Houle, Stephanie; MacDonald, Ryan

Working Paper

Identifying nascent high-growth firms using machine learning

Bank of Canada Staff Working Paper, No. 2023-53

Provided in Cooperation with:
Bank of Canada, Ottawa

Suggested Citation: Houle, Stephanie; MacDonald, Ryan (2023) : Identifying nascent high-growth firms using machine learning, Bank of Canada Staff Working Paper, No. 2023-53, Bank of Canada, Ottawa,
https://doi.org/10.34989/swp-2023-53

This Version is available at:
https://hdl.handle.net/10419/297438

BANK OF CANADA
BANQUE DU CANADA

Staff Working Paper/Document de travail du personnel—2023-53

Last updated: October 16, 2023

Identifying Nascent High-Growth Firms Using Machine Learning

by Stephanie Houle[1] and Ryan Macdonald[2]

[1] Canadian Economic Analysis Department
Bank of Canada
SHoule@bankofcanada.ca

[2] Economic Analysis
Statistics Canada
Ryan.Macdonald@statcan.gc.ca

DOI: https://doi.org/10.34989/swp-2023-53 | **ISSN** 1701-9397

Acknowledgements

The authors would like to thank Louise Earl, Maryam Haghighi and Lin Shao as well as the audiences at the Society for Economic Dynamics Conference 2021, the Canadian Economics Association 2022 annual conference and members of the Central Performance and Impact Assessment Unit within Treasury Board Secretariat for their helpful comments. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Bank of Canada.

Abstract

Predicting which firms will grow quickly and why has been the subject of research studies for many decades. Firms that grow rapidly have the potential to usher in new innovations, products or processes (Kogan et al. 2017), become superstar firms (Haltiwanger et al. 2013) and impact the aggregate labour share (Autor et al. 2020; De Loecker et al. 2020). We explore the use of supervised machine learning techniques to identify a population of nascent high-growth firms using Canadian administrative firm-level data. We apply a suite of supervised machine learning algorithms (elastic net model, random forest and neural net) to determine whether a large set of variables on Canadian firm tax filing financial and employment data, state variables (e.g., industry, geography) and indicators of firm complexity (e.g., multiple industrial activities, foreign ownership) can predict which firms will be high-growth firms over the next three years. The results suggest that the machine learning classifiers can select a sub-population of nascent high-growth firms that includes the majority of actual high-growth firms plus a group of firms that shared similar attributes but failed to attain high-growth status.

Topics: Firm dynamics; Econometric and statistical methods
JEL codes: C55, C81, L25

Résumé

Les chercheurs étudient depuis des décennies la question de savoir comment prédire quelles entreprises vont connaître une croissance rapide et pour quelles raisons. Les entreprises qui croissent rapidement sont susceptibles d'introduire des innovations, de lancer de nouveaux produits ou de mettre en place des processus inédits (Kogan et autres, 2017), de devenir des entreprises phares (Haltiwanger et autres, 2013) et d'avoir une incidence sur la part globale du travail (Autor et autres, 2020; De Loecker et autres, 2020). Nous explorons l'utilisation de techniques d'apprentissage automatique supervisé pour cerner un groupe d'entreprises émergentes à forte croissance en nous basant sur des données administratives sur les entreprises canadiennes. Nous appliquons un ensemble d'algorithmes d'apprentissage automatique supervisé (modèle du filet élastique, forêt aléatoire et réseau neuronal) pour déterminer si un vaste éventail de variables concernant des entreprises canadiennes – données sur la situation financière et l'emploi tirées des déclarations fiscales, données descriptives (p. ex., secteur, emplacement géographique) et indicateurs de complexité (p. ex., activités industrielles multiples, détention par des étrangers) – peuvent prédire celles qui afficheront une forte croissance au cours des trois années à venir. Les résultats obtenus donnent à penser que les classifieurs d'apprentissage automatique peuvent sélectionner un sous-groupe d'entreprises émergentes composé d'une majorité d'entreprises qui enregistrent réellement une croissance élevée et d'autres qui, même en ayant des attributs similaires, n'ont pas réussi à atteindre un fort niveau de croissance.

Sujets : Dynamique des entreprises, Méthodes économétriques et statistiques
Codes JEL : C55, C81, L25

1 Why is predicting high-growth firms so important?

Firms that grow rapidly capture widespread interest as their dynamic growth propels them to generate significant employment growth in a short period of time (Kogan et al. 2017; Dixon and Rollin 2014; Daunfeldt and Halvarsson 2015; Henrekson and Johansson 2010; Bravo-Biosca and Westlake 2009; Storey 2016). They also have the potential to become superstar firms (Haltiwanger, Jarmin, and Miranda 2013) and are associated with the reduction in the aggregate labour share (Autor, Dorn, Katz, Patterson, and Van Reenen 2020; De Loecker, Eeckhout, and Unger 2020).

As such, these firms represent a class of firms that garner considerable interest from academics, policy makers and providers of firm financing, such as venture capitalists or lending institutions (Cote and Rosa 2016; Coad, Daunfeldt, Hölzl, Johansson, and Nightingale 2014). These firms have the potential to produce significant returns on investment and to impact tax revenues and labour markets. They often start small and young but then rapidly scale up (Daunfeldt and Halvarsson 2015; Haltiwanger et al. 2013; Davis, Haltiwanger, and Schuh 1996; Henrekson and Johansson 2010).

Traditional econometric approaches have demonstrated limited success in predicting high-growth firms (HGFs) (Storey 2016; Hölzl 2009; Coad 2009). The limited explanatory power of regressions, coupled with the intermittent, or sporadic, nature of high-growth periods, suggest that contemporaneous identification of HGFs is not feasible (Coad et al. 2014; Marsili 2001).

The latest advancements in computing capacities, combined with machine learning approaches to big data, allow us to better approach these topics. Machine learning methods have recently been applied in economics and have been found to be particularly useful when it comes to prediction problems (Athey 2018; Mullainathan and Spiess 2017).

Researchers have looked at using firm balance sheet data to predict firm outcomes, such as default on loans or worker productivity in hiring decisions (Chalfin et al. 2016). A paper similar to ours by Coad and Srhoj (2020) uses a Lasso on Croatian firm-level data and finds that firms with low inventory, higher previous employment growth and higher reported liabilities are more likely to become HGFs. Another paper on a panel of European firms finds that firm size, age, past growth, sales per employee, fixed asset ratio and debt ratio are quite important for predicting firm growth (Weinblat 2018). Finally, Miyakawa et al. (2017) use Japanese firms' credit reporting data and find the best predictor of firm growth is a composite index of firm creditworthiness built by the Japanese credit rating agency.

We examine whether, and how, high-performing firms can be identified contemporaneously using supervised machine learning techniques on Canadian administrative tax filing data.[1] Our paper combines a rich set of firm tax filing and attribute data with machine learning techniques to tease out important factors that are associated with HGFs prior to them becoming so. Supervised machine learning approaches have advantages for prediction over traditional econometric approaches.

We demonstrate that these algorithms can identify nascent HGF populations. However, the degree of performance needs further refinement. The supervised machine learning algorithms

[1] These firm-level data are typically available with a three-year lag in order to incorporate late or modified tax filings. Hence, this paper tries to predict high-growth firms three years in the future to reflect the typical lag in data availability from statistical institutions.

change the objective function from explaining the variation in the dependent variable (e.g., minimizing the squared error) to minimizing the prediction error (often the mean squared prediction error when working with continuous variables). By changing the objective function, additional techniques become available (decision trees, neural networks), and the way the data are employed changes.

The goal of our analysis is then twofold. First and foremost, we determine whether contemporaneous and lagged information, of the type that would be included in a firm's information circular or annual report, contains information about whether or not a firm is a candidate for the high-growth designation. Our results indicate this approach is fruitful, but further use and development of these data sources for prediction are warranted.

Second, we examine which types of supervised machine learning algorithms provide the strongest signal for predicting high growth. Results suggest that random forests and simple neural nets perform best for predicting which firms will be high performers; however, this is by only a small margin over a logistic model with elastic net regularization. Both top-performing models suggest that the variables that play the largest role in prediction are industry and size variables. This suggests that being in the right industry with room to grow is more important than most tax filing variables from firms.

Within the data set, roughly 12–13% of firms are high growth. This means that a firm selected at random from the firm population would have an approximately one in eight chance of being high growth. Using the supervised machine learning classifiers, we can select a population of nascent high-growth firms. We find that, based on the best performing algorithms from the study (random forest/neural network), a firm in the predicted population has a roughly one in four chance of being high growth. Moreover, the predicted nascent HGF population is considerably smaller than the total population. In effect, the algorithms seem to be able to help find the proverbial needle in a haystack, while simultaneously reducing the size of the haystack.

However, the magnitude of our results should not be taken as an overwhelming success. Our results indicate that three out of four predicted high-growth firms are false positives. So, while the likelihood is doubled that a firm selected at random from the population of interest includes a high-growth firm, the nascent HGF population continues to encompass a majority of non-high-growth firms. This could be due to features in the input data that may be missing and that we do not have data on, such as attributes of the business owner or CEO. In addition, some investments may take longer to come to fruition (Meulbroek et al. 1990), so our three-year window may not be enough for the factors that contribute to long-term high growth.

