provides multipurpose tools from computer science and artificial intelligence. Supervised machine learning, a branch of ML, deals mostly with prediction problems, where learning algorithms are trained with historical data to identify complex relationships, which can be used to predict unknown outcomes. ML methods have proven capable of simultaneously fitting highly flexible functional forms to data, identifying previously unknown but generalizable patterns from it, and perform well in unseen data samples. ML has been applied, e.g., in vehicle steering, speech and image recognition, text classification and predicting stock exchange indices. Among other fields, ML is believed to have a significant impact on the field of economics in the near future. (Athey, 2018; Mullainathan & Spiess, 2017.)

ML methods are quite different from the ones used in econometrics. In economics, many applications involve estimating parameters for a set of variables affecting the outcome of interest. Moreover, the econometric approach is usually driven by some theoretical reasoning. While some parametric models exist in ML too, the main purpose of ML methods is solely to determine the outcome by letting data speak for itself. In fact, due to model complexity, parameters in ML models rarely have easily interpretable features such as in econometric models. (Mullainathan & Spiess, 2017.)

Another major difference lies in applying the methods. Regression models, for example, require numerous decisions on issues that significantly affect the outcome. These include the number and nature of variables to include, their transformations and whether to include interactions between some of them. ML methods are less sensitive to these issues and provide a rather consistent framework, where these decisions are made in a more transparent and autonomous manner. For instance, ML methods are designed in a way that the most significant variable interactions are learned independently from the data. Overall, ML methods provide potentially quality predictions at the cost of interpretability.[3] (Mullainathan & Spiess, 2017.)

[3] See, e.g., Varian (2014) for an introduction of applying some ML methods in economics.

In this master's thesis, I study, whether commonly applied ML methods can be applied to improve on HGF predictions provided by standard regressions and which predictors contribute the most. Therefore, my main research questions are the following:

1. Are ML algorithms able to improve prediction accuracy of HGFs compared to basic econometric models?
2. Which predictors matter the most and how are they related to the outcome?

Moreover, in my auxiliary analyses, I examine whether altering the HGF definition, adding predictors proxying expert information or studying a sample of only young firms will affect predictive performance.

My strategy for an ML analysis is the following. I compile and preprocess a large dataset of private Finnish firms filling the criteria for analysis, totaling 13,602 unique firms for training and 9,975 for testing. Moreover, the training sample consists of observations from 2008 to 2012, leaving the year 2013 as a single observation point for validation. This firm-level register data consists of 24 predictors based on the firm growth literature, which are used to predict a binary outcome of a firm experiencing high-growth in employment[4] or not. Importantly, the outcome is considered three years out of observation based on the OECD and Eurostat (2007) definition of HGFs, and some predictors three years back as growth rates. Next, I choose a set of well-working decision algorithms, which I train and tune the models with. Finally, I assess predictive performance in the separate test sample. My hypothesis is that ML methods provide more predictive power compared to conventional regressions, given the various ML applications where these algorithms have worked well.

The contributions of this master's thesis are the following. First, I add on to the scarce field of predictive studies identifying HGFs. Moreover—inspired by Weinblat (2018) and Sharchilev et al. (2018)—I present and implement an ML-based predictive paradigm similar to a real forecasting scenario, which is potentially capable of tackling the difficult task of identifying HGFs. As my second contribution, I provide reliable and diverse results with discussion to my research questions, which are useful for policy. Finally, as my last contribution, I propose a guideline for future research.

The rest of this master's thesis is organized as follows. In Chapter 2, I briefly review the relevant literature concerning firm growth, followed by describing the data and variables I use for analysis in Chapter 3. I develop and discuss the empirical framework in Chapter 4. Next, I report the prediction results in Chapter 5 and discuss the implications and limitations of the study in Chapter 6. Finally, I conclude this master's thesis in Chapter 7. Additional descriptive statistics, technical summaries and results are provided in Appendices A, B and C.

[4] I use 'personnel' as a synonym for employment throughout this thesis.

2 LITERATURE REVIEW

The literature on firm growth is vast and makes the topic undoubtedly one of the most researched in economics (for reviews see, e.g., Coad, 2007b; Davidsson, Achtenhagen, & Naldi, 2010; Henrekson & Johansson, 2010; Storey, 1994; Delmar, 2006; Machado, 2016; Wiklund, 1998). In this chapter, I do not intend to review the extensive firm growth literature comprehensively but will focus on the following relevant topics. First, to identify potential predictors of HGFs, I briefly review the empirical literature considering factors hindering firm growth. Next, I cover and summarize the recent literature on characteristics of HGFs to understand how my target of interest is expected to be represented and behave in the data. Third, the scattered literature on conceptual frameworks of firm growth is briefly revisited in order to verify what kind of predictions the theories provide. Finally, I provide a more detailed overview of the few studies that have taken some care to identify potential HGFs.

2.1 Determinants of Firm Growth

A wide variety of firm growth determinants have been studied in previous literature, which can be categorized into *internal* and *external* factors (Davidsson et al., 2010, p. 97). Popular subcategorization by Storey (1994) divides the internal factors to ensembles regarding *the firm*, *the entrepreneur* and the firm's *strategy*.

There is empirical support for many variables in the categories of internal factors by Storey (1994) to have influence on firm growth. Considering *the firm* category, according to evidence the author compiled from UK studies, *firm age, size* and *legal form* are related to firm growth. While most subsequent empirical studies confirm these findings for firm age and size, highlighting their negative relationship (Davidsson et al., 2010, p. 101–102) and overruling the Gibrat's law discussed in Chapter 2.3, a more recent study by Haltiwanger, Jarmin and Miranda (2013) finds that there is no clear relationship between size and growth when the firm age is controlled.

TABLE 1 Internal Factors of Firm Growth (Storey, 1994).

The Strategy of the Firm	The Firm	The Entrepreneur
Workforce training	Age	Motivation
Management training	Sector	Unemployment
External equity	Legal form	Education
Technological sophistication	Location	Management experience
Market positioning	Size	Number of founders
Market adjustments	Ownership	Prior self-employment
Planning		Family history
New products		Social marginality
Management recruitment		Functional skills
State support		Training
Customer concentration		Age
Competition		Prior business failure
Information and advice		Prior sector experience
Exporting		Prior firm size experience
		Gender

Variables in *the entrepreneur* category such as *motivation, education, management experience, number of founders and functional skills* have all positive effect on firm growth according to evidence compiled by Storey (1994). Many subsequent studies have been conducted that support these variables and suggest alternatives closely related to them. Alongside with the owner-manager's motivation, there is evidence of association to firm growth from the entrepreneur's goals and visions (Delmar & Wiklund, 2008), although most entrepreneurs have only modest ambitions towards growth (Human & Matthews, 2004). Evidence of other variables like the gender of the manager and prior sector experience influencing firm growth is mixed. (Davidsson et al. 2010, p.98-99.)

The overall evidence is not as robust and consistent in the category of the firm's *strategy* than in the two other ones of internal factors. Nevertheless, Storey's (1994) survey finds variables indicating *technological sophistication, market positioning* and *new product introduction* having a positive relationship on firm growth used more frequently than many other suggested variables. Truly, many other variables have been used. A complete listing of generally used internal factors based on Storey's (1994) survey is provided in TABLE 1 above.

Studies using models based on *external factors* of firm growth have emphasized, for example, *industrial and regional factors* as drivers for firm growth (Capon, Farley, & Hoenig, 1990; Davidsson, & Delmar, 2006). Moreover, support policies to firms in the form of *innovation grants* (Wallsten, 2000), *access to internal and external finance* (Becchetti & Trovato, 2002; Beck & Demirguc-Kunt, 2006; Carpenter & Petersen, 2002), *networking* and *alliances* (Barringer, Jones, & Neubaum, 2005) but also factors related to *market and demand conditions* (Coad & Tamvada, 2012; Kangasharju, 2000) have all been considered with supportive results. I have provided a listing of most commonly used external factors used in previous studies in TABLE 2 below. (Machado, 2016.)

TABLE 2 External Factors of Firm Growth (Machado, 2016).

External Factors of Firm Growth
Market and supply-demand conditions
Dynamism of the sector and entrance impairments
Investors and venture capital
Universities and mechanisms of transference of technology
Availability and access facility to resources
Availability of human resources and prime matter
Importance of stakeholders
Importance of family ties
Networks, alliances and firms´ network
Public policies and national or local support policies to firms

Altogether, the evidence on *external factors* of firm growth supports growth of the industry and increase in the dynamism of its region to have modest positive effects on firm growth. However, there are no unambiguous results on the effects of other environmental variables. While external factors have an evident role in firm growth, most of them are contextual and yield different results depending on the setting. (Davidsson et al., 2010, p. 107.)

Concluding the extent literature on factors hindering firm growth, some generalizations can be made. Internal factors seem to explain most of firm growth, but external factors have their undeniable role as well. However, Davidsson et al. (2010) argue that including the long list of potential factors of firm growth as explanatory variables with their interactions might be "…beyond the capacity of any researcher, or even the statistical software used" (p. 111). However, in the past decade, better machine learning techniques and more readily available computational capacity have alleviated the issue. All in all—for my predictive study—I have identified a massive list of potential variables for predictors.

2.2 Characteristics of High-Growth Firms

There is a considerable amount of literature studying characteristics of HGFs. Most of these studies are reviewed by Henrekson and Johansson (2010). Evidence for several research questions is mixed. However, there is robust evidence for seven distinct HGF characteristics or results (Coad et al., 2014), that are reported as stylized facts in TABLE 3 and described below.

The first characteristic deals with the distribution of HGFs. Several authors, including Bottazzi & Secchi (2006), have considered the heavy-tailed distribution of firm growth. HGFs have attained much attention at the right tail of the distribution, but high-decline firms have not received corresponding interest. The Second characteristic addresses the main motivation of this master's thesis: HGFs create a large share of new jobs. There is loads of evidence for this result for various countries (see, e.g., Acs & Mueller, 2008; Birch & Medoff, 1994; Davidsson & Henrekson, 2002; Delmar, Davidsson, & Gartner, 2003) and also for Finland (Littunen & Tohmo, 2003). In numbers, HGFs represent on 3-6% of the total firm population when growth is measured by employment using the OECD definitions for HGFs (Hoffman & Junge 2006).[5]

According to the third result, HGFs tend to be young but not necessarily small (Acs, Parsons, & Tracy, 2008; Daunfeldt, Elert, & Johansson, 2014), whereas the fourth one revokes the previously prevailing idea of HGFs being more common in high-tech industries (Henrekson & Johansson, 2010). In fact, service industries appear to be more HGF intensive. The fifth characteristic states that high growth is not persistent over time (Delmar et al. , 2003; Hölzl, 2013), however, there is some evidence opposing this finding (see, e.g., Acs et al., 2008). Related to the previous characteristic, there is some evidence denying the existence of the so-called 'survivorship bias' (Weinblat, 2018), where HGFs have gained their status by excessive risk-taking and are also likely to fail. The sixth result argues that there is a clear trade-off between defined HGFs regarding different growth measures (Daunfeldt et al., 2014). Finally, the seventh characteristic addresses the hard predictability of future HGFs (Coad et al., 2014), which is another essential motivator for this master's thesis.

TABLE 3 Characteristics of high-growth firms as stylized facts (Coad et al., 2014).

Stylized Fact		Evidence
(1)	Growth rate distributions are heavy-tailed.	Bottazzi and Secchi (2006)
(2)	HGFs create a large share of new jobs.	Henrekson and Johansson (2010)
(3)	HGFs tend to be young but not necessarily smaller than average.	Acs et al. (2008)
(4)	HGFs are not overrepresented in high-tech industries.	Henrekson and Johansson (2010)
(5)	Persistence of high-growth depends on the measure of growth.	Delmar et al. (2003) Hölzl (2013)
(6)	Using different growth indicators selects different sets of firms.	Daunfeldt et al. (2014)
(7)	Future HGFs are difficult to predict.	Coad et al. (2014)

While it has become a policy goal to directly target HGFs for policy and develop environments to enable firms to reach the transitory phase of high growth, not very much is known about the qualities of these firms or the determinants of high growth, not to mention their theoretical features (OECD, 2010). It is rather apparent that high growth must depend on multiple factors and their interactions. In fact, many of the suggested determinants for high-growth are similar to Storey's (1994) factors of firm growth. After the overview above, I

[5] The OECD definitions (*OECD-Eurostat Manual on Business Demography Statistics* 2007) are revised in Chapter 1 of this master's thesis.

review a few other studies providing more evidence towards a comprehensive view of HGFs and using a particular set of variables when predicting HGFs.

Barringer et al. (2005) aim to identify features of HGFs using a qualitative approach, analyzing narrative descriptions of 50 HGFs compared to descriptions of 50 non-HGFs.[6] Examining attributes based on founder characteristics, firm attributes, business practices and HRM practices, they find that HGF founders have higher education, more experience and compelling story than non-HGF counterparts. In firm attributes, HGFs have credibly committed to growth compared to non-HGFs and in business practices, they add unique value and know their customers better than non-HGFs. Finally, in HRM practices, HGFs underline continuous learning and financial solutions significantly more than their counterparts.

In OECD's (2010) report, a set of links between high-growth firms and suggested high-growth factors are investigated through multiple studies. Motivated by the success stories of high-tech firms, these links include the firm's ability to innovate, manage intellectual assets, its networking activities and business practices, and finally its access to finance. The links are studied through seventeen ad-hoc studies in various countries, Finland included. Main findings indicate that high-growth is a temporary phase in the firm's life cycle that is, quite surprisingly, not dependent on its age, size or sector. Some country studies find a correlation between innovative activities and high growth, but results are neither universal nor in the scope of causal inference. Finally, the financing needs of innovative high-growth firms are different from the ones of an average firm, but there is not a credit rationing problem among these firms. Concisely, from the investigated perspectives it still seems difficult to identify high-growth firms based on the listed determinants.

Considering the study on Finland in the same OECD's (2010) report, especially the innovation activities and business practices were studied using telephone surveys of 170 firms. No significant relationship between the level of innovation and growth rates were found. On managing intellectual assets, Finnish research found that firms have adopted numerous ways to minimize losses due to damage on intellectual property. These were categorized as intellectual property rights, contracts and informal protection methods. However, the research found that firms are not very experienced in practicing these activities. Neither networking nor barriers to finance were examined for Finland in this study.

A subsequent review of high-growth determinants is provided by Audretsch (2012). For the most part, the author leans on traditional firm growth literature but points out the factors that seem to be associated with higher growth rates according to the evidence. The determinants are categorized by the author to ones at the firm level and at the locational level. Moreover, in the first one, the author considers characteristics of the entrepreneur and the founding team, entrepreneur's gender, market orientation, access to resources, human and social capital, financial capital and finally intellectual property the most

[6] The authors define an HGF as a firm experiencing at least 80% growth in sales over a three-year period. The data of their analysis consisted of winning firms of the Ernst & Young LLP Entrepreneur of the Year award competition.

meaningful for HGFs. Locational characteristics here include geographical clusters as one big factor of unanswered questions due to the paucity of research in the field. Nevertheless, evidence suggests that location has a reasonable role in the process of high growth.

Entrepreneurial quality has also been recognized as an important factor for the success of firms. However, defining and measuring entrepreneurship is not an easy task. Guzman and Stern (2015a) contribute to this challenge by developing two indices of entrepreneurial quality: EQI (Entrepreneurial Quality Index) and RECPI (Regional Entrepreneurship Cohort Potential Index) and use them for placecasting and nowcasting quality and growth events. With 1988–2014 data from firms of the state of Massachusetts, the authors implement a series of regressions where the number of growth outcomes per year is considered as a dependent variable. The results are promising, with both indices being statistically significant but also with high elasticities: "Doubling RECPI is associated with more than a 50% increase in the number of expected growth events in a region-cohort-year." (p. 37).

2.3 Conceptual Frameworks

Theoretical firm growth literature is widely scattered. Moreover, theoretical predictions have been of little use in understanding firm growth. Also, theoretically deductive reasoning is not too relevant for an ML based analysis such as the one carried out in this master's thesis (as discussed in Chapter 1). Nevertheless, for a general view and understanding of how different frameworks relate to HGFs, I briefly review a few theoretical concepts discussed by Coad (2007b).

A traditional theoretical discussion on firm growth is about Gibrat's law (Gibrat, 1931), which in its simplest form states that the firm's expected growth rate is independent of its size at the beginning of the period at hand. According to empirical evidence discussed previously in this chapter, it turns out however, that the evidence for Gibrat's law is mixed but generally not supported.

The neoclassical foundations of growth in the context of firm growth states a prediction that firms are attracted towards some optimal size through profit maximization (Viner, 1952). Therefore, growth is seen as the means towards a goal, not as the goal itself. However, the concept lacks empirical support and is therefore of little use in understanding firm growth.

Penrose's seminal book on the theory of the firm (Penrose, 1959) introduced new concepts called 'economies of growth' and the 'Resource-based view' of the firm. The first one implies that firms have strong incentives to grow, which are generated by a process where productivity increases automatically due to managers' increasing expertise over time. In parallel with a faster rate of firm growth, the operating costs evolve. The second concept considers that firm's performance depends on its continuous capabilities of creating and managing resources. Therefore, Penrose's firm grows because of its ability to adapt along the dynamic process of growth. While Penrose's contribution has mostly been confined in the industrial organization literature, its ideas are intuitive in economics as well. In the context of HGFs, Penrose's theory suggests that high growth requires lots of resources, but also abilities to manage them. Generally, growth is generated through incentives through learning-by-doing, which implies that faster learners are in a more probable position to grow faster.

In the managerial approach by Marris (1963, 1964), the manager is seen in a fundamentally important part as maximizing the utility function with respect to firm growth and profit. Moreover, the manager reaches for the highest possible growth rate of the firm subject to the constraint of earning a sufficient profit appealing enough to shareholders. In this approach, the growth of young small firms is in line with profit maximization, whereas in the case of other larger firms, the manager has to balance between the two objectives. The managerial approach, therefore, suggests that high-growth is more probable in a sample of young small firms.

Evolutionary theory and the principle of 'growth of the fitter' is based on Schumpeter's vision of 'creative destruction', borrowing the notions of diversity creation and selection to explain economic development. Developed by Downie (1958), the theory argues that fitter firms survive and grow, whereas weaker firms leave the market. However, the theory assumes that firms grow by reinvesting their earnings, and therefore growth rates rise alongside profitability. This is empirically a problematic assumption since no such relationship between profits and growth in data is usually found.

The last framework considered here is the population ecology approach based on the work of Hannan and Freeman (1977). In this approach, organizations require resources that are unique and scarce at each niche market. Each niche has a carrying capacity, which was to get full, growth opportunities would cease to exist. Finding a new niche with abundant resources would result in a lot of growth. After initial discovery, new firms will enter the niche and through competition, the resources and opportunities for growth will equally run out for each firm in that niche. Empirically speaking this theory alone is not directly supported. While there is some evidence on different growth rates between industries, the growth rates differ significantly inside industries as well, questioning the direct implications of the theory. Averagely speaking though, population ecology gives an intuitive idea for the mechanism of how some newly founded small firms become HGFs or how large firms can achieve or maintain high growth by searching new sources of niches.

2.4 Identifying High-Growth Firms

The literature on identifying HGFs is scarce and struggling with the difficulty of the task in mostly regression-based studies. However, some recent applications have demonstrated the usefulness of ML in predicting HGFs and therefore provide meaningful benchmarks for this master's thesis. The related studies are summarized in TABLE 4 below.

Starting with a few regression-based studies, Sampagnaro and Lavadera (2013) contribute to the literature on predictors of HGFs. Motivated by the inverse rationale of the theoretical prediction of the credit scoring model, the authors examine balance sheet ratios as predictor candidates of HGFs in three regression models: quantile regression and Tobit model with random and fixed effects. Using Italian AIDA data of 21,182 firms from 2001 to 2008, they find that in addition to the apparent firm size and age, internal cash flows is the most relevant predictor across models. These conclusions are based on statistical tests on the model coefficients and using a distributional high-growth definition of a firm belonging to the top 10% of its industry measured by sales growth. However, no analysis on prediction performance is reported.

Megaravalli and Sampagnaro (2018) have a similar goal of identifying the most important predictors of HGFs from balance sheet ratios with a probit model. They develop Sampagnaro and Lavadera's (2013) analysis with a more recent (2010–2014) data set of Italian firms considering only family businesses totaling 45,000 firms in the analysis. Also, their definition of HGF is based on 20% p.a. sales growth for two consecutive years after a year without such growth. Predicting HGFs with the previous year's observations, results imply that the most important financial indicators are liquidity ratio, solvency ratio, firm age, cash flows and working capital. The model's predictive performance is also reported with an AUC of the ROC curve of 0.7078. However, the model is assessed only in-the-sample, which is typical for a variable importance analysis but not reliable for assessing and comparing model performance.

A few recent studies have implemented machine learning algorithms to predict HGFs alongside other relevant outcomes. One example is provided by Miyakawa, Miyauchi and Perez (2017), where the authors predict firm's exit, sales growth and profit growth using a weighted random forest algorithm with data of over 1.7 million Japanese firms from 2006–2014. With predictors based on firm characteristics, geography and industry, supply-chain network and a solvency score, they are able to reach an out-of-sample area under the ROC curve (AUC) of 0.68 and identify 25% of high growth firms with a fixed probability threshold. This approach clearly outperforms a model with just the solvency score as a predictor. Their target variable is based on a high-growth definition of a firm exceeding the average growth of the forecast period plus one standard deviation. However, the authors do not provide any benchmark results with conventional methodology to mirror these results against. Their paper supports a concept of using an ML method in firm performance prediction, but the results provide only little internally comparable value.

Weinblat (2018) provides a relevant ML-based approach. The author uses a random forest algorithm with 15 structural and financial predictors to forecast European high-growth firms and determine the most relevant predictors for them in nine different countries, including Finland, covering 179,970 firms in total. With a recent (2004–2014) data set from Amadeus -database, the author reports the best out-of-sample prediction results for Great Britain with an area under the ROC curve (AUC) of 0.8110. The author confirms this result with Venkatraman's statistical test, comparing the ROC curves between the studied countries. The results for Finland are not as glamorous but still fair with an out-of-sample AUC of 0.6439. Weinblat (2018) uses a distributional high-growth definition of a firm belonging to the high-growth class if its Birch-Schreyer growth indicator of employment is within the top 10% of the sample. The Birch-Schrayer indicator considers both the absolute and relative components of employment growth.

In a supplementary analysis, Weinblat (2018) doesn't find clear differences in predictability over different size groups of firms. The random forest algorithm also provides a tool to assess variable importance, on which the author's results are in line with the literature. The most important predictors across countries are the firm's size, past growth and age. Concluding, Weinblat (2018) notes that out-of-sample predictions of HGFs are not outside of our capabilities anymore but the predictability of HGFs varies across countries, and therefore results should not be generalized across them. Also, many country-specific model improvements can and should be made in terms of included features and algorithms.

Sharchilev et al. (2018) predict startup success during the early stages of their life cycles by successfully including web-based information combined with a highly sophisticated ensemble ML framework. Moreover, the algorithm suggested by the authors, named WBSSP, combines logistic regression, neural networks and CatBoost, a high-performance boosting algorithm. The authors model the prediction task by classifying whether a company that has already received initial funding will obtain another round of investment in a given period. They conduct their analysis with international data of 21,947 privately-owned companies with basic features and combined with web mentions from Crunchbase, LinkedIn and open web sources up until 2017. The authors report impressive out-of-sample performance results with statistically significant AUC of 0.854, precision of 0.626 and F-score of 0.383. These are 6.75%, 131.9% and 83.3% higher than with the current state-of-the-art ensemble algorithm benchmark. In addition to presenting an outperforming ML framework, Sharchilev et al. (2018) show clear evidence of a performance boost including web mentions compared to models with structured data only.

TABLE 4 Summary of studies related to predicting HGFs.

Study	Data	Method	Main Results
Sampagnaro and Lavadera (2013)	Basic features of 21,182 firms from 2001–2008	Quantile regression, Tobit model (random and fixed effects)	Firm size, age and internal cash flows are the most relevant HGF predictors
Megaravalli and Sampagnaro (2018)	Financial data of 45,000 Italian family businesses from 2010–2014	Probit model	In-sample AUC of 0.7078 with most important financial predictors being liquidity ratio, solvency ratio, cash flows and working capital in predicting HGFs
Miyakawa, Miyauchi and Perez (2017)	Basic features of over 1.7 million Japanese firms from 2006–2014	Weighted random forest algorithm	Out-of-sample AUC of 0.68 and in predicting HGFs
Weinblat (2018)	Basic features of 179,970 firms (for Finland, France, Germany, Italy, Portugal, Spain, Great Britain, Poland and Sweden) from 2004–2014	Random forest algorithm with SMOTE resampling	Out-of-sample AUC of 0.8110 for UK and 0.6439 for Finland with most important predictors being related to size, its variation and firm age in predicting HGFs
Sharchilev et al. (2018)	Basic features of 21,947 firms combined with web mentions from LinkedIn, Crunchbase and open source sites up until 2017.	WBSSP algorithm (a combination of CatBoost, neural networks and logit)	Out-of-sample AUC of 0.854 in predicting start-up success

3 DATA AND VARIABLES

In this chapter, I describe the data and variables used in this master's thesis starting with data sources. I continue with defining HGFs and describing their distribution in the data. Finally, I define a set of predictors that will be used in the predictive models. Tables listing all the variables and providing descriptive statistics are reported at the end of this chapter.

3.1 Data Sources

I have compiled a 2005–2016 panel data set of firms in Finland by combining four data sources: the official Business Register by Statistics Finland[7], the financial statements database of Suomen Asiakastieto Oy[8], Business Finland's (formerly Tekes) on its public R&D grant recipients [9] and the Finnish Venture Capital Association's (FVCA) records on companies that have attracted private equity investments.[10] I have limited my full data set to include observations of privately owned limited liability companies, which (a) employ at least ten persons, (b) are in the national Value Added Tax register, and (c) are included in the Tax Administration's Employer Register (at the time of observation). Following these definitions, I have longitudinal data of 16,333 firms, totaling almost 60,000 observations when including the longitudinal dimension. After omitting observations with missing data and after preprocessing[11], I am left with 14,714 unique firms to work with.

For my ML analysis, I divided the data into a learning sample (LS) for training and a left-out test sample (TS) to assess predictive performance. In my approach, the learning sample consists of values of predictive variables from 2005 to 2012 and observations of high-growth from 2008 to 2015. The test sample consists of more recent predictor values from 2010 to 2013, which are used

[7] The Finnish public authority established for statistics https://www.stat.fi/index_en.html.
[8] A Finnish information services company https://www.asiakastieto.fi/web/fi/.
[9] The Finnish public authority financing innovation https://www.businessfinland.fi/en/do-business-with-finland/home/.
[10] Association for venture capitalists in Finland http://paaomasijoittajat.fi/.
[11] The preprocessing of data among other methodological considerations are addressed in Chapter 4.3.

to predict the outcomes between 2013 and 2016. As a result, I have created a predictive scheme similar to a real forecasting scenario at the end of the year 2013. Here, the learning sample contains 81% of the data. For the last auxiliary analysis, I compile an additional data set, based on the full data, but including only young firms that are ten years old or younger.

3.2 Variable Definitions

3.2.1 Defining High-Growth Firms

Studying high-growth as a binary outcome, the definition of high-growth plays a significant role. Delmar et al. (2003) point out how researchers should acknowledge the fundamental differences and possibly different results based on the definition and measure of growth used. There is not a universal approach for determining HGFs up to date, but some definitions have gained ground in research. Among the most popular ones, OECD and Eurostat (2007, p. 61) define HGEs as "All firms with average annualized growth greater than 20% per annum, over a three-year period". This trajectory is equal to total growth of 72.8% over the three-year period. The growth can be measured in employment or turnover. In addition, they recommend a size threshold, such as firms with at least ten employees at observation, to be set in order to reduce distortion due to small firm growth. I employ the Eurostat-OECD definition of HGFs measured in employment with the ten employees size threshold for my baseline model. The turnover measure is applied in the first auxiliary analysis.

Daunfeldt, Johansson and Halvarsson (2015) note that it would be meaningful to standardize the high-growth measures in the HGF literature and the Eurostat-OECD definition is probably the closest to a standard measure due to its popularity. However, the authors give a cautionary note on using the definition. They find that using the definition will exclude approximately 95% of survived firms and 39% of the created jobs on Swedish data. Therefore, policy based on this definition might be misleading or counterproductive if the applier is not aware of the details of this definition. The possible pitfall, in their opinion, concerns the threshold of including firms with at least ten employees, which seems rather high. Of course, including the smallest firms would create bias in growth results and the very smallest firms are less innovative too. Perhaps the definition recommendation should be reconsidered by the authors, nevertheless.

Similar to Daunfeldt et al.'s (2015) finding, defining HGFs as above yields rather imbalanced class distributions for Finnish data, as reported in TABLE 5 below. The imbalance composes some methodological challenges discussed in Chapter 1. In the baseline model, the proportion of HGFs is under 5% in LS and TS. The measure in turnover classifies approximately 7.5% of firms as HGFs, and finally, the corresponding proportion is about 10% for young firms using the growth measure in employment. These distributions are reported alongside with numbers of unique firms in LS and TS for different models in TABLE 5.

TABLE 5 Number of unique firms in learning- and test samples and proportion of HGFs across different models.

Model	Learning sample		Test sample	
	Firms	High-growth firms	Firms	High-growth firms
Personnel	13,602	4.54 %	9,975	4.93 %
Turnover	13,602	7.56 %	9,975	7.49 %
Young	3,792	9.50 %	2,004	10.93 %

Notes: The personnel model is the baseline model of this thesis. The turnover model is provided as an auxiliary analysis. Finally, the young model uses the personnel growth definition of high growth but for a smaller dataset of only young (≤ 10 years old) enterprises. The model for expert information uses the same dataset and has the same definition of HGEs as the baseline model.

3.2.2 Predictors

I approach model selection by looking at the literature on factors of firm growth, mostly relying on Storey's (1994) categorization of internal factors and Machado's (2016) listing of external factors hindering growth. In a data-driven ML analysis, such as this one, as many relevant predictors should be included as possible. For the baseline analysis, I include 24 predictors in total, which are defined below. Descriptive statistics and summarizing table of targets and predictors are provided in TABLE 6 and TABLE 7 below. The statistics for preprocessed data and the sample of young firms are reported in Appendix A.

The age of a firm (*Age*) is measured from the founding date of its first establishment. The size of a firm is measured by the number of its full-time equivalent workers (*Personnel*), sales in euros (*Revenue*), and productivity is proxied by its sales divided by personnel (*Productivity*). For *Personnel*, *Revenue* and *Productivity*, I also include *lagged 3-year growth rates,* using the 'Davis, Haltiwanger & Schuh' definition of centralized growth for *Personnel* and in logarithmic differences for the last two. To control for branch and location, I add categories for twenty sectors (*Industry*) and sixteen regions (*Ely*). I also include a binary indicator for foreign ownership (*ForeignOwned*).

Furthermore, I consider a few predictors related to business strategy. These include the number of places of business (*NumOfPos*) and a binary variable for being part of a group of businesses (*PartOfAGroup*). To control for the intensity of competition in the industry, I include the top decile (*TopDecGrossMargin*; a higher value indicates a less competitive industry) and the above-median (*MedGrossMargin*; a higher value indicates a more competitive industry) cumulative gross margins in a firm's three-digit industry. The role of innovation is captured by a firm's cumulative count of *EPO* patents (*PatCount*) in the past three years. In addition, I also include indicators for having foreign subsidiaries (*ForeignSubsidiaries*) and being an exporter (*Exporting*).

I have a few predictors related to the entrepreneur or the top manager. Among the predictors, I include the age of the person listed as the CEO of the company (*CEOAge*) and that person's gender (*CEOGender*).

A few common financial indicators are included as predictors: profitability, measured by operating result-% (*Profit*) and its difference in the past three years (*ProfitGrowth*); financial strength, defined as the amount of equity divided by assets (*Solidity*); and capital intensity (tangible assets divided by revenue, *TangAssetsPerRev*). I also include a firm's credit rating (*Rating*).

To study any further information employed in private and public investors' decisions, I have two final variables. The first is a binary variable, which indicates that a firm received private venture capital finance (*Vc*) at or prior to the year of observation. The second variable indicates that a firm received a public R&D grant at or prior to the year of observation (*Tekes*).

Therefore, I have a rather extensive set of predictors to consider, although some—often survey-based—measures suggested in earlier literature could not be included. My list of predictors in *the firm* category of internal factors is particularly comprehensive. Disappointingly, I have only two predictors in *the entrepreneur* category, which seems to be one of the most important for HGFs in the literature. Fortunately, I have a fair set of predictors in the category of the firm's *strategy*. The external factors are slightly underweighted, but their role in the literature is also more or less in the background.

TABLE 6 Descriptive statistics for the full data set before preprocessing.

Variable	n	mean	sd	min	max
HighPersonnelGrowth	59,915	0.05	0.21	0.00	1.00
HighRevenueGrowth	59,915	0.08	0.26	0.00	1.00
Personnel	59,915	67.08	295.79	10.00	15,976.00
LagPersonnelGrowth	59,880	0.19	0.56	-5.54	7.94
Revenue	59,476	21,335,539.70	228,165,626.51	0.00	26,940,000,000
LagRevenueGrowth	59,117	0.23	0.64	-8.19	10.07
Productivity	59,476	250,274.60	1,075,062.60	0.00	173,543,824.00
LagProductivityGrowth	59,082	0.04	0.44	-7.52	7.31
Profit	59,579	1.95	90.18	-9,033.3	3,666.70
ProfitGrowth	59,352	-0.05	123.36	-8,083.3	9,536.50
Age	59,915	21.58	11.26	1.00	113.00
NumOfPos	59,915	2.72	13.00	1.00	770.00
PartOfAGroup	59,915	0.36	0.48	0.00	1.00
ForeignOwned	59,915	0.10	0.30	0.00	1.00
Exporting	59,915	0.26	0.44	0.00	1.00
Solidity	59,910	38.64	52.41	-3,700.00	100.00
Rating	58,191	19.47	18.15	3.00	99.00
TangAssetsPerRev	59,447	0.33	24.35	-0.01	5,897.37
TopDecGrossMargin	59,915	23.55	7.96	-43.70	200.00
MedGrossMargin	59,915	8.08	9.15	-1,996.15	33.40
PatCount	59,915	0.40	23.10	0.00	2,421.00
ForeignSubsidiaries	59,915	0.15	0.36	0.00	1.00
CEOAge	55,811	48.88	9.11	20.00	89.00
CEOGender	56,506	0.09	0.29	0.00	1.00
Vc	59,915	0.04	0.20	0.00	1.00
Tekes	59,915	0.11	0.31	0.00	1.00

Notes: Observations with missing values: 6408.

TABLE 7 Listing and descriptions of all variables used in this master's thesis by category.

Variables	Description	Change (Δ) / Value (t_0) / Categorical (C)
Target variables		
HighPersonnelGrowth	Binary: Leaded 20% p.a. growth for 3 years (OECD).	Δ+
HighRevenueGrowth	Binary: Leaded 20% p.a. growth for 3 years (OECD).	Δ+
Predictors		
Internal Factors		
The Firm		
Age	Age of the firm from the first location.	t_0
Personnel	Number of Personnel.	t_0, Δ-
Revenue	Turnover in euros.	t_0, Δ-
Productivity	Revenue/Personnel.	t_0, Δ-
Industry	Categorical variable for 20 different industries.	C
Ely	Categorical variable for 16 Ely regions.	C
ForeignOwned	Binary variable for foreign ownership.	C
Strategy		
NumOfPos	Number of places of business.	t_0
PartOfAGroup	Binary variable for being part of a group.	C
TopDecGrossMargin	Top decile gross margin by tol3 industry.	t_0
MedGrossMargin	Median gross margin by tol3 industry.	t_0
PatCount	Cumulative number of patents from the past three-year period.	t_0
ForeignSubsidiaries	Binary variable for having foreign subsidiaries.	C
Exporting	Binary variable for any international exporting.	C
The Entrepreneur		
CEOAge	CEO's age at the time of observation.	t_0
CEOGender	Binary variable for the CEO being a woman.	C
Financial Key Figures		
Profit	Business profit in percentages.	t_0, Δ-
Solidity	Equity/Assets.	t_0
TangAssetsPerRev	Tangible Assets/Revenue.	t_0
Rating	Rating points (1-100, descending order).	t_0
External Factors		
Vc	Binary variable for received venture capital finance.	C
Tekes	Binary variable for received innovation grants lagged 3 years back.	C

Notes: Δ+ (Δ-) stands for leaded (lagged) change three years forward (back).