The remainder of our paper is organized as follows. Section 2 describes the data set employed, as well as the definitions used to identify HGFs. Section 3 provides a statistical summary of the data set and how it relates to high-growth firms, while Section 4 discusses the methodology. Section 5 provides the results of the machine learning exercise and Section 6 concludes.

2 Data and data set preparation

2.1 Data

The data set we employ is derived from two sources. We collect firm-specific data from the the National Accounts Longitudinal Micro File (NALMF) that is maintained by Statistics Canada. Industry-level information is taken from publicly available industry-level multifactor productivity accounts (Statistics Canada 2018).

The NALMF is an enterprise-level data set that can be used to construct longitudinal firms for studying different phenomenon. For the purposes of examining high-growth firms, we extract a subset of firms for the 2010 to 2015 period. This subset includes incorporated firms that are both foreign-controlled and Canadian-controlled private corporations (CCPCs).

We define high-growth firms based on growth rates of employment and revenue in the 2012 to 2015 period. We also collect a set of financial and firm characteristics necessary for prediction. NALMF financial variables are populated from administrative data sources, including the General Index of Financial Information, Goods and Services Tax (GST) and T2 and T4 firm tax files (summary in Table 1). NALMF firm structure variables, such as location, industry activity or foreign ownership status, are based on Statistics Canada's Business Register. A full list of variables is available in Appendix Table A-1.

Table 1: Data sources and variables

Data source	Variables
NALMF	Revenue, profits, payroll, total costs, employment, assets, liabilities, shareholder equity, revenue, retained earnings, total intangible assets, sales and profits (one- and two-year lags of these variables).
Derived	Labour productivity, one- and two-year change in retained earning, one- and two-year change in assets, debt-to-equity, sum of R&D in the last three years, intermediate inputs and gross output.
Business Register	Industry, firm age and years since incorporation. Indicators for: NAICS, province, multiple province firms, multiple establishement firms, multiple NAICS firms, foreign ownership.

2.2 Defining high-growth firms

Traditionally, classification of HGFs takes one of two forms (Coad et al. 2014; Henrekson and Johansson 2010). In one case, a relative measure, such as the top 5% of firms in the growth rate distribution, is employed. The alternative is to choose an absolute threshold that a firm must cross to be considered high growth. We employ the latter approach in this paper.

A firm is classified as high growth if it:

- has a compound annual growth rate for employment or revenue of at least 20% from 2012 to 2015,
- has positive revenue in both those years, and
- has at least 10 employees in 2015.

This measure is similar to the definition used by the Organisation for Economic Co-operation and Development (OECD 2007), but includes a different minimum size measure to limit the effect of micro-firm growth.

The presence of micro-firms (those firms with fewer than 10 employees) creates a challenge for defining high-growth firms because micro-firms constitute the majority of enterprises in the data set. These micro-firms can exhibit significant growth rates due to, for example, changing from one to two employees, but this does not coincide with the notion of high growth that is typically intended when researchers look at firm behaviour (Cote and Rosa 2016; Coad et al. 2014). This leads to a small size class bias when measuring high-growth firms. It is therefore desirable to have a method for distinguishing between high growth related to important increases in the number of jobs in a firm, and increases that produce high growth rates for a firm but do not represent an important change in the number of jobs.

A number of approaches exist to limit the effect of the small size class bias. One is to exclude all firms with fewer than 10 employees in the first year. This is the approach followed by the OECD-Eurostat guidelines (OECD 2007). Alternatively, Clayton et al. (2013) propose a kink point method for firms with fewer than 10 employees that only include those that grow by at least eight jobs over the three years. The strength of this "kink point approach" is that it allows for all firms to be included in the definition. However, it imposes a more stringent requirement on small firms than on large ones.

We employ a hybrid approach. All HGFs are classified based on their growth rates and the number of employees they have in their final year. As a result, micro-firms that pass the 10 employee threshold are counted as high-growth firms as long as their growth rate is sufficiently high. This definition allows for the inclusion of a more complete firm population for the training data set used to parameterize models, while simultaneously limiting the effect of the small firm size class bias.

Alternative definitions of firm growth could be explored. Firms could be categorized as negative, slow and fast growth. Persistently high-growth firms could also be examined. Given the focus of this paper and the cross-sectional use of data to predict high-growth firms three years ahead, these alternative definitions are outside the scope of this paper.

2.3 Data preparation

The data extraction from NALMF requires additional manipulations to form the longitudinal units in order to correct for simple (one-to-one) changes in administrative data structure. Firms with more complex changes, such as one-to-many or many-to-one, are not included. These tend to be larger, more complex entities. They are relatively few in number, and despite constituting an important share of economic activity, they represent older, larger and more established firms that rarely reach the high-growth threshold. As such, their exclusion does not importantly affect the data set constructed for high-growth firm prediction. The subsequent data set includes only those firms that survive from 2012 to 2015 for which a longitudinal structure can be created.

The data set also contains missing data values for a number of enterprises. For example, a firm may not have a reported value for profits in one year. In a longitudinal data set, there are two options for dealing with missing values. One is to filter the data to remove firms with missing values. However, this can lead to important information loss, particularly as different variables can exhibit different patterns for missing values, which can lead to a significant number of firms being removed. The alternative is to impute values for the missing observations. This preserves information in the data set at the cost of adding noise (or measurement error) to the variables.

Given a large number of explanatory variables that can be included in machine learning prediction, the goal here is to avoid dropping observations due to any of the variables containing missing values. To impute values, multiple imputation with cascading equations (MICE) is applied to the data set (Van Buuren and Groothuis-Oudshoorn 2011; Burgette and Reiter 2010).

This data set is subsequently filtered to remove all enterprises from public sectors (Education – NAICS 61, Health care and social assistance – NAICS 62 and Public administration – NAICS 91). These are sectors that have an important degree of government involvement, may have different profit maximizing schemes and are typically excluded in these types of analyses. Firms are also filtered to include only those with a minimum of one employee in the initial year, and to remove firms with zero revenue in the initial or final year. This leads to a little over 130,000 firms for analysis, split into approximately 104,000 in the training set and 26,000 in the testing set (80%/20% split).

From this population, high-growth firms are identified based on the minimum growth threshold (20%) and the minimum number of employees in 2015 (10 employees). The common requirement for high-growth identification ensures that the same set of firms is used for examining both employment high-growth firms and revenue high-growth firms. This increases the comparability of resulting outputs as the input data set has the same set of conditions for both types of high-growth firms.

3 Who are these high-growth firms?

Before reporting the results from machine learning algorithm predictions, we discuss a number of salient features of the data. We explore these features for the 2012 cross-section of firms and compare firms that end up being high growth by 2015 versus those that do not. We provide this basic information about important characteristics of high-growth firms to help inform the discussion of how the models perform for prediction.

3.1 Growth characteristics

High-growth firms are rare in the data set. From our sample, 11.5% of firms are employment high-growth, 12.9% of firms are revenue high-growth and 6.6% meet both high-growth criteria. These firms occupy the right-hand tail of the firm growth distributions (Chart 1, Chart 2), which exhibit kurtosis and skewness to the right. The right skewness occurs as firms' growth rates are bounded by -100% on the negative side (it is not possible to have negative employment and we eliminate firms with zero employment or revenue) but do not face a similar constraint for positive changes.

Figure 1: Density of firms' average annual employment growth rate from 2012 to 2015