4 EMPIRICAL FRAMEWORK

In this chapter, I describe the strategy and methodology used to train and tune the classifiers, approaches to validate and assess their performance and tools to evaluate variable importance. My discussion here is only a very concise summary of the applied methodology based on Hastie, Tibshirani and Friedman (2009), which is adequate for implementation purposes.[12] For a technical summary of the ML algorithms, see Appendix B.

The predictive paradigm applied here uses a validation set approach, dividing the full data set into a learning sample (LS), which will be used for training and a test sample (TS) for assessing predictive performance. The prediction results reported in Chapter 5 are acquired in the TS, which makes them reliable estimates of the true out-of-sample performance. I employ five different machine learning algorithms to train classifiers, which are the following:

- classification and regression trees (CART),
- bootstrap aggregation (bagging),
- boosting,
- random forests and
- artificial neural networks (ANN).

Also, I combine the predictions of ML classifiers to create a simple ensemble classifier. The linear probability- and logit models are used as benchmarks.[13]

4.1 Decision Algorithms

4.1.1 Classification and Regression Tree (CART)

Classification and regression tree (CART) is a tree-based algorithm for regression and classification, however, I describe only the classification version here.[14]

[12] For an introductory approach, see, James, Witten, Hastie and Tibshirani (2013).

[13] All the models are trained and evaluated using the caret package by Kuhn (2018) in the statistical software R version 3.5.1 by R Core Team (2018).

[14] I implement the CART algorithm by the 'rpart' method (Therneau & Atkinson, 2018) in the caret package.

Breiman (2017) provides a full and modern description of the CART algorithm.[15]

CART entails two main steps to make predictions. The first step is to use recursive binary splitting to stratify the predictor space into K distinct regions, $R_1, R_{2, \ldots,} R_K$. Beginning at the top of the tree, the goal is to split the predictor space into subsamples by a set of decision rules, which determine how the splits are made and when the tree is finished. In classification, the algorithm chooses predictor x_j and a cut point s for a split to minimize a measure of node purity, the gini index, at each node. The splitting continues recursively until a stopping rule of minimum node size is reached. The final nodes are referred to as the terminal nodes. A classification tree can be illustrated in the form of a tree chart presented FIGURE 1 below. Based on the classification task of this master's thesis, I have presented a chart of a fully-grown decision tree in Appendix C.

The second step is a simple one. By using the most common class of the response values in each of the regions $\mathrm{R}_1, \mathrm{R}_{2, \ldots,} \mathrm{R}_K$, one can make predictions of the target variable for any test observation. In other words, the same prediction is made for each observation in the same region.

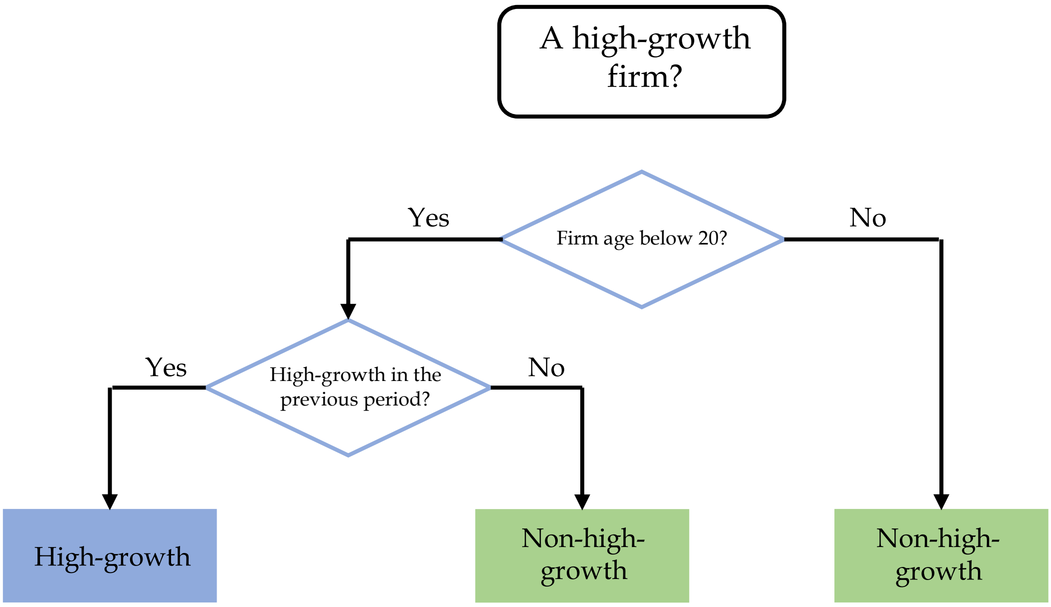

FIGURE 1 An imaginary tree chart example of a simple decision tree.

Tree-based methods don't require any distributional assumptions, which makes them rather safe to implement. In addition, simple tree models are easy to interpret with a single chart. On the other hand, simple tree algorithms like CART are known to be sensitive to changes in predictor space and hyperparameter values. High variance usually leads to unreliability and poor performance in the test sample. The CART algorithm works for both regression and classifica-

[15] This subchapter is based on Breiman (2017) and Hastie et al. (2009, p. 305–312).

tion problems. However, the same tools cannot be used to assess model performance for the two tasks. I will describe the preferred measures to evaluate predictive classification models later in this chapter.

4.1.2 Bootstrap Aggregating Predictor (Bagging)

Bootstrapping, a powerful statistical tool can be used to improve the performance of decision trees, such as the CART. This applied method is generally called the bootstrap aggregating predictor, or just bagging, as proposed by Breiman (1996). Bagging leads to reduced variance and enhanced predictive performance compared to single CARTs.[16]

The idea of bagging[17] is to create several training sets from the initial sample, train separate CARTs for each set, and finally aggregate the outcomes for prediction. Since several training sets are usually not available, the training sample is bootstrapped. In other words, repeated, equal-sized resamples are taken from the learning sample and are then used separately for model training. After training, predictions for classification problems are generated in a voting process, where each bootstrapped tree has a single vote. The majority class is the final prediction for a test observation. A schematic figure of the bagging algorithm in classification tasks is provided in FIGURE 2 below.

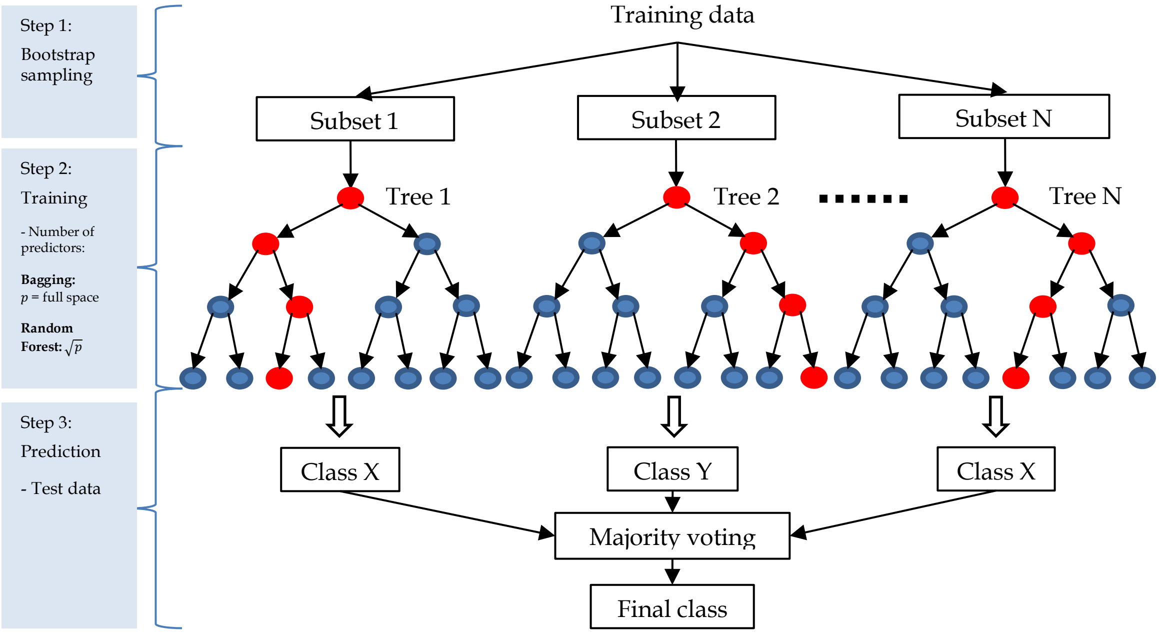

FIGURE 2 Schematic of bagging and random forest classifiers.

Notes: The figure design is inspired by He, Chaney, Schleiss and Sheffield (2016, p. 8220).

[16] This subchapter is based on Breiman (1996) and Hastie et al. (2009, p. 282–283).

[17] I implement a bagged CART algorithm by the 'treebag' method in the caret package. Packages ipred (Peters & Hothorn, 2018), plyr (Wickham, 2011) and e1071 (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2018) are employed.

4.1.3 Random Forest

Random forest (RF) is a modification of a bagged decision tree that further improves the method by decorrelating the bootstrapped trees, as proposed by Breiman (2001).[18]

RF is implemented by allowing the algorithm to use only a small random sample of predictors as candidates for a split at each node of the tree. A new sample of m predictors is considered at each split, and the sample is usually chosen to be equal to square root of the total number of predictors p. After training the trees with this decision rule modification, predictions can be made similarly as described for bagging, both for regression and classification problems.[19] The schematic of the random forest predictor is similar to the one of bagging and is therefore combined in FIGURE 2 above. The differing part is pointed out in step two, where RF uses a subset of predictors instead of the whole predictor space at each split.

The random forest algorithm provides a significant advantage over bagging regarding reduction in variance, especially if there is a dominating predictor in the model. The bagging predictor is likely to choose the strongest predictor for the top split in all its bootstrapped trees, growing a number of similar trees. Aggregating several, almost equal trees does not result in reducing variance by a lot. The random forest algorithm, however, considers smaller subsamples of predictors at each split, and therefore, it is not as likely for the algorithm to choose a dominating predictor for the first split. This modification decorrelates the underlying trees, and aggregating those trees reduces variance, improving the overall prediction performance.

4.1.4 Boosting

Like bootstrap methods, boosting is a general method that can be applied to various statistical learning contexts. It is based on the same fundamental idea of aggregating several weak learners into a strong learner. Freund and Schapire (1997), Friedman (2001) and have developed applications of the boosting algorithm compatible with classification and regression trees. Here, I describe boosting as it can be used with decision trees in a classification setting.[20]

In boosting, the training sample is modified for each tree to grow on. As in bagging and random forests, there is a large set of small decision trees, weak learners, which are combined to create a strong learner. However, this time, learning is performed in sequences, not independently. Boosting algorithms learn slowly, using information gathered on the way.

There are three steps to train a model with a boosting algorithm, of which the last two are repeated for a sequence of decision trees. First, a small classification tree, a base learner, is fitted, giving equal attention to each observation.

[18] This subchapter is based on Breiman (2001) and Hastie et al. (2009, p. 587–602).

[19] I implement a random forest algorithm by method 'ranger' in the caret package. Packages ranger (Wright & Ziegler, 2017), dplyr (Wickham, François, Henry, & Müller, 2018) and e1071 (Meyer et al., 2018) are employed.

[20] This subchapter is based on Freund and Schapire (1997), Friedman (2001), Chen and Guestrin (2016) and Hastie et al. (2009, 337–380).

Next, the algorithm assigns more weight to misclassified observations. Finally, another tree is grown with the newly assigned weights. This iterative process is repeated for the last two steps, slowly improving the model in weakly performing areas until growing the last decision tree, determined by a stopping rule. In my analysis, I use a gradient boosting algorithm by Chen and Guestrin (2016), which gradually minimizes a loss function using a gradient descent method.[21] Prediction in the test sample is carried out through an output function. A simplified schematic of a binary gradient boosting classifier is provided in FIGURE 3 below.

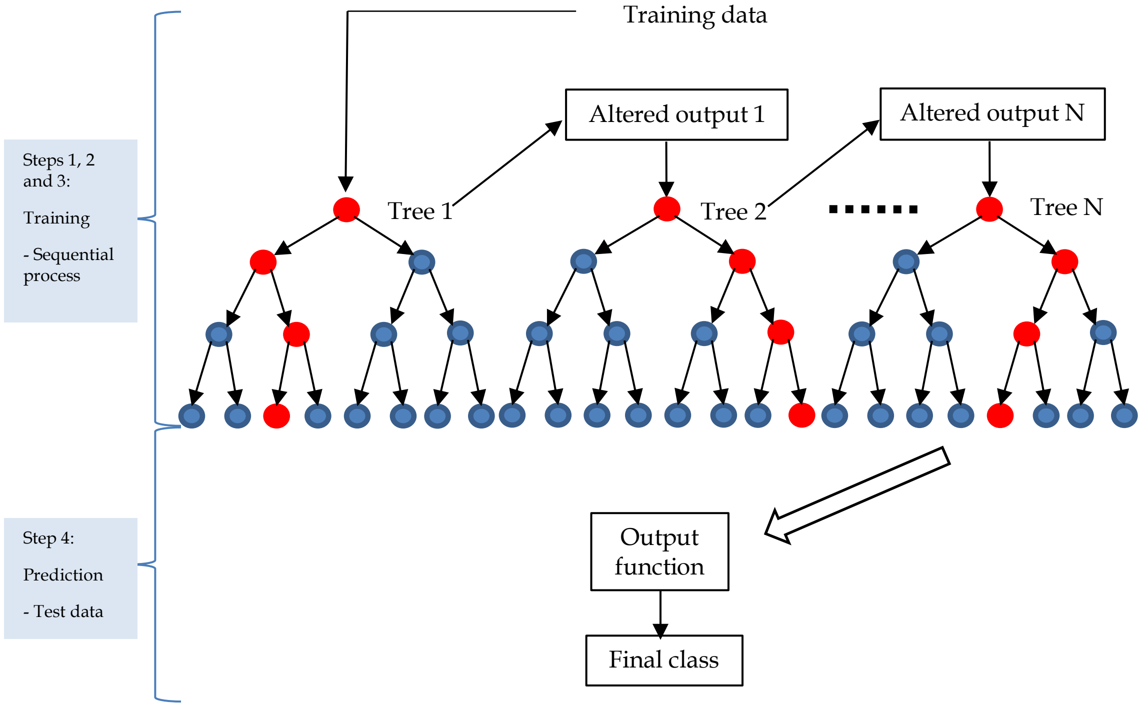

FIGURE 3 Schematic of a gradient boosting classifier.

Notes: The figure design is modified from HE et al.'s (2016, p. 8220) random forest graph.

4.1.5 Artificial Neural Network (ANN)

A large group of methods fall under neural networks. In general, they are complex nonlinear parametric statistical models motivated by how biological neural networks, such as the human brain, work. Here, I briefly describe a single-layered neural network in a classification setting similar to the one presented by Hastie et al. (2009, p. 389–416). A simple schematic figure of a single-layered ANN is provided in FIGURE 4 below.

A neural network[22] is implemented in two stages for classification. The

[21] I implement the extreme gradient boosting algorithm from the package xgboost (Chen et al., 2018), which goes by the 'xgbDART' method in the caret package. The plyr package (Wickham, 2011) is also needed.

[22] I implement a neural network by the 'nnet' method in the caret package. Package nnet (Venables & Ripley, 2002) is required.

first stage includes creating derived features, called hidden units, which are linear combinations of the predictor variables. In the second stage, an output function is used to link the linear combinations of predictors to the target variable with another set of linear combinations. To fit the training data, unknown parameters are estimated for the first and second stages by minimizing cross-entropy, a popular loss function used with neural networks. Moreover, the fitting follows a process called back-propagation to avoid overfitting. After training the model, an output function is used to predict outcomes with test data.

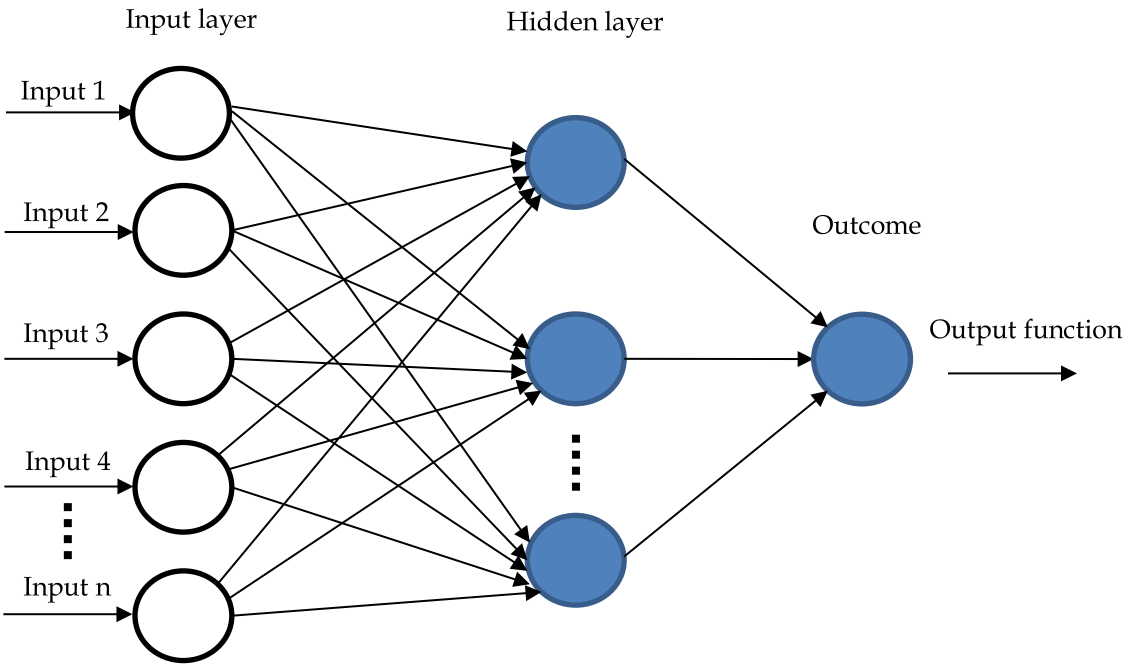

FIGURE 4 Schematic of a single-layered neural network.