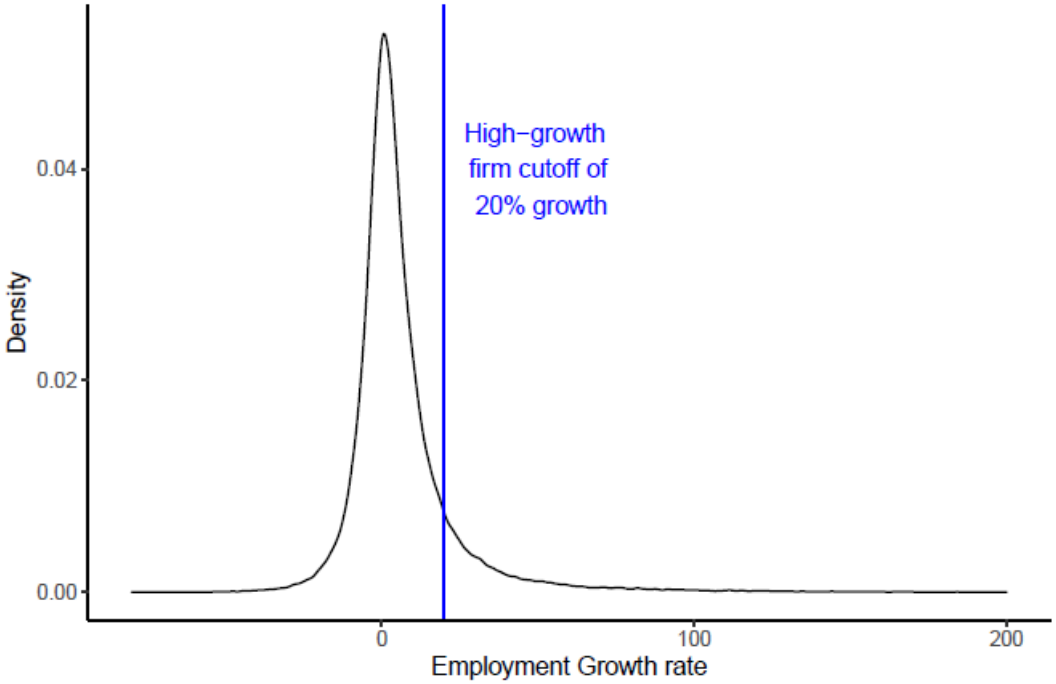
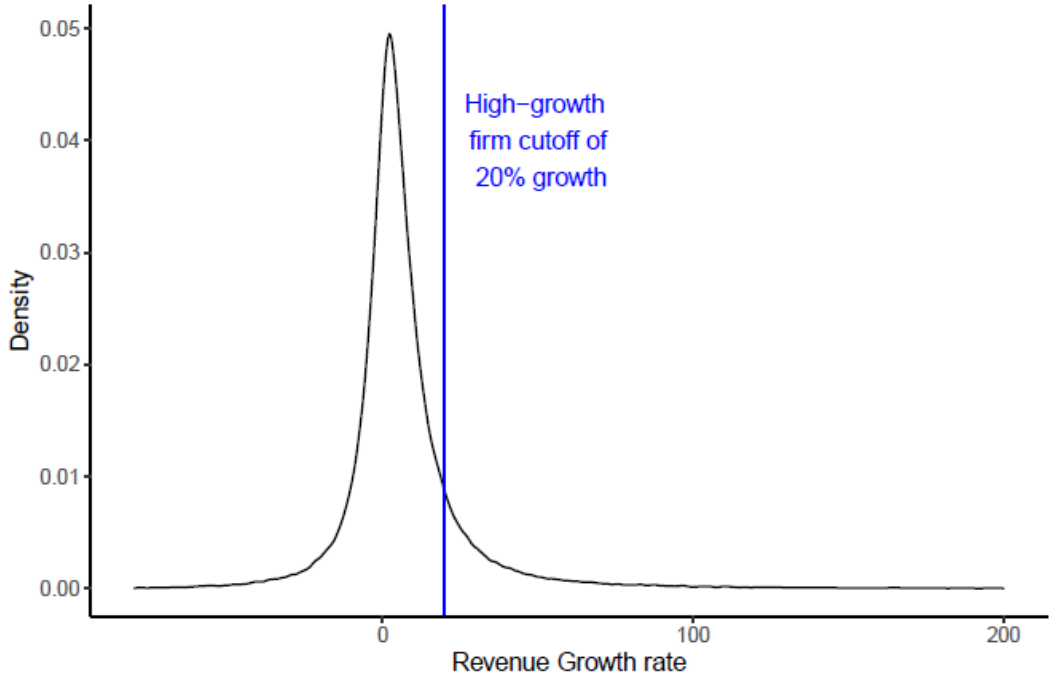


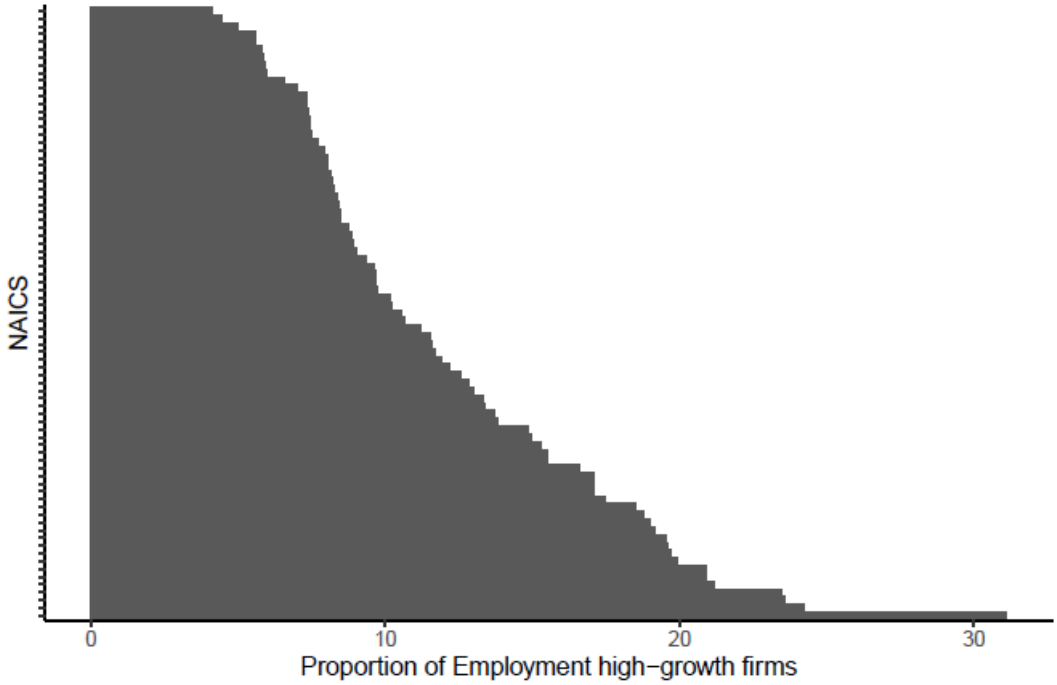
Figure 2: Density of firms' average annual revenue growth rate from 2012 to 2015



3.2 They are in the right industry at the right time

High-growth firms are present regardless of the way the data are categorized. In this sense, they are a general phenomenon that appears across the economy (Henrekson and Johansson 2010). However, the number and proportion of high-growth firms can differ importantly depending on the industry or level of firm complexity. When high-growth firms are divided among NAICS industries, high-growth firms are found in all industries regardless of whether employment or revenue is used as the variable of interest (Chart 3, Chart 4). The proportion of high-growth firms ranges from a high of over 30% to a low of less than 5%.

Figure 3: Every industry contains high-growth firms but some contain more[2]



The top 10 industries with the highest proportion of high-growth firms represent a broad range of industries and are not concentrated only in information technology industries (Tables 2 and 3). Moreover, the industries containing the largest number of employment high-growth firms are not always the same industries as for revenue high-growth firms.

This does not mean that revenue high-growth firms and employment high-growth firms are unrelated. Within the data set we find a positive correlation between the proportion of employment high-growth firms and revenue high-growth firms (Chart 5). Only 11.5% are employment high-growth firms and 12.9% are revenue high-growth firms. Altogether, 6.6% of firms fulfill both high-growth definitions. As a general rule across industries, the higher the proportion of employment high-growth firms, the higher the proportion of revenue

[2]We abstract from listing the NAICS numbers because some some industries contain too few high-growth firms. This is to maintain confidentiality and comply with the *Statistics Act*.

[3]We abstract from listing the NAICS numbers because some some industries contain too few high-growth firms. This is to maintain confidentiality and comply with the *Statistics Act*.

Figure 4: Every industry contains high-growth firms but some contain more[3]

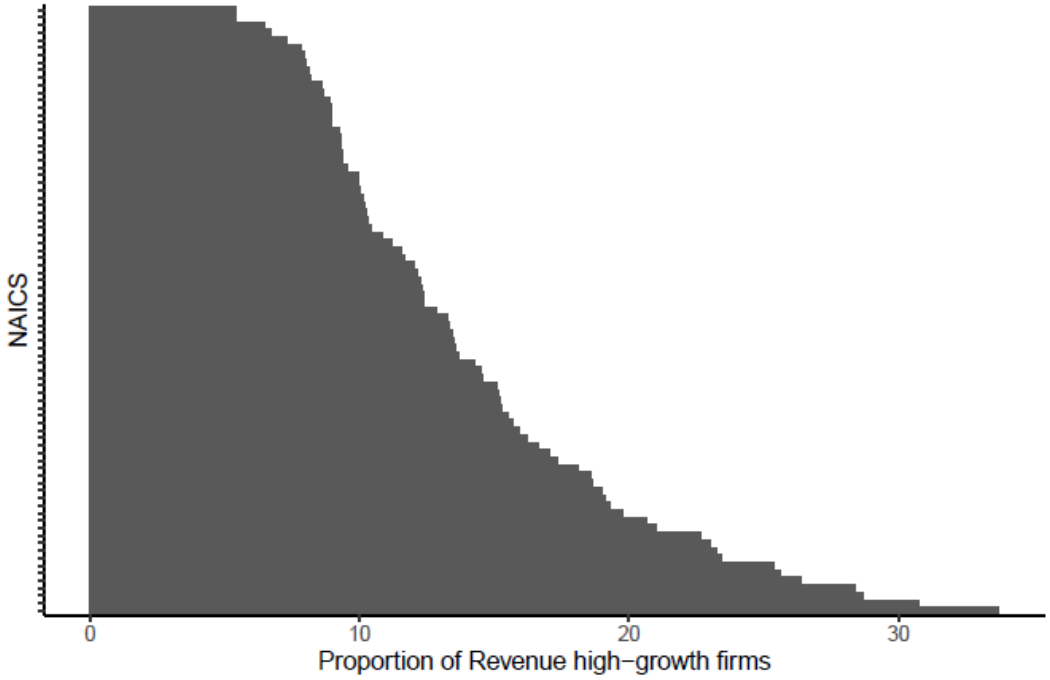


Table 2: Top 10 industries for proportion of employment high-growth firms: 2012–2015

Rank	Employment NAICS	NAICS description
1	519	Other information services
2	512	Motion picture and sound recording industries
3	211	Oil and gas extraction
4	518	Data processing, hosting, and related services
5	312	Beverage and tobacco product manufacturing
6	492	Couriers and messengers
7	114	Fishing, hunting and trapping
8	213	Support activities for mining, and oil and gas extraction
9	236	Construction of buildings
10	562	Waste management and remediation services

Table 3: Top 10 industries for proportion of revenue high-growth firms: 2012–2015

Rank	Revenue NAICS	NAICS description
1	211	Oil and gas extraction
2	518	Data processing, hosting, and related services
3	519	Other information services
4	312	Beverage and tobacco product manufacturing
5	114	Fishing, hunting and trapping
6	512	Motion picture and sound recording industries
7	112	Animal production and aquaculture
8	517	Telecommunications
9	526	Funds and other financial vehicles
10	523	Securities, commodity contracts, and other financial investment

high-growth firms. This suggests that the two are interrelated, but the way in which they interact requires further investigation. It may be the case that revenue growth precedes employment growth such that a lag exists between the two measures, as is found in OECD (2017). They may appear contemporaneous here due to the use of annual data and a three-year time span for the high-growth definition.

It may equally be the case that the two must increase in concert as firms take advantage of positive demand changes, or as firms overcome managerial challenges and develop internal mechanisms for controlling the flow of information. Industry heterogeneity, such as the degree of labour intensity for an industry, is likely also contributing to the differences in correlation between the two growth rates.

3.3 Complexity and room to grow

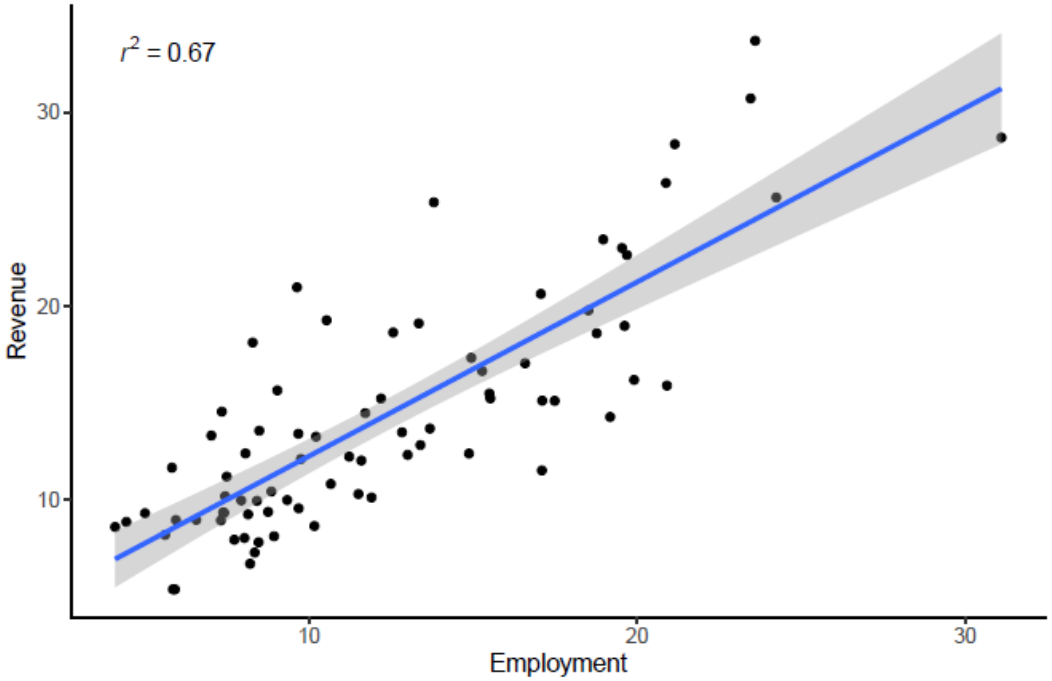
For continuous variables, contemporaneous relationships suggest little if any correlation between levels of firm assets, revenues, payrolls, R&D or debts with whether or not a firm becomes high growth. In effect, Gilbrat's law appears to hold, and the explanatory value of these variables is often weak. The exception is age, which is known to be an important determinant of firm growth (Haltiwanger et al. 2013).

The data set also contains a number of indicators for firm complexity summarized in Chart 6, and these illustrate a number of important features of high-growth firms. Within the data set, it is possible to distinguish among:

- firms with a single establishment and with multiple establishments,
- firms that operate within a single province and firms that operate in multiple provinces,
- firms that engage in multiple activities (NAICS industries) and
- whether a firm is majority foreign-controlled or not.

Across indicators of complexity, the proportion of high-growth firms is always higher for single indicator firms than for multiple indicator firms (darker colour bars in Chart 6). Single firms, be they single activity, single establishment or single province, are smaller, less complex firms with more room to grow. In effect, the relative scarcity of more complex high-growth firms is suggestive of a process whereby high growth may be dependent more

Figure 5: Proportion of high-growth firms by NAICS grouping (%)



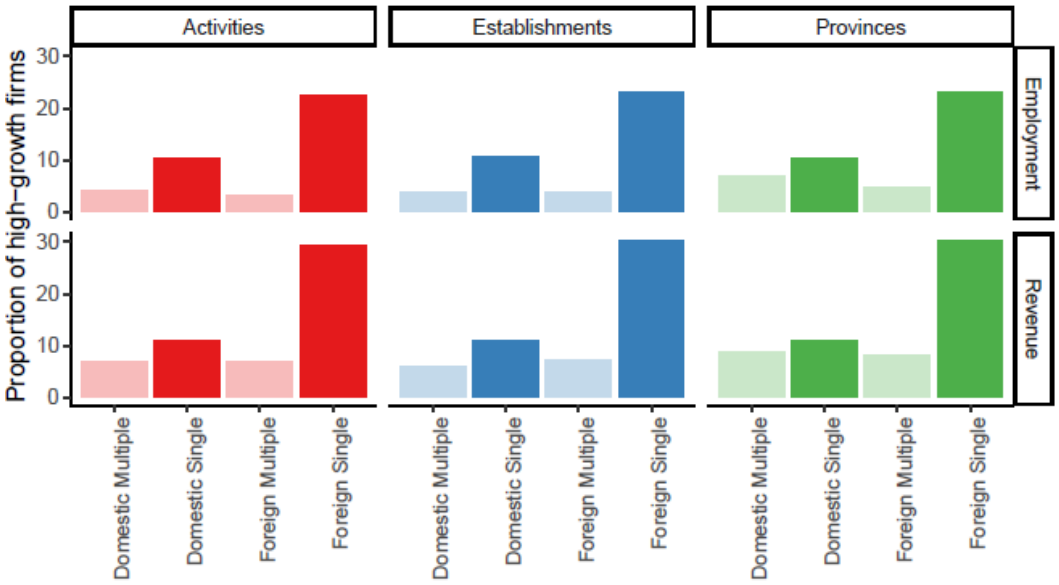
on a maturation process or on a firm's ability to overcome some form of obstacle. These obstacles may be related to management (e.g., HR strategies or strategic direction), funding (e.g., accessing financing, reinvesting revenue growth), or overcoming informational hurdles related to recognizing opportunities, marketing or control of expanding firm functions.

Another salient feature of HGFs is the role of foreign control in producing a higher proportion of high-growth firms. Across all measures of complexity, foreign-owned enterprises are more likely to be high-growth firms than are domestic firms (most right-hand-side bars for each indicator). This likely reflects that foreign-controlled firms represent components of much larger firm structures, and that, while they are high-growth in their Canadian operations, they may not be high-growth overall. Moreover, these firms have likely addressed issues associated with growth (management functions, cash flow, financing, business functions, etc). As such, their propensity for high growth likely reflects competitive strategies (e.g., the acquisition of market share) as opposed to a maturation process. So, while they are high-growth firms in Canada, they likely represent a different form of high-growth firm than do domestic firms. However, further work is necessary to identify differences and similarities.

4 Supervised machine learning models

To assess the presence of high-growth firms and the ability of supervised machine learning algorithms to predict these firms, we use the growth rates for employment or revenue to classify firms as high growth (1) or not high growth (0).

Figure 6: Proportion of high-growth firms by complexity indicator



The data display a heavy class imbalance due to the existence of many more non-high-growth firms than high-growth firms. Only 11.5% and 12.9% of firms meet the definition of high-growth firms for each employment and revenue, respectively. To account for this, weights are used in training the model when re-sampling each fold.