Notes: The figure design is inspired by Hastie et al. (2009, p. 393).

4.2 Assessing Predictive Performance

The performance of the classification models is assessed in a held-out test sample with various measures. There is no single gold standard measure for classification performance, which is why I report and evaluate several. Most numerical measures of predictive performance for classification problems are calculated based on the confusion matrix illustrated in TABLE 8 and summarized in TABLE 9 below.

Sensitivity and specificity calculate the probabilities of a predicted positive value given that there is an observed positive value and a predicted negative value when there is an observed negative value.[23] Intuitively, in my case, sensitivity stands for the proportion of HGFs a classifier is able to identify. Symmetrically specificity is the proportion of non-HGFs a classifier is able to identify.

[23] A positive value stands for a binary outcome of 1, and a negative value stands for binary outcome of 0. In the context of this master's thesis, a positive value in the response

Out of these two metrics, only sensitivity is reported because of the interest in the positive outcome. However, specificity is needed to calculate and plot the ROC curves. Furthermore, the false positive rate (FPR) is reported, which is defined as the probability of assigning a falsely positive prediction to a negative outcome. The list of reported measures continues with positive predictive value (PPV), which calculates the correct positive predictions over the total positive predictions. The F-score, a harmonic mean of sensitivity and PPV, is also reported. Finally, overall accuracy is reported, which is popular in the literature. (Fawcett, 2005, p. 862.)

TABLE 8 A confusion matrix

		Observed Class		
		Positive	Negative	Total
Predicted Class	Positive	True Positive	False Positive	Predicted condition positive
	Negative	False Negative	True Negative	Predicted condition negative
	Total	Condition positive	Condition negative	

One crucial instrument for the analysis is the receiver operating characteristics curve (ROC), which is essentially a graphical tool. The ROC curve plots the false positive rate (1-specificity) against sensitivity over all probability thresholds, drawing an increasing concave graph, ideally hugging the top left corner. (Fawcett, 2005, p. 862–866.)

TABLE 9 Predictive performance measures for binary classifiers.

Name	Definition
Sensitivity	True positives/ Condition positive
Positive Predictive Value (PPV)	True positives / Predicted condition positive
F-score	2 * (Sensitivity * PPV) / (Sensitivity + PPV)
Specificity	True negatives / Condition negative
False Positive Rate (FPR)	False positives / Condition negative
Accuracy	(True positives + True negatives) / (Total obs.)

ROC curves entail various advantages in comparing algorithms. They do not require fixing a cutoff value for the prediction probability, nor do they depend on

stands for the firm experiencing high growth, while a negative value indicates that high growth is not achieved.

misleading measures of performance due to class imbalances.[24] The ROC curve, however, is only illustrative. Acquiring a numerical value for analysis requires calculating the area under the curve (AUC), which is commonly used in the literature. The AUC receives values between 0.5 and 1, where 0.5 indicates a random walk classifier and values near 1 suggest good predictive performance with high sensitivity and low false positive rate. In the context, the AUC can be interpreted as the probability of a model ranking higher propensity to an HGF than to a non-HGF. (Fawcett, 2005, p. 868; Weinblat, 2018, p. 265.)

Another graphical tool, precision-recall curve (PR curve), is used for illustration as well. The PR curve plots sensitivity (recall) against PPV (precision) over all probability thresholds. Ideally, the curve should be up high, close to the top right corner of the graph. Like ROC curve, PR curve is a useful tool comparing algorithms' performance in an overall sense. However, where the ROC curves capture prediction models' ability to correctly identify outcomes based on both classes, the PR curves focus solely on the positive class, which is of interest. Disappointingly though, PR curves depend on measures dependent on class distribution, which will make the results not externally comparable. Both of these tools provide unique but non-comprehensive informative value and therefore are considered side by side. (Fawcett, 2005, p. 865.)

It is clear that high values below one are aimed for all performance measures described above but FPR, for which as small as possible values near zero are desired. It should also be noted that all measures except AUC are dependent on the chosen probability threshold for prediction. In addition, it must be underlined that all measures except AUC, sensitivity and FPR are sensitive to class imbalances, which makes them unreliable to generalize (Fawcett, 2005, p. 864).

Furthermore, Venkatraman's statistical test (Venkatraman, 2000) is used in a paired setting to formally compare algorithms. Venkatraman's test is a two-sided permutation test comparing two ROC curves. The null hypothesis states that the two compared curves are equal. For descriptive value and to increase interpretability, I use variable importance to identify the most meaningful predictors. Finally, partial dependence plots (PDP) are presented for bagging, boosting, and random forest algorithms to obtain a sense of how the values of the most meaningful predictors are associated with the probability of classifying a firm as high-growth, given all other predictors.[25]

4.3 Methodological Considerations

After describing the nature of the prediction methodology and tools to assess performance, a few issues still require attention. A crucial phase of ML applications is the preprocessing of data for training. I implement a typical approach,

[24] ROC curves use measures from a single column of the confusion matrix, which makes the curves independent of class distribution (Fawcett 2006, p.864).

[25] For methodological notes and examples on variable importance and partial dependence plots, see Hastie et al. (2009, p. 367–384).

centering and scaling all the numerical predictors. In addition, I have omitted all the observations with missing values. Since their proportion is relatively low—approximately 10% —I do not see a need for imputing missing data. By centering, I mean subtracting the mean of the predictor's data from its actual values. Scaling divides the predictor's value by its standard deviation. It is important to note that the test data are preprocessed as well, but using distributional information only from the training sample, avoiding any information leakage and therefore unrealistic results.

The second issue concerns the vast imbalance of classes in the target variable. Defining high-growth firms as described in Chapter 3.2.1 yields only 4.5% of all firms classified as high-growth in the training sample. ML algorithms are known to be sensitive to class imbalances, producing undesirable results.[26] However, there are various techniques, such as resampling and threshold optimization, to mitigate this issue (Sun, Wong, & Kamel, 2009, p. 700–710). I experimented with several approaches[27] and proceeded with optimizing the probability threshold for prediction using the F-score, since it returned the most promising improvements in performance. The basic idea of threshold optimization with the F-score is to choose the probability threshold that maximizes the F-score in the training sample and use it for prediction in the test sample.[28] The procedure is commonly used in ML literature, and its properties have been studied, for example, by Lipton, Elkan and Naryanaswamy (2014).

The third issue concerns the specification of the algorithms. Most of the ML algorithms implemented in this thesis have several hyperparameters, which require assigned values for training. A typical approach is to cross-validate training results to obtain a combination of hyperparameters that yields the best in-sample results (Hastie et al., 2009, p. 241–257). I apply 10-fold cross-validation in the training phase to tune the most important parameters with predetermined grids of parameter values. The final set of hyperparameter values assigned for the algorithms are reported in TABLE 10 below.

The fourth issue concerns the overfitting of data due to the complex learning patterns of ML algorithms. Overfitting results in a good fit with the training data, but poor performance out-of-sample, and is more likely with more predictors and with less data. Again, several approaches exist to address this issue, including cross-validating the training sample and several regularization schemes (see, e.g., Friedman, Hastie, & Tibshirani, 2010; Hastie et al., 2009, p. 139–181, 219–257). As described in the previous paragraph, I implement a cross-valida-

[26] In my case, most of the ML algorithms predict all firms belonging to the class of no high growth with the original dataset if the imbalance of class distribution is not mitigated.

[27] Moreover, I tried the following approaches: downsampling, upsampling, SMOTE (synthetic oversampling) and optimized threshold in prediction using AUC and F1. In addition, I tried several mixed strategies using resampling in training and optimized thresholds in prediction. However, the other benchmark, Lpm, does not support resampling schemes.

[28] Classifying a firm as high-growth requires defining a probability threshold at which an observation is assigned to the positive class.

tion scheme in the training phase and therefore mitigate the possible overfitting. Combined with a relatively large dataset and few predictors, I am confident in my approach.

The fifth issue of the analysis considers the predictors used for analysis. Despite the vast literature, there is no consensus on the drivers of firm growth. In addition, there are several approaches in the ML literature and no agreement on how to choose the most meaningful predictors from the full predictor space to reduce noise in prediction (see, e.g., Guyon & Elisseeff, 2003). Furthermore, ML approaches to choose a subset of variables entail instability over iterations if variables are correlated (Mullainathan & Spiess, 2017, p. 96–98). Nevertheless, algorithms such as random forest and boosting are known to perform well with large predictor spaces due to their built-in variable selection. Also, the total number of predictors is relatively small, making it questionable to shrink the set of predictors any further. For the abovementioned reasons, I do not implement any feature selection scheme in my approach.

Finally, it should be mentioned that the applied ML algorithms (mostly bagging, boosting and random forest) are computationally very costly to implement. Depending on the size of the data set and approach in the learning phase, without professional hardware, it can take even days to train a single classifier, even with parallel computing[29]. Some of the methodological decisions (such as the number of different ML algorithms, resampling schemes, number of iterations in cross-validation or size of the hyperparameter grids for cross-validation) have to be considered from this point of view as well. Computational costs raise another dilemma too. In this master's thesis, the primary interest is in predictive performance, but perhaps a computational scientist would be interested in questions on computational cost-efficiency relative to predictive performance as well. Nevertheless, issues related to computational efficiency are disregarded here.

TABLE 10 Tuned hyperparameter values of the machine learning algorithms.

CART		**Boosting**	
Complexity parameter	0.00084	Number of iterations	150
		Maximal tree depth	2
Random Forest		Shrinkage parameter	0.3
Number of randomly selected predictors	2	Minimum loss reduction	0.5
Spliting rule	gini	Subsample percentage	0.5
Minimal node size	5	Subsample ratio	0.8
		Fraction of trees dropped	0.2
Artificial Neural Network		Probability of skipping drop-out	0.2
Number of hidden units	3	Minimum sum of instance weight	0.5
Weights decay	0.1		

Notes: The bagging algorithm applied does not entail any tuning parame-

[29] Parallel computing refers to harnessing multiple computer cores to calculate a single task simultaneously, which results in faster execution. I parallelized all computational tasks in the training phase with all algorithms.

5 RESULTS

In this chapter, I report and evaluate the results obtained using the empirical framework described in Chapter 4. Starting with the baseline model, the results are assessed and compared with various performance metrics and mirrored against the relevant previous literature. I also provide auxiliary prediction results for an alternative measure of firm growth, additional predictors of expert information and a sample of young firms.

5.1 The Baseline Model

5.1.1 Out-of-Sample Predictive Performance

TABLE 11 provides the baseline results in the test sample. Underlying confusion matrices and prediction results with alternative resampling schemes are provided in Appendix C. Compared to the better benchmark (Logit), the best machine learning technique, random forest (RF), with an out-of-sample AUC of 0.6422, provides a 0.055-point AUC improvement (equivalent to a 9.4% improvement) over a simple logit classifier, which is a somewhat modest but nevertheless relevant improvement. Considering AUC's interpretation, the RF algorithm ranks a higher propensity for a random Finnish firm to be an HGF than a non-HGF with a probability of 64.22%.

TABLE 11 Out-of-sample prediction results with F1 optimized thresholds in prediction for the baseline model.

Classifier	AUC	Threshold	Sensitivity	PPV	F-score	FPR	Accuracy
Lpm	0.5727	0.1002	0.3272	0.1607	0.2155	0.0887	0.8825
Logit	0.5872	0.1071	0.3049	0.1868	0.2317	0.0689	0.9003
CART	0.5221	0.1290	0.2053	0.2028	0.2040	0.0419	0.9210
Bagging	0.6109	0.3600	0.0528	0.2680	0.0883	0.0075	0.9462
Boosting	0.6190	0.1125	0.3191	0.2052	0.2498	0.0641	0.9055
RF	0.6422	0.1415	0.1707	0.2877	0.2143	0.0219	0.9382
ANN	0.5889	0.1198	0.2703	0.2031	0.2319	0.0550	0.9117
Ensemble	0.6326	0.1577	0.1728	0.2833	0.2146	0.0227	0.9376

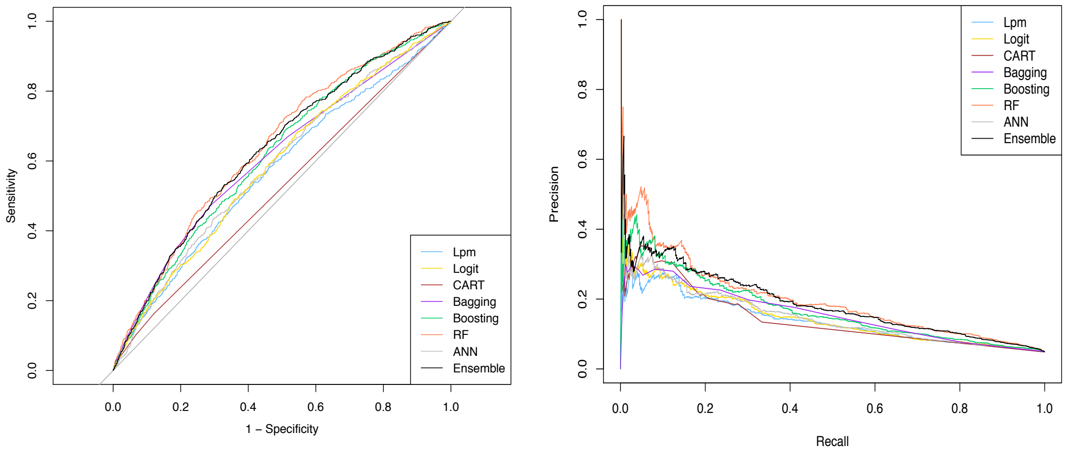

FIGURE 5 ROC curves (left pane) and PR curves (right pane) in the test sample for the baseline model.

The RF classifier's AUC is far below the highest AUCs in the previous literature; Weinblat (2018) achieved an AUC of 0.8110 for the UK, and Sharchilev et al. (2018) achieved an AUC of 0.854 with international data, but close to Miyakawa et al.'s (2017) AUC of 0.68 for Japanese firms and almost equal to Weinblat's (2018) AUC of 0.6439 for Finland.30 Differences across these values can have various explanations since choices in methodology and variables vary substantially. However, since the best result in terms of AUC is in line with Weinblat's (2018) result for Finland, the country of interest seems to play a role in predictive performance.

Observing the out-of-sample prediction results in TABLE 11, all but the CART outperform the baselines and can be ranked in descending order based on the value of AUC as follows: RF, Ensemble, Boosting, Bagging, ANN, Logit, Lpm, and CART. The absolute differences in AUC are rather small, as one can notice in FIGURE 5, left pane, where the overlapping ROC curves are difficult to distinguish from each other. However, at least the curves of RF and Ensemble seem to arch somewhat closer to the top left corner compared to the curves of other classifiers. Considering performance only in the positive class, the PR curves illustrate a similar standing in FIGURE 5, right pane, where the curves of RF, Ensemble, and Boosting seem to be slightly closer to the top right corner than the rest of the curves.

Comparing the ROC curves, the ordering is mostly supported by Venkatraman's (2000) permutation test. The p-values for the test are reported in TABLE 12. For all classifiers except ANN and Bagging, the null hypothesis of two ROC curves being equal is rejected at the 95% confidence level compared to the baselines, Lpm and Logit. The RF classifier differs significantly from all others except the Ensemble classifier. Besides, the null hypothesis is not rejected for the following pairs of ROC curves: Bagging and Boosting, Baggingand ANN, and Boosting and Ensemble.

[30] Actually, I was able to achieve the same AUC (0.6439) as Weinblat (2018) when applying class weights in the training phase. However, overall performance with the class weight approach works poorly across all other algorithms, which is why I report results without any weights and with threshold optimization as my baseline setting.

TABLE 12 Venkatraman's test p-values for the baseline model.

Classifier	Lpm	Logit	CART	Bagging	Boosting	RF	ANN	Ensemble
Lpm	1	0.0000	0.0000	0.0070	0.0000	0.0000	0.0425	0.0000
Logit	0.0000	1	0.0000	0.0570	0.0000	0.0000	0.7085	0.0000
CART	0.0000	0.0000	1	0.0000	0.0000	0.0000	0.0000	0.0000
Bagging	0.0070	0.0570	0.0000	1	0.0775	0.0035	0.0820	0.0035
Boosting	0.0000	0.0000	0.0000	0.0775	1	0.0015	0.0020	0.0665
RF	0.0000	0.0000	0.0000	0.0035	0.0015	1	0.0000	0.1595
ANN	0.0425	0.7085	0.0000	0.0820	0.0020	0.0000	1	0.0000
Ensemble	0.0000	0.0000	0.0000	0.0035	0.0665	0.1595	0.0000	1

Although AUC is the most common and reliable overall measure of predictive performance in this context, sensitivity and false positive rate (FPR) are also useful measures for interpretation and comparison when a probability threshold for prediction is defined. The RF classifier correctly identifies 17.07% (sensitivity) of the high-growth firms with only a 2.19% (FPR) chance of misclassifying a non-high-growth firm as a high-growth firm. Compared to Weinblat's (2018) corresponding values of 26.45% (sensitivity) and 14.42% (FPR) for Finland, the RF classifier is clearly more cautious, identifying HGFs correctly almost 10 percentage units less, but doing so with a fraction of the risk of misclassification.[31]

Considering the nature of other classifiers, CART and Ensemble are also cautious with low sensitivities and FPRs, where the Bagging classifier is the most conservative, with a sensitivity of just 0.0528 and FPR of 0.0075. The baselines, Boosting and ANN, represent more liberal classifiers with sensitivities ranging between 0.27 and 0.33 and FPRs varying from 0.055 to almost 0.09.

The rest of the reported performance measures in TABLE 11 are not directly comparable to those of previous studies due to differences in class distributions (as discussed in Chapter 4.2). Additionally, similar to sensitivity and FPR, they depend on the probability threshold used in prediction, which in my case is determined separately for each classifier using the F-score optimization in training. However, these measures entail information on the positive class and are internally comparable. In particular, the F-score—the harmonic mean of sensitivity and PPV—is of particular interest since the positive class is a more interesting one. PPV (precision) and sensitivity (recall) should be considered together since their variation is usually observed in a trade-off, as shown in FIGURE 5. A balance between the two is more meaningful than extreme values in one or another. Of course, depending on the purpose and preferences of the entity forecasting HGFs, one could put more weight on precision to have more confidence in picking a few most potential HGFs (tight budget) or on sensitivity to identify more potential HGFs with decreased certainty (loose budget).