To assess model performance, a number of out-of-sample prediction metrics can be employed. These differ from measures of explanatory power (e.g., AIC) in that they focus on out-of-sample prediction accuracy as opposed to explaining in-sample variation. Here, model performance is evaluated by maximizing sensitivity (sometimes referred to as recall).

The sensitivity metric takes the number of high-growth firms that are accurately predicted by the model and divides them by the total number of high-growth firms in the sample (true positives and false negatives). This biases selection toward a model that is more likely to predict a firm to be high growth if it is actually a high-growth firm, but at the expense of increasing the number of false-positive predictions. The use of sensitivity helps mitigate the effect of the class imbalance by giving more weight to information attached to high-growth firm observations.

The algorithms are run using the Caret package from the R programming language (Kuhn et al. 2018). Three supervised machine learning algorithms are utilized:

1. Logistic model with elastic net regularization (Friedman, Hastie, and Tibshirani 2013)
2. Random forests (Wright and Ziegler 2017)
3. Neural net (Venables and Ripley 2002)

4.1 Logistic model with elastic net regularization

Logistic regression is applied to the binary $y = [0, 1]$ indicator variable for whether or not a firm is high growth. The full variable list is used as the set of inputs to the model. Then, we use the elastic net to shrink certain coefficients and select variables. The elastic net uses a linear combination of the Lasso (L1-normalization) and Ridge regression (L2-normalization) to improve on the performance of Lasso while producing parsimonious models. It encourages grouping of correlated inputs for inclusion or exclusion in the model, and improves on Lasso when the number of predictors is greater than the number of observations (Zou and Hastie 2005).

We implement a logistic regression with elastic net by minimizing the negative of the logistic likelihood function and including the shrinkage penalty (Hastie, Tibshirani, and Wainwright 2015):

$$-L(\beta|X) = \min_{\beta} \left(-\frac{1}{N} \left(y_i X\beta - log\left(1 + e^{X\beta}\right)\right) + \lambda((1-\alpha)\|\beta\|^2 + \alpha\|\beta\|)\right)$$

In this case, the penalty is the weighted average of the L1 and L2 norms, with the parameter α determining the relative weighing between the two. As with Lasso and Ridge regression, the parameter λ determines the degree of shrinkage.

Values for the hyperparameters α and λ are set iteratively within each fold by maximizing sensitivity during the k-fold cross validation. If α is found to equal 0, the equation reduces to the Ridge regression; if $\alpha = 1$, it reduces to the Lasso.

However, general linear models, such as logistic regression, can suffer from poor performance if the decision variable cannot be separated by a linear decision boundary. The following two non-linear methods are applied to allow for non-linear dissections of the prediction space. The first is a prediction method based on a forest of decision trees, while the second is a neural network.

4.2 Random forest

Decision trees are known to produce noisy, but unbiased, predictions (James, Witten, Hastie, and Tibshirani 2013). Researchers have therefore applied different methods of aggregating decision trees to produce forests with lower error variance than comes from individual trees. Here, we apply the fast implementation of random forests (Ho 1995) from the Ranger package (Wright and Ziegler 2017) through the Caret package (Kuhn et al. 2018). The forest is then used for classification through majority vote. By using the majority vote, the noisy individual trees produce aggregated predictions with lower error variance.

The Ranger algorithm allows for a comparison between the original implementation of random forest (Ho 1995) and the *ExtraTrees* (sometimes referred to as extreme randomized forest) algorithm (Geurts et al. 2006). The original random forest algorithm creates random tree structures by randomizing a sub-sample of the available units, a sub-sample of the available variables and the order of the sub-sampled variables. This process for creating an independent and identically distributed (i.i.d.) forest of predictors is referred to as bootstrap aggregating (bagging), and has been demonstrated to improve the prediction performance of tree classifiers that are noisy on their own (Ho 1995). The *ExtraTrees* algorithm uses random

variable selection and random variable splitting to similar effect, but does not sub-sample available units. Rather, the random ordering of variables, coupled with randomly choosing where to split variables, produces a forest of i.i.d. trees.

To implement the tree-based predictor through Caret, we tune hyperparameters for the type of randomization algorithm (random forest vs *ExtraTrees* – *splitrule*) and for the number of random variables selected from which a tree is grown (*mtry*).

4.3 Neural network

A neural network is a two-stage classification model where a series of inputs enter into the model, pass through a collection of nodes in a number of intermediate steps and then produce a predicted output. The intermediate steps, referred to as hidden layers, are derived as linear combinations of the inputs. These hidden layers are then fed into an activation function (here the commonly applied sigmoid function is used) to produce the final output. A simple neural network with a single hidden layer and up to 19 nodes is estimated. This bias term captures the intercepts of the two stages. To estimate the neural network, two hyperparameters are tuned to determine the number of nodes in the hidden layer (*size*) and the rate at which information in the model decays as it passes through the different layers (*decay*).

5 Results

Results are presented in four steps. First, we explain the pre-processing done and report the hyperparameter values. These speak to the complexity of the models used for prediction and describe the form of the models ultimately employed. Second, we report the results of the prediction exercise. We highlight the predictions made during the training phase, which are used to select the best model based on the chosen performance metric (sensitivity), as well as the out-of-sample predictions made using the holdout data set. Third, we present the most important variables for prediction for each model. And, fourth, we present the results from a one-year-ahead prediction exercise. In Appendix 2 we also include a variety of robustness tests, including running the elastic net without weights, the removal of the industry variable, changing our cutoffs for the HGF definition from 20% to 15% and 25%, exploring more hyperparameters for our neural net and using a different training/testing split.

5.1 Pre-processing and hyperparameter tuning

In our data pre-processing, we one-hot encode our categorical variables and scale and centre the numerical variables. Prior to one-hot encoding, we have 130,607 observations and 90 features. We also include the current, one- and two-year lags for many of the variables. A full list of all our variables is included in Appendix 1.

To estimate the models, the data set is divided into two sub-data sets: a training/validation data set and a testing/holdout data set. The training data set comprises 80% of observations and is used to parameterize the models. For each model, we use tenfold cross-validation, based on partitioning the training data set as 75% training and 25% validation, to tune the hyperparameters. The testing/holdout data set comprises 20% of the observations and is employed to assess the final models' performance.

Hyperparameter values for predicting employment high-growth firms and revenue high-growth firms are presented in Table 4.

Table 4: Parameter tuning for HGF predictions

Parameters	Employment	Revenue	Description
Elastic Net			
alpha	1	1	mixing between Ridge and Lasso
lambda	0.000433	0.0724	shrinkage parameter
Random Forest			
mtry	2	2	number of randomly selected predictors at each node
splitrule	extratrees	extratrees	splitting rule
Neural Net			
size	3	1	number of nodes in hidden layers
decay	0.00024	0.01778	regularization parameter (it avoids overfitting)

For the elastic net, α puts a higher weight on a Lasso-type penalty for employment and revenue, and λ is close to zero, which leads to the coefficient size to be assigned a low penalty. This is not unexpected since the more observations are in the model relative to predictors, the smaller the penalty will be (Zou and Hastie 2005).

For the random forest, we generate a total of 500 trees to produce the forest. Parameter tuning for the number of variables selected for comparison at each node, *mtry*, is selected over the range $[2, p]$, where p is the number of predictors. Higher values for the number of variables reflects greater similarity across trees, while lower values produce a forest with greater variability across trees (Probst, Wright, and Boulesteix 2019). The minimum number of samples (observations) at end node is set to one. Finally, we select the *ExtraTrees* method as opposed to a traditional random forest algorithm for randomization since it maximizes the performance of the model.

For the neural network, we use the Ripley (2008) algorithm for model selection and parameterization of the feed-forward network. Implementation of the neural networks for employment high-growth firms and revenue high-growth firms suggest a use of models that are nearly linear for both employment (three nodes in the hidden layer) and revenue (one node in the hidden layer). The values for the second hyperparameter, decay, are small. Decay serves a similar role to λ in the elastic net to penalize overfitting. This mimics the types of models found using logistic regression with elastic net where there is minimal shrinkage and parsimonious parameterization through Lasso. Other hyperparameters were not calibrated for in this model (learning rate, dropout, regularization, activation function choice, network weight initialization, momentum, batch size and epochs). More complexity, such as a larger neural net with more hidden layers, could be explored in future work; however, this would take away from interpretability as we would no longer be able to extract important variables (Section 5.3).[4]

[4]We tested other models that allow for more layers; however, when hyperparameter tuning these models, the parameters for the number of nodes in the second and third hidden layers yielded an optimal number of zero nodes in both of these. Given the capacity to interpret important variables that comes with only having one neural net layer, we chose a model with only one hidden layer.

5.2 Prediction results

We present prediction results in two ways. For results based on the in-sample cross-validation exercise, we report the sensitivity, specificity and balanced accuracy metrics. These measures are based on the ability of the models to accurately predict which firms are high growth. Sensitivity, the metric maximized during model selection, is defined as the number of accurately predicted high-growth firms divided by the total number of high-growth firms in the sample. Conversely, specificity is defined as the number of accurately predicted non-high-growth firms divided by the total number of non-high-growth firms in the sample. Balanced accuracy is the average of sensitivity and specificity. Hence, it is a measure of accuracy that corrects for class imbalance by giving less importance to accurately classifying the larger class (non-high-growth firms) and more importance to accurately classifying the small class (high-growth firms).