[31] It is worth noting here that Weinblat's (2018) application for Finland has a class distribution of about 10% HGFs in the training and test samples using the Birch-Schrayer indicator for the high-growth definition. Furthermore, a resampling scheme called SMOTE is used in training and a classical 0.5 probability threshold in prediction. These differences influence how a classifier performs.

Using the F-score as the performance measure of interest, the ranking of the classifiers (Boosting, ANN, Logit, Lpm, Ensemble, RF, CART, and Bagging) is notably different from the AUC ordering. Where RF is the best classifier when the probability threshold is not determined, considering performance also based on the F-score with the best possible probability thresholds for each classifier, the Boosting classifier is the only classifier to perform better than the baselines.

The measure of overall accuracy in the last column of TABLE 11 is not a very meaningful one in the case of unbalanced class distribution since a classifier could achieve over 95% accuracy just by predicting all observations as non-HGFs. This result is an undesirable outcome, and I have dealt with the issue (cf. Chapter 4.3). As a result, however, the overall accuracy will usually end up (as in my case) remaining under the "no-information rate" (as the proportion of the major class is called) in the test sample. Nevertheless, the measure is reported due to its popularity in the literature.

It is difficult to explicitly state which method assessed here is the tool of choice in forecasting HGFs, since alternative measures of predictive performance lead to somewhat different conclusions. Based on my results, it can be summarized, that ML techniques provide slight improvements over baselines in predictive performance when a specific probability threshold for prediction is not predetermined. In this case, the RF or Ensemble classifiers seem to be the tools of choice in terms of AUC. With an optimized prediction threshold, however, it seems possible for a conventional classifier to produce better forecasts than an ML classifier, depending on the measure of performance. Using the F-score as the measure of interest, the Boosting classifier seems to win the race, where most of the simple ML techniques cannot beat the baselines. It all comes down to the choice of measure, which should arise from the needs of policy-makers and investors.

5.1.2 Evidence on the Most Meaningful Predictors

The results on variable importance for tree-based classifiers (CART, Bagging, Boosting, and Random Forest) are presented in FIGURE 6, where relative importance is on the horizontal axis, and the 20 most important predictors in descending order from the top are on the vertical axis. In RF, the two most important predictors are past personnel and past sales growth; the third and fourth most important predictors are past productivity growth and initial productivity level, respectively. The three most important predictors after these four are firm age, capital intensity, and size (as proxied by sales). The top three predictors are the same for RF and Bagging, and the top two for Boosting as well (although the ordering of the top two is flipped for Boosting). The CART classifier values especially the age of the firm, age of the CEO and the indicator of being part of a group.

It seems that the aggregation of trees in different ways moves the focus towards past growth and current level indicators of the target or predictors closely related to it. My findings on the most important predictors are in line with Weinblat's (2018) findings for several countries (including Finland) and the previous literature on the matter, summarized as follows. In addition to the importance of firm size and its variation, based on regression studies, size and age contribute more than financial or sectoral predictors that do not seem to play a significant role (p. 279–280).

Partial dependence plots (PDP) provide another useful interpretation mechanism for some tree-based ML techniques. I present these plots for the ten most important predictors (based on the RF ranking in FIGURE 6) using the Bagging (blue), Boosting (green), and RF (red) classifiers in FIGURE 7. Here, the horizontal axis represents centered and scaled values of the predictors, and the vertical axis represents the probability of assigning a firm as an HGF, given all other predictors. The plots are scaled to the distribution, illustrated by decile tick marks.

ML techniques' ability to capture nonlinear relationships truly stands out in PDPs. My main finding here is that—in the case of RF—there appears to be a slight nonlinear rising in the probability of high growth when considering increasing values in the first four predictors, which are the past growth of personnel, revenue and productivity and finally the initial productivity level. The observation is intuitive and in line with most of the previous literature (see, e.g., Coad, 2007a).

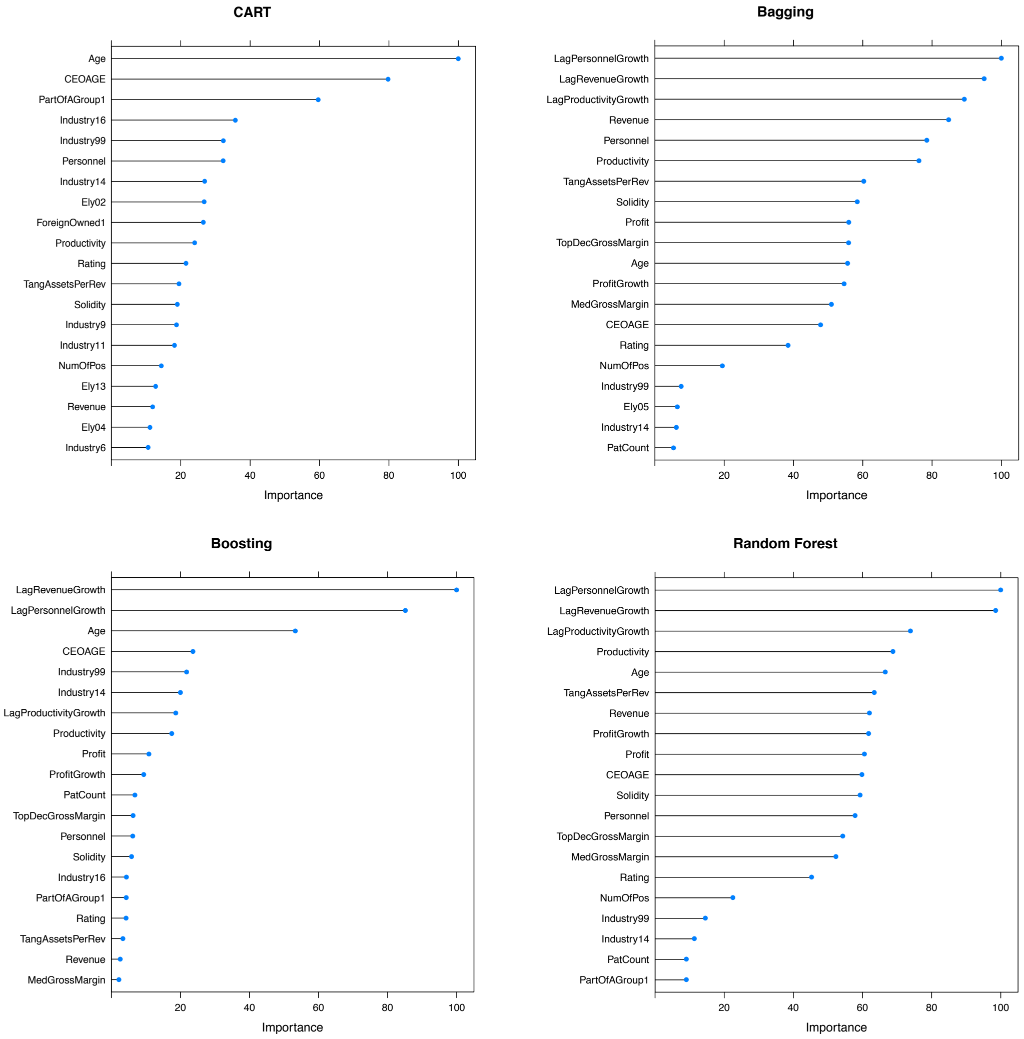

FIGURE 6 Variable importance in the learning sample for the baseline model.

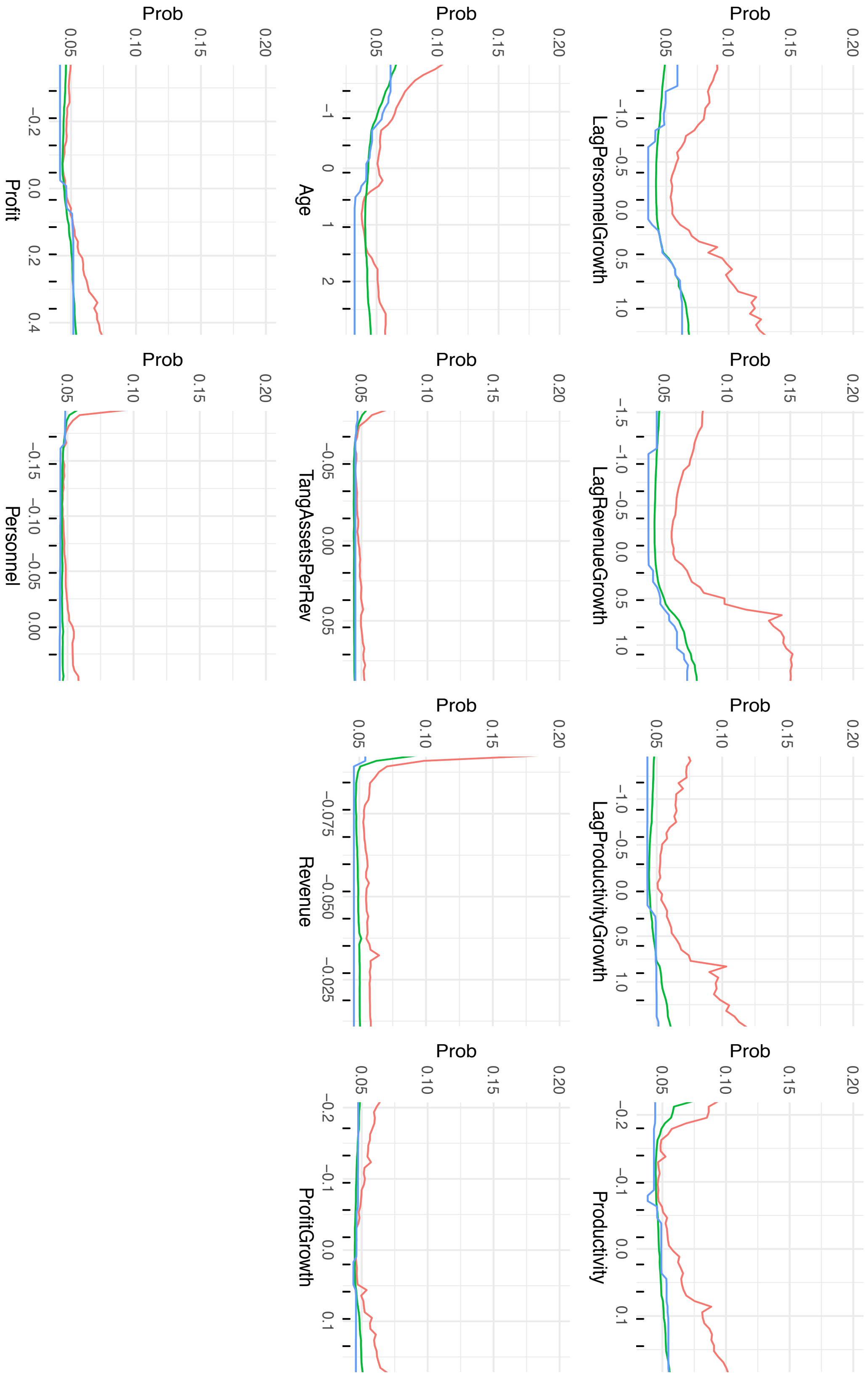

FIGURE 7 Partial dependence plots in the learning sample for the baseline model.

Notes: The 10 most important predictors based on the ordering implied by the RF classifier are presented with a probabilistic scale. Red: Random Forest, green: Boosting and blue: Bagging.

Combined with the variable importance analysis, past growth indicators seem to contribute the most and exhibit the largest conditional variation in partial dependence. However, this association is not as evident with Bagging and Boosting. Also, there seems to be some evidence on younger firm age being associated with higher probability of fast growth, which is in line with the literature (Acs et al., 2008).

5.2 Auxiliary Analyses

5.2.1 Growth in Turnover

I used growth in turnover as an alternative measure in the definition of HGFs to mirror my baseline results against. The out-of-sample prediction results with this modification are presented in TABLE 13 and FIGURE 8 below.

Overall—in terms of AUC—there are major improvements (ranging between 0.1 and 0.15 points) in predictive performance compared to the baseline model. This result is observable in FIGURE 8 (left pane), as the ROC curves lie much closer to the top left corner compared to the baseline model in FIGURE 5. Nonetheless, the ranking and point differences of AUC values across classifiers remain similar to the baseline model. Venkatraman's test (not reported but available upon request) was also carried out, where similarly to the baseline, the RF and Boosting stand out from other classifiers. ANN and Bagging remain the only classifiers not being able to reject the null hypothesis against the baselines.

TABLE 13 Out-of-sample prediction results with the OECD's high-growth definition by turnover and F1 optimized thresholds in prediction.

Classifier	AUC	Threshold	Sensitivity	PPV	F-score	FPR	Accuracy
Lpm	0.7281	0.1296	0.3829	0.1942	0.2577	0.1286	0.8348
Logit	0.7260	0.1029	0.5408	0.1761	0.2657	0.2048	0.7761
CART	0.6192	0.1563	0.0402	0.1364	0.0620	0.0206	0.9091
Bagging	0.7188	0.3600	0.1111	0.3430	0.1678	0.0172	0.9175
Boosting	0.7447	0.1421	0.3722	0.2299	0.2843	0.1009	0.8596
RF	0.7769	0.1511	0.2811	0.2778	0.2794	0.0592	0.8914
ANN	0.7370	0.1422	0.3788	0.2039	0.2651	0.1197	0.8427
Ensemble	0.7661	0.1753	0.2718	0.2827	0.2771	0.0558	0.8938

Slight improvements in the F-score are also observed; however, the underlying class distribution has changed, making the results not directly comparable. Misleadingly, the PR curves in FIGURE 8 (right pane) have been raised up a notch, partially because the proportion of HGFs has increased from approximately 5% to 7.5% in the sample.

Based on these results, a considerate interpretation could be made that the overall task of identifying future HGFs in Finland is a somewhat easier task when considering high growth in turnover rather than in employment. This result is in slight controversy with the findings of Daunfeldt, Elert and Johansson, (2014), which do indicate that the set of HGFs is different depending on the growth measure, but the measures in sales and employment are not sensitive to the same issue. Nevertheless, my observation agrees with the challenge previously recognized in the literature: comparing predictive results and executing policy based on them is difficult and questionable if the results can differ based on the HGF definition (Delmar et al., 2003).

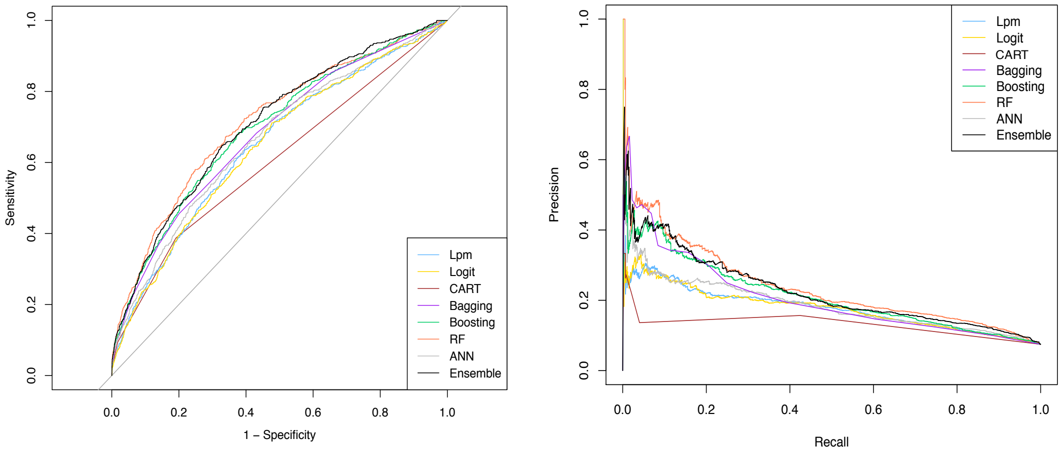

FIGURE 8 ROC curves (left pane) and PR curves (right pane) in the test sample using the OECD turnover definition of high growth as a target.

5.2.2 Expert Information

The out-of-sample prediction results including indicators of private and public expert information (venture capital finance and innovation grants) as predictors in the model are provided in TABLE 14; illustrative ROC and PR curves are provided in FIGURE 9.

TABLE 14 Out-of-sample prediction results with expert information and F1 optimized thresholds in prediction.

Classifier	AUC	Threshold	Sensitivity	PPV	F-score	FPR	Accuracy
Lpm	0.5741	0.0916	0.3740	0.1410	0.2048	0.1182	0.8567
Logit	0.5910	0.0941	0.3455	0.1577	0.2166	0.0958	0.8767
CART	0.5222	0.1122	0.2012	0.2045	0.2029	0.0406	0.9220
Bagging	0.6023	0.3600	0.0671	0.3000	0.1096	0.0081	0.9463
Boosting	0.6122	0.1491	0.2500	0.2500	0.2500	0.0389	0.9260
RF	0.6420	0.1351	0.1850	0.2716	0.2201	0.0257	0.9353
ANN	0.6018	0.0866	0.3963	0.1237	0.1885	0.1457	0.8317
Ensemble	0.6320	0.1563	0.1646	0.2784	0.2069	0.0221	0.9377

Prima facie evidence suggests that both private and public investors have some ability to predict and/or to nurture firm growth: the *ex-post* shares of high-growth firms among both Venture- and Tekes-backed companies (6.94% and

6.63%, when measured in personnel growth, respectively) are approximately two percentage points higher than in the overall sample. However, when considered in tandem with all the other predictors, including this expert information does not yield improvements in predictive performance. The results are almost identical to the baseline in terms of AUC and supported by Venkatraman's test (not reported here but available upon request). There are no significant improvements in F-scores either. However, this time, the RF classifier is able to slightly outperform the baselines together with Boosting.