For results based on out-of-sample prediction, we report metrics of accuracy, sensitivity, specificity and balanced accuracy.[5] The out-of-sample exercise uses our testing set of the data (20% of original sample). The testing set is data the models did not see when training the models using cross-validation. Hence, out-of-sample metrics on the testing set are a better measure of how our models perform.

5.2.1 Cross-validation performance

Table 5 shows the results from our cross-validation exercise on the training set. Based on maximizing sensitivity, the elastic net produces the best model for predicting in-sample employment high-growth firms, whereas the neural net performs best for revenue high-growth firms. However, the differences in performance between the models are not large, suggesting limited benefit to using more computationally intensive non-linear models, such as the random forest. This does not mean that machine learning cannot yield improvements on previous linear prediction exercises since our logit model uses elastic net regularization for variable selection.

When doing the cross-validation exercise, we apply weights to give more importance to training on high-growth-firm observations over non-high-growth ones. In Table A-2 of the appendix, we show how drastically the results change when we do not apply weights in training.

Table 5: Results for HGF in-sample predictions

Model	Sensitivity	Specificity	Balanced Accuracy
Employment			
Elastic Net	70.8%	69.7%	70.3%
Random Forest	66.7%	72.1%	69.4%
Neural Net	70.3%	72.3%	71.3%
Revenue			
Elastic Net	67.6%	63.6%	65.6%
Random Forest	62.4%	75.0%	68.7%
Neural Net	70.7%	59.0%	64.9%

[5] Accuracy is defined as the proportion of correctly predicted high-growth and non-high-growth firms.

5.2.2 Testing data set accuracy

The confusion matrices on the out-of-sample testing set are presented in Table 6 and the performance metrics for the different models are further summarized in Table 7.

Table 6: Confusion matrix for HGF predictions

Predicted	Employment HGF Actual		Revenue HGF Actual	
Models	Non HGF	HGF	Non HGF	HGF
Elastic Net				
Non HGF	61.2%	3.4%	54.8%	4.3%
HGF	27.2%	8.1%	32.3%	8.7%
Random Forest				
Non HGF	63.6%	3.7%	65.6%	4.5%
HGF	24.8%	7.8%	21.4%	8.5%
Neural Net				
Non HGF	63.8%	3.4%	61.9%	5.1%
HGF	24.7%	8.1%	25.1%	7.9%

For employment high-growth firms, the neural network performs best overall with a sensitivity of 70.6% and a balanced accuracy of 71.4%. The higher balanced accuracy comes from the relatively stronger ability of the neural network to correctly predict simultaneously the non-high-growth firms (specificity) and the high-growth firms (sensitivity).

For revenue high-growth firms, the random forest performs best with a sensitivity of 65.1% and a balanced accuracy of 70.2%. The random forest is much better at predicting the non-high-growth firms (specificity) than the other models. Even though it is not as good as the elastic net model at predicting the high-growth firms (sensitivity), the higher specificity leads to the highest balanced accuracy of all the models.

Table 7: Results for HGF out-of-sample predictions

Model	Accuracy	Sensitivity	Specificity	Balanced Accuracy
Employment				
Elastic Net	69.3%	70.2%	69.2%	69.7%
Random Forest	71.4%	67.7%	71.9%	69.8%
Neural Net	71.9%	70.6%	72.1%	71.4%
Revenue				
Elastic Net	63.5%	67.2%	62.9%	65.1%
Random Forest	74.0%	65.1%	75.4%	70.2%
Neural Net	69.8%	61.0%	71.2%	66.1%

Relative to the in-sample metrics, the out-of-sample metrics have similar values, but the ranking of the models can differ. These, however, represent small changes as the numerical values are not much different than the metrics reported from the training data. This suggests

that the models do not contain too much in-sample bias, and that they generalize well to new observations.

5.3 Teasing out the important variables

We list the 10 most important variables for prediction in each model in Table 8 (employment high-growth firms) and Table 9 (revenue high-growth firms). It is important to note two things. First, although these variables are determined by each model to be important in explaining the predictions, it does not indicate whether that relationship is positive or negative. Second, if co-linear variables are present, the models may randomly attribute importance to one or the other in prediction. Hence, these important variables should not be taken as independently important or causal. In our models, the logit model uses elastic regularization for variable selection, while all variables are used for the random forest and neural net.

Table 8: List of important variables for employment growth[6]

Variable
Elastic Net
Average Employment 2012
Province of Nunavut
Food services and drinking places
Other information services
Primary metal manufacturing
Forestry and logging
Random Forest
New firm indicator (< 4 years)
Firm age since birth
Foreign ownership indicator
Firm age since incorporation
Food services and drinking places
Multi-location indicator
Multi-establishement indicator
Professional, scientific and technical services
Province of Quebec
Province of Alberta
Neural Net
Average Employment 2012
Average Employment 2011
Average Employment 2010
Payroll 2012
Payroll 2011
Payroll 2010
Food services and drinking places
Multi-location indicator
Province of Quebec

In calibrating the α parameter for the elastic net, we find an optimal value of one, indicating that the elastic net penalty coincides with Lasso. As a result, only a subset of variable coefficients take a non-zero value. In the case of the revenue high-growth firms, the only non-zero coefficients are for firm age and whether a firm is a recent entrant. These correspond to firm characteristics that are generally found to be indicators of firm dynamism. For the employment high-growth firms, the list of important variables also includes variables related to employment, province and industry indicators.

Table 9: List of important variables for revenue growth[7]

Variable
Elastic Net
Firm age since birth
New firm indicator (< 4 years)
Random Forest
New firm indicator (< 4 years)
Foreign ownership indicator
Firm age since birth
Firm age since incorporation
Food services and drinking places
Multi-location indicator
Professional, scientific and technical services
Multi-establishment indicator
Construction of buildings
Province of British Columbia
Neural net
Firm age since birth
Firm age since incorporation
New firm indicator (< 4 years)
Foreign ownership indicator
3-year industry GDP growth rate
3-year industry price growth rate
1-year industry GDP growth rate
Food services and drinking places
Province of Quebec
Multi-location indicator

The random forest important variables again contain age and entry. However, since this model allows for a non-linear decision boundary, more discrete variables are captured as important variables. Foreign ownership and measures of firm complexity (multiple NAICS, multiple locations) are also important predictors, as are some province and industry indicators.

The neural network for employment predictions selects all different types of variable, but emphasizes the employment and payroll variables and their lags. For the revenue predictions,

[6]Some industry variables have been removed because they contain too few high-growth firms. This is to maintain confidentiality and comply with the *Statistics Act*.

[7]Some industry variables have been removed because they contain too few high-growth firms. This is to maintain confidentiality and comply with the *Statistics Act*.

it select variables for firm age and entry as well as complexity and foreign ownership indicators and industry GDP and deflator growth rates.

In Appendix B, we test removing the industry variable from the machine learning models and find that the models rely more on some firm balance sheet financial variables such as capital and labour costs, one- and two-year changes in asset and prior sales. However, the level of importance accorded to these variables remains small (importance scores of less than 2 out of 100) and removing the industry variable comes at the cost of decreasing the models' performance (see Table B-2).

Overall, the random forest puts more emphasis on discrete variables, whereas both the elastic and the neural net models emphasize a mix of continuous and discrete variables. This is likely due to the random forest being a better non-linear separator for discrete variables along different dimensions. These discrete variables turn out to be important in high-growth revenue prediction since the random forest model performs best in terms of balanced accuracy for this classification. In turn, the continuous variables may be more important for employment high-growth prediction since the neural net is the best performer in terms of balanced accuracy.

5.4 Predictions one year ahead

As a final test for the models, we conduct a pseudo one-step-ahead prediction exercise for the year 2016. For the prediction exercise, we use the data from the 2013 NALMF cross-section to predict which firms, given they exist in 2013 and 2016 and their data in 2013, will be high growth over the 2013 to 2016 period. We then compare the predictions with the actual growth patterns exhibited by the firms.

The predictions metrics (Table 10) for comparing model outcomes illustrate that the best-performing supervised machine learning classifiers are, once again, the neural net for employment high-growth and the random forest for revenue high-growth. These algorithms can predict approximately two-thirds of high-growth firms for 2016. When compared with the results from the holdout data set (Table 7), the best score for balanced accuracy falls only 0.4 percentage points for employment and 0.4 percentage points for revenue, meaning that they perform well on data one year ahead. They also perform better when it comes to predicting employment high-growth rather than revenue high-growth firms.

Table 10: Results for HGF predictions on future data

Model	Accuracy	Sensitivity	Specificity	Balanced Accuracy
Employment				
Elastic Net	69.5%	69.7%	69.4%	69.6%
Random Forest	70.5%	66.5%	71.1%	68.8%
Neural Net	72.0%	69.7%	72.3%	71.0%
Revenue				
Elastic Net	64.1%	66.1%	63.8%	64.9%
Random Forest	71.9%	62.2%	73.4%	67.8%
Neural Net	69.8%	60.1%	71.3%	65.7%

6 Conclusion

High-growth firms garner considerable interest due to their dynamic nature. Research on these firms shows that small, young firms are an important source of business dynamics, and that they are the most likely source of high-growth firms. This stylized fact is strongly reflected in the prediction models selected by the supervised machine learning algorithms where firm age and new entrant are key variables for prediction.

The predictions here are based on features and characteristics of firms that are externally visible (e.g., age, industry, size, industry growth rate, etc). We provide evidence that visible firm characteristics can yield important information for making high-growth predictions, particularly when we can control for sample imbalances in training our models and when we can let the model choose from a large number of variables. Moreover, the use of a logistic model (which has a linear boundary) produces results that are not greatly different from non-linear approaches (random forest or neural network). Consequently, it appears possible to train a relatively simple, parsimonious model for prediction.

The predictions themselves produce two types of information. The first is a predicted population of nascent high-growth firms that is about 12–13% of the size of the overall sample. As a result, the algorithms appear to be able to reduce the scale of the prediction problem by weeding out unlikely candidates. Within the predicted population, the probability of finding a high-growth firm is approximately twice that of the general population. In effect, the smaller predicted population has a higher concentration of target firms than the overall population.

Finally, the results presented here suggest that supervised machine learning models can find some predictive power from commonly observed firm characteristics. However, a note of caution is warranted. The nascent high-growth firm population, while containing a higher concentration of high-growth firms than the overall population, still has three out of four firms that are false positive predictions. The results here should therefore be viewed as a first attempt at using machine learning for high-growth firms prediction using Canadian administrative data. The results have not been proven to generalize beyond the range of years employed, and the input data set contains numerous indicator variables. A richer input data set, with derived business owner and employee features, may add to the models' predictive ability. Nevertheless, as a first step, the results show promise.

Appendix

A: Variable list

This is a list of all the variables used in our models. We include all the variables as-is and also include the one- and two-year lags or the one- and two-year changes in some, as indicated in Table A-1.

Table A-1: Variable list

Variable Name	Type	Other Specifications	Description
PD7 AvgPay NonZero	numeric	1- and 2-year lags	Average payroll reported from PD7s recording a value of one or more. Calculated by taking the mean of all non-zero monthly payroll submissions.
Total assets	numeric	1- and 2-year lags. 1- and 2-year changes	Total of all current, capital, long-term assets, and assets held in trust. (T2 Schedule 100)
Total liabilities	numeric	1- and 2-year lags	Total of all current and long-term liabilities. (T2 Schedule 100)
Total shareholder equity	numeric	1- and 2-year lags	Sum of all shareholder equity amounts. (T2 Schedule 100)
GSTModeledSales	numeric	1- and 2-year lags	Sales as derived from GST and recorded. (Business Register)
Total cost	numeric	1- and 2-year lags	Sum of cost of sales and total operating expenses. (Derived from T2 Schedule 125)
KL costs	numeric	1- and 2-year lags	Sum of labour income and capital costs. (Derived from T2 Schedule 125)
Profits nozero	numeric	1- and 2-year lags	Gross output derived from income statement minus total costs. (Derived from T2 Schedule 125)
Gross output	numeric	1- and 2-year lags	Sum of capital income, payroll and intermediate inputs. (Derived from T2 Schedule 125)
Value added nozero	numeric	1- and 2-year lags	Capital income plus payroll. (Derived from T2 Schedule 125)
Total long term liabilities	numeric	1- and 2-year lags	Total long term liabilities. Sum of all long term liabilities reported. (T2 Schedule 100)

Table A-1: Variable list *(continued)*

Variable Name	Type	Other Specifications	Description
Total tangible assets	numeric	1- and 2-year lags	Total tangible capital assets. Sum of all tangible capital assets reported. (T2 Schedule 100)
Total intangible assets	numeric	1- and 2-year lags	Total intangible capital assets. Sum of all intangible capital assets reported. (T2 Schedule 100)
Retained earnings	numeric	1- and 2-year lags.	Retained earnings/deficit at end of period minus retained earnings/deficit at start of period. (T2 Schedule 100)
SRED Expenditures	numeric	1- and 2-year lags. 3-year sum.	Expenditures qualifying for the Scientific Research and Experimental Development (SR&ED) expenditure claim. (T2 Schedule 32)
Capital cost	numeric	1- and 2-year lags	Total cost of capital related to amortization, royalties, expenses, leases and interest payments. (Derived from T2 Schedule 125)
Intermediate inputs	numeric	1- and 2-year lags	Total cost minus sum of labour income and capital cost. (Derived from T2 Schedule 125)
Total cost of sales	numeric	1- and 2-year lags	Total cost of sales. Sum of all cost of sales amounts reported. (T2 Schedule 125)
Capital cost allowance	numeric	1- and 2-year lags	Capital Cost Allowance. Portion of capital cost permitted to deduct from income it earned. (T2 Schedule 125)
Sales goods and services	numeric	1- and 2-year lags	Total sales of goods and services. (T2 Schedule 125)
Age birth	numeric	NA	Age since birth. (Derived from Business register)
Age inc	numeric	NA	Age since incorporation. (Derived from Business register)
OPAddressProvince	factor with 14 levels	NA	Operating Province. (Business Register)

Table A-1: Variable list *(continued)*

Variable Name	Type	Other Specifications	Description
EntMultiEstablishFlag	binary	NA	Indicator for enterprise with more than one establishment under it. (Business Register)
EntMultiLocationFlag	binary	NA	Indicator for enterprise with more than one location under it. (Business Register)
EntMultiProvinceFlag	binary	NA	Indicator for enterprise operating in more than one province. (Business Register)
EntMultiActivityFlag	binary	NA	Indicator for enterprise operating in more than one industry. (Business Register)
Foreign owned	binary	NA	Indicator that country of residence of the ultimate shareholder or group of shareholders for the enterprise is non-Canadian. (Derived from AFTS)
Naics main	factor with 91 levels	3-digit NAICS code.	Mode of non-missing NAICS code between 2009 and 2016. (Derived Business Register)
IFPA3 gr	numeric	1- and 3-year growth rates	3-digit industry-level price growth rate. (Derived from KLEMS)
IFQA3 gr	numeric	1- and 3-year growth rates	3-digit industry-level GDP growth rate. (Derived from KLEMS)
New firm	binary	NA	Indicator for entreprise born in the last 3 years. (Derived from Business Register)
Debt to equity	numeric	Derived	Total shareholder equity divided by non-missing and non-zero values of total long term liabilities. (Derived from T2 Schedule 100)

B: Robustness

We perform five robustness checks of our results: we remove weights in k-fold cross validation hyperparameter calibration; we remove the industry variable from the features; we change our cutoffs for the HGF definition from 20% growth over three years to 15% and 25%; we explore more hyperparameters for our neural net; and finally, we use a different training/testing split for the data. We compare these with our baseline results in Table 7.

We use sample weights in our cross-validation exercise to give more weight to the observations that are high-growth firms in our training stage. This use of weights in re-sampling allows us to account for the imbalanced presence of high-growth firms in the sample (only 12–13% of sample). When we remove the weighted folds, we are allowing our models to learn equally from all observations in the training sample.

The elastic net with unweighted folds (Table B-1) have a higher accuracy than our baseline results, 89%, meaning that it best predicts both non-high-growth and high-growth firms, but it has a very low sensitivity and correctly predicts only 7.6% of the high-growth firms in the sample. Hence, the elastic net with weighted folds performs better when using balanced accuracy as a performance metric (the average of sensitivity and specificity—corrects for class imbalance). For the prediction of employment high-growth firms, the balanced accuracy is 69.7% when using weighted folds in training, relative to the balanced accuracy of unweighted folds of 53.6%.

Table B-1: Results for HGF predictions without using weights in k-fold cross-validation

Model	Accuracy	Sensitivity	Specificity	Balanced Accuracy
Elastic Net - employment	89.0%	7.6%	99.6%	53.6%
Elastic Net - revenue	87.5%	12.1%	98.8%	55.4%

Next, we remove the industry variable to evaluate its impact on our results (Table B-2). We do this to ensure our model does not over-rely on the industry the firm is in to make predictions. The random forest is the only model that becomes significantly worse at predicting high-growth firms than in the baseline. This is in line with the results found in the important variables section that the random forest model relies heavily on the industry variable for prediction, given its non-linear decision boundary. The elastic and neural nets perform only slightly worse than the baseline results when it comes to sensitivity.

When we run this model without industry-level data, we do get some firm balance sheet variables such as total capital and labour costs for the elastic net; asset changes (one and two years prior) for the random forest; sales (one and two years prior) and total cost of sales (one year prior) for the neural network, when predicting employment HGFs. For revenue HGFs, important predictors are asset changes (one and two years prior) for the random forest; sales (one and two years prior) and total cost of sales (one year prior) for the neural network. However, all the importance scores for all of these variable are low (<2 out of a 100), so they do not play a dominant role in prediction.