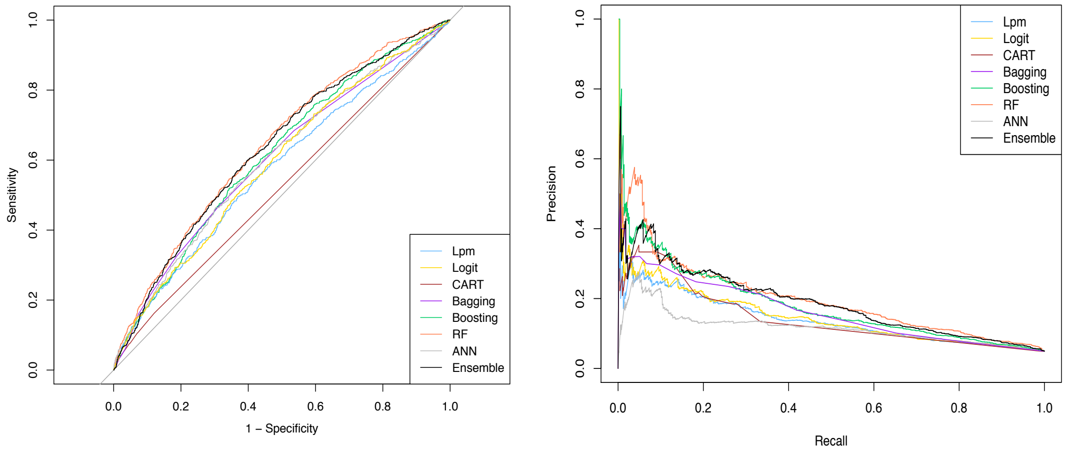

FIGURE 9 ROC curves (left pane) and PR curves (right pane) in the test sample including predictors of expert information.

In a slight contradiction to my result, investor features play a substantial role in the ML application of Sharchilev et al. (2018). However, their set of firms consists of young startup companies with data of various investor features across time, whereas I use a set of firms of various sizes and ages. In addition, investor features might be more relevant as predictors to identify ventures based on Sharchilev et al.'s (2018) choice of target (securing another round of equity funding).

5.2.3 Young Firms

My final analysis considers a model with the same predictor space and HGF definition as in the baseline model but with a dataset with only young (≤ 10 years old) firms. The out-of-sample prediction results are provided in TABLE 15 and FIGURE 10.

As observable in TABLE 5, the share of HGFs is over twice higher (with a share of about 10% of total firms) among young firms than in the overall sample. Despite the more balanced HGF distribution among young firms, the results reflect major impairment in out-of-sample predictive performance. However, the ranking and point differences of the algorithms remain roughly the same. The weakest AUC values are close to the random walk level of 0.5 (Lpm and Logit), and the strongest are slightly under 0.6 (RF). This time, however, all algorithms seem to outperform baselines in absolute terms, but according to Venkatraman's test (not reported here but available upon request), the ROC of the Bagging classifier still does not significantly differ from the baselines. The ROC curves in FIGURE 10, left pane, illustrate the poor performance by overlapping each other near the 45-degree random walk line.

TABLE 15 Out-of-sample prediction results with a dataset of young (≤ 10 years old) firms at the time of observation and F1 optimized thresholds in prediction.

Classifier	AUC	Threshold	Sensitivity	PPV	F-score	FPR	Accuracy
Lpm	0.5110	0.1494	0.3653	0.2192	0.2740	0.1597	0.7884
Logit	0.5118	0.1560	0.3151	0.2396	0.2722	0.1227	0.8159
CART	0.5617	0.1420	0.2146	0.2474	0.2298	0.0801	0.8428
Bagging	0.5537	0.3600	0.0868	0.2969	0.1343	0.0252	0.8777
Boosting	0.5812	0.1537	0.3196	0.2405	0.2745	0.1238	0.8154
RF	0.5945	0.1714	0.1918	0.3559	0.2493	0.0426	0.8738
ANN	0.5199	0.1531	0.4155	0.2382	0.3028	0.1630	0.7909
Ensemble	0.5707	0.2012	0.1689	0.3058	0.2176	0.0471	0.8673

Quite misleadingly, the overall poor performance does not seem to affect the F-score compared to the baseline model. Again, this is due to a change in the underlying class distribution in the target variable. In fact, in the sample of young firms, it is over twice as probable to pick an HGF from the sample at random. This finding is also indirectly visible in FIGURE 10, right pane, where the PR curves have remained roughly on the same level of the vertical axis, although at first thought, they should shift similar to the ROC curves but in the opposite direction.

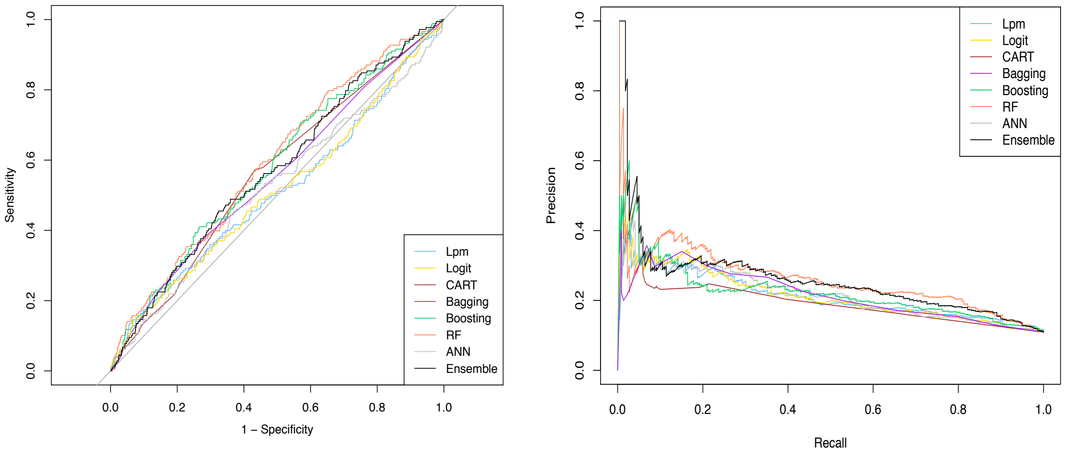

FIGURE 10 ROC curves (left) and PR curves (right) in the test sample using a dataset of young firms.

My results with young firms imply that the ML techniques implemented in this master's thesis function similarly relative to the benchmarks (Lpm and Logit) as in the baseline model. However, identifying HGFs from a sample of only young firms turns out to be a more difficult task than from a representative sample of all firms.

6 DISCUSSION

In this chapter, I discuss the implications of my findings and address the limitations of this study. My results evoke several notable topics for discussion and relevant implications for policy. The baseline analysis confirms that ML methods provide improvements in predictive performance in the context of identifying HGFs. The classifiers behave differently, but generally they identify relatively few HGFs but do so with relatively high accuracy. All in all, my results are in line with the general finding that ML approaches can be useful for prediction policy problems similar to many previous applications (see, e.g., Mullainathan & Spiess, 2017).

However, the improvements in performance with ML methods are modest compared to conventional techniques and not always unambiguous. Therefore, it seems, that not even ML methods are able to find superior relationships from data for a prediction problem that is genuinely hard, as the task at hand is considered to be (see, e.g., Coad et al., 2014). Moreover, the evidence on the most meaningful predictors is in line with the previous literature (see, e.g., Weinblat, 2018) and therefore confirms what is already known on the matter. With the type of data available, predictors related to firm size, its past change and firm age contribute the most.

The auxiliary analyses provide several interesting findings essential for policymakers and investors. First, based on my finding from a simple robustness check, it seems that predictive accuracy is sensitive to the high-growth definition used. More accurately, the prediction problem at hand seems to be an easier one, when measuring growth in turnover rather than in employment. While not formally tested, this is clearly observable in absolute terms. Agreeing with the previous literature (see, e.g., Daunfeldt et al., 2014, 2015; Delmar et al., 2003), it is essential to acknowledge the differences of growth measures and HGF definitions and how they can result in different sets and distributions of HGFs. Although the academic literature would favor a standardized HGF definition for increased comparability—for practical applications—the definition should always be chosen based on the purpose of the identification task.

In the second auxiliary analysis of additional predictors, the initial observation is that the share of HGFs is higher in venture capital and public innovation grant backed companies than in the overall sample. However, adding these proxies of expert information as predictors in the baseline model does not seem to provide improvements in out-of-sample predictive performance. This finding is related to literature on forecast encompassing and combination. If the predictors of expert information are thought of as forecasters of HGFs, it is said that the baseline forecaster encompasses the additional ones, since predictive performance is not improved by a combination of the forecasters (Clements & Harvey, 2009). However, this outcome is not formally tested here. My findings imply that while private and public investors might be able to identify HGFs with some accuracy, including proxies of their investment decisions is not meaningful in a predictive model since all the useful information is already identifiable from the initial data.

Finally, based on my results, identifying HGFs from a sample of young firms seems to be close to an impossible task even with ML classifiers. The results here are not completely surprising considering the baseline model's analysis on variable importance. The results reveal that the predictors on firm size and its past growth contribute the most in predictions with increasing partial dependences to probability of high growth. Although the firm age does play a significant role in prediction, its value is not dominant. Moreover, my findings here are in line with the literature arguing that young small firms grow unexpectedly compared to old large firms with steady growth trajectories and long planning horizons (Coad, 2007b).

However, my findings on the predictability of young HGFs are not obvious and ignoring them would probably lead to counter-effective outcomes in practice. Observational evidence in the literature suggests that HGFs tend to be young (see, e.g., Acs et al., 2008). The same finding is visible in this study as well, where the mean age of HGFs is close to 10 years (using the baseline HGF definition), which is less than half of the overall mean age of approximately 21 years. Moreover, the share of HGFs in the sample of young firms is over twice as large (with a share of approximately 10%) than in the overall sample (with a share of approximately 5%). Therefore, the observational conclusion from a policymaker's (or a private investor's) point of view would be that targeting young firms will probably lead to the most effective results. Based on my findings, however, this is not the most effective approach since HGFs are almost impossible to predict from a sample of only young firms. The main message to policymakers and investors following this auxiliary analysis is therefore the following: targeted policy measures should be aimed at potential HGFs that are predictable, but as of yet, the task is practically close to impossible for young firms, at least with the historical data available.

Should a policymaker or a private investor then employ ML methods for only minor expected advantages in predicting HGFs? The answer depends on what the applier is aiming for. If the interest is solely in predictive performance, ML methods provide slight but relevant improvements and they should be used. However, as discussed in Chapter 4.3, ML techniques require a lot of computing power, and can therefore be time consuming. Another disadvantage lies in interpretation, which is considerably more challenging for ML based models than conventional econometric ones. Moreover, the choice of a specific classifier depends on the applier's preferences. As pointed out in Chapter 5.1.1, the trained classifiers behave quite differently out-of-sample, with some identifying a relatively large share of HGFs but with less accuracy, and vice versa. Overall, the choice of a classifier, whether an ML based or not, depends on many considerations, which should be addressed before applying.

Following the discussion above, two interesting questions arise: why is predicting HGFs so difficult and are there any possibilities whatsoever to alleviate the issue. While the first one is not in the direct scope of this master's thesis, the same problem has been identified in the previous literature (see, e.g., Coad et al., 2014). As discussed in Chapter 2, the reasoning for the hardship lies mostly in the heterogeneity of firms and how they grow. In addition, there are multiple factors known to be associated with firm growth but with no high-dimensional data available for.

For the second question above, the possibilities are three-folded. Following the approach of this master's thesis, additional methodological improvements can be made to potentially enhance predictive accuracy. Most promising improvements, however, are expected from increased quantity and quality of data. Finally, if neither of these options yield further improvements, policymakers are forced to shift their focus from targeted policy to enabling generally favorable environmental conditions for firm growth. In the same case, private investors would need to solely rely on their superior views and other tools of analysis of firms' prospects as a basis for investment decisions.

I consider the ideas and topics in the previous two paragraphs as part of the general guideline for future work related to identifying HGFs. However, the analysis carried out in this master's thesis entails a few limitations, which can be turned into specific propositions for future research. From a methodological point of view, as discussed in Chapter 4.3, ML approaches require a set of decisions regarding the training phase. If time was not a constraint, one could probably improve predictive performance by trial and error (e.g., by trying different algorithms, resampling schemes and tuning parameter values). Methodological tuning usually results in limited improvements only, however. Most promising results here might be achieved by implementing some off-the-shelf ML methods that better utilize panel techniques familiar from econometrics.

A more effective approach could be to enhance the quantity and quality of data and variables. The data sample used in this master's thesis is relatively large with 24 predictors in the baseline model, but small compared to what ML algorithms are capable of handling. Four areas of improvement can be identified here, based on ideas throughout the book of Hastie et al. (2009) and discussed in ascending order based on the expected yield in predictive performance. First, the data sample at hand could be improved by removing outliers and trying to reduce noise in other ways. Second, by creating features like additional lags and interactions of existing variables combined with feature selection approaches, might further enhance the complex learning processes of ML. Third, more data of the same variables could be gathered to enable the best possible learning environment for data-hungry ML algorithms. Fourth, including additional relevant predictors showing univariate correlation to the outcome would most probably improve predictive performance.

Out of the propositions above, I consider the last one particularly promising in the context of identifying HGFs. Given ML methods' ability to handle unconventional data formats such as text, I see solid possibilities in utilizing news articles in HGF prediction, for example. As reviewed in Chapter 2.4, a similar approach has already been carried out by Sharchilev et al. (2018) with encouraging results. Out of conventional variables, my models do not entail predictors of the quality of leadership, which is known to have significant influence on the firm's prospects (see, e.g., Guzman & Stern, 2015b).

In the big picture, more studies enhancing predictive performance of future HGFs are needed. However, if a satisfactory level of predictive accuracy is reached and any targeted policy measures are wished to be carried out, the question on how to optimally allocate resources and with what tools, needs to be addressed to the same degree. Moreover, this requires studies on interference rather than pure prediction (see, e.g., Athey, 2017).

7 CONCLUSION

High-growth firms (HGF) have attracted recent attention as job creators, because of which policymakers are in need of a robust mechanism identifying future HGFs for targeted policy measures. However, the task has proven difficult with conventional methodology.

In this master's thesis, I have answered the need stated above by conducting a predictive scheme similar to a real forecasting scenario, where I have observable past values available to predict unknown future outcomes with. Using advanced but commonly used machine learning (ML) algorithms and a broad set of predictors, I have trained several classifiers in a 2005–2012 learning sample of Finnish firms to predict HGFs. Predictive performance of these classifiers is then assessed in a truly out-of-sample test window of 2013–2016, putting the classifiers to a hard test against benchmarks.

Overall, the results of this master's thesis conclude that ML methods provide modest but statistically significant improvements over simple regressions in predicting HGFs, which answers the first research question stated in Chapter 1. My best performing classifier—random forest (RF)—offers a 0.055-point (which corresponds to 9.4%) out-of-sample improvement over the better benchmark in terms of AUC (area under the ROC curve), the most common performance measure in the context. Depending on the measure of interest, however, the classifiers behave differently, and the ML methods applied are not uniformly capable of beating the benchmarks. Therefore, some preferential decisions are needed to be made when choosing a classifier for the task.

Considering variable importance, mostly predictors of firm's current and past growth of size indicators and firm age seem to contribute the most to predictions. Moreover, larger size and higher growth during the previous period seems to be associated with a higher probability of high-growth in the following period. In addition, younger firms are more probable HGFs than older ones. These findings are in line with the previous literature and answer the second research question stated in Chapter 1.

I also studied, whether using an alternative high-growth definition, including information on private and public investors' investment decisions or studying a subsample of young firms affect predictive performance. My findings imply that identifying HGFs is a considerably easier task when growth is measured in turnover rather than in employment. Despite the observation of *ex-*

post shares of HGFs being higher in investment backed firms, including predictors of expert information does not yield improvements in predictive performance. Finally, predicting HGFs in a sample of young firms is a notably harder task than in a sample with no age restrictions.

The empirical framework of this master's thesis entails a few limitations considering the quantity and quality of data and further improvements in methodological choices. Where future research enhancing the predictive scheme applied in this thesis is needed, the question of how to optimally allocate resources for potential HGFs and with what tools, needs to be addressed through causal studies.

Nevertheless, in this master's thesis, I have provided a robust ML-based predictive scheme with useful results for policy. Altogether, I find that the best ML methods are useful but not overpowering in predicting HGFs, with the data available. That is, if the interest is exclusively in prediction accuracy. Therefore, ML should be considered in the context, if computational costs or model opacity are not concerned.

REFERENCES

Acs, Z. J., & Mueller, P. (2008). Employment effects of business dynamics: Mice, gazelles and elephants. Small Business Economics, 30(1), 85–100.

Acs, Z. J., Parsons, W., & Tracy, S. (2008). High-impact firms: Gazelles revisited. Washington, DC, 1–82.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. Science, 355(6324), 483-485.

Athey, S. (2018). The impact of machine learning on economics. In The Economics of Artificial Intelligence: An Agenda. University of Chicago Press, 1–31.

Audretsch, D. B. (2012). Determinants of high-growth entrepreneurship. In Report prepared for the OECD/DBA International Workshop on High-Growth Firms: Local Policies and Local Determinants, Copenhagen, 1–37.

Barringer, B. R., Jones, F. F., & Neubaum, D. O. (2005). A quantitative content analysis of the characteristics of rapid-growth firms and their founders. Journal of Business Venturing, 20(5), 663–687.

Becchetti, L., & Trovato, G. (2002). The determinants of growth for small and medium sized firms. The role of the availability of external finance. Small Business Economics, 19(4), 291–306.

Beck, T., & Demirguc-Kunt, A. (2006). Small and medium-size enterprises: Access to finance as a growth constraint. Journal of Banking & Finance, 30(11), 2931–2943.

Birch, D. L., & Medoff, J. (1994). Gazelles. Labor Markets, Employment Policy and Job Creation, 159–167.

Bottazzi, G., & Secchi, A. (2006). Explaining the distribution of firm growth rates. The RAND Journal of Economics, 37(2), 235–256.

Breiman, L. (1996). Bagging predictors - Springer. Machine Learning, 140, 123–140.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Breiman, L. (2017). Classification and regression trees. Routledge.

Capon, N., Farley, J. U., & Hoenig, S. (1990). Determinants of financial performance: a meta-analysis. Management Science, 36(10), 1143–1159.

Carpenter, R. E., & Petersen, B. C. (2002). Is the growth of small firms constrained by internal finance? Review of Economics and Statistics, 84(2), 298–309.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785–794.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., … Li, Y. (2018). xgboost: Extreme Gradient Boosting.

Clements MP, Harvey DI. (2009). Forecast combination and encompassing. In Palgrave Handbook of Econometrics: Volume 2 Applied Econometrics, Mills TC, Patterson K (eds). Palgrave Macmillan: Basingstoke, U.K., 169–198.