Table B-2: Results for HGF predictions without the industry variable

Model	Accuracy	Sensitivity	Specificity	Balanced Accuracy
Employment				
Elastic Net	68.3%	69.2%	68.2%	68.7%
Random Forest	92.5%	52.6%	97.7%	75.1%
Neural Net	76.6%	66.9%	77.8%	72.4%
Revenue				
Elastic Net	68.7%	62.2%	69.6%	65.9%
Random Forest	88.6%	40.8%	95.7%	68.2%
Neural Net	74.5%	57.7%	77.0%	67.4%

Next, we test different cutoffs for our definition of high-growth firms (Table B-3). Our baseline result classifies a firm as being high growth if it achieves an average growth rate of at least 20% over three years based on the OECD definition (OECD 2007). Since this is an ad-hoc definition, it may arbitrarily exclude or include some firms that may bias the results. Hence, we test results for cutoffs of 15% and 25% growth over three years.

For a cutoff of 15%, we find that while the number of correctly predicted HGFs increases, our measure for sensitivity (of the HGFs, how many of them are correctly predicted) decreases by about 0.2 pp. We might expect the performance to go up since we use a cutoff that includes more HGFs, and hence more information on them. However, this signals our modelling weights are effective in extracting good information given our small sample of HGFs.

Table B-3: Results for HGF out-of-sample predictions using different cutoffs for high-growth firms

Model	Accuracy	Sensitivity	Specificity	Balanced Accuracy
Employment				
Cutoff 15%				
Elastic Net	67.9%	68.3%	67.8%	68.1%
Random Forest	68.5%	65.6%	69.1%	67.3%
Neural Net	73.3%	65.1%	74.9%	70.0%
Cutoff 25%				
Elastic Net	72.1%	73.0%	72.0%	72.5%
Random Forest	74.5%	67.2%	75.1%	71.2%
Neural Net	76.5%	70.5%	77.0%	73.7%
Revenue				
Cutoff 15%				
Elastic Net	68.0%	66.1%	68.4%	67.2%
Random Forest	69.0%	62.5%	70.5%	66.5%
Neural Net	70.7%	62.3%	72.6%	67.5%
Cutoff 25%				
Elastic Net	71.7%	68.8%	72.0%	70.4%
Random Forest	76.4%	64.7%	77.7%	71.2%
Neural Net	76.2%	67.3%	77.2%	72.3%

Then we investigate the structure we impose on the neural net model. We explore adding more hidden layers as a hyperparameter. This hyperparameter allocates the optimal number of nodes in each hidden layer for up to three hidden layers. Our baseline model imposes one hidden layer in order to keep the interpretability of the model by allowing us to extract the important variables. Our k-fold cross-validation yields that multiple nodes are optimal only for the first hidden layer and that zero nodes are optimal for hidden layers two and three. Hence, we choose to use only one hidden layer, given that it allows us to extract important variables for the model if there is only one layer.

Finally, we experiment with the use of a 95/5% training/testing split (Table B-4) as opposed to the traditional 80/20% split. Given that the total size of our data set includes over 130,000 observations and very few high-growth firms, allocating more data to training may yield benefits. However, the improvements are minimal. The balanced accuracy for revenue HGFs is only 4 pp for the elastic and neural nets, but it is actually reduced for the random forest. There are virtually no improvements for predicting employment high-growth firms. Though there is a slight gain in balanced accuracy for the random forest, this gain comes from improvements in prediction of non-HGF (specificity) and worse prediction of HGFs (sensitivity).

Table B-4: Results for HGF out-of-sample predictions using a different training/testing split

Model	Accuracy	Sensitivity	Specificity	Balanced Accuracy
Employment				
Elastic Net	70.5%	70.2%	70.5%	70.4%
Random Forest	72.6%	68.8%	73.1%	70.9%
Neural Net	77.3%	67.3%	78.6%	73.0%
Revenue				
Elastic Net	70.8%	66.9%	71.4%	69.2%
Random Forest	73.9%	61.0%	75.8%	68.4%
Neural Net	74.7%	62.7%	76.5%	69.6%

References

Athey, S. (2018). The impact of machine learning on economics. In *The Economics of Artificial Intelligence: An Agenda*, pp. 507–547. National Bureau of Economic Research, Inc.

Autor, D., D. Dorn, L. F. Katz, C. Patterson, and J. Van Reenen (2020, 02). The Fall of the Labor Share and the Rise of Superstar Firms*. *The Quarterly Journal of Economics 135*(2), 645–709.

Bravo-Biosca, A. and S. Westlake (2009). The vital 6 per cent: How high-growth innovative businesses generate prosperity and jobs. *London: NESTA*.

Burgette, L. F. and J. P. Reiter (2010, 09). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology 172*(9), 1070–1076.

Chalfin, A., O. Danieli, A. Hillis, Z. Jelveh, M. Luca, J. Ludwig, and S. Mullainathan (2016, May). Productivity and selection of human capital with machine learning. *American Economic Review 106*(5), 124–27.

Clayton, R. L., A. Sadeghi, D. M. Talan, and J. Spletzer (2013). High-employment-growth firms: defining and counting them : Monthly Labor Review: U.S. Bureau of Labor Statistics.

Coad, A. (2009). *The Growth of Firms: A Survey of Theories and Empirical Evidence*. New perspectives on the modern corporation. Edward Elgar.

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

Coad, A. and S. Srhoj (2020, October). Catching Gazelles with a Lasso: Big data techniques for the prediction of high-growth firms. *Small Business Economics 55*(3), 541–565.

Cote, S. and J. Rosa (2016, February). Comparing different measures of high-growth enterprises: A canadian case study. *Innovation, Science and Economic Development*.

Daunfeldt, S.-O. and D. Halvarsson (2015). Are high-growth firms one-hit wonders? evidence from sweden. *Small Business Economics 44*(2), 361–383.

Davis, S. J., J. Haltiwanger, and S. Schuh (1996). Small business and job creation: Dissecting the myth and reassessing the facts. *Small Business Economics 8*(4), 297–315.

De Loecker, J., J. Eeckhout, and G. Unger (2020). The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics 135*(2), 561–644.

Dixon, J. and A.-M. Rollin (2014). The distribution of employment growth rates in canada: The role of high-growth and rapidly shrinking firms. *Economic Analysis Research Paper Series 2014091*.

Friedman, J. H., T. Hastie, and R. Tibshirani (2013). *Regularization Paths for Generalized Linear Models via Coordinate Descent*, Volume 33.

Geurts, P., D. Ernst, and L. Wehenkel (2006, April). Extremely randomized trees. *Machine Learning 63*(1), 3–42.

Haltiwanger, J., R. S. Jarmin, and J. Miranda (2013, 05). Who Creates Jobs? Small versus Large versus Young. *The Review of Economics and Statistics 95*(2), 347–361.

Hastie, T., R. Tibshirani, and M. Wainwright (2015, May). *Statistical Learning with Sparsity: The Lasso and Generalizations*. New York: Chapman and Hall/CRC.

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

Ho, T. K. (1995). Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition Volume 1*, ICDAR '95, USA, pp. 278. IEEE Computer Society.

Hölzl, W. (2009, June). Is the R&D behaviour of fast-growing SMEs different? Evidence from CIS III data for 16 countries. *Small Business Economics 33*(1), 59–75.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning*. Springer.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman (2017, 03). Technological Innovation, Resource Allocation, and Growth*. *The Quarterly Journal of Economics 132*(2), 665–712.

Kuhn, M., cre, J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, R. C. Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan, and T. Hunt (2018, August). caret: Classification and Regression Training.

Marsili, O. (2001). The anatomy and evolution of industries: technological change and industrial dynamics. In *The Anatomy and Evolution of Industries*. Edward Elgar Publishing.

Meulbroek, L. K., M. L. Mitchell, J. H. Mulherin, J. M. Netter, and A. B. Poulsen (1990). Shark repellents and managerial myopia: An empirical test. *Journal of Political Economy 98*(5), 1108–1117.

Miyakawa, D., Y. Miyauchi, and C. Perez (2017). *Forecasting firm performance with machine learning: Evidence from Japanese firm-level data*. RIETI.

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

OECD (2007). Eurostat-oecd manual on business demography statistics.

OECD (2017). *The growth of Canadian firms: Evidence using different growth measures*.

Probst, P., M. N. Wright, and A.-L. Boulesteix (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery 9*(3), e1301. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1301.

Ripley, B. (2008, 01). *Pattern Recognition And Neural Networks*, Volume 11.

Statistics Canada (2018). Multifactor productivity, value-added, capital input and labour input in the aggregate business sector and major sub-sectors, by industry.

Storey, D. J. (2016). *Understanding the small business sector*. Routledge.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software 45*(3), 1–67.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Statistics and Computing. New York, NY: Springer.

Weinblat, J. (2018, September). Forecasting European high-growth Firms - A Random Forest Approach. *Journal of Industry, Competition and Trade 18*(3), 253–294.

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

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 67*(2), 301–320.