Coad, A. (2007a). A closer look at serial growth rate correlation. Review of Industrial Organization, 31(1), 69–82.

Coad, A. (2007). Firm growth: A survey. Max Planck Institute of Economics, Papers on Economics and Evolution No. 0703, Jena, 1–72.

Coad, A., Daunfeldt, S.-O., Hölzl, W., Johansson, D., & Nightingale, P. (2014). High-growth firms: introduction to the special section. Industrial and Corporate Change, 23(1), 91–112.

Coad, A., & Tamvada, J. P. (2012). Firm growth and barriers to growth among small firms in India. Small Business Economics, 39(2), 383–400.

Daunfeldt, S.-O., Elert, N., & Johansson, D. (2014). The economic contribution of high-growth firms: do policy implications depend on the choice of growth indicator? Journal of Industry, Competition and Trade, 14(3), 337–365.

Daunfeldt, S.-O., Johansson, D., & Halvarsson, D. (2015). Using the Eurostat-OECD definition of high-growth firms: a cautionary note. Journal of Entrepreneurship and Public Policy, 4(1), 50–56.

Davidsson, P., Achtenhagen, L., Naldi, L., & others. (2010). Small firm growth. Foundations and Trends in Entrepreneurship, 6(2), 69–166.

Davidsson, P., & Delmar, F. (2006). High-growth firms and their contribution to employment: The case of Sweden 1987–96. Entrepreneurship and the Growth of Firms. Cheltenham: Elgar, 156–178.

Davidsson, P., & Henrekson, M. (2002). Determinants of the prevalance of start-ups and high-growth firms. Small Business Economics, 19(2), 81–104.

Delmar, F., Davidsson, P., & Gartner, W. B. (2003). Arriving at the high-growth firm. Journal of Business Venturing, 18(2), 189–216.

Delmar, F., & Wiklund, J. (2008). The effect of small business managers' growth motivation on firm growth: A longitudinal study. Entrepreneurship Theory and Practice, 32(3), 437–457.

Downie, J. (1958). The Competitive Process, Duckworth, London.

European Commission. (2010). Europe 2020: A strategy for smart, sustainable and inclusive growth: Communication from the commission. Publications Office of the European Union.

Fawcett, T. (2005). An introduction to ROC analysis Tom. Irbm, 35(6), 299–309.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences, 55(1), 119–139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 1189–1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of statistical software, 33(1), 1–22.

Gibrat, R. (1931). Les inégalits économiques. Sirey.

Gompers, P., Gornall, W., Kaplan, S. N., & Strebulaev, I. A. (2016). How do venture capitalists make decisions? (No. 22587). National Bureau of Economic Research, 1–63.

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3(3), 1157–1182.

Guzman, J., & Stern, S. (2015a). Nowcasting and placecasting entrepreneurial quality and performance (No. 20954). National Bureau of Economic Research, 1–68.

Guzman, J., & Stern, S. (2015b). Where is silicon valley? Science, 347(6222), 606–609.

Haltiwanger, J., Jarmin, R. S., & Miranda, J. (2013). Who creates jobs? Small versus large versus young. Review of Economics and Statistics, 95(2), 347–361.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data Mining, Inference and Prediction (Vol. 2). New York: Springer.

He, X., Chaney, N. W., Schleiss, M., & Sheffield, J. (2016). Spatial downscaling of precipitation using adaptable random forests. Water Resources Research, 52(10), 8217-8237.

Hannan, M. and Freeman, J. (1977). The Population Ecology of Organisations, American Journal of Sociology, 82 (3), 929-964.

Henrekson, M., & Johansson, D. (2010). Gazelles as job creators: a survey and interpretation of the evidence. Small Business Economics, 35(2), 227–244.

Hölzl, W. (2013). Persistence, survival, and growth: a closer look at 20 years of fast-growing firms in Austria. Industrial and Corporate Change, 23(1), 199–231.

Hoffman, A. N., & Junge, M. (2006). Documenting data on high-growth firms and entrepreneurs across 17 Countries. Fora.

Human, S. E., & Matthews, C. H. (2004). Future expectations for the new business. Handbook of Entrepreneurial Dynamics: The Process of Business Creation, 386–400.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer.

Kangasharju, A. (2000). Growth of the smallest: Determinants of small firm growth during strong macroeconomic fluctuations. International Small Business Journal, 19(1), 28–43.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction Policy Problems. American Economic Review, 105(5), 491–495.

Kuhn, M. (2018). caret: Classification and Regression Training. R package version 6.0-80.

Lipton, Z. C., Elkan, C., & Naryanaswamy, B. (2014). Optimal thresholding of classifiers to maximize F1 measure. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 225–239.

Littunen, H., & Tohmo, T. (2003). The high growth in new metal-based manufacturing and business service firms in Finland. Small Business Economics, 21(2), 187–200.

Machado, H. P. V. (2016). Growth of small businesses: a literature review and perspectives of studies. Gestão & Produção, 23(2), 419–432.

Marris, R. (1963). A Model of the Managerial enterprise. Quarterly Journal of Economics, 77 (2), 185-209.

Marris, R. (1964). The Economic Theory of Managerial Capitalism. Macmillan: London.

Megaravalli, A. V., & Sampagnaro, G. (2018). Predicting the growth of high-growth SMEs: evidence from family business firms. Journal of Family Business Management.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2018). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.

Miyakawa, D., Miyauchi, Y., & Perez, C. (2017). Forecasting Firm Performance with Machine Learning: Evidence from Japanese firm-level data. RIETI DP.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), 87–106.

OECD. (2010). High-Growth Enterprises: What governments can do to make a difference OECD, Studies on SMEs and Entrepreneurship.

OECD, & Eurostat. (2007). Eurostat − OECD Manual on Business Demography Statistics.

Penrose E. T. (1959); The Theory of the Growth of the Firm, Oxford: Basil Blackwell; and New York: Wiley.

Peters, A., & Hothorn, T. (2018). ipred: Improved Predictors.

R Core Team. (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria.

Sampagnaro, G., & Lubrano Lavadera, G. (2013). Identifying High Growth SMEs Through Balance Sheet Ratios. SSRN 2207550.

Sharchilev, B., Roizner, M., Rumyantsev, A., Ozornin, D., Serdyukov, P., & de Rijke, M. (2018). Web-based Startup Success Prediction. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2283–2291.

Storey, D. J. (1994). Understanding the small business sector. Routledge.

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of Imbalanced Data: A Review. International Journal of Pattern Recognition and Artificial Intelligence, 23(04), 687–719.

Therneau, T., & Atkinson, B. (2018). rpart: Recursive Partitioning and Regression Trees.

Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2), 3–28.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York.

Venkatraman, E. S. (2000). A permutation test to compare receiver operating characteristic curves. Biometrics, 56(4), 1134–1138.

Viner, J. (1952). Cost curves and supply curves. Chapter 10 in: AEA readings in price theory, 198–232, 1952, edited by George Stigler and Kenneth Boulding. (Originally published in Zeitschrift für Nationaloekonomie, Vol III, 1931, 23–46.)

Wallsten, S. J. (2000). The effects of government-industry R&D programs on private R&D: the case of the Small Business Innovation Research program. The RAND Journal of Economics, 82–100.

Weinblat, J. (2018). Forecasting European high-growth Firms-A random forest approach. Journal of Industry, Competition and Trade, 18(3), 253–294.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1–29.

Wickham, H., François, R., Henry, L., & Müller, K. (2018). dplyr: A Grammar of Data Manipulation.

Wiklund, J. (1998). Small Firm Growth and Performance Entrepreneurship and Beyond. Doctoral Thesis - Jönköping International Business School, 1–361.

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in {C++} and {R}. Journal of Statistical Software, 77(1), 1–17.

APPENDIX A – DESCRIPTIVE STATISTICS

TABLE 16 Descriptive statistics for preprocessed learning (left pane) and test (right pane) data sets based on the full data set.

Variable	n	mean	sd	min	max	Variable	n	mean	sd	min	max
HighPersonnelGrowth	43,532	1.95	0.22	1.00	2.00	HighPersonnelGrowth	9,975	1.95	0.21	1.00	2.00
HighRevenueGrowth	43,532	1.92	0.27	1.00	2.00	HighRevenueGrowth	9,975	1.93	0.26	1.00	2.00
Personnel	43,532	0.00	1.00	-0.20	49.69	Personnel	9,975	-0.02	0.85	-0.20	32.85
LagPersonnelGrowth	43,532	0.00	1.00	-10.41	14.17	LagPersonnelGrowth	9,975	0.07	0.97	-4.74	9.76
Revenue	43,532	0.00	1.00	-0.09	106.38	Revenue	9,975	-0.01	0.69	-0.09	44.08
LagRevenueGrowth	43,532	0.00	1.00	-11.31	15.38	LagRevenueGrowth	9,975	0.04	0.92	-9.46	12.59
Productivity	43,532	0.00	1.00	-0.22	148.84	Productivity	9,975	0.00	0.75	-0.22	54.80
LagProductivityGrowth	43,532	0.00	1.00	-16.32	16.53	LagProductivityGrowth	9,975	-0.03	0.96	-12.45	14.68
Profit	43,532	0.00	1.00	-92.44	2.29	Profit	9,975	-0.04	1.70	-90.48	2.02
ProfitGrowth	43,532	0.00	1.00	-39.28	111.13	ProfitGrowth	9,975	0.00	1.23	-44.45	58.44
Age	43,532	0.00	1.00	-1.85	8.05	Age	9,975	0.08	1.06	-1.85	8.14
NumOfPos	43,532	0.00	1.00	-0.13	55.36	NumOfPos	9,975	-0.01	0.88	-0.13	47.93
PartOfAGroup	43,532	1.39	0.48	1.00	2.00	PartOfAGroup	9,975	1.38	0.48	1.00	2.00
ForeignOwned	43,532	1.11	0.30	1.00	2.00	ForeignOwned	9,975	1.10	0.30	1.00	2.00
Exporting	43,532	1.28	0.44	1.00	2.00	Exporting	9,975	1.26	0.44	1.00	2.00
Solidity	43,532	0.00	1.00	-75.48	1.22	Solidity	9,975	-0.03	0.96	-23.24	1.21
Rating	43,532	0.00	1.00	-0.91	4.58	Rating	9,975	0.06	1.10	-0.91	4.58
TangAssetsPerRev	43,532	0.00	1.00	-0.08	154.79	TangAssetsPerRev	9,975	-0.01	0.37	-0.08	28.29
TopDecGrossMargin	43,532	0.00	1.00	-8.42	22.01	TopDecGrossMargin	9,975	-0.17	0.96	-5.88	5.92
MedGrossMargin	43,532	0.00	1.00	-57.16	6.20	MedGrossMargin	9,975	-0.20	0.85	-7.78	5.57
PatCount	43,532	0.00	1.00	-0.02	95.88	PatCount	9,975	0.00	0.82	-0.02	76.31
ForeignSubsidiaries	43,532	1.16	0.36	1.00	2.00	ForeignSubsidiaries	9,975	1.16	0.36	1.00	2.00
CEOAge	43,532	0.00	1.00	-3.18	4.43	CEOAge	9,975	0.01	1.02	-3.18	4.09
CEOGender	43,532	1.09	0.29	1.00	2.00	CEOGender	9,975	1.10	0.29	1.00	2.00
Vc	43,532	1.04	0.20	1.00	2.00	Vc	9,975	1.04	0.21	1.00	2.00
Tekes	43,532	1.12	0.31	1.00	2.00	Tekes	9,975	1.10	0.31	1.00	2.00

Notes: Observations with missing values are omitted.

TABLE 17 Descriptive statistics for the full data set of young (≤ 10 years old) firms before preprocessing.

Variable	n	mean	sd	min	max
HighPersonnelGrowth	12,027	0.10	0.30	0.00	1.00
HighRevenueGrowth	12,027	0.14	0.34	0.00	1.00
Personnel	12,027	33.63	114.58	10.00	6,958.00
LagPersonnelGrowth	12,008	0.46	0.79	-5.54	6.65
Revenue	11,900	8,425,740.61	113,680,378.89	0.00	7,692,391,966
LagRevenueGrowth	11,748	0.48	0.90	-7.03	8.32
Productivity	11,900	240,104.79	1,739,021.51	0.00	173,543,824
LagProductivityGrowth	11,729	0.03	0.60	-7.31	7.31
Profit	11,942	-3.37	159.14	-9,033.30	3,666.70
ProfitGrowth	11,841	5.48	227.02	-8,083.30	9,536.50
Age	12,027	6.87	2.16	1.00	10.00
NumOfPos	12,027	1.50	1.87	1.00	52.00
PartOfAGroup	12,027	0.31	0.46	0.00	1.00
ForeignOwned	12,027	0.09	0.28	0.00	1.00
Exporting	12,027	0.15	0.35	0.00	1.00
Solidity	12,023	28.17	62.15	-2,667.60	99.90
Rating	11,575	26.07	19.84	3.00	99.00
TangAssetsPerRev	11,883	0.75	54.24	0.00	5,897.37
TopDecGrossMargin	12,027	24.18	8.64	-43.70	200.00
MedGrossMargin	12,027	7.61	18.89	-1,996.15	31.20
PatCount	12,027	0.25	10.98	0.00	771.00
ForeignSubsidiaries	12,027	0.11	0.32	0.00	1.00
CEOAge	11,098	45.28	8.73	20.00	79.00
CEOGender	11,117	0.10	0.31	0.00	1.00
Vc	12,027	0.05	0.21	0.00	1.00
Tekes	12,027	0.11	0.31	0.00	1.00

Notes: Observations with missing values: 1,610.

TABLE 18 Descriptive statistics for preprocessed learning (top pane) and test (bottom pane) data sets of young (≤ 10 years) firms.

Variable	n	mean	sd	min	max
HighPersonnelGrowth	8,413	1.91	0.29	1.00	2.00
HighRevenueGrowth	8,413	1.87	0.34	1.00	2.00
Personnel	8,413	0.00	1.00	-0.21	58.87
LagPersonnelGrowth	8,413	0.00	1.00	-7.68	8.04
Revenue	8,413	0.00	1.00	-0.08	65.71
LagRevenueGrowth	8,413	0.00	1.00	-8.34	8.10
Productivity	8,413	0.00	1.00	-0.13	85.44
LagProductivityGrowth	8,413	0.00	1.00	-12.17	12.33
Profit	8,413	0.00	1.00	-54.18	1.66
ProfitGrowth	8,413	0.00	1.00	-16.38	57.74
Age	8,413	0.00	1.00	-2.62	1.44
NumOfPos	8,413	0.00	1.00	-0.28	27.32
PartOfAGroup	8,413	1.33	0.47	1.00	2.00
ForeignOwned	8,413	1.09	0.29	1.00	2.00
Exporting	8,413	1.16	0.37	1.00	2.00
Solidity	8,413	0.00	1.00	-45.83	1.20
Rating	8,413	0.00	1.00	-1.17	3.85
TangAssetsPerRev	8,413	0.00	1.00	-0.06	85.68
TopDecGrossMargin	8,413	0.00	1.00	-7.77	20.01
MedGrossMargin	8,413	0.00	1.00	-51.63	5.19
PatCount	8,413	0.00	1.00	-0.02	70.60
ForeignSubsidiaries	8,413	1.12	0.32	1.00	2.00
CEOAge	8,413	0.00	1.00	-2.91	3.88
CEOGender	8,413	1.10	0.30	1.00	2.00
Vc	8,413	1.05	0.22	1.00	2.00
Tekes	8,413	1.11	0.32	1.00	2.00

Variable	n	mean	sd	min	max
HighPersonnelGrowth	2,004	1.89	0.31	1.00	2.00
HighRevenueGrowth	2,004	1.84	0.36	1.00	2.00
Personnel	2,004	-0.01	1.20	-0.21	49.80
LagPersonnelGrowth	2,004	0.09	0.92	-3.67	6.45
Revenue	2,004	0.00	1.18	-0.08	52.55
LagRevenueGrowth	2,004	0.04	0.91	-4.72	8.71
Productivity	2,004	-0.03	0.17	-0.13	2.82
LagProductivityGrowth	2,004	-0.07	0.98	-7.59	10.95
Profit	2,004	-0.08	2.20	-63.78	1.04
ProfitGrowth	2,004	-0.03	1.14	-23.14	30.35
Age	2,004	0.12	0.90	-2.62	1.44
NumOfPos	2,004	0.00	1.06	-0.28	18.12
PartOfAGroup	2,004	1.32	0.47	1.00	2.00
ForeignOwned	2,004	1.07	0.26	1.00	2.00
Exporting	2,004	1.14	0.34	1.00	2.00
Solidity	2,004	-0.02	1.07	-19.39	1.19
Rating	2,004	-0.01	1.07	-1.17	3.85
TangAssetsPerRev	2,004	-0.02	0.16	-0.06	2.76
TopDecGrossMargin	2,004	-0.17	0.91	-1.85	4.54
MedGrossMargin	2,004	-0.17	0.73	-1.55	4.90
PatCount	2,004	0.01	1.37	-0.02	61.17
ForeignSubsidiaries	2,004	1.12	0.32	1.00	2.00
CEOAge	2,004	-0.02	1.02	-2.79	3.07
CEOGender	2,004	1.11	0.32	1.00	2.00
Vc	2,004	1.04	0.20	1.00	2.00
Tekes	2,004	1.10	0.30	1.00	2.00

Notes: Observations with missing values are omitted.

APPENDIX B – TECHNICAL SUMMARY OF THE MACHINE LEARNING ALGORITHMS

In this appendix, I provide some mathematical background for the machine learning algorithms employed in this master's thesis. The presentations here are abridged versions. Please, see the references for full treatment.

The CART Algorithm

The CART algorithm is a tree-based nonparametric method for regression and classification. Following Breiman's (2017, p. 27–36)[32] notation, the CART algorithm in a classification setting can be presented as follows.

In the learning sample $\mathcal{L}(x_1, x_2, \dots, x_n)$ with a predictor space of $x_1, x_2, \dots, x_i$, and for a $\mathcal{J}$ class problem, let's denote N_j as the number of units in class j. The prior probabilities are taken as proportions as follows: $\pi(j) = N_j/N$, where N represents the total number of units in $\mathcal{L}$. Let $N(t)$ stand for the number of units in node t, for which holds that $x_n \in t$. Finally, a few probability estimates can be derived. First, $p(j,t) = \pi(j)\, N_j(t)/N_j$ stands for the resubstitution estimate of the probability of a unit being in class j and falling into node t. Second, $p(t) = \sum_j p(j,t)$ is the definition for the resubstitution estimate of the probability that any unit falls into node t. Third, the resubstitution estimate of the probability that a unit is in class j given its existence in node t is given by $p(j|t) = p(j,t)/p(t)$.

Growing a classification tree entails four components:

1. Generating a set Q of binary questions asking whether $x \in A$, where $A \subset X$,
2. The goodness of a split criterion $\phi(s,t)$,
3. A stopping rule,
4. A rule for assigning a class for each terminal node.

The set Q of binary questions of the form $\{Is\ x \in A?\}$ creates a set S of s splits at each node of the tree. If the answer is positive in node t, the unit is assigned to left descendant node t_L. In the case of a negative answer the unit goes to the right descendant node t_R. Given an impurity function ϕ, the measure of impurity is defined as $i(t) = \phi(p(1|t), \dots, p(\mathcal{J}|t))$. Therefore, the decrease in impurity in a split s is given by $\Delta i(s,t) = i(t) - p_R i(t_R) - p_L i(t_L)$, which is the criterion for the goodness of a split. At each node t, the split s is chosen which maximizes a measure of the goodness of a split. Of a few options for the functional form of $i(t)$, the gini index is chosen to be applied in this master's thesis.[33] The gini index is given by: $i(t) = 1 - \sum_{j=1}^{\mathcal{J}} p^2(j|t)$.

[32] Breiman's (2017) book on classification and regression trees was originally published in 1984.

[33] Alternative measures of impurity include simple classification error and cross-entropy.

The change in overall impurity is denoted by $\Delta I(s,t) = \Delta i(s,t)p(t)$. Now, the simplest stopping rule is the following: set a threshold β, for which it must hold that $max_{s\in S}\Delta I(s,t) < \beta$, for the tree to keep growing. When the condition doesn't hold, a terminal node is assigned. The growing continues until there are only terminal nodes left. The final tree is denoted by T and the terminal nodes by $\widetilde{T}$. Finally, the class assignment rule $j^*(t)$ assigns a class to each terminal node $t \in \widetilde{T}$ by minimizing the resubstituition estimate of the probability of misclassification given that a unit falls into node t, which is given by $\sum_{j\neq j(t)} p(j|t)$. Including altered misclassification costs for different classes, the class assignment rule $j^*(t)$ is modified to assigning classes based on minimizing the expected misclassification cost, given by $\sum_j c(i|j)p(j|t)$, where $c(i|j)$ is the cost of misclassifying class j as class i. With the assigned classes, predictions can be made for any new observation by predicting the outcome of the assigned class.

Bagging and Random Forest

Bagging and random forest (RF) algorithms are based on the idea of aggregating several CART trees to reduce variance for improved out-of-sample performance. The algorithms are constructed similarly despite one difference in strategy when considering predictors for a split, as noted below, and therefore will be considered here in tandem.[34] Following Hastie et al.'s (2009, p. 588) notation, bagging and RF can be summarized for classification as follows:

1. For $b = 1$ to B:
    (a) Pull a bootstrap sample $\boldsymbol{Z}^*$ of size N from the training sample.
    (b) Grow a CART tree T_b to $\boldsymbol{Z}^*$ until the given minimum node size n_{min} is attained with the following modifications:
        i. Bagging: consider all predictors from the p variables at each split. RF: Select m variables at random from the p variables at each split. (Usually $m = \sqrt{p}$.)
        ii. Pick the best predictor and split point among p (bagging) or m (RF) and split the node into two child nodes following the CART algorithm.
        iii. Continue growing the tree until the stopping rule by CART is reached.
2. Output the ensemble of CART trees $\{T_b\}_1^B$.
3. Prediction at a new point x by a majority vote:

$$\hat{C}^B_{rf/bagging}(x) = majority\ vote\ \{\hat{C}_b(x)\}_1^B.$$

[34] For original sources, see, Breiman (1996) for bagging and Breiman (2001) for random forest.

Gradient Boosting

Boosting algorithms grow decision trees in sequences, enhancing the learning process in areas where it does not perform well. In this master's thesis, I employ a gradient boosting algorithm by Chen and Guestrin (2016) in a classification setting, which can be summarized in a generic manner for a binary classification problem following Hastie et al.'s (2009, 359–387) notation below.

For inputs in gradient boosting machines, a prediction rule $f_M(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$ is set, where $L(y, f(x)) = -(y_i f(x_i) - \log{(1 + \exp{(f(x_i))}}))$ is a binomial loss function and I is the event indicator function. Therefore, the algorithm is trained and used through the following process.

1. Initialize $f_0(x)$.
2. For $m = 1$ to M:
   (a) Compute elements of the negative gradient
   $$r_{im} = -\left[\frac{\partial \sum_{i=1}^{n} L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}, \quad i = 1, 2, \dots, n.$$
   (b) Fit a regression tree to targets r_{im}, $i = 1, 2, \dots, n$, resulting in terminal regions R_{jm}, $j = 1, 2, \dots, J_m$.
   (c) Compute updates
   $$y_{jkm} = argmin_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma), \quad j = 1,2, \dots, J_m.$$
   (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.
3. Output $\hat{f}(x) = f_M(x)$.
4. Prediction in a new point x using $\hat{f}(x)$.

Single-layered Artificial Neural Network

Artificial neural networks estimate parameters for complex linear combinations, which can be used for prediction. Following Hastie et al.'s (2009, p. 392–397) notation, a single-layered neural network for K-class classification can be presented as follows.

With a set of inputs $X = (X_1, \dots, X_p)$, derived hidden units $Z = (Z_1, \dots, Z_M)$ as linear combinations of the input variables and target Y_k are further derived as a function of linear combinations of the Z_m using an activation function σ. The output $f_k(X)$, which is used for prediction in the test sample, is determined by the *softmax* function $g_k(T)$ using the vector of derived features $T = (T_1, \dots, T_K)$:

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), \quad m = 1, \dots, M, \quad \sigma = 1/(1 + e^{-v}),$$

$$T_k = \beta_0 + \beta_k^T Z, \quad k = 1, \dots, K,$$

$$f_k(X) = g_k(T) = \frac{e^{T_k}}{\sum_{l=1}^{K} e^{T_l}}, \quad k = 1, \dots, K.$$

Estimating the set of parameters θ for the linear combinations is carried out by minimizing cross-entropy given by $R(\theta) = -\sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} log f_k(x_i)$ and applying a process called back-propagation to avoid overfitting.

APPENDIX C – ADDITIONAL RESULTS

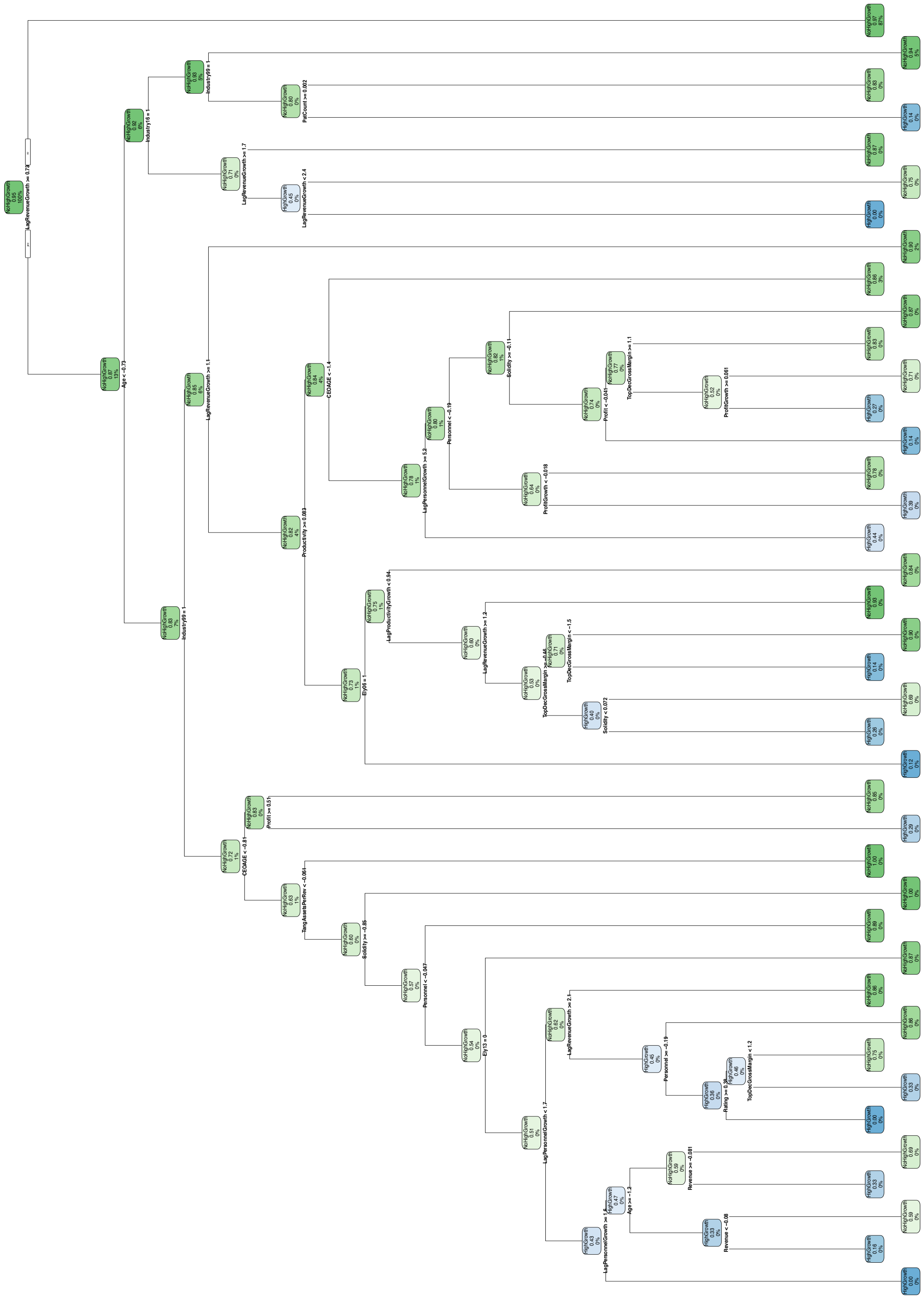

FIGURE 11 A tree chart of a trained CART algorithm in the learning sample for the baseline model.

TABLE 19 Confusion matrices for the baseline model's prediction results in the test sample for different classifiers.

Predicted Class

	Observed Class			Observed Class	
Lpm			**Logit**		
	High Growth	Non-High Growth		High Growth	Non-High Growth
High Growth	161	841	High Growth	150	653
Non-High Growth	331	8642	Non-High Growth	342	8830
CART			**Bagging**		
	High Growth	Non-High Growth		High Growth	Non-High Growth
High Growth	101	397	High Growth	26	71
Non-High Growth	391	9086	Non-High Growth	466	9412
Boosting			**Random Forest**		
	High Growth	Non-High Growth		High Growth	Non-High Growth
High Growth	157	608	High Growth	84	208
Non-High Growth	335	8875	Non-High Growth	408	9275
ANN			**Ensemble**		
	High Growth	Non-High Growth		High Growth	Non-High Growth
High Growth	133	522	High Growth	85	215
Non-High Growth	359	8961	Non-High Growth	407	9268

Notes: Rows represent predictions and columns the observed response, as in numbers of firms belonging to the class. The concept of a confusion matrix is described in Chapter 4.2. Total number of firms in the test sample: 9,975.

TABLE 20 Out-of-sample prediction results using class weights (top pane) and downsampling (bottom pane) in training.

Classifier	AUC	Threshold	Sensitivity	PPV	F-score	FPR	Accuracy	Threshold	Sensitivity	PPV	F-score	FPR	Accuracy
Lpm	0.5727	0.5000	0.0000	NaN	NaN	0.0000	0.9507	0.1002	0.3272	0.1607	0.2155	0.0887	0.8825
Logit	0.5872	0.5000	0.0020	0.3333	0.0040	0.0002	0.9506	0.1071	0.3049	0.1868	0.2317	0.0689	0.9003
CART	0.5221	0.5000	0.0264	0.3023	0.0486	0.0032	0.9490	0.1290	0.2053	0.2028	0.2040	0.0419	0.9210
Bagging	0.5969	0.5000	0.0142	0.2500	0.0269	0.0022	0.9493	0.3600	0.0752	0.3058	0.1207	0.0089	0.9460
Boosting	0.5000	0.5000	1.0000	0.0493	0.0940	1.0000	0.0493	0.5000	0.0000	NaN	NaN	0.0000	0.9507
RF	0.6439	0.5000	0.0000	NaN	NaN	0.0000	0.9507	0.1283	0.2195	0.2634	0.2395	0.0318	0.9312
ANN	0.5929	0.5000	0.0000	NaN	NaN	0.0000	0.9507	0.1150	0.2927	0.1887	0.2295	0.0653	0.9031
Ensemble	0.6274	0.5000	0.0000	0.0000	NaN	0.0003	0.9504	0.2371	0.1565	0.2884	0.2029	0.0200	0.9393

Classifier	AUC	Threshold	Sensitivity	PPV	F-score	FPR	Accuracy	Threshold	Sensitivity	PPV	F-score	FPR	Accuracy
Lpm	0.5727	0.5000	0.0000	NaN	NaN	0.0000	0.9507	0.1002	0.3272	0.1607	0.2155	0.0887	0.8825
Logit	0.5916	0.5000	0.6687	0.0917	0.1613	0.3437	0.6569	0.7451	0.2683	0.1777	0.2138	0.0644	0.9027
CART	0.5589	0.5000	0.6585	0.0825	0.1466	0.3800	0.6219	0.8193	0.3516	0.1187	0.1774	0.1355	0.8392
Bagging	0.6123	0.5000	0.7012	0.1033	0.1801	0.3157	0.6851	0.8400	0.1301	0.2540	0.1720	0.0198	0.9382
Boosting	0.5972	0.5000	0.6667	0.1034	0.1791	0.2998	0.6985	0.7364	0.2785	0.1908	0.2264	0.0613	0.9062
RF	0.6180	0.5000	0.6992	0.1140	0.1960	0.2820	0.7171	0.6553	0.2967	0.2068	0.2437	0.0591	0.9092
ANN	0.5922	0.5000	0.6789	0.0917	0.1616	0.3488	0.6525	0.7345	0.3211	0.1648	0.2178	0.0845	0.8862
Ensemble	0.6108	0.5000	0.6768	0.1036	0.1797	0.3039	0.6951	0.7193	0.2886	0.2107	0.2436	0.0561	0.9116

Notes: Final model chosen by the largest cross-validated area under the ROC curve in the training sample. Classical 0.5 (left pane) and F1 optimized (right pane) thresholds in prediction. Lpm does not support weights or downsampling.

TABLE 21 Out-of-sample prediction results using upsampling (top pane) and SMOTE resampling (bottom pane) in training.

Classifier	AUC	Threshold	Sensitivity	PPV	F-score	FPR	Accuracy	Threshold	Sensitivity	PPV	F-score	FPR	Accuracy
Lpm	0.5727	0.5000	0.0000	NaN	NaN	0.0000	0.9507	0.1002	0.3272	0.1607	0.2155	0.0887	0.8825
Logit	0.5890	0.5000	0.6667	0.0948	0.1660	0.3303	0.6696	0.7156	0.3089	0.1745	0.2230	0.0758	0.8938
CART	0.5557	0.5000	0.5894	0.0992	0.1698	0.2777	0.7158	0.7325	0.4248	0.1403	0.2109	0.1351	0.8432
Bagging	0.5547	0.5000	0.1423	0.2389	0.1783	0.0235	0.9353	0.9600	0.0000	NaN	NaN	0.0000	0.9507
Boosting	0.6064	0.5000	0.6504	0.1148	0.1952	0.2601	0.7354	0.7302	0.2703	0.2131	0.2384	0.0518	0.9148
RF	0.6074	0.5000	0.3252	0.2073	0.2532	0.0645	0.9054	0.6340	0.0935	0.3538	0.1479	0.0089	0.9469
ANN	0.5819	0.5000	0.6646	0.0968	0.1690	0.3216	0.6777	0.7498	0.2967	0.1899	0.2316	0.0657	0.9029
Ensemble	0.6028	0.5000	0.4533	0.1585	0.2349	0.1249	0.8543	0.6718	0.1646	0.2755	0.2061	0.0225	0.9374

Classifier	AUC	Threshold	Sensitivity	PPV	F-score	FPR	Accuracy	Threshold	Sensitivity	PPV	F-score	FPR	Accuracy
Lpm	0.5727	0.5000	0.0000	NaN	NaN	0.0000	0.9507	0.1002	0.3272	0.1607	0.2155	0.0887	0.8825
Logit	0.5834	0.5000	0.5650	0.1114	0.1861	0.2339	0.7562	0.6891	0.2703	0.1762	0.2133	0.0656	0.9017
CART	0.5566	0.5000	0.5163	0.1041	0.1733	0.2305	0.7570	0.7171	0.4472	0.1142	0.1820	0.1799	0.8017
Bagging	0.5783	0.5000	0.4512	0.1333	0.2057	0.1523	0.8282	0.8000	0.0610	0.2778	0.1000	0.0082	0.9459
Boosting	0.5821	0.5000	0.4939	0.1494	0.2295	0.1458	0.8364	0.5854	0.3333	0.1667	0.2222	0.0865	0.8849
RF	0.6047	0.5000	0.4736	0.1657	0.2455	0.1237	0.8564	0.6075	0.2093	0.2384	0.2229	0.0347	0.9280
ANN	0.5744	0.5000	0.5203	0.1196	0.1945	0.1987	0.7875	0.7112	0.2947	0.1688	0.2147	0.0753	0.8936
Ensemble	0.5904	0.5000	0.4959	0.1411	0.2197	0.1566	0.8263	0.6703	0.2114	0.2306	0.2206	0.0366	0.9263

Notes: Final model chosen by the largest cross-validated area under the ROC curve in the training sample. Classical 0.5 (left pane) and F1 optimized (right pane) thresholds in prediction. Lpm does not support upsampling or SMOTE